文化部「建置國家語言資料庫先期規劃研究」勞務採購案修正後期末報告

執行單位：國立臺灣大學

計畫主持人：高照明教授

協同計畫主持人：翁聖賢律師　黃子桓博士

研究助理：陳祐萱、陳蓓怡、呂曉鈞、沈瑞恩、郭瑋星、林哲宇、黃子育

中華民國 109 年 8 月 10 日

目次

圖目錄

表目錄

前言

民國 108 年 1 月，《國家語言發展法[1]》公布，其中第一條提到：「為尊重國家多元文化之精神，促進國家語言之傳承、復振及發展，特制定本法。」。第八條更提到：「政府應定期調查提出國家語言發展報告，建置國家語言資料庫」。而文化部「建置國家語言資料先期規劃研究」勞務採購案需求說明書中也提到：「國家語言資料庫除應含國家語料庫外，亦應納入各國家語言史料、統計調查等相關資料，以作為國家語言傳承、復振及發展之基石。」

依據《國家語言發展法》第三條，「本法所稱國家語言，指臺灣各固有族群使用之自然語言及臺灣手語」，臺灣手語為此法明列之國家語言，但未述明自然語言是哪些語言。此概括立法為的是將語言名稱之命名權交予語言使用者，亦讓未來能夠將更多語言納入保障(“國家語言發展法摘要說明”)。參考民國 106 年文化部舉辦之公聽會 (國家語言發展法相關法令研究, 2017)，即可了解國家語言名稱之爭議，公聽會上共提出甲、乙兩案版本，甲案為最終公告施行之版本，但甲案中提及「華語」一詞之使用，避免以「國語」稱呼所造成之混淆。此外，因為手語的語言特性，為化解對其常見的誤解，認為手語不是一個語言，特別將臺灣手語獨立列出。其實臺灣手語亦屬於自然語言，是在長期的語言使用與發展下，自然而成的一種語言，即臺灣手語（Taiwanese Sign Language, TSL），而非由聽人主導、套用中文文法之文字手語（Signed Chinese, SC）。(Smith, 2005)

[1] https://www.moc.gov.tw/content_275.html

基於上述原因，本報告以《國家語言發展法》及相關重要法源作為依據，將臺灣手語、原住民族語、客語、閩南語、閩東語、與華語納入先期研究的討論。民國 106 年公布施行之《原住民族語言發展法[2]》與 107 年《客家基本法[3]》修正案皆明定「原住民族語言」及「客語」為國家語言，並以「原住民族語言」（簡稱「族語」）及「客語」稱呼。若參考國教院修訂課綱所列之語言，則國家語言亦包括閩南語及閩東語。(“國家語言 111 年列國高中部定課程含手語及閩東語”, 2020)

若以上述這幾點為基礎，目前我國建置國家語言資料庫的大方向應為，將現有瀕危語言相關資源納入典藏、並廣收各個國家語言的相關資料以保存臺灣的語言多樣性。因此，未來成立的國家語言資料庫，其定位除了收錄由書面、口語語言資料所構成的平衡語料庫外，亦應包含各典藏資料、語言史料、語言統計調查、連結資源等各種語言相關資料。

以下，本報告將從世界各國家語言資料庫的現況開始，逐一介紹各國國家語料庫、手語語料庫、群眾外包計畫、瀕危語言典藏計畫等的設計理念、使用的工具與技術、收集方式、應用層面等資訊；接著，再彙整本國各國家語言現有的相關資料庫及語料庫，並探討其語料的轉寫、標記與工具等議題；最後，本報告提出國家語言資料庫的用途以及使用對象、國家語言資料庫的完整架構、招標與需求說明書的內容項目、以及各種授權書及授權機制草案等規劃與建議。

[2] https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=D0130037

[3] https://law.moj.gov.tw/LawClass/LawAll.aspx?pcode=D0140005

壹、　世界各國家語料庫的現況分析

本章對世界各國家語料庫，包括書面及口語的語料的現況作介紹。首先將針對英國、美國、日本、中國等較具代表性，或是與我國建置國家語言資料庫相關性較高的語料庫提供詳細的文獻探討，接著再補充歐洲和其他地區的國家語料庫介紹。本章特別針對英國、美國、日本、中國國家語料庫作詳細說明，其理據如下：

首先，選擇英國是因為英國國家語料庫是世界上的第一個國家語料庫。該語料庫的特色包括語料較為平衡（包含書面語和口語），規模達到一億詞，有詳細的標記，應用廣泛等特點，因此對於我國建置國家語料庫具有指標性的參考意義。此外，英國威爾斯和蘇格蘭的地方語言受到強勢官方語言英語的影響而有式微趨勢，威爾斯和蘇格蘭政府和學界透過建立威爾斯和蘇格蘭語語料庫來保存並復振他們的語言的經驗對我們而言非常寶貴。本章所介紹的威爾斯語語料庫和蘇格蘭語語料庫雖沒有國家語料庫之名，卻非常值得我們借鏡。

選擇美國是因為在語料庫的科技與應用，一直以來英國和美國都是領先世界的國家。選擇開放美國國家語料庫主要是因為其題材選擇較為平衡，大部分屬於開放資料，且語料的標記與應用較為全面。雖然除了開放美國國家語料庫之外，另外還有不少頗具代表性且大型的美式英語語料庫（如，當代美式英語語料庫 COCA），考量到這些語料庫並不是開放資料，沒有國家語料庫之名，也沒有類似英國國家語料庫的組織來營運，因此這些語料庫最後並沒有被納入本章作介紹。另外，雖同為英語語料庫，美國國家語料庫的語料年代選擇（1990 年以後）也和英國國家語料庫（1960～1990 年代）不大相同，可以相互作比對，提供我國參考。

選擇日本主要原因有二：第一，日本的國家型語料庫主要是由國立國語研究所這個專責機構所完成的，這點和我國即將成立的國家語言研究發展中心的規劃剛好一致，因此日本的模式值得我國參考。第二，目前多數的國家語料庫都是以平衡語料庫（balanced corpus）作為定位，內容收錄各種能代表該國國家語言的書面和口語資料；而日本國立國語研究所除了日本現代書面語平衡語料庫之外，另外還收錄了像是瀕危語言的語料庫、歷史語料庫、將日語作為第二語言的語料庫、語言地圖等等，不同類型的語言資料，這點也和我國即將成立的國家語言資料庫定位相似，因此日本國立國語研究所的語言資料庫非常值得我們參考。

選擇中國國家漢語語料庫，是因為臺灣的華語、閩南語、客家語、閩東語皆同屬漢語語系，因此該語料庫的內容設計也值得我國參考。

在整理本章資料的過程中本團隊也發現，目前各國的國家語料庫主要都是收錄該國最強勢語言的資料，關於瀕危語言與手語的資料則相對稀少，甚至未收錄，目前發現只有日本和澳洲的國家語料庫有收錄瀕危語言的資料。威爾斯語雖然並非瀕危語言，但其語料收集機制對於本國正面臨傳承危機，而且資源又相對不足的國家語言來說，頗具參考價值。客觀而言，目前各國的國家語料庫現況，明顯與我國期望建置可以保存語言多樣性的國家語料庫之目標有所落差，為了彌補這些不足，本團隊針對手語主題撰寫了第二章節的介紹（國外手語語料庫、資料庫的現況分析），還有針對瀕危語言主題撰寫了第四章的介紹（國外相關數位典藏計畫、資料格式、與工具的分析）。

1.1. 英國國家語料庫—British National Corpus (BNC)

1.1.1. 1994 年之版本—BNC1994

英國是第一個建置國家語料庫的國家，在 1991～1994 年建立了世界第一個國家語料庫後，其它各國才漸漸開始建立自己的國家語料庫。英國國家語料 British National Corpus (BNC)是目前最具代表性的大型國家語料庫之一，其設計有許多地方值得我們參考。英國國家語料庫網址為：http://www.natcorp.ox.ac.uk/。

英國國家語料庫是個單語（Monolingual）、共時（Synchronic）、樣本（Sample）的一般語料庫（General Corpus），語料庫不限定任何特定的主題，目前包含約 1 億詞。該語料庫從 1991 年開始建置，1994 年完成，主要收錄了 20 世紀下半葉（1964 年以後）的各類的書面（90%）和口語（10%）資料。目前最新版本（第三版）是 BNC XML 版本，於 2007 年發布。書面資料包括各類報紙、期刊、學術書籍、小說、信件、備忘錄、學校論文等等；口語資料包括即興非正式訪談的轉寫檔、還有企業、政府會議、廣播節目、電話等各種情境的口語檔案。

英國國家語料庫最初是由牛津大學出版社（OUP）所領導的 BNC 聯盟（BNC Consortium）所創建，原始的成員包括牛津大學出版社、Longman 出版商、Chambers Harrap 出版商、牛津大學計算服務中心（OUCS），蘭卡斯特大學計算機語言學研究中心（UCREL）和大英圖書館研究與創新中心（British Library Research and Development Department）。該語料庫計畫主要是由英國貿易及工業部門（Department of Trade and Industry）、英國科學與工程研究委員會（Science and Engineering Research Council）所資助，而大英圖書館

（British Library）、英國國家學術院（British Academy）也提供了額外的資金。 語料庫創建團隊一旦找到了合適的文本並且確認獲得授權後，就會將資料轉換成機器可讀模式，並且添加各類標記，其分工模式如下：首先，牛津大學出版社、Longman 出版商和 Chambers Harrap 出版商會分工收集各類文本資料，牛津大學出版社負責收集書面資料，Longman 出版商負責收集口語資料，Chambers Harrap 出版商則負責收集其餘未出版的雜項資料。牛津大學出版社與 Longman 出版商在收集資料時，會依照標準程序取得資料授權，其詳細內容可參考以下網址：http://www.natcorp.ox.ac.uk/corpus/permletters.html 。收集到各類資料後，這三家出版社會使用掃描、鍵盤打字、轉換現有電子資料格式等方式，將資料轉成機器可讀模式，完成後便會將這些資料傳送至牛津大學計算服務中心。該中心接著會進行標準化的編碼處理、文本語義檢查等工作，編碼所採用的規範為「語料庫文檔交換格式（Corpus Document Interchange Format，簡稱 CDIF）」，關於編碼格式的內容與應用可以參考 BNC Users Reference Guide（網址：http://www.natcorp.ox.ac.uk/docs/URG/）和 Encoding the British National Corpus（網址：http://www.natcorp.ox.ac.uk/docs/Burnage93a.htm）的說明。當牛津大學計算服務中心完成上述工作後，便會將資料傳送至蘭卡斯特大學計算機語言學研究中心進行詞性標註。蘭卡斯特大學的 Roger Garside 開發了一款名為 CLAWS4 自動標註工具，可以自動將資料分割成和句子差不多的單位並標上詞性，其所採用的詞性集（簡稱為 C5 Tagset）有 65 個類別，標註結果的正確率達 99.3%。關於詞性集的內容與自動標註過程可以參考以下兩個網頁：http://ucrel.lancs.ac.uk/papers/coling.html 、http://www.natcorp.ox.ac.uk/docs/garside_allc.html。完成詞性標記的資料

最後會再次傳送回牛津大學計算服務中心加上標頭（header）並進行最後的檢查，若無誤就會直接加入語料庫中。此外，上述語料庫處理的各階段相關資訊都會紀錄在由牛津大學計算服務中心所維護的資料庫中。

BNC 共收集 4,049 篇文章，總詞數為 96,986,707 詞（orthographic word），但經過詞性標記後的總詞數為 98,363,783 詞（w-unit）。BNC 之目標為能夠代表英式英語在各情境使用的情形，因此包含各種文類以及主題的書面語料及口語語料。然而，由於收集及處理口語語料需要的時間與經費相當多，本語料庫內的口語語料只占約 10%（共 10,409,851 w-units），書面語料占約 90%（共 87,953,932 w-units）。

書面語料依照抽樣方式組成，每篇文章抽出之語料樣本不超過 45,000 詞，平均詞數為 40,000 詞。該部分以三種不同的標準而挑選：即「領域」、「時代」、以及「媒介」。「領域」之標準代表每筆語料之主題，主要分成兩類：「想像性」（imaginative；含小說及其他虛構作品，占約 20%）以及「資訊性」（informative；更細分成以下共 8 類：自然科學、應用科學、社會科學、國際事務、貿易及金融、藝術、思想及信仰、以及閒暇，占約 80%）。「時代」之標準，基於英國國家語料庫是共時（synchronic）的語料庫，因此語料文字的出版日期不應早於 1975 年，但針對「想像性」類別的語料則放寬標準至 1965 年，因為該語料文字一直都很流行，且對該語言有影響力。「媒介」之標準指各筆語料出版的類型，例如圖書（57.9%）及刊物（29.8%）。每筆語料經過挑選後更細分成其他描述性的分類，例如作者資料、目標觀眾、及出版地區。口語語料內容分成兩部分：以說話者的性別、年齡、以及社會階級均衡挑選的自然會話（demographic）、以及各種場

合及語境（context-governed）（含教育或資訊性、商業、政府或其他制度、及閒暇）之語料。

本語料庫附有詞性標記，使用的標記系統及方法為 CLAWS4 C5。目前能夠下載的語料庫有 BNC-XML 完整版、BNC Baby（400 萬詞的子語料庫；XML 版的初版）、以及 BNC Sampler（200 萬詞的子語料庫）。本語料庫不採用多媒體資料、也不提供口語語料的音檔。由於本語料庫的平衡性及代表性相當理想，許多其他國家也依照 BNC 的設計原則建立自己的國家語料庫，例如下述的美國國家語料庫（ANC）以及波蘭國家語料庫（NKJP）。

使用對象：BNC 的三大用途為學術，商業，及教育。主要的使用者是編輯英語學習者辭典的出版業者以及研究自然語言處理及語料庫語言學的學者。

維運管理：BNC 由政府及民間共同出資，維運管理由英國國家語料庫聯盟負責，主要由三家出版商（牛津大學出版社，Longman 和 Chambers Harrap），兩所大學（牛津大學和蘭卡斯特大學）和大英圖書館的合作。

應用推廣：從一開始設計，BNC 即朝向語料提供大眾使用的目標，因此 BNC 語料授權方面也採取相對應的措施，使用者可以免費下載全部的語料。大多數的應用集中於自然語言處理，英語教學，和語言學。具體來說，編纂字典和同義詞典（thesauri）、編寫語言教材時，可借助語料庫引用自然生成（naturally occurring）的例子，學習者學會操作語料庫後，亦可用於自學英文；應用於自然語言處理時，可提供訓練或測試集資料、開發標記器（tagger）與剖析器（parser）等。

1.1.2. 2014 年之版本—Spoken BNC2014

1994 年英國建立了世界上第一個國家語料庫，事隔二十年，於 2014 年推出國家口語語料庫。(McEnery et al., 2017) 雖然 1994 年版本的 BNC 語料庫已有 420 萬字的口語語料，2014 年的版本擴增至 1150 萬字，從 124 位增加至 668 位發音人參與貢獻語料，語料錄音時間從 2012 年至 2016 年，於 2017 年上線，歷時五年的語料蒐集與處理，平均每年增加一百萬詞的口語語料，再加上一年的時間最終完成語料庫的建置，但目前尚無網站檢索介面，可從蘭卡斯特大學的 CQPWeb 平台註冊會員後，進入並下載 XML 的檔案，網址為：https://cqpweb.lancs.ac.uk。

因口語語料蒐集不易，團隊透過各式管道與潛在發音人取得聯繫，包括：

(1) 邀請有興趣者參加「公眾參與科學研究（Public Participation in Scientific Research, PRSR）」，由發音人上網填寫問卷，待團隊回覆並安排錄音。

(2) 與大型實體活動合作，像是由大學或研究機構主辦的活動，實際到場詢問與介紹。

(3) 2014 年至 2015 年間，為求大範圍蒐集各地的語料，放送全國性的廣告，內容除了招募訊息外，例如：針對特定群體的標題「媽媽說了算…亙古不變（“Mum’s the word…both then and now”）」，更將 Spoken BNC2014 的初步發現撰寫成新聞吸引更多人的參與，例如：十年間的用字變化標題「還在說 Cheerio Marvellous 嗎？你已經不夠 Awesome 了！（“Cheerio Marvellous… You’re No Longer Awesome”）」。

在錄音報酬上，每次錄音以一小時18英鎊計算，以錄音時間而非對話人數為準，主發音人須找到2至4位參與者、讓所有參與者填寫相關同意書（speaker consent forms）、提供完整後設資料、由主發音人負責錄製有品質、可供後續轉寫的語料，並於整個過程結束後的月底才獲得報酬。

Spoken BNC 2014 將採樣標準區分為兩類，分別是篩選標準（selection criteria）及描述性標準（descriptive criteria），前者包括發音人的性別、年齡、社經地位與所在區域（“gender, age, socio-economic status and region of the speakers”），後者包括談論主題及口語類型（“domain and type of speech”）。篩選標準是預先決定好的，也是蒐集語料時希望能夠取得平衡的比例設定，而描述性標準較隨機，須邊蒐集邊計算。另外，考量到口語語料不若書面蒐集速度快，如果有願意提供語料的發音人，通常不會停收，而是在蒐集語料的過程中，追蹤比例變化是否拉大、需要多收其他語料，而從 Spoken BNC1994 的發音人組成來看，男性、60 歲以上的年長者及 14 歲以下的幼童語料最為缺乏。

回顧 Spoken BNC2014 的建置過程，針對口語語料庫相對於書面語料庫的建置，有以下的優勢與困難：

(1) **蒐集與轉寫的成本懸殊：**Burnard (2002:6; in Love, Dembry, Hardie et al., 2017) 提到，欲增添一百萬字的口語語料，其蒐集與轉寫的成本比同等規模的書面語料多了至少十倍的成本。Love, Dembry, Hardie et al. (2017) 也表示，科技發展隨著時間推進，口語與書面語料的成本差異會越來越大，但人們日常溝通其實是以口語為主。

(2) **若不開源，使用者須負擔龐大費用：**以 BNC1994 的口語資料為

例，該資料開源、開放，可以將語料庫推廣並觸及更廣大的使用族群。不少語料庫在基於商業考量下無法讓外界取得資料，或必須支付一筆授權費用。

(3) **語料規模與時代代表性與現今語言的落差逐漸加大：**口語語料的規模必須夠大，才能夠代表語言，且須考量該語言使用的時代背景，BNC 1994 的資料已無法很全面地反映英式英語今日的樣貌，但許多使用者仍因為其開源的優勢而持續採用該語料庫語料。

(4) **可考慮捨棄「將語境作為控制變因（context-governed）」的採樣標準：**採樣標準主要有「基於地域範圍的採樣（demographically-sampled）」以及「將語境作為控制變因（context-governed/task-oriented part）(Leech et al., 2001:2) 」的採樣兩種，Spoken BNC2014 捨棄後者，該語料庫的建置者認為民間已有許多學者針對研究目的與問題建置了較小型且專門的語料庫，但較少有一般語境的大量語料可供檢索與利用。

(5) **口語語料必須經過更多去識別化的處理：**由於發音人的個人資料與錄音內容中的敏感內容須去識別化，卻又須顧及發音人的背景資料是否完整一致（consistent）、有用（helpful），建置過程中須對後設資料的分類和定義說明清楚。此外，隨著《歐盟一般資料保護規範》（General Data Protection Regulation, GDPR）正式上路，若有多位發音人參與對話錄音，可請每位發音人透過自己的行動裝置填寫、傳送個人資料，而非由其中一位發音人代填所有的資料，此填寫方式與 BNC 1994 不同。

綜合上述的原因，國家語料庫能夠發揮統合的效益，匯聚各方的語料，以其團隊力量進行語料的轉寫、標記，並在蒐集語料的過程中解決著作權的問題，盡可能將語料庫開源，讓使用者毋須負擔費用即可檢索或下載語料。然而，由於國家語料庫的規模龐大，在處理發音人、語料中的敏感資訊方面，需要更為審慎，這也是 Spoken BNC2014 尚未開放音檔下載的原因，因為音檔的匿名處理尚未完成。

1.2. 美國國家語料庫—American National Corpus (ANC)

「美國國家語料庫」（American National Corpus；ANC）於 1998 年建立，第二版於 2005 年發行。由於美式英語與英式英語有許多明顯的差異，美式英語的研究不適合使用英國國家語料庫（BNC）之語料，因此促成美國國家語料庫（ANC）的誕生，ANC 的目標為建立如 BNC 同樣龐大但專注在美式英語之語料庫。第二版 ANC 具有的語料約有 2200 萬個詞，與其他英語的語料庫相比這個規模不算大，例如，現代美式英語語料庫(COCA)的總詞數已多達 5 億 6 千萬詞。ANC 網址為：http://www.anc.org/。

ANC 採用書面語料（1853 萬詞，佔全部語料的 82.7%）及口語語料（386 萬詞，佔全部語料的 17.3%）。書面語料的來源包含許多不同的文類，含刊物、報紙、部落格、查閱技術資料等。口語語料包含電話錄音、面對面會話、以及學術場合的用語等。ANC 另外與 BNC 不一樣的是，ANC 只收集 1990 年之後的語料，因此能夠納入許多線上語料如郵件、網頁、以及聊天室之語料，BNC 不包含這些較新的語料。

雖然 ANC 的語料庫規模不大，但是該語料具有較多標記，不但附有不同詞性標記方法（含 Penn、CLAWS C5/C7、Biber），並且有原

形詞的訊息（lemma）、名詞組標示（noun chunk）、動詞組標示（verb chunk）、以及其它種類的標記。

ANC 是個協作開發計畫（Collaborative Development Project）類型的語料庫。該語料庫的語料仰賴語言學學者和一般民眾等主動提供或進行加註整理，若學者或民眾想要提供語料或針對語料進行編註的話，可以遵從網站上關於貢獻者身分、語料年代和類型、資料格式、著作權等等指示，對語料庫提供貢獻。這方面非常值得我們借鏡。

ANC 目前收錄了 1990 年以來各種體裁的書面和口語轉寫語料，而且網站上所有的語料和註釋都是完全對外開放的，任何使用都不受限制。語料庫又可分為 OANC 和 MASC 兩個子語料庫。

(1) OANC：包括總數達 1,500 萬個英語詞的當代美語語料，並針對某些語言現象採取自動標記，包括文章段落、節、句子、詞、名詞組、動詞組、人名、地名、組織名、及時間的自動標記。

(2) MASC：平均收錄了 19 種體裁的語料(包括法庭筆錄、辯論轉寫檔、電子郵件、文章、小說、政府文件、刊物、信件、報紙、非虛構作品、口語資料、技術類、旅遊指南、推特、部落格、Ficlets、電影劇本、垃圾郵件、笑話)，總共約有 500,000 個詞，另外這些資料的注釋都是經過人工添加或驗證的。所有 MASC 註釋，都被轉換為 ISO TC37 SC4 語言註釋框架。

(3) 資料來源：OANC 和 MASC 是協作開發資源（Collaborative Development Project），主要依賴語言學家或社會大眾提供 1990 年以後的各種類型的書面、口語轉寫資料，或是語料庫等，並且鼓勵資料提供者完全開放資料使用權限，使大家都能夠使用資料。此外，該網站也鼓勵使用者能主動為 OANC 和 MASC 添

加註釋，以利更多人使用。關於協作開發資源的進一步介紹，可以參閱「3.2 群眾外包—美國」一節。

1.3. 日本国立国語研究所（こくりつこくごけんきゅうしょ；National Institute for Japanese Language and Linguistics, NINJAL）

1.3.1. NINJAL 與 KOTONOHA 計畫

日本國立國語研究所（日語：国立国語研究所/こくりつこくごけんきゅうしょ）（英語：National Institute for Japanese Language and Linguistics，簡稱 NINJAL）是隸屬於日本「大學共同利用機關法人」（大学共同利用機関法人）的日語研究機關，旨在研究、調查、推廣日語，並且發布正確的日語用法。該機關成立於 1948 年 12 月 20 日，現址位於日本東京的立川市。

日本國立國語研究所的官方網站網址為 https://www.ninjal.ac.jp/。網站內收錄了包括各種語料庫、線上字典、語言地圖、貴重古書的掃描圖檔、論文掃描圖檔、語言分析工具、圖書和研究資料的資料庫、還有機構公開的語言調查等資料。是目前各國家語言資料庫中設計原則最值得我們參考的。

其中，日本國立國語研究所對方言研究（dialectology）的投入工作成果展示於其官方網站英文版的「資料庫（Databases）」之下，尤其是「研究主題 》日本方言與語言多樣性（Research Subjects > Japanese Dialects and Language Diversity）」以及「語言地圖（Linguistic Maps）」的部分，可參考網址：

https://www.ninjal.ac.jp/english/database/type/maps/ 以 及 https://www.ninjal.ac.jp/english/database/subject/diversity/。

日本國立國語研究所的主要配置為一位所長、兩位副所長，底下再分為研究部門、研究資源中心、語料庫開發中心、管理部門等四個部門。其中，研究部門底下可再依照領域細分成五個子部門，包括理論與類型學部門（Theory & Typology Division）、語言變異部門（Language Variation Division）、語言變遷部門（Language Change Division）、口語部門（Spoken Language Division）與日語教育部門（Japanese as a Second Language Research Division）。而管理部門底下則可再細分成總務、財務與研究推廣三個子部門。詳細的組織配置訊息可以參考該網站提供組織配置圖（英語版：https://www.ninjal.ac.jp/english/info/aboutus/organization-chart/img/organization-chart_en.png；日語版：https://www.ninjal.ac.jp/info/aboutus/organization-chart/img/organization-chart_jp.png）。

日本國立國語研究所所收錄的語料庫主要都是由語料庫開發中心主導整理、維護或開發。官方網站所提供的語料庫資源，除了語料庫開發中心所開發的語料庫外，其他有的是直接連結到別的現有語料庫的網站（如 A Glossed Audio Corpus of Ainu Forklore），也有部分是語料庫開發中心與研究部門共同合作的成果（如，Learner-Corpus Study of Acquisition of Japanese as a Second Language（http://lsaj.ninjal.ac.jp/），該語料庫的計畫主持人迫田久美子就是日語教育部門底下的研究員）。目前，語料庫開發中心正在進行一項名為 KOTONOHA 的計畫，該計畫內容為廣收從日本平安時代到現今的各種

日語書面和口語資料，並且將這些資料整理開發成各種類型的語料庫，英語版和日語版的計畫簡圖如下：

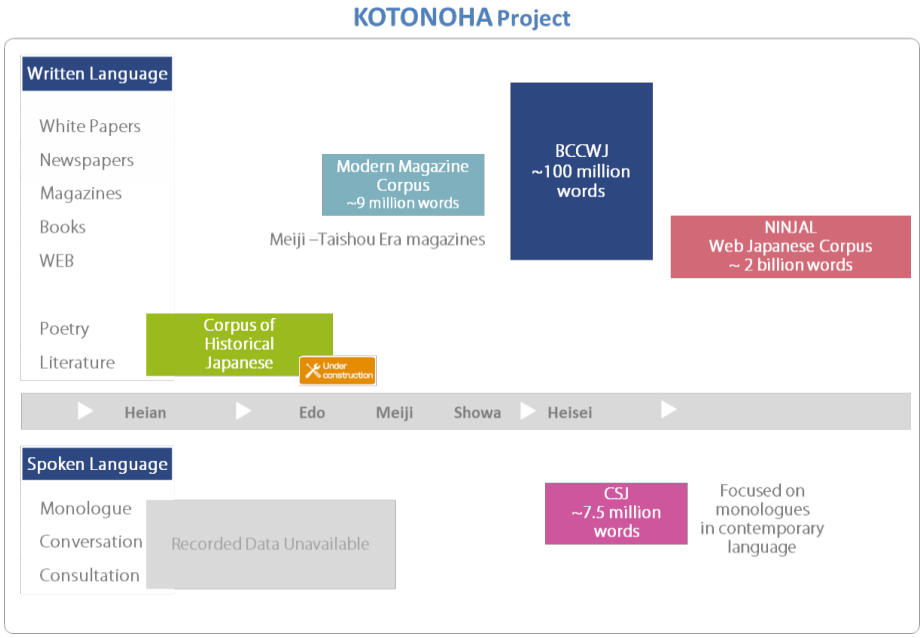


圖 1. KOTONOHA 計畫英語版和日語版簡圖之一（圖片來源：https://pj.ninjal.ac.jp/corpus_center/en/kotonoha.html）

圖 2. KOTONOHA 計畫英語版和日語版簡圖之二（圖片來源：https://pj.ninjal.ac.jp/corpus_center/en/kotonoha.html）

目前日本國立國語研究所的網站總共收錄了 14 個語料庫，包括：Balanced Corpus of Contemporary Written Japanese（https://pj.ninjal.ac.jp/corpus_center/bccwj/en/）、Shonagon（http://www.kotonoha.gr.jp/shonagon/）、NINJAL-LWP for BCCWJ (NLB)（http://nlb.ninjal.ac.jp/）等 3 個日本現代書面語平衡語料庫相關的語料庫；Corpus of Spontaneous Japanese (http://pj.ninjal.ac.jp/corpus_center/csj/en/）1 個自發性口語日語的語料庫；Corpus of Historical Japanese（http://pj.ninjal.ac.jp/corpus_center/chj/）1 個歷時語料庫；NINJAL Web Japanese Corpus（http://pj.ninjal.ac.jp/corpus_center/nwjc/）1 個定期收錄線上日語文本的語料庫；Learner-Corpus Study of Acquisition of Japanese as a Second Language（http://lsaj.ninjal.ac.jp/）1 個探討將日語作為第二語言的語料庫；Corpora of Modern Japanese（http://pj.ninjal.ac.jp/corpus_center/cmj/）1 個收錄了明治和大正時代日語的語料庫；Chunagon

（https://chunagon.ninjal.ac.jp/）1 個可對日本國立國語研究所開發的語料庫進行三向搜索的語料庫；A Glossed Audio Corpus of Ainu Folklore（http://ainucorpus.ninjal.ac.jp/en/）1 個收錄了阿伊努族民間故事的語料庫；Learners' L1-Japanese Contrastive Databases（https://db3.ninjal.ac.jp/contr-db/）1 個比較日語學習者的日語和其母語的語料庫；The NINJAL Parsed Corpus of Modern Japanese (NPCMJ)（http://npcmj.ninjal.ac.jp/?lang=en）1 個現代日語的語法、語義標註語料庫，裡面同時收錄有書面和口語語料；Nagoya University Conversation Corpus（https://mmsrv.ninjal.ac.jp/nucc/）1 個收錄了 129 個自然日語對話的語料庫； Oxford-NINJAL Corpus of Old Japanese（http://oncoj.ninjal.ac.jp/?lang=en）1個上古日語語料庫。整體來說，日本國立國語研究所的語言資料庫類型豐富，這和臺灣未來想成立的國家語言資料庫定位最相似；此外，不同於多數國家的國家語料庫，日本國立國語研究所還收錄有瀕危語言阿伊努語的相關資料。以下 1.3.2 小節將會針對日語的平衡語料庫--現代書面日語平衡語料庫(BCCWJ)--作進一步介紹；而 1.3.3 小節也會詳加介紹阿伊努語口傳文學語料庫（A Glossed Audio Corpus of Ainu Folklore）。

除了由語料庫開發中心主導整理、維護或開發的語料庫之外，日本國立國語研究所還有一系列的合作計畫，這些計畫有的是內部團隊的研究計畫，有的則是對外進行招標的計畫。這些計畫的主題類型包括：基於機構的研究計畫（Institute-based Project）、跨學科的合作研究計畫（Multidisciplinary Collaborative Projects）、網路型研究計畫（Network-based Projects）、語料庫基礎研究（Basic Research for Corpus Development）。基於機構的研究計畫（Institute-based Project）底下又可分為核心（Core Research Projects）、特定領域（Topic-

specific Projects）、新領域（New Frontier Projects）、共同利用（Joint Usage Projects）等四個子項。除了特定領域、新領域、共同利用這三個主題的計畫是對外向各大專院校進行招標外，其餘剩下的研究計畫皆是由國立國語研究所團隊所主持。詳細的計畫一覽可以參考以下網址介紹：https://www.ninjal.ac.jp/english/research/project-3/。

1.3.2. 現代書面日語平衡語料庫（現代日本語書き言葉均衡コーパス；Balanced Corpus of Contemporary Written Japanese, BCCWJ）

現代書面日語平衡語料庫（BCCWJ）為 2011 年公開的平衡語料庫。使用者可以選擇直接在線上檢索語料庫，或是向該機構購買 DVD 的版本。線上版又分成可直接使用的簡易版「少納言」，和需要註冊的完整版「中納言」兩種。該語料庫的成立目的是為了能夠呈現當代書面日語的多樣性，因此其語料取材都是盡量採用隨機抽樣的方式，語料來源為 1976～2008 年的各種書面資料，包括一般的出版品(如報紙、雜誌、書籍等)、政府出版品、及網路公開文章或留言等。截至 2012 年 3 月，語料庫規模已達一億五百萬詞。

在「少納言」版本中所收錄的語料總共有 11 個類別，每個類別的收錄時間略有不同，分別如下：書籍(1971~2005 年)、雜誌(2001~2005 年)、新聞(2001~2005 年)、白皮書(1976~2005 年)、教科書(2005~2007 年)、文宣(2008 年)、Yahoo 奇摩知識+(Yahoo!知恵袋)(2005 年)、雅虎部落格(Yahoo!ブログ)(2008 年)、韻文(1980~2005 年)、法律條文(1976~2005 年)、國會會議記錄(1976~2005 年)。各類別的字數與所佔比例如下表：

表 1.「少納言」各文類所佔比例

類別	字數	比例
書籍	6270,0000	59.7%
雜誌	440,0000	4.2%
新聞	140,0000	1.3%
白皮書	490,0000	4.7%
教科書	90,0000	0.9%
文宣	380,0000	3.6%
Yahoo 奇摩知識+	1030,0000	9.8%
雅虎部落格	1020,0000	9.7%
韻文	20,0000	0.2%
法律條文	110,0000	1.0%
國會會議紀錄	510,0000	4.9%

日本國家語料庫的語料標記方式，是把詞條(lexical item)分成兩種單位來分析標記：「短單位」和「長單位」。「短單位」是透過一個或多個「最小單位」所組成，而所謂的「最小單位」是指最小的、有意義的單位。如果用漢語分析來類比，「短單位」就相當於漢語的「詞」，「最小單位」則相當於「語素」。日語的「最小單位」可再根據來源或是性質來細分，如下表。不同類別的「最小單位」可能需要透過不同的規則才能構成一個「短單位」，例如和語(Native Japanese)與漢語(Sino-Japanese)的「最小單位」通常需要兩個組合在一起才能構成一個「短單位」；而外來語(Borrowing)的「最小單位」通常是自己一個就能構成「短單位」。因此在分析語料時，必須先了解每個「最小單位」是屬於哪個類別。在「短單位」的分析標記上，日本國立國語研究所是採用與日本千葉大學共同開發的 UniDic 系統，來自動分析語料。目前最新版的 UniDic 可透過以下網址來取得：https://unidic.ninjal.ac.jp/。「長單位」則是由「短單位」組合而成，有點類似短語(phrase)的概念。在分析句子時，主要是透過分析每個「長單位」所具有的語法意義來進行的。

表 2. 日語「最小單位」分類表

分類	例子
一般	和語：豊か、大、雨... 漢語：国、語、研、究、所... 外來語：コール、センター、オレンジ...

數字		一、二、十、百、千...
其他	詞綴	前綴：相、御、各... 後綴： 兼ねる、がたい、的...
	助詞 / 助動詞	う、だ、ます、か、から、て、の...
	人名/地名	星野、仙一、大阪、六甲...
	記號	Ａ、Ｂ、ω、イ、ロ、ア、JR...

日本國家語料庫採用 XML 格式。此外，為了設計出能夠客觀呈現當代書面日語多樣性的語料庫，日本國立國語研究所在收錄語料時，採用了一套很特殊的隨機抽樣方法，這套方法有兩種模式，分別為「等長樣本」和「不定長度樣本」。「等長樣本」所抽取的語料樣本比較短，抽取方法為先隨機選取一本書(或一份報紙、一份雜誌等)的某一頁，然後把該頁面的長與寬分成 9 格，如此會得到 81 個格子與 100 個交叉點，

接著，再隨機選取 100 個交叉點的其中一個，找到後，算出距離該交差點最近的字(廣告、圖表、照片、插圖的字不列入計算)，然後再把那個字週遭不包括標點符號的 1000 個字選取起來，如此就可以得到「等長樣本」。不過，因為「等長樣本」的開頭和結尾有時候會坐落

在句子的中間，所以為了句子的完整性，通常會再多收錄幾個字，把完整的內容都納進來。

圖 3.「等長樣本」抽取示意圖（圖片來源：https://pj.ninjal.ac.jp/corpus_center/bccwj/images/sampling/sashie3.png）

「不定長度樣本」的抽樣方式則相對簡單，只要隨機在書籍、報紙或雜誌等等中抽取一個章節或一個段落即可。不過，不同語料的章節或段落字數有時候可能相差很大，比如說 A 語料的某章節只有幾百字，但 B 語料的卻有上萬字，為了避免落差太大造成分析上的偏差，日本國家語料庫特別限制「不定長度樣本」以一萬字為上限。

1.3.3. 阿伊努(愛奴)語口傳文學語料庫（アイヌ語口承文芸コーパス；A Glossed Audio Corpus of Ainu Folklore）

阿伊努語口傳文學語料庫主要是由知名阿伊努語研究學者中川裕、日本國立國語研究所研究員 Anna Bugaeva，還有兩位日本國立國語研究所兼職研究員小林美紀、吉川佳見所建置完成，語料庫網站分為日文版與英文版，網址為：https://ainucorpus.ninjal.ac.jp/。該語料庫所收錄的語料來自兩位阿伊努耆老的口述故事，兩位耆老分別為 Kimi Kimura（木村きみ）女士還有 Ito Oda（小田イト）女士，口述故事的音檔分別是在 1977 年至 1983 年，還有 1999 至 2000 年所錄製完成。目前該語料庫總共包含 30 個音檔，其中 23 個檔案為散文體民間故事（uepeker，即用接近口語的方式來講述的故事），另外 7 個檔案則為英雄敘事詩（kamuy yukar，描述神話或英雄冒險題材的敘事詩），音檔時長總計約為 8 個小時。另外，為了能夠安全且永久典藏珍貴的阿伊努語相關資料，研究員 Anna Bugaeva 在 2007 到 2009 年時曾經將部分音檔交予倫敦大學亞非學院的瀕危語言典藏機構（Endangered Language Archive of SOAS, University of London）來保管，這些音檔主要為 Kimi Kimura（木村きみ）女士的口述故事，包括 20 個散文體民間故事（uepeker）與 3 個英雄敘事詩（kamuy yukar），共計 23 個音檔，音檔時長約為 7 小時、字數約有 44717 字，存放連結為：https://elar.soas.ac.uk/Collection/MPI124799。在阿伊努語口傳文學語料庫中每個音檔文本被切割成以句為單位，並標上編號，每句語料的轉寫格式如下：

(1) 第一行為用日文片假名所轉寫的阿伊努語文本。

(2) 第二行為用羅馬拼音所轉寫的阿伊努語文本。

(3) 第三行也是用羅馬拼音所轉寫的阿伊努語文本，不過會再依照單詞結構作切割。

(4) 第四行為以詞素為單位的英文版釋義（English glosses）。

(5) 第五行為以詞素為單位的日文版釋義（Japanese glosses）。

(6) 第六行為日文全句翻譯。

(7) 第七行為英文全句翻譯。

語料庫的截圖如下，紅色框框處為上面所述的七行轉寫格式：

圖 4. 阿伊努語口傳文學語料庫頁面截圖

1.4. 中國國家漢語語料庫—语料库在线

中國國家語料庫是由中國大陸教育部的語言文字應用研究所所建置的，裡面又分成現代漢語的「國家語委現代漢語通用平衡語料庫」(以下簡稱現代漢語語料庫)，與古代漢語的「古籍語料庫」(以下簡稱古代漢語語料庫)兩個子語料庫。現代漢語語料庫的部分，其成立宗旨是為了能夠真實反映現代漢語的全貌，因此裡面收錄了 1919 年以後的

各類漢語語料，但大多數的語料還是以 1977 年以後的語料為主。之所以著重採用 1977 年以後的語料是因為，中國在 1910 年代曾發起白話文運動，而其影響是漸進式的，從 1910 年代到今日的 21 世紀，各時期使用的「現代漢語」可能還是有很多不一樣的地方，因此為了能客觀呈現今日的現代漢語，才會著重採用 1977 年以後的語料。目前現代漢語語料庫的規模已超過一億字符(包括漢字、字母、數字、標點等)，算是現代漢語頗具代表性的語料庫。網址為：http://corpus.zhonghuayuwen.org/。

現代漢語語料庫的語料以書面為主，口語為輔，語料大致上分為五大類別，即教材、人文與社會科學、自然科學、報紙及刊物、應用文(包括各類政府公文、書信、說明書、廣告等)。這五大類別的語料取材時間不盡相同，例如教材和自然科學以選取共時性的資料為主；人文與社會科學則是按照現代漢語脫離文言文的程度，分成五個時期來取材，分別是 1919~1925 年(5%)、1926~1949 年(15%)、1950~1965 年(25%)、1966~1976 年(5%)、1977 年之後(50%)。這五大類別所佔比例如表 3，不過，該表格的字符數據是語料庫剛成立時的數據，這是因為網站上目前並沒有提供之後加入新語料的更新數據。如下表是中國現代漢語語料庫標計的五大類別。

表 3. 中國現代漢語語料庫各大類別比例

(大)類別	字符數(約略值)	比例
教材	2000,0000	28.57%

人文與社會科學	3000,0000	42.86%
自然科學	300,0000	4.28%
報紙及刊物	1300,0000	18.57%
應用文	400,0000	5.71%

在標記方面，現代漢語語料庫把詞類分成 13 個一級類，16 個二級類；切分單位也被分成其他 7 個一級類，13 個二級類。完整的標記如表 4，語料標記圖例如圖 5。

表 4. 中國現代漢語語料庫標計類別

標記代碼		**類別名稱**
一級類	**二級類**	
a		形容詞
	aq	性質形容詞
	as	狀態形容詞

c		連詞
d		副詞
e		嘆詞
f		區別詞
g		語素詞
	ga	形容詞性語素詞
	gn	名詞性語素詞
	gv	動詞性語素詞
h		前接成分
i		慣用語
	ia	形容詞性慣用語
	ic	連詞性慣用語
	in	名詞性慣用語

	iv	動詞性慣用語
j		縮略語
	ja	形容詞性縮略語
	jn	名詞性縮略語
	jv	動詞性縮略語
k		後接成分
m		數詞
n		名詞
	nd	方位名詞
	ng	普通名詞
	nh	人名
	ni	機構名
	nl	處所名詞

	nn	族民
	ns	地名
	nt	時間名詞
	nz	其他專有名詞
o		擬聲詞
p		介詞
q		量詞
r		代詞
u		助詞
v		動詞
	vd	趨向動詞
	vi	不及物動詞
	vl	聯繫動詞

	vt	及物動詞
	vu	能願動詞
w		其他
	wp	標點符號
	ws	非漢字字符串
	wu	其他未知符號
x		非語素詞

样本编号：BF29701101
样本名称：鸟的世界
类别：文学·散文
作者：杨栋
出版时间：1997-12-11
书刊名称：人民日报

鸟/n 的/u 世界/n
杨栋/nh
鸟/n ,/w 是/vl 大自然/n 的/u 歌手/n ,/w 鸟语/n 就是/vl 大自然/n 的/u 音乐/n 和/c 诗歌/n 了/u 。/w
山村/n 里/nd 的/u 鸟/n 除了/p 麻雀/n ，/w 就/d 数/v 燕子/n 多/a 了/u 。/w 村/n 人/n 对/p 燕子/n 很/d 爱护/v ，/w 说/v 它/r 吃/v 庄稼/n 的/u 害虫/n ，/w 常/a 吓唬/v 孩子/n 们/k 不要/vu 去/v 玩/v 燕子/n ，/w 会/vu 坏/v 自己/r 的/u 眼睛/n 。/w 有时/r 光/v 屁股/n 的/u 小/a 燕/n 掉/v 下来/vd ，/w 也/d 要/vu 送回/v 燕/n 窝/n 里/nd 去/v 。/w

圖 5. 中國現代漢語語料庫語料標記圖例

在技術層面，語料數據庫採用 Access 數據庫格式(.MDB)，語料文本則採用(.TXT)格式。除了一般查詢，網站也提供 3 種語料分析處理的查詢功能，分別為分詞和詞性標註、漢語拼音自動標註、字詞頻率統計。

1.5. 歐陸地區的國家型語料庫

1.5.1. 保加利亞國家語料庫（Bulgarian National Corpus）

保加利亞國家語料庫成立於 2009 年，為一共時語料庫；除了單語的保加利亞語料庫部份外，另外還設置了 47 個平行語料庫，包括英語，德語，法語，大多數斯拉夫語和巴爾幹語，以及許多其他歐洲和非歐洲的語言。目前保加利亞語部分收錄了約 12 億單詞，語料收錄了自 1945 年以來的各種書面資料（97.35%），還有演講、議事程序、字幕的口語資料（2.65%）。大部分的語料都是電腦自動或人工手動從網路上下載下來的（97.5%），而剩下的 2.5%則是由作者或出版商提供。而在平行語料庫的部分，語料部分僅保留具有和保加利亞語相互對應的原文和翻譯文本。截至目前（2013 年 1 月底），平行語料庫的總規模已達 42 億單詞，其中以保加利亞語/英語規模最大，約有 2.6 億詞，而最小的保加利亞語/日語語料庫，則約有 5 萬詞。另外在著作權部分，該語料庫完全遵守保加利亞和歐盟有關著作權及相關權的法律；在一般情況下，該語料庫的資料只能用於非商業科學研究，教育或私人學習等情況，除非資料提供者有其他特定要求，不然一般的語料都會提供語料來源和作者等信息，供大眾據此引用。另外因為部份資料受著作權保護，因此使用者們並無法完全下載保加利亞語料庫的資料。語料庫網址為：https://dcl.bas.bg/bulnc/en/。

1.5.2. 捷克國家語料庫（Czech National Corpus）

該語料庫為捷克國家語料庫研究所和各方合作並建置的，包括研究人員和學生、270 個出版商、國家和國際研究項目合作等；其中，研究人員和學生主要進行口語資料的收集，出版商則提供書面資料。語料庫總共包含五大部分，分別為書面、口語、平行語料庫、歷時語料庫以及專門的語料庫。書面語料庫收錄了 20 和 21 世紀（尤其是最近 20 年）的資料，目前規模超過 22 億個單詞；書面語料庫每 5 年會更新出版一次由小說、專業文學和報紙組成的平衡語料庫，另外也包括僅由新聞語料組成的大型語料庫。口語語料庫僅收錄自發性、非正式、對話式的語料，語料轉寫也致力保存與音檔一致，目前規模約為 480 萬詞，為世界上數一數二大的自發語音資料庫。平行語料庫共有 30 多種語言的對應，資料來源包括手動對齊和校對的小說文本，以及來自各個領域的自動處理文本，並且有註釋，目前規模近 10 億詞。歷時語料庫收錄包括 14 世紀以後的書面資料，其長期目標是創建一個涵蓋 1850 年至今的大型監控語料庫（monitor corpus）用來比較並研究語言的變遷。專門的語料庫的語料是針對特定研究目的來進行收集的，例如 DIALEKT（方言），CzeSL（由捷克籍非母語學習者撰寫的文字），DEAF（由聾人撰寫的捷克文字），或 Jerome（已翻譯和未翻譯的捷克文）。語料庫網址為：https://ucnk.ff.cuni.cz/cs/。

1.5.3. 希臘國家語料庫（Hellenic National Corpus）

希臘國家語料庫(The Hellenic National Corpus)是由希臘教育與宗教事務部(Ministry of Education and Religious Affairs)的語言與語音處理研究所(The Institute for Language and Speech Processing, ILSP)所設立，於 2000 年 10 月正式啟用。語料庫中收錄了 1976 年至現今的書面語料，

語料選取以使用標準希臘語、而且流覽量高的資料為主。目前語料庫的規模已超過五千一百萬詞，並且持續增加中，為現代標準希臘語最具代表性的語料庫。網址為：http://hnc.ilsp.gr/。

希臘國家語料庫所收錄的書面語料，可根據資料來源再分成五大類別：書籍、網路資源、報紙、雜誌、其他，各類別資料數量語所佔比例如表 5。

表 5. 希臘國家語料庫各類語料所佔比例

語料來源	資料筆數	百分比
書籍	252	0.45%
網站/網路	5485	9.77%
報紙	46649	83.06%
雜誌	2127	3.79%
其他	1647	2.93%
總共	56160	100%

希臘國家語料庫是一個監控型的語料庫(monitor corpus)，每天都會不斷更新，而且舊的語料也不會遭到刪除，語料庫採用的格式為 application/octet-stream。語料庫中的每筆資料都有兩種註記，即詮釋資料註記，和參考了 PAROLE Corpus Encoding Standard (PAROLE : 1995) 計畫格式的內文註記。PAROLE Corpus Encoding Standard (PAROLE : 1995)是歐洲的一項語言計畫，目標是希望能統合 12 種歐洲語言的語言

資源，該計畫採用了 TEI (Sperberg-McQueen and Burnard, 1994)和 EAGLES guidelines 的格式來做註記。語料庫以書面文字檔為主，從查詢與結果頁面來看，並沒有額外採用其他多媒體或影片等。

此外，希臘國家語料庫提供三種查詢方式，分別為單詞(word)、詞目(lemma)、詞類(part of speech)查詢。除了在這三者中擇一單獨查詢之外，也可以將之隨意疊加組合來做查詢，如單詞+詞類、單詞+詞目+詞目等，最多可以疊加到三層，如圖 6。

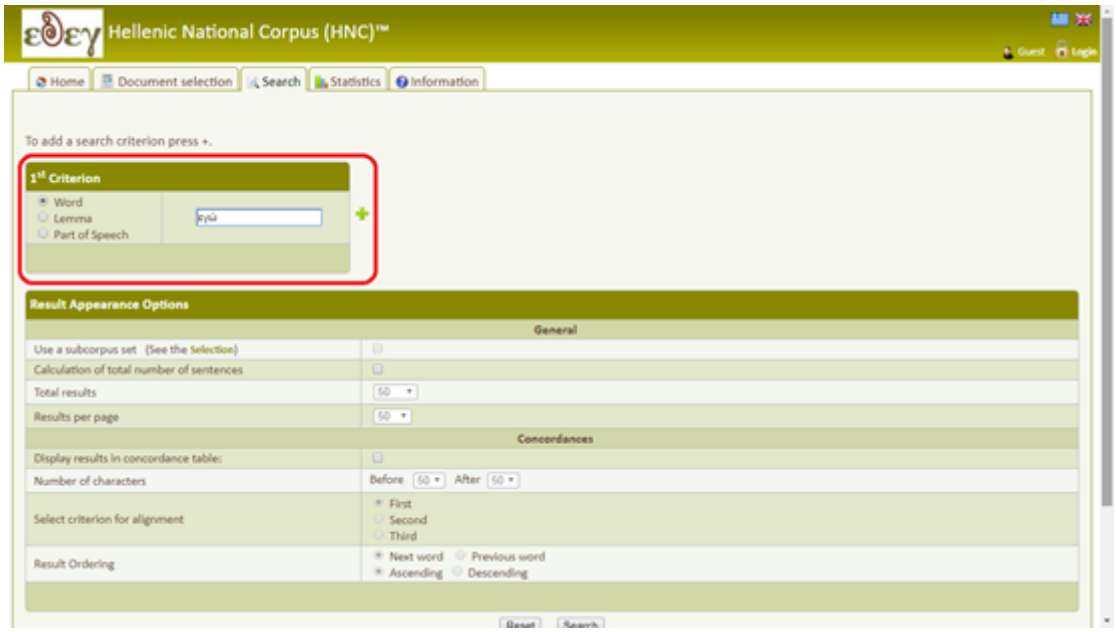


圖 6. 希臘國家語料庫查詢頁面之一

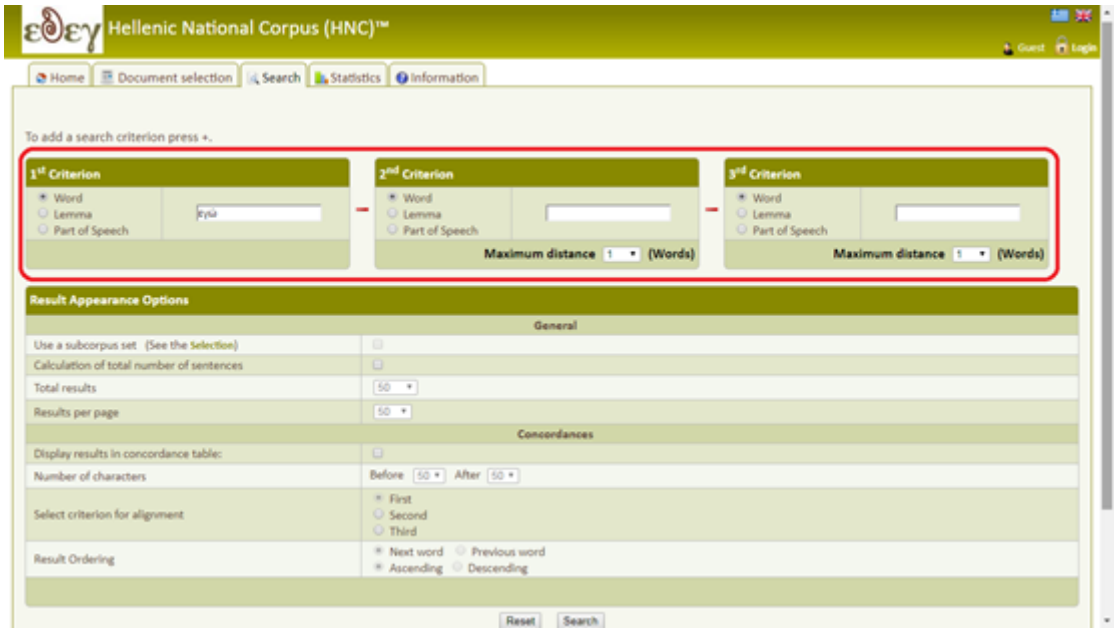


圖 7. 希臘國家語料庫查詢頁面之二

1.5.4. 匈牙利國家語料庫（Hungarian National Corpus）

匈牙利國家語料庫於 1998 年開始建置，對每個人開放，目的是創建一個規模約一億詞的平衡參考語料庫。不過從 2002 年開始，語料庫建置團隊開始將數據收集的範圍擴展到整個喀爾巴阡盆地所使用的匈牙利語，並於 2005 年 11 月時釋出。該語料庫具有詳細的句法註釋自動分析，具有 97.5%精準度，剩下的 2.5%則由人工修正。目前包含 1.876 億詞。按地區語言變體分為五個子集（匈牙利、斯洛伐克、亞喀爾巴阡山脈、特蘭西瓦尼亞、伏伊伏丁那），也按文本類型分為五個子集（新聞媒體、文學、科學、官方、個人資料）。在語料收集方面，匈牙利國家語言所主要負責匈牙利境內的語料收集與註釋，其他變體的語料則分別由 Gramma 語言辦公室（斯洛伐克）、Hodinka Antal 研究所（亞喀爾巴阡山脈）、SzabóT. Attila 語言學院（特蘭西瓦尼亞）、Vojvodina 匈牙利語言所（伏伊伏丁那）來負責收集與註釋。語料庫網址為：http://corpus.nytud.hu/mnsz/index_eng.html。

1.5.5. 當代威爾斯語國家語料庫（Corpws Cenedlaethol Cymraeg Cyfoes, CorCenCC; National Corpus of Contemporary Welsh）

威爾斯語雖為英國境內使用人口第二多的語言（Office for National Statistics, 2011），但相較於英國國家語料庫，還有其他歐洲語言的語料庫，威爾斯語的語料庫相關資源與發展都明顯落後許多；現有的一些威爾斯語語料庫要不取材單一（只有書面或只有口語），要不規模很小，要不並非平衡語料庫等（Knight, Loizides, Neale, et al., 2020）。為了解決這些問題，威爾斯國家語料庫利用開發 app 的群眾外包方式來廣收語料（詳見 3.5 章節介紹）。以下將簡介威爾斯國家語料庫的運作模式，其語料庫開發機制對於本國正面臨傳承危機，而且資源又相對不足的國家語言來說，極具參考價值。

威爾斯國家語料庫是一個社區驅動（community driven）的計畫，來自各種背景，各種程度的威爾斯語使用者皆可參與該計畫。這個計畫主要是收集威爾斯語使用者的語言使用，計畫約從 2016/3/1 開始，預計花 3.5 年，內容包括 4,000,000 詞的口語語料，4,000,000 詞的書面語料，以及 2,000,000 詞的 E-language；E-language 像是包括 email、網站等等的語言使用。另外，該計畫也會收集各國語言的例子，像是從書籍、新聞、廣播、電視台節目等收集。該計畫最大的特色是，威爾斯語使用者可依照個人意願自行下載計畫的 app（含安卓和 ios 版本），填寫一些個人資料（如性別、地點等），記錄自己的日常的威爾斯語使用（如自己錄音對話、自己的 email 文字檔、或是書面和影音等語言資料），然後把語言資料傳給計畫來幫助建置威爾斯語語料庫，或者針對使用者的語言使用地點來繪製語言地圖等。個人資料不會被隨意公開，語料部份則會進行匿名處理。語料庫網址為：

http://www.corcencc.org/。有關語料群眾外包的部分爾斯語語料庫非常值得借鏡。

1.6. 其他國家型語料庫

1.6.1. 俄羅斯國家語料庫（Russian National Corpus）

俄羅斯國家語料庫是由俄羅斯科學院(Russian Academy of Sciences)的俄語研究所(Institute of Russian language)所設置，於 2004 年 4 月 29 日啟用。語料庫裡面包含 18 世紀中到 21 世紀初的俄語語料，目前的規模至少超過 3 億個詞彙，為俄語最具代表性的語料庫。網址如下：http://ruscorpora.ru/en/。

俄羅斯國家語料庫由一個主語料庫和幾個子語料庫所構成。主語料庫收錄標準俄語語料，又可分為三個部分，包括從 1950 年代至現今的現代書面文本、真實口語語料，還有從 18 世紀中到 21 世紀的早期文本。子語料庫則包括，收錄了句子完整構詞和句法註記的「深度註解語料庫」(The Deeply Annotated corpus)、提供俄語語料與英語、德語、烏克蘭語、白俄羅斯語等語言的相互對照翻譯的「平行語料庫」(The Parallel Corpora)、收錄了俄羅斯各地方言的口語語料音檔的「方言語料庫」(The Dialectal corpus)、還提供詩詞韻律和押韻等查詢的「詩詞語料庫」(The Poetry corpus)、語料採用統一規範的同音異義詞，因此可做為學校教學參考的「教育語料庫」(The Educational corpus) 、包括公開且自發性的俄語口語語料錄音檔，還有 1930 到 2007 年間的俄語電影轉寫資料的「俄語口語語料庫」(The Corpus of Spoken Russian) 。因為語料庫收錄的語料種類繁多，所以在計算時大概可以將語料分成三大類別：內容為虛構故事的「Fiction」，Fiction 以外的其他書面資料

「Non-fiction」，還有口語語料「Oral presentation」。表 6 為網站所附的各類別語料所佔比例，可以看到九成以上都是書面語料，口語語料只佔了 3.9%。

表 6. 俄羅斯國家語料庫各類語料所佔比例

文類 **Text type**	**文本數** **Number of texts**	**詞數** **Number of tokens**	**詞佔比** **Percentage of tokens**
虛構類 Fiction	3893	5854,7176	39.7%
非虛構類 Non-fiction	37249	8321,8964	56.4%
口語 Oral presentation	1245	581,0482	3.9%

俄羅斯國家語料庫目前採用的格式是參考 the XMLized TEI scheme 和 the EAGLES guidelines。在主語料庫的部分，每筆語料都會有詮釋資料和構詞的註記(meta tagging and morphological tagging)，而且構詞註記幾乎都由電腦自動分析所完成，只有碰到同音異義詞(Homonym)的時候才會採取人工分析來解決歧異。目前整個語料庫約有 500 萬詞已完成人工分析，未來還會持續增加。構詞註記的資料除了可以應用到現代俄羅斯語構詞學研究之外，還可以作為構詞學的分析和自動處理

所使用到的，搜尋演算法與程式的測試平台。此外，如果使用者單純只想使用已經經過人工分析的語料的話，可以參考「深度註解語料庫」(The Deeply Annotated corpus)這個子語料庫。該子語料庫僅包含已完成人工分析的語料，而且語料庫的每個句子都附有依存關係樹（dependency trees）的分析註解，句子結構樹的每個結點對應到句中各個單詞，旁邊會標上其語法關係（syntax relationships）。

俄羅斯國家語料庫的構詞註記格式主要是參考 Zalizniak (1977; 4th ed., 2003)的 Grammatical dictionary of the Russian Language 一書。語料庫的每個詞形（wordform）都會被標上四種構詞的相關資訊，包括詞位（Lexeme，含其原形、詞性）、該詞位的各種 word-classifying 語法特徵（如，名詞的性別、動詞的及物性）、該詞位的各種 word-altering 語法特徵（如，名詞的格位、動詞的格式要符合主詞的數量）、該詞位的其他非正式形式或者拼寫變體等。詳細的標記內容可以參考以下網址：http://www.ruscorpora.ru/old/en/corpora-morph.html。

在語義方面，俄羅斯國家語料庫是利用 Semmarkup program (by A. E. Poliakov)，讓電腦自動分析語料庫裡的語義詞典（Semantic dictionary），來完成語義註解的。語料庫裡的語義詞典是以，Zalizniak 的 Grammatical dictionary of Russian 一書，所製成的 DIALING system 構詞學詞典作為基礎。不過，目前語料庫上的同音異義詞(Homonym)還沒有經過人工分析排除語義的歧異，系統只會在這些詞上標出多種可能的語義分析。語義標註的格式則是參考 E. V. Paducheva and E. V. Rakhilina 在 1992 年時，為了建置 Lexicograph 資料庫而設計的分類系統，然後再依俄羅斯國家語料庫的需要作增修。語料庫所採用的語義標註大概可以分成六大類，包括 Taxonomy（即語義角色，主要應用在名詞、動詞、形容詞、副詞上）、Mereology（如 part – whole、element

– aggregate 的關係，主要應用在具體和抽象名詞上）、Topology（主要應用在具體名稱（concrete names）上）、Causation（主要應用在動詞上）、Auxiliary status（主要應用在動詞上）、Evaluation（主要應用在具體和抽象名詞、形容詞、副詞上）。詳細的標記內容可參考以下網址：http://www.ruscorpora.ru/old/en/corpora-sem.html。

除了詮釋資料註記、構詞註記和語義註記(semantic annotation)，俄羅斯國家語料庫還有重音註記(accentual annotation)，未來也會再加上語法註記(syntactic annotation)。

1.6.2. 澳洲國家語料庫（Australian National Corpus）

澳洲國家語料庫（Australian National Corpus）是澳洲國家資料服務計畫（Australian National Data Services, ANDS）的一部分，該計畫由澳洲政府資助與推動，將多方來源的語料蒐集，以提供檢索，目前已建置完成，且由非營利組織澳洲國家語料庫小組（Australian National Corpus Incorporated）擁有與維護，該國家語料庫網址為：http://www.ausnc.org.au/。

該語料庫並非從無到有地創建起來，而是集結各個來源的語料，並訂定一致的原則與技術規範，這一點非常值得參考，例如：後設資料的項目、標記原則與形式等，讓資料更有系統地保存。該語料庫的資源已很豐富，資料形式包括文字、轉寫文字（transcription）、音檔或影音檔等，口語及書面語料庫皆有，亦涵納文學作品、19 世紀澳洲英文等，惟語種以英文為主。雖然原住民語言的語料目前看來還未成為焦點，但因為其為政府推動之計畫，目前正逐漸擴增中，這樣的建置架構與內容值得擁有多種國家語言的臺灣來學習。

從 2012 年起，澳洲國家語料庫挑選了 6 至 10 個現有的、已被作為研究資源的語料庫，進行語料收錄 (Peters, 2009)，像是：

(1) 澳洲英語語料庫（Australian Corpus of English, ACE）：其語料庫組成方式與美國布朗語料庫相似。

(2) 奧茲早期英語語料庫（Corpus of Oz Early English, COOEE）：時間橫跨澳洲被殖民前期至 19 世紀末期，雖然 COOEE 語料庫的標記訊息可能不夠完整，但為一珍貴收藏，且其年代資訊亦是重要後設資料之一。

(3) 澳洲文學資料庫（Australian Literature Resource, AusLit）：如果全部收錄進國家語料庫的話，將有約莫 75 萬個文本，但由於著作權因素使得最終收錄文本數目不及此數字。

(4) 澳洲國際英語語料庫（Australian Component of the International Corpus of English, ICE）：擁有豐富的口語語料，並經過轉寫與標記，提供口語中的說者發言交替（speaker turns）、同時發言（overlaps）等資訊，但語音檔案亦因著作權問題無法公開。

(5) 蒙納許口語英語語料庫（Monash Corpus of Spoken English）：其資料較 ICE 少，但其採用 Lerner (2004) 的言談標記原則。

(6) 格理菲斯澳洲英語口語語料庫（Griffith Corpus of Australian Spoken English）：其資料較 ICE 少，但其採用 Lerner (2004) 的言談標記原則。

(7) Mitchell 與 Delbridge 語料庫（Mitchell and Delbridge Corpus）：為 1960 年代蒐集的語音資料，特別的是該語料庫已經將音檔切分成以單詞為單位，適合作為語音變遷的研究材料。

(8) Braided 頻道（Braided Channels）：其語料受訪對象是澳洲女性，為一部長達 70 小時的紀錄片，附有影片檔、逐字稿以及一

些照片與音樂資料，其中逐字稿的文字標有說話者，但沒有其他的標記。

從上述(1)到(8)的語料庫內容，可以看到搜羅現有語料庫的過程中可以期待的優勢，以及可能會遇到的難處。由於每個語料庫不可能使用相同的規格建置，會有部分資料缺失，但也有部分資料是該語料庫所特有的，值得被保留，不應在尋求一致的規範中被省去，例如：

(1) 後設資訊：較早期的語料若有已知的年份資訊，應被保留。

(2) 口語語料已採用國際通用的標記方法時，可省下重新整理標記的人力，進而邁向確認標記正確與否的階段。

(3) 語料若已經過細緻的處理，尤其是比起整合語料庫的規格更細的時候，宜將該處理資料保留，甚至在後設資料上特別提出該子語料庫的特色，像是已切分完成的音檔即為一例。

然而，在整合各語料庫的過程中，也有可能會遇到困難，例如：

(1) 由於著作權的關係，使得原語料庫能夠公開的部分資料變得不可公開，尤其常發生在口語語料庫的原始音檔取得時；著作權問題在 Lampert (2009) 提供澳洲電子郵件語料給澳洲國家語料庫時也有提到，因為電子信件會有收寄信者的個人資料、前幾封信件的引用文字、附件檔案等的顧慮，因此很難一一詢問各方取得授權。面對如此情況，解決辦法便是從政府方發起相關活動，並推廣至潛在的對象，像是此例中，澳洲國家語料庫與動力博物館（Powerhouse Museum）及 NineMSN 企業合作，在一定時間內徵求電子郵件語料，除了作為展覽的內容，也讓國家語料庫變得更加充實。

(2) 有些語料庫的規模較小，僅以 WORD 檔紀錄，如何將其轉為一致的格式，是需要耗費時間與人力的任務，但有了整合語料庫的構想與推動，亦讓小規模語料庫的建置者有更大的動力。

關於澳洲國家語料庫的統一規格，Cassidy et al. (2012) 提到，他們將蒐集到的語料先轉成純文字（plain text）的格式，僅保留文字與標點符號，刪去各種標記資訊，但同時發言（overlaps）的文字部分則不刪去；如果語料是影音檔，則將多媒體檔案與文字檔分開保存。至於後設資料的部分，則先訂定一套原則，如能與國際標準相符更理想，像是都柏林後設資料標準（Dublin Core）及開放語言典藏社群（Open Language Archives Community）的標準，以期最終存成 XML 檔案，以此方式將原始的資料格式進行處理。但是，即便已有國際標準能夠遵循，仍可能與政府資料庫採用的標準不同，澳洲國家資料服務計畫即是採用廣泛的、常使用於 XML 資料的「資料交換的蒐集與服務格式（Registry Interchange Format——Collections and Services, RIFCS）」，使得語言學界的共同原則可能需要取捨，或是調整成符合自身情況的格式規範。Cassidy (2013) 在跨語言的整合資料庫設計下，為了檢索的時候更方便、更精確，澳洲國家語料庫特別注重各個子語料庫的分類，尤其是各子語料庫的建置時間、語料代表的年代或年份、是否有地理資訊、是否有說話者的年齡或性別資訊等，並針對各語料庫的類型（genre）作細緻的區分，並非只是附上其為口語或書面語料庫的類型資訊，而是提供像是流行文學、新聞、廣播等的細緻類型資訊。

在標記資料的部分，則是遵循國際語言學標記框架標準（International Organization for Standardization - Linguistic Annotation Framework, ISO-LAF），每個單位都被轉成資源描述框架（Resource

Description Framework, RDF）的形式，且可以有多個標記，但必須與文字檔或影音檔區分開來。

除了語料庫的資料與後設資料本身，其使用者介面的設計以及內容管理系統（Content Management Systems, CMS）也是整合語料庫的關鍵之一，如此一來使用者才能進行跨語料的全文檢索（full text search），例如：使用者可以從多個語料庫中尋找「20 歲以下的女性為受訪者的口語轉寫文字」。然而，囿於經費不足的關係，澳洲國家語料庫在最初建置的時候未能以標記資訊（search based on annotation data）進行檢索。這些檢索的結果需要網頁介面的支持，甚至是能夠讓使用者下載部分的資料，進行離線檢索與語料處理（offline search and processing），但下載這些資料會使得資料不再受到嚴謹的規範，因此很多時候資料僅供網頁檢索，但至少已是公開資料。

1.6.3. 韓國國家語料庫

1.6.3.1.韓國語意網絡研究中心（Semantic Web Research Center, SWRC）

韓國語意網絡研究中心（Semantic Web Research Center）致力於開發韓文的自然語言處理（Natural Language Processing, NLP）工具，其前身為 1998 年創立的韓語術語學語言與知識工程研究中心（Korea Terminology Research Center for Language and Knowledge Engineering），網址為：http://semanticweb.kaist.ac.kr/home/index.php/Home。

該研究中心主力為與自然語言處理相關的知識本體（ontology）、機器翻譯（machine translation）與資訊檢索（information retrieval）等，先前已完成韓文的句法剖析器（synatactic parser），在資源方面則包含

各式自然語言處理所需的語料庫，例如：代名詞字典、複合詞字典、複合名詞論元結構字典、可離線使用的韓文字母圖像庫與手寫韓文圖像庫、單一音節名詞詞表、標有詞素與句法的語料庫、各領域的術語語料庫、新聞語料庫、韓中英日等的平行語料庫、語音語料庫等，是自然語言處理導向的研究中心，但也在研究過程中建置了各種資料庫或語料庫，充分顯現語言資料庫與自然語言處理的互動。

1.6.3.2.世宗語料庫（세종 말뭉치；Sejong Corpus）

韓國設有國立國語院（National Institute of the Korean Language），致力於推廣韓語以及韓文研究與相關資源的開發。1998年起，韓國展開了為期十年的「21世紀世宗計畫（21st Sejong Project，紀念世宗發明了韓文文字）」，其中的一項主要目標是建立韓國國家語料庫。韓國國家語料庫分成兩大類，一大類是一般性的語料庫，例如：附有語意或句法標記的語料庫，另一大類是特殊的語料庫，例如：口語語料庫、韓英平行語料庫、韓日平行語料庫、歷史語料庫、北韓與海外韓語語料庫（corpus of Korean used by the North and overseas Korean），並建立韓文電子辭典。(Kim, 2006)

韓國國語語料庫採用的是文本編碼規範（Text Encoding Initiatives, TEI），並在檔案開頭附上標題（header）、電子化與修正的歷史紀錄。截至2006年，一般性的語料庫共有8億字（89,830,015字），其中1千5百萬字（15,226,186字）附有句法標記，另有1千萬字（10,132,348字）附有語意標記。此外，為求語料庫的文類平衡，訂定各文類比例如表 7：

表 7. 韓國國家語料庫各文類比例 (Kim, 2006)

文類	**比例**
新聞（newspapers）	20%
雜誌（magazines）	10%
學術文章（academic works）	35%
文學作品（literary works）	20%
半口語語料（quasi-spoken data）	10%
其他	5%
總計	**100%**

此外，特殊語料庫則含有 2 千 3 百萬字（23,394,220 字）。(Kim, 2006) 但根據國教院的考察報告 (2016)，因為韓國族群組成的同質性較高，因此語料庫著重在韓文資料的蒐集，值得一提的是，該語料庫將韓語學習者納入對象，因此在詞典的編撰上，另編有學習者字典，並區分讀懂韓文新聞報紙的程度、以及一般日常生活對話所需的詞彙，前者共需約 5 萬個詞單詞量，而後者僅需 2 至 3 千個詞彙即可，網址為：https://www.korean.go.kr/front_eng/main.do。

1.6.4. 其它的歐亞國家語料庫

除了上述已介紹過的各國國家語料庫之外，另外還有不少國家設有、或正在建置國家語料庫，以下僅列出這些國家的語料庫與其網址供參考。

(1) 阿布哈茲國家語料庫(The Abkhaz National Corpus)
http://clarino.uib.no/abnc/page

(2) 阿爾巴尼亞國家語料庫(Albanian National Corpus)

(3) http://web-corpora.net/AlbanianCorpus/search/

(4) 克羅埃西亞國家語料庫(Croatian National Corpus)
https://web.archive.org/web/20060424031437/http://hnk.ffzg.hr/

(5) 喬治亞國家語料庫(Georgian National Corpus)
http://gnc.gov.ge/gnc/page

(6) 北奧塞提亞共和國國家語料庫 (Ossetic National Corpus)
http://corpus.ossetic-studies.org/search/index.php?interface_language=en

(7) 波蘭國家語料庫(National Corpus of Polish)
http://nkjp.pl/index.php?page=0&lang=1

(8) 新加坡國家口語語料庫(National Speech Corpus, NSC)：
https://www2.imda.gov.sg/programme-listing/digital-services-lab/national-speech-corpus

(9) 斯洛伐克國家語料庫(Slovak National Corpus)
https://korpus.sk/index_en.html

(10) 韃靼斯坦共和國國家語料庫(Tatar National Corpus)
http://tugantel.tatar/?lang=en

(11) 泰國國家語料庫(Thai National Corpus)
http://www.arts.chula.ac.th/ling/tnc/

(12) 土耳其國家語料庫(Turkish National Corpus (TNC))

https://www.tnc.org.tr/

最後，以下 1.6.5 和 1.6.6 兩個小節將介紹蘇格蘭語和愛爾蘭語語料庫。雖沒有國家語料庫之名，但與威爾斯語語料庫一樣都屬於英語系國家中的地方語言，這些語言的處境與臺灣大多數的國家語言一樣受到強勢官方語言的影響漸漸式微，當地政府和學術機構的相關語料庫建設，非常值得我們參考。

1.6.5. 蘇格蘭文本與語音語料庫（Scottish Corpus Of Texts & Speech, SCOTS）

蘇格蘭文本與語音語料庫（SCOTS）是由格拉斯哥大學（University of Glasgow）與愛丁堡大學（The University of Edinburgh）合作建置的多媒體語料庫，語料庫計畫在 2002 至 2004 年時得到工程與物理科學研究委員會（Engineering and Physical Sciences Research Council, EPSRC）的資助，另外也分別在 2004 至 2007 年，還有 2007 至 2010 年得到藝術與人文研究委員會（Arts and Humanities Research Council）的資助。語料庫網址為：https://www.scottishcorpus.ac.uk/。

蘇格蘭文本與語音語料庫（SCOTS）收錄超過 1300 筆書面和口語文本，共計超過 470 萬詞（Kopaczyk, 2016），其中 77%書面語料，23%口語語料，書面語料收錄 1945 年至今的語料，口語語料則是 2000 年之後錄製的語料，口語語料的影音檔皆被轉寫成文字檔，文字檔和原本的影音檔皆作同步化處理。語料庫所收錄的文本體材包括散文、小說、詩歌、商務和個人書信、宗教文字、議會和行政文件、電子郵

件、對話、訪談等。雖然語料庫團隊在收錄語料會盡可能收錄來自不同地區、背景、年齡、性別、職業的語料，不過因為版權因素，目前語料庫的語料來源以相對取得容易的詩歌、散文等為大宗，因此該語料庫並不是真正的代表性語料庫（Representative Corpus）。

蘇格蘭文本與語音語料庫（SCOTS）最初的設計方向偏社會語言學，語料來源不限於蘇格蘭語的母語使用者，這是因為建置團隊認為任何在蘇格蘭生活過一段時間的人都可能會受到蘇格蘭語或蘇格蘭英語的影響（Anderson, Beavan, Kay, 2007）。也因為如此，該語料庫的語言變異度很高，除了蘇格蘭語語料外，還包括蘇格蘭英語、蓋爾語等語料，而蘇格蘭語部分又包括 Doric、Lallans、Glaswegian、Insular Scots 等不同方言別的語料。在收錄語料時，語料庫團隊也會盡可能收錄完整版的語料，以便於研究話語特徵（Discourse features）。也因為該語料庫設計方向偏社會語言學、強調真實語料，因此一定也會收到帶有冒犯或不雅的語料內容，對於這部分的語料語料庫團隊會以 lesser -（冒犯）和 serious -（非常冒犯）作標記。此外，該語料庫並沒有語法標註，不過在轉寫口語語料時會用到語言重疊（Overlap）、False Start、Truncation、聽不清楚、不確定內容、非詞彙聲音（如笑聲、噴嚏聲、咳嗽等）、非語言事件的聲音（如電話鈴響）等相關標記，標記用法與範例可參考以下連結：https://www.scottishcorpus.ac.uk/corpus-details/。

蘇格蘭文本與語音語料庫（SCOTS）轉寫口語語料所採用的工具為 Praat 軟體，另外在建置語料庫時，語料庫團隊也使用了像是 Collocate Cloud、compare、MI Concordance 等工具，這些工具皆可在網站上取得 https://www.scottishcorpus.ac.uk/tools/。

蘇格蘭文本與語音語料庫（SCOTS）有兩點值得我國作參考。首先，在口語轉寫部分，該語料庫透過收錄完整版語料、將語料標上各

種話語標記，來盡可能保留原語料的語境，這對於未來進行篇章分析、口語語意分析等領域非常有幫助。另外，該語料庫也廣收各種語言變異度、各種地區、背景、年齡、性別、職業等的語料，如此便能收集到豐富而多面相的蘇格蘭語語料，而非單一或某種比較官方或比較主流的語料，這點對我國未來在收錄各國家語言的各方言相關資料時也頗具參考性。

1.6.6. 愛爾蘭語料庫

Gaois 是都柏林城市大學（Dublin City University，DCU）人文與社會科學學院 Fiontar & Scoil na Gaeilge 科系的一個研究小組，其目標是透過開發創新和值得信賴的資源來維持和改變愛爾蘭的語言和文化。從 2004 開始 Gaois 團隊陸續獨自或與各單位合作開發了各項愛爾蘭語相關資源，這些資源主要可分為 Gaois.ie (https://www.gaois.ie/)、Téarma.ie (http://www.tearma.ie/)、Logainm.ie (https://www.logainm.ie/)、Ainm.ie (https://www.ainm.ie/) 與 Dúchas.ie (https://www.duchas.ie/en)這五大類別。除了 Téarma.ie 網站目前僅提供愛爾蘭語頁面因此無法作介紹外，以下將簡單介紹上述各項資源。

1.6.6.1.Gaois.ie

Gaois.ie 是由都柏林城市大學 Gaois 研究團隊所建置的資源，經費是由愛爾蘭政府的文化、遺產與愛爾蘭語區部門（Department of Culture, Heritage and the Gaeltacht, Government of Ireland）、國家彩卷（National Lottery）、愛爾蘭語委員會（Foras na Gaeilge）、國家民俗基金會（National Folklore Foundation）等機構所贊助，官網於 2014 年上線。Gaois.ie 網站上目前的成果包括當代愛爾蘭語料庫（Corpus of

Contemporary Irish）、英語-愛爾蘭語立法平行語料庫（Parallel English-Irish Corpus of Legislation）、愛爾蘭語姓氏資料庫（Database of Irish-language surnames）、PEADAR Ó LAOGHAIRE 成語資料庫（PEADAR Ó LAOGHAIRE IDIOM COLLECTION）、術語資料庫（Terminology Database）等五項，另外 Gaois 研究小組也在網站上釋出了相關技術資源與軟體。

1.6.6.2.Gaois.ie 當代愛爾蘭語料庫（Corpus of Contemporary Irish）

當代愛爾蘭語料庫（Corpus of Contemporary Irish）是愛爾蘭語的單語語料庫，語料組成為 21 世紀以來至今的各種經過編輯的文本（詳見 https://www.gaois.ie/en/corpora/monolingual/），目前規模達到 3090 萬詞。該語料庫原本是 Gaois 和 Fiontar & Scoil na Gaeilge 所使用的內部術語資源（internal terminological resource），不過在 2016 年的時候開始對外免費開放。語料庫目前僅提供特定和模糊搜尋功能，語料尚未經過標註（Loingsigh, Raghallaigh and Cleircín, 2017），資料儲存格式為 XML。

1.6.6.3.Gaois.ie 英語-愛爾蘭語立法平行語料庫（Parallel English-Irish Corpus of Legislation）

英語-愛爾蘭語立法平行語料庫（Parallel English-Irish Corpus of Legislation）原本是 Fiontar & Scoil na Gaeilge 的 Gaois 研究小組成員要用來發展 LEX 計畫的，不過後來 Gaois 團隊將相關資料整理成這個語料庫。該語料庫主要收錄各種法規的愛爾蘭語和英語的翻譯對照資料，法規的細項詳見 https://www.gaois.ie/en/corpora/parallel/sources/。目前語料庫的規模達到 5480 萬詞，其中 2810 萬詞為愛爾蘭語料，2670 萬詞為

英語語料。關於語料下載，由於版權因素，目前網站上僅開放下載愛爾蘭法規，歐盟法規的部分不提供下載。愛爾蘭法規可在以下連結以 TMX（Translation Memory eXchange）的格式來進行下載：https://www.gaois.ie/en/corpora/parallel/data/，下載的資料只能用於個人目的，不可以用任何形式進行重製與散佈。

1.6.6.4.Logainm.ie

Logainm.ie 是由都柏林城市大學 Fiontar & Scoil na Gaeilge 科系和愛爾蘭政府文化、遺產與愛爾蘭語區部門（Department of Culture, Heritage and the Gaeltacht, Government of Ireland）底下的地名分會（Placenames Branch）所合作建置的資源，並由國家彩卷（National Lottery）所贊助，目前成果包括愛爾蘭地名資料庫（The Placenames Database of Ireland）、Meitheal Logainm.i 群眾外包計畫、地名分支聲音典藏（Sound Archive of the Placenames Branch）這 3 項。愛爾蘭地名資料庫（The Placenames Database of Ireland）最初的建置目的是為了提供正確且標準化的愛爾蘭語地理名稱，自從愛爾蘭在 2003 年通過了 Official Languages Act 2003，加上愛爾蘭語在 2007 年也變作歐盟的官方語言之一後，這項需求便與日俱增（Easpaig, 2009）。該資料庫計畫的第一階段從 2007 年 4 月開始進行，現在已達到第七階段並且持續進行中，計畫各階段細項詳見以下連結：https://www.logainm.ie/en/inf/proj-about。透過建置該語料庫，使用者們便得以欣賞愛爾蘭豐富的地名相關文化資產。

除了愛爾蘭地名資料庫，Logainm.ie 團隊也設置了 Meitheal Logainm.i 群眾外包計畫，民眾只要在網站上進行註冊（https://meitheal.logainm.ie/en/signup/），並依規定填寫相關資料，便

可貢獻自己所知的次要地名（minor placename）相關資訊。次要地名有別於縣、鄉、鎮、村、街道等地理名稱，主要是指人們對於物理特徵（例如，湖泊，河流，海灣，岬角，島嶼，山脈，丘陵）和人造特徵（例如，堡壘，教堂，修道院，墓地，橋樑，十字路口）所給予的地理名稱。民眾在貢獻次要地名時，要先在地圖上點選正確地理位置，接著再填寫該地愛爾蘭語/英語地名、地名類型、訊息來源、其他資訊等，最後還要錄下自己對該地地名發音的錄音檔。透過收集愛爾蘭語的次要地名，人們便得以用愛爾蘭人的觀點來了解愛爾蘭的本土文化、生活方式等資訊。

地名分支聲音典藏（Sound Archive of the Placenames Branch）收錄了 1960 至 1970 年代 24 個愛爾蘭鄉鎮超過 4000 人的地名發音錄音檔，音檔規模達到 1200 小時，並且在 2009 年被數位化作成資料庫。該資料庫是由 Cáit Nic Fhionnlaoich 在爭取碩士研究獎學金計畫（MA Research Fellowship）時的成果之一，另外也在 2010 到 2011 年得到 Department of Community, Equality and Gaeltacht Affairs 的贊助。

1.6.6.5.Ainm.ie

Ainm.ie 計畫，又名愛爾蘭語傳記國家資料庫（National Database of Irish-Language Biographies），收錄了自 1560 年代至今 1774 份傳記，共 130 萬詞的愛爾蘭語語料，語料主要來源自 Diarmuid Breathnach 和 MáireNíMhurchú 的《Beathaisnéis》系列叢書，目前該資料庫規模還在持續擴增中。該計畫是由都柏林城市大學 Fiontar & Scoil na Gaeilge 科系自 2009 年起與 Cló Iar-Chonnacht 編輯，還有 Diarmuid Breathnach 和 MáireNíMhurchú 兩位作者所合作建置的，經費來源為愛爾蘭政府文化、

遺產與愛爾蘭語區部門（Department of Culture, Heritage and the Gaeltacht, Government of Ireland），還有國家彩卷（National Lottery）。

1.6.6.6.Dúchas.ie

Dúchas.ie 是由都柏林城市大學 Fiontar & Scoil na Gaeilge 科系、都柏林城市大學國家民俗收藏機構（National Folklore Collection）、愛爾蘭政府文化、遺產與愛爾蘭語區部門（Department of Culture, Heritage and the Gaeltacht, Government of Ireland）所合作建置的資源，目前成果包括 Dúchas 國家民俗收藏 UCD 數位化計畫（Dúchas, the National Folklore Collection UCD Digitization Project）、Meitheal Dúchas.ie 群眾外包計畫、愛爾蘭姓氏索引（Irish Surname Index）這 3 項。

Dúchas 國家民俗收藏 UCD 數位化計畫（Dúchas, the National Folklore Collection UCD Digitization Project）收錄 19,760 件手稿，252,771 件來自學校收藏（1937–38）的故事和 12,271 件來自攝影收藏品的圖像，計畫網址為：https://www.duchas.ie/en/。

除了上述計畫，Dúchas.ie 團隊也設置了 Meitheal Dúchas.ie 群眾外包計畫，民眾只要在 Dúchas 計畫網站註冊為會員（https://www.duchas.ie/en/meitheal/in），並依規定填寫相關資訊，就可以貢獻語料或者幫忙轉寫語料，群眾外包方法流程詳見 https://www.duchas.ie/en/meitheal/。

1.6.6.7.小結

從上述介紹可見，Gaois 研究團隊參與建置的成果十分多樣化且豐富，雖然部分成果可能還不是很完美（例如當代愛爾蘭語料庫（Corpus of Contemporary Irish）的語料還沒有經過標註），不過還是

有很多值得參考的地方。例如，從英語-愛爾蘭語立法平行語料庫（Parallel English-Irish Corpus of Legislation）與術語資料庫（Terminology Database）可見，Gaois 研究團隊在愛爾蘭語與英語的雙語語料平行對照發展十分成熟，除了建置雙語語料庫外，還建置了能進一步了解專業術語語意、用法等的資料庫，這樣的設計對於本國未來建置相關平行語料庫時，頗具參考價值。另外，Gaois 研究團隊建置的愛爾蘭地名資料庫（The Placenames Database of Ireland）構想也很值得作參考，自從我國在民國 108 年 1 月公布《國家語言發展法》後，各國家語言地位一律平等；然而目前本國仍有不少地理名稱、景點名稱皆採用華語，而非當地居民原本所使用的名稱。例如，蘭嶼的東北角有個名叫「雙獅岩」的景點，該景點為火山活動中噴出的熔岩冷卻後形成的自然景觀，遠看就像是兩隻獅子彼此對望一樣；然而蘭嶼本身並沒有「獅子」這種動物，可見「雙獅岩」名稱並非當地達悟族人所取，而是外人來到蘭嶼島後所加上的。其實蘭嶼當地達悟族人都稱呼該景點為 Jipanatosan，意為「東清」和「朗島」兩個部落之間的分界線。因此倘若國家未來也能收集這些資料，並建置成相關資料庫，就能幫助國人更進一步地理解當地文化。

貳、　國外手語語料庫、資料庫的現況分析

本章延續上章的內容，針對手語（sign language）語料庫及資料庫進行現況分析。手語不只是聾人社群的溝通方式，更是具有完整架構的語言，《國家語言發展法》明定臺灣手語為國家語言之一，而手語語料庫、資料庫因為手語語言特性，在建置上須考量圖片、影像檔及文字說明等特殊處理。關於臺灣手語的語言資源、以及自然手語與文法手語的差異，可參見「5.6 臺灣手語」一節。

手語語料庫的建置有其必要性，McEnery & Ostler (2000) 呼籲語料庫語言學領域應擴大研究與應用範圍，顧及更多語言的語料庫建置，也表示手語語料庫的建置是其中一項艱難的任務。不過，因為有了技術支援，語料的時間訊息處理及標記變得相對可行，亦開啟了手語語料庫開源、線上的嘗試。此外，手語能讓我們對語言多樣化更加了解。

澳洲手語語料庫建置者 Trevor Johnston (2009) 提到，(1) 手語是少數族群使用的語言，沒有書寫系統，許多手語打法亦較少從使用社群發展而成，(2) 手語面臨跨世代的語言傳承危機（generational transmission），以手語為母語的人口數稀少（few native speakers），(3) 建置手語語料庫時，若僅以文字描述手語打法而無原始影像，會造成日後語料檢審與利用的困難。因此，藉由手語語料庫的建置，蒐集自然的手語語料，了解手語社群所使用的手語為何，且手語語料庫的建置方式並非手語語料的轉寫，而是手語語料的標記 (“The aim is to create an annotated SL [=Sign Language] corpus and not a body of SL texts which have been transcribed to a greater or lesser degree of detail.”)。雖然轉寫語料對於語言分析來說仍很重要，但若技術上能夠處理原始影像，則影

像較轉寫文字更能完整紀錄手語的表達；不過，轉寫文字若是能夠與影像檔在時間上達到對應，則有助於語言分析。

以下介紹美國、英國、荷蘭、澳洲與瑞典的手語語言資源，美國設有國家單位美國國立手語與手勢中心（National Center for Sign Language and Gesture Resources, NCSLGR），亦有 ASL SignBank 的建置計畫，並與 ASL-LEX 計畫合作，前者蒐集各年齡層的自然手語語料，後者則屬詞彙資料庫（lexical database），以心理語言學為基礎，設計不同的實驗，蒐集相關手語詞彙及實驗資料，並以資料視覺化的方式呈現手語詞彙，反映手語語言象似性（iconicity）很高的特色。英國手語語料庫及資料庫從社會語言學著手，邀請不同語言與社會背景的手語使用者擔任發音人，蒐集多元化的手語語料。澳洲手語資料庫的 ID 表概念廣為其他手語資料庫採用，並藉由主題式的研究計畫擴充澳洲手語辭典，從生活需求切入，充實語言資源的內容。荷蘭手語資料庫扮演著技術共享與跨團隊合作的串連角色。瑞典手語語料庫於近期釋出網頁版的手語語料庫，使用者毋需仰賴離線軟體檢索語料，且以手語教育現場為考量設計，使用者可上傳語料並使用相關功能。有關上述手語語料庫或資料庫，以下分述之。

2.1. 美國手語資料庫與語料庫

2.1.1. 美國國立手語與手勢資源中心手語與手勢資源語料庫——National Center for Sign Language and Gesture Resources (NCSLGR) Corpus

美國國立手語與手勢資源中心手語與手勢資源語料庫（National Center for Sign Language and Gesture Resources Corpus, NCSLGR）為該中心建置之美國手語語料庫，由布朗大學執行該中心的美國手語研究

計畫（American Sign Language Linguistic Research Project, ASLLRP），NCSLGR 語料庫的網址為：https://www.bu.edu/asllrp/ncslgr-for-download/download-info.html。

NCSLGR 語料庫的語料來源為以手語為母語的發音人，已有 1,866 個手語或手勢詞形（sign type），若計算重複的詞形，則有 11,854 個詞（sign token），不含手勢的話則有 1,278 個手語詞形及 10,718 個手語詞；以句子為單位來看，在 19 篇短篇敘事語料中，納入此語料庫的部分有 1,002 句（utterances），另有 885 句獨立的句子，以搜羅不同句構的手語表達方式。語料呈現方式為影像檔，且顧及不同的拍攝角度，如側拍、正面近拍等，網頁上可選擇顯示不同的角度截圖，但目前看不到影像檔，僅有截圖。語料標記方面，以 XML 形式儲存，網址為：https://www.bu.edu/asllrp/ncslgr-for-download/signstream-xmlparser.zip ，包括手語的起迄時間（start and endpoint）、頭部動作、詞性等。

奠基於上述的語料庫資源，該中心進而建立了美國手語影像詞典（American Sign Language Lexicon Video Dataset, ASLLVD），網址為：http://www.bu.edu/asllrp/av/dai-asllvd.html，收錄了超過 3,300 個詞彙，每個詞彙由最多六名手語母語者錄製，因此總計有高達一萬筆資料。如果是複合詞（compound words），則在標記上標有詞素（morpheme）的訊息。除此之外，因為手語資料呈現在網頁上時，需要顧慮到畫質及檔案大小的取捨，因此該語料庫提供網頁呈現版本及下載版的影像檔。

2.1.2. 美國手語資料庫——ASL (American Sign Language) SignBank

美國手語資料庫（ASL SignBank）為手語習得、標記、典藏與資料分享（Sign Language Acquisition, Annotation, Archiving and Sharing,

SLAAASh）計畫的成果之一，由美國高立德大學（Gallaudet University）負責執行，資料庫託管於美國耶魯大學，網址為：https://aslsignbank.haskins.yale.edu。

目前，ASL SignBank 已收錄超過 3300 個詞彙 (Becker et al., 2020)，由 4 對聾人父母與聾人小孩提供，小孩年齡介於 1 歲 4 個月至 4 歲 1 個月。(Hochgesang et al., 2018) 詞彙的認定採用 Fenlon et al. (2015) 的詞彙原形（lemmatization）原則，由高立德大學 Julie Hochgesang 教授負責詞彙 ID 表的管理，如果標記人員需要新增 ID，須在標記系統提出申請，錄製並上傳該新詞彙的打法影片，並將此筆資料標註為需要檢審的詞彙。若該詞彙通過審核成為 ASL SignBank 的新詞彙，再由聾人老師錄製示範影片。除了標記人員，ASL SignBank 的註冊會員亦可提出增修、刪減申請，該研究團隊也會定期公告 ID 表誌（ID Gloss Digest）。

詞彙 ID 表的設計最先由澳洲手語學者 Johnston (2009) 提出，考量到手語尚無一致的書寫系統，而將各個手語表達方式的最小單位編號。除了 ASL SignBank，荷蘭奈梅亨拉德堡德大學（Radboud University）的 NGT 語料庫（Nederlandse Gebarentaal Corpus）、倫敦大學學院（University College London）的 BSL 資料庫（British Sign Language SignBank）以及芬蘭于韋斯屈萊大學（University of Jyväskylä）的 FinSL 資料庫（Finnish SignBank）將採納此概念建置手語資源。ASL SignBank 的資料可直接使用 ELAN 軟體處理，標記系統以荷蘭 NGT SignBank 的基礎為底開發而成。

ASL SignBank 計畫與 ASL-LEX 計畫合作，使用相同的詞彙 ID 表、在詞彙分析方面採用相同的原則標記 (Becker, 2020)，以便跨資源搜尋，例如：

(1) 手勢（handedness）：若手語打法為雙手時，通常會有對稱或交替的動作，標記為[SymmetricalOrAlternating]，若無則為[AsymmetricalSameHandshape]，單手打法則標記為[OneHanded]。另一種打法有主副手之分，標記為[AsymmetricalDifferentHandshape]。若不符合上述對稱原則（symmetry condition）或主副原則（dominance condition），則標記為[Other]。

(2) 主要身體部位（major location）：臂腕（arm, including wrist）、軀幹（body, signer's torso）、手（hand）、頭及臉部（head, including face）、無特定部位（neural, signing space in front of singer's body）等。

(3) 主手（dominant hand）、使用的手指（selected finger）、以及手指彎曲狀態（flexion）：使用的手指是依手指彎曲或伸直的狀態而定，而手指彎曲的標記分為 9 類（categorical），而非給予連續性的數值（numerical）。

(4) 尚未包含手語打法的動作方向（direction of movement），因此有些不同的手語詞彙在兩個資料庫的標記資料是相同的。

雖然 ASL SignBank 和 ASL-LEX 的資料共享，但兩者建置目的不同，ASL SignBank 的任務是蒐集實際使用的手語語料，包括該計畫蒐集之語料及其他採用相同 ID 系統蒐集到的語料，因此 ASL SignBank 的 ID 不是預先訂定的，最初的 ID 表亦無經過詞形還原（lemmatization），而是在詞彙量足以進行相關分析後才調整的；ASL-LEX 則為了涵蓋高低頻率、高低象似性的手語詞彙，而以實驗設計的方法蒐集不同的手語詞彙。

此外，由於部分語料的蒐集是在其他計劃下完成的，而非 SLAAASh 計畫開始後才著手語料蒐集，須重新向當時的受試者取得同意，讓這些語料得以納入 ASL SignBank 與延伸使用。該計畫未來亦希望能透過焦點團體（focus group）的訪談蒐集更多語料，將參與者擴及至聾人家屬與社群，以奠定更深厚的社群基礎。

2.1.3. 美國手語詞庫——ASL-LEX (A Lexical Database of American Sign Language)

美國手語詞庫（ASL-LEX）是美國聖地牙哥州立大學（San Diego State University）、塔夫茨大學（Tufts University）及波士頓大學（Boston University）三校合作的研究計畫，該計畫的特色是打造一個資料視覺化的網頁，將相似的手語詞彙以節點（node）串連起來，該網頁雖然不是語料庫，但若以語言資料庫的角度來看，其呈現相當具有特色，網址為：http://ase.tufts.edu/psychology/psycholinglab/asl-lex/visualization.html，如下圖。

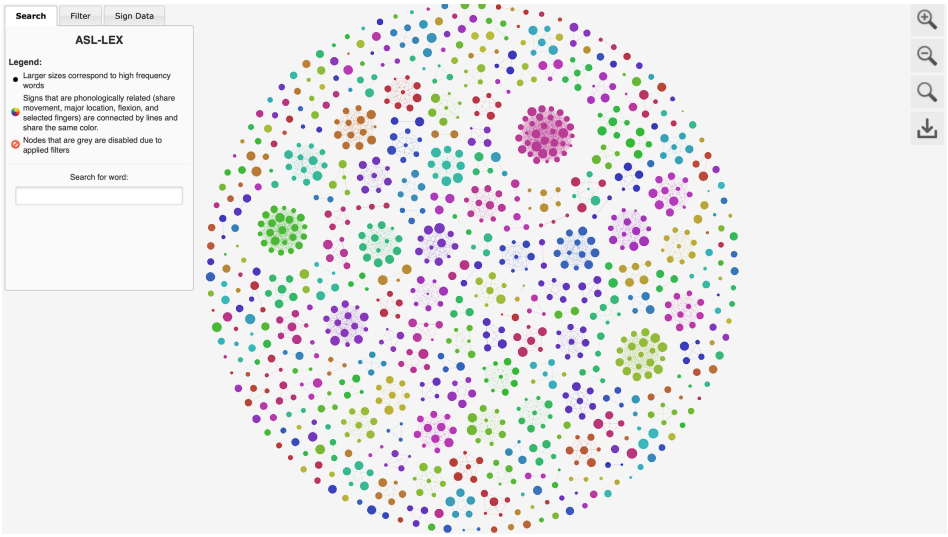


圖 8. ASL-LEX 將手語詞彙視覺化，每個圓點代表一個詞彙。

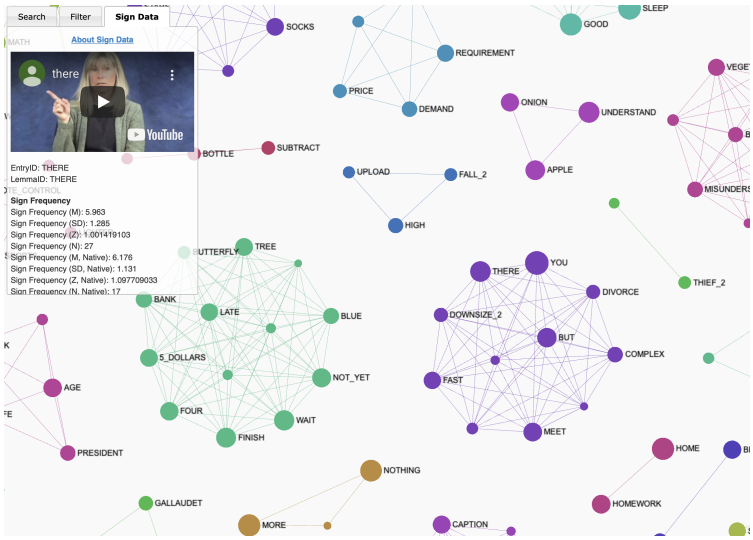


圖 9. 在 ASL-LEX 中檢索「 there 」一字的結果畫面。

ASL-LEX 詞庫先後釋出 1.0 及 2.0 的版本，該計畫最初以心理語言學為出發點，聚焦在象似性（ iconicity ）的主題上，以實驗的方法蒐集相關數據，而受試者的各項評分可以讓我們更了解人類語言處理的機制。象似性指的是詞彙的形式（ form ）和意義（ meaning ）展現出高度的相關性，這個特徵在手語特別明顯，許多詞彙都是建立在象似性之上（ iconically-movitivated ），而口說語言的例子則有擬聲詞。該資料庫的象似性標記是由群眾外包的方式，在亞馬遜 MTurk 平台上讓受試者評分（ 相關介紹可參考第三章 ）。

ASL-LEX 詞庫目前共蒐集了 1,000 個手語打法，在 1.0 版時附有 (1) 受試者主觀評分該詞彙的使用頻率（ subjective frequency ratings ）(2) 受試者對該詞彙的象似性評分（ iconicity ratings ）(3) 影像檔的時間長度 (4) 是否為複合詞等標記資訊等。

最新 2.0 版本的 ASL-LEX 詞庫網頁上，詞彙數目並無增加，但正逐步在顯示結果中加入更細項的資訊，像是除了主要的英文翻譯外，亦列出其他可能的英文翻譯（ alternative english translations ）、譯者間對此主要翻譯的一致程度（ Agreement ）、聽人對該手語詞彙的象似性

評分、聾人對該手語詞彙的象似性評分、手指拼寫的手語打法數目（Fingerspelled loan sign，意指直接將英文字母比畫出來的手語打法）、使用的手指與身體部位為何（selected finger、location）等，不僅有原先的心理語言學實驗資料，也將手語本身的資訊標記出來。在檢索設計上，使用者可從上述資訊中搜尋、篩選出特定結果。此外，整個資料庫的資料可至開放科學中心平台（Open Science Framework, OSF）瀏覽與下載，包括所有詞彙截圖的 pdf 檔、英文翻譯、鄰近詞、評分實驗結果等細項 csv 檔。

在資料視覺化方面，每個節點代表一個詞彙，圓點越大表示該詞彙的使用頻率越高，鄰近點的詞彙可能有相同的打法，包括使用的手指（selected fingers）、手指是否彎曲（flexion）、動作（movement）及該手語打在身體的哪個部位上（location），這些資訊是手語詞彙常見的資訊，在檢索時除了以各項資訊篩選顯示結果，亦可點擊特定圓點查看該詞彙的詳細資料及鄰近詞。

2.2. 英國手語語料庫與資料庫——BSL (British Sign Language) Corpus & BSL SignBank

英國手語語料庫（BSL Corpus）主要由英國倫敦大學學院聾人認知與語言研究中心（Deafness Cognition and Language Research Centre, University College London）於 2008 年至 2011 年建置，並有威爾斯班戈大學（Bangor University）、蘇格蘭赫瑞瓦特大學（Heriot-Watt University）及北愛爾蘭貝爾法斯特女王大學（Queen's University Belfast）及布里斯托大學（University of Bristol）的研究人員參與，經費來源為英國經濟和社會研究理事會（UK Economic and Social Research Council, ESRC），網址為：http://www.bslcorpusproject.org，其影像及

後設資料可於CAVA（人類溝通影音典藏平台，Human Communication: An Audio-Visual Archive）瀏覽或下載，網址為：https://bslcorpusproject.org/cava/。BSL Corpus 的語料來源為 249 位聾人手語使用者（以手語為母語者、達母語程度或流利程度之手語使用者，且學習手語的時間不晚於 7 歲），其中 31%（76 位）發音人以手語為母語，來自英國 8 個地區（倫敦、布魯斯托、卡地夫、伯明罕、新堡、曼徹斯特、格拉斯哥與貝爾法斯特），性別、年齡、語言學習背景不一，男女比為 48%（120 位）:52%（129 位），年齡組成為 18-35 歲 59 人（24%）、36-50 歲 62 人（25%）、51-64 歲 68 人（27%）、65 歲以上 60 人（24%）。(Schembri et al., 2013) 語料蒐集方式為 30 分鐘的雙人對話、102 個附圖的英文詞彙的手語打法[4]及與語言使用和態度相關的訪談[5]，並以三個角度、藍色布幕為背景錄製語料（兩個發音人各一、及整體的影像）。雖然 BSL Corpus 邀請不同語言及社會背景的手語使用者擔任發音人，但由於尚未有英國手語語言使用的詳細調查，此計畫僅能盡可能控制各變因的比例，超出比例則不再蒐集，往後亦希望進行更多相關的計畫，考慮更多語言及社會的因素。 (Schembri et al., 2013)

在語料標記方面，由兩個相關計畫負責，分別為音韻及詞彙變異（phonological and lexical variation）計畫及詞彙頻率（lexical frequency）計畫，前項計畫在音韻分析上，已完成 211 位發音人 6,330 個手語（sign）的手形及位置標記，其中 2,110 個詞（token）還有詞性及使用

[4] 詞彙語料蒐集表可參考：https://bslcorpusproject.org/wp-content/uploads/BSLCP-Vocabulary-task-rights-cleared.pptx

[5] 問卷內容可參考：https://bslcorpusproject.org/wp-content/uploads/BSLCPInterviewQuestionnaire.pdf

單或雙手的標記，而在詞彙分析上，已完成 249 位發音人的 7,332 個手語的標記。後者計畫已完成約 25,000 個手語的標記（50 個發音人各標 500 個手語）。若要檢索語料庫內容，BSL Corpus 提供兩個子網站，一個是前述的 CAVA 平台，以資料庫形式檢索後設資料，另一個是讓使用者選擇地區與年齡等選項，再隨機推薦語料庫中的手語影片，但僅有影片，並無相關敘述或標記資訊，因此目前 BSL Corpus 還未釋出可機讀版本的語料庫。(Schembri et al., 2013)

英國手語資料庫（BSL SignBank）於 2011 年至 2015 年進行建置，網址為：https://bslsignbank.ucl.ac.uk。BSL SignBank 收錄 2,528 個詞彙，在這些詞彙中，出現在 BSL Corpus 的詞形（sign type）約有 1,700 個，重複的詞彙（sign token）約有 50,000 個。在語料標記上，BSL Corpus 標有 295 種與顏色、數字、國家相關的詞彙。

針對手語辭典的編纂，Fenlon et al. (2015) 提到，因為手語語料庫資源稀少，加上手語沒有一定的書寫系統，難以系統性地從語料庫語料找出可新增至辭典的手語詞彙。過往以來的一種作法是建立手語詞彙表（wordlist），但每個手語詞彙與翻譯可能僅是一對一的對應關係，無法代表手語在詞彙層次的全貌。不過，線上手語辭典的建置模式可以解決一些紙本辭典對手語造成的麻煩與限制，如辭典的編排呈現及搜尋頁面/方式不受限於頁面排版、手語打法的影像亦可不完全仰賴文字敘述說明及截圖，因此手語辭典的編纂和手語語料庫的建置在目前技術的幫助下，是很值得投入且寶貴的。

與手語語料庫建置相關的詞彙處理議題，Fenlon et al. (2015) 認為可將翻譯與檢索 ID 看成兩套不同的系統，像是目前收錄於澳洲資料庫（Auslan SignBank）的澳洲辭典是以英文翻譯檢索手語詞彙，使得不同的手語詞項（lemma）被編成同一詞條（entry）的多義（polysemous）

詞彙，而檢索 ID 的設計可讓使用者針對不同的手語詞項檢索相關內容，對手語詞彙數量等統計或語言分析也較以手語為本。在 BSL SignBank 裡，使用者可根據「手形（handshape）」、打法「位置（location）」與英文關鍵字檢索，便會列出符合的一或多項結果，並附上手語打法的影像。

2.3. 荷蘭手語語料庫及資料庫——Corpus NGT (Nederlandse Gebarentaal; Dutch Sign Language) & NGT SignBank

荷蘭手語語料庫（Corpus Nederlandse Gebarentaal, Corpus NGT）是荷蘭科學研究組織（Nederlandse Organisatie voor Wetenschappelijk Onderzoek; Netherlands Organization for Scientific Research, NWO）為時兩年的計畫，於 2008 年由荷蘭奈梅亨拉德堡德大學（Radbound University）的研究團隊建置完成。NGT 語料庫的語料來源為 92 位學語前失聰（prelingually deaf）的荷蘭手語使用者，共有 145,000 筆資料（gloss），線上版本的 NGT 語料庫已包括 370,000 個標記資料，以看圖說故事、看影片說故事、以及談論失聰、聾人教育、手語等主題的方式錄製手語語料，而標記項目則從以句為單位的分析到手勢皆有，目前 NGT 語料庫的資料存放於語言典藏（Language Archive）平台，網址為：https://archive.mpi.nl/tla/islandora/object/tla%3A1839_00_0000_0000_000A_00E1_2，亦可直接從 NGT 語料庫網站下載標記資料，網址為：https://www.ru.nl/corpusngtuk/methodology/annotation/，方便使用者在 ELAN 軟體上進行離線的語料標記。

與 NGT 語料庫相關的計劃為 NGT 資料庫（NGT SignBank），希望將 NGT 語料庫的語料轉換成詞彙資料庫（lexical database），因為語

料庫蒐集到的語料能夠擷取出更豐富的資訊，尤其是較細緻的語意（semantics）及語用（pragmatics）資訊。在語料處理工具方面，NGT 資料庫使用 ELAN 軟體，以 EVC（externalcontrolled vocabulary）檔來管理資料庫的詞彙，標記人員在標記語料時便從此 EVC 檔案尋找符合的詞條。目前，NGT 資料庫共收錄了 3,200 個詞彙，並附有 (1) 各詞彙在 NGT 語料庫的出現頻率及各發音人使用此詞彙的頻率，以區分此詞彙在語言社群中的使用程度，(2) 與時間對應的口形動作（mouth action）標記，此項標記不一定只與一個詞彙對應，口形動作與手勢的互動關係是 NGT 團隊主要的研究主題之一。

NGT 團隊將 NGT 資料庫的程式碼託管於 GitHub 網站，作為手語語言資源開發的交換平台，並取名為全球手語資料庫（Global SignBank），網址為：https://github.com/Signbank (Crasborn et al., 2018)，目前除了 NGT 資料庫，亦包括澳洲手語、英國手語、芬蘭手語資料庫。此外，NGT 團隊於 2018 年發表了全球手語資料庫的操作手冊，指引使用者如何新增詞彙至該資料庫。首先，以 ID 查詢表確認該詞彙並未收錄在資料庫中，再確認該詞彙是否多義，是否可能已被翻譯成其他意思，此時僅需加入新詞意即可。在標記上，如果有多個選擇，以使用頻率（frequency）和象似性（iconicity）為原則，標記項目包括：複合詞的切分及對應詞彙、單手或雙手（handedness）及是否雙手動作對稱（symmetry）、手形變化（handshape changes）、動作方向（movement direction）及是否重複（repetition）、手勢打在身體的哪個部位上（location），此外亦提供實名（name entity）及語意欄（semantic field）的標記，其中語意欄的分類係根據澳洲手語資料庫（Auslan SignBank）發展而成。有了這些標記資料，該資料庫的系統可即時更新最小對立體（minimal pair）的資料。

在跨團隊合作方面，NGT 語料庫於 2008 年完成階段性的建置任務，緊接著開始手語語料庫網（Sign Linguistics Corpora Network, SLCN）的計畫，2008 年至 2011 年獲得荷蘭科學研究組織的經費支持，2014 年至 2016 年進一步與英國藝術與人文研究委員會（Arts & Humanities Research Council）合作，使用 ELAN 軟體作為手語語料庫的工具，SLCN 計畫集結了歐洲相關的手語語料庫建置工作，共同提升手語的語言權利，保存珍貴的手語語言資料以共同面對手語的傳承危機，並探討手語語言資源建置的技術層面，藉由舉辦主題式的工作坊，分階段性地解決相關問題，參與國家如瑞典、芬蘭、斯洛維尼亞、波蘭、比利時及英國等。

2.4. 澳洲手語資料庫——Auslan (Australian Sign Language) SignBank

澳洲手語資料庫（Auslan SignBank）是澳洲麥覺理大學（Macquarie University）教授 Trevor Johnston 所建置之手語資源，該網站包括澳洲手語語料庫（Auslan Corpus）及澳洲手語辭典（Auslan SignBank Dictionary），網址為 http://www.auslan.org.au。Auslan 語料庫的語料來源是澳洲手語典藏計畫（Auslan Archive），此一系列計畫期望藉由澳洲手語典藏資源的建立，達到瀕危語言紀錄與保存的目標，而澳洲手語語料庫的建立，將典藏的內容轉成可機讀的語料，以供語言學習與研究之用，特別是基於語料庫的研究方法（corpus-based approach）。(Johnston, 2009) Auslan 典藏計畫於 2008 年收錄於瀕危語言典藏網站（Endangered Languages Archive, ELAR），Auslan 語料庫的建置則早在 2004 年便已開始進行，並於 2012 年開放公開使用，但目前該網站無法查看到語料庫的內容。

Auslan 語料庫的語料為 50 位以手語為母語或達母語程度的聾人，錄製共 150 小時的語料，每個語料為 3 小時的雙人對話，包括：訪談、問卷問題簡答、敘事、自由對話、以及一些常用語料蒐集方式，諸如：看圖說故事、看影片說故事等，這些影像檔被切分成約莫 1100 個檔案，其中 130 個檔案有不同程度的標記。(Johnston, 2009)

Auslan 語料庫的語料除了時間對應，在語料標記上從基本的詞彙單位開始，依各個研究主題逐步增加不同的標記項目。(Johnston, 2009) 不過，Johnston (2009) 也提到詞彙標記的困難處，在於同樣的打法（sign）可能會有不同的英文詞彙來標記，造成無法從英文詞彙找到該打法的語料，因此 Auslan 語料庫在此項標記上是採用 ID 檢索（ID-gloss）的標記，並使用大寫英文字母標示 gloss，以及將「不喜歡」標為「LIKE-NOT」，而非「DON'T-LIKE」。雖然 gloss 部分以英文詞彙標示，但不代表該詞彙被使用在任何特定的語境下（not context-dependent），從 Johnston (2009) 的敘述來看，ID-gloss 的用意為查詢同樣的打法（sign），並讓後續的標記能夠奠基於此手語詞彙上。此外，Auslan 語料庫其中一欄標記為詞彙化程度（fully-lexical/partly-lexical signs），因為有些手語詞彙仍在發展、成形當中，尚未在使用社群中取得較有共識的打法，而越是完全詞彙化的打法，越可能被收錄至辭典中。

在一篇談論 Auslan 語料庫的文章中，澳洲樂卓博大學教授 Adam Schembri 提到，如果有機會，手語語料庫的建置亦可與科技技術結合，例如：手語辨識（sign language recognition）以及虛擬的手語員（virtual signer/signing avatar）(“The Auslan Corpus”, 2011)

除了語料庫資源，該網站的辭典目前可供搜尋。Auslan 辭典以影像及釋義的方式呈現，可輸入英文關鍵字檢索手語詞彙，或是選擇主

題查看相關詞彙，但並無提供官方統計資料，若以 49 個主題，每個主題平均 75 個詞彙計算，Auslan 辭典約莫有 3750 個可重複的詞彙。特別的是，該網站近期與醫療手語資料庫計畫（Medical Signbank Project）合作，醫療與教育主題的詞彙是目前辭典擴增的重點。根據該計畫的目標敘述，聾人使用手語的場合越來越多，手譯員有越來越多機會需要將接收到的訊息譯成手語，所需的詞彙亦隨之改變，希望將相關詞彙列入 Auslan 辭典裡，從中獲得使用者的使用回饋，並作內容上的調整。該計畫(1) 透過焦點團體（focus group）的對話，蒐集醫療相關語彙，(2) 針對聾人、手譯員、醫療人員發放線上問卷，了解相關需求，(3) 了解現有的手語詞彙有哪些；對於無手語詞彙、在手語中無法直接以詞彙單位表達的概念，錄製手語影片解釋此概念，如此一來在網站上輸入英文關鍵字後，仍可找到對此概念的手語表達方式。

從澳洲手語資料庫的例子，可透過典藏、語料庫與辭典的建置規劃，充實手語的語言資源，此系列計畫已蒐集大量手語語料，但還無法線上使用該語料庫，不過已可取得該語料庫的標記原則說明書。(Johnston & De Beuzeville, 2014) 在辭典擴增方面，除了以英文關鍵字及主題的方式搜尋，尚無提供其他方式檢索手語詞彙，但該辭典計畫以聾人、手譯員與相關需求人員為中心的語料蒐集方式值得參考。

2.5. 瑞典手語語料庫——STS-korpus (Svenskt Teckenspråk Korpus; Swedish Sign Language Corpus)

本節介紹的瑞典手語語料庫，指的是 STS-korpus (Svenskt teckenspråk Korpus) (Öqvist et al., 2020)，而 Swedish Sign Language Corpus (SSLC，Wallin & Mesch, 2015) 是 2009 年至 2011 年由瑞典斯德哥爾摩大學 Johanna Mesch 教授所建置之語料庫，兩者之間有時間前後

的關係，STS-korpus 是基於 SSLC 的網頁版語料庫，但 STS-korpus 還包含了其他語料庫計畫的語料，例如：歐洲文化遺產計畫（European Cultural Heritage Online Project, ECHO Project）中的瑞典手語語料，未來亦希望納入瑞典手語學習者語料庫（Corpus of Swedish Sign Language as Second Language, SSLC-L2）及瑞典觸覺手語語料庫（Tactile Swedish Sign Language Corpus, TSSL，觸覺手語是聾人和視障者間的溝通語言。）(Öqvist et al., 2020)

STS-korpus 為了解決手語教學現場對語料庫使用的需求，方便手語老師、學生及對學習手語有興趣的人士使用，近期剛推出線上版的語料庫，網址為 https://teckensprakskorpus.su.se/。由於該線上語料庫的建置是基於教育的需求，除了上述的語料庫之外，教師可註冊成為會員並上傳自己的語料，該語料會被歸類為「課程（kurs）語料」。(Öqvist et al., 2020)

STS-korpus 的主體是 SSLC 語料庫，SSLC 語料庫計畫於 2009 年至 2011 年獲得瑞典銀行百年紀念基金會（Bank of Sweden Tercentery Foundation）的經費支持，希望能建置一個以篇章（discourse）為單位的瑞典手語語料庫，同時作為瑞典手語辭典擴充的參考。SSLC 語料庫的語料來源為 42 位 20 至 82 歲、以手語為母語或達母語程度的發音人，共錄製 24 小時的語料，切分成 300 個檔案。(Wallin & Mesch, 2015) 語料內容為自由對話、敘事、以青蛙故事（Frog, where are you?）及雪人故事（The snow man）為基礎錄製的看圖說故事內容。(“Swedish Sign Language Corpus Project”, Apr. 25, 2018) 其中，約有 23%（69 個檔案；4 小時 48 分鐘）的語料有瑞典詞彙（gloss）標記及翻譯，17%（52 個檔案；4 小時 46 分鐘）的語料僅有瑞典文翻譯。(Wallin & Mesch, 2015)

觸覺手語語料方面，約有 4 小時、55 個檔案的語料，由 8 位 38 至 77 歲的聽視障發音者提供。手語學習者語料方面，約有 9.5 小時、164 個檔案的語料，由 17 位 38 至 77 歲的聽人發音者提供，為斯德哥爾摩大學的手語二語學習者。(Wallin & Mesch, 2015)

在 STS-korpus 網站上，使用者可輸入瑞典語，亦支援星號（*）等萬用字元的搜尋方式，例如：「korpus:sslc rad:“glosa_dh *” pek* 」指的是在 SSLC 語料庫中，搜尋語料標記分別為任一主手（dominant hand）及指物（pointing）的手語語料。目前 STS-korpus 的標記項目有：手語影像提供者 1 的主手（Glosa_DH S1）、手語影像提供者 2 的主手（Glosa_DH S2）、手語影像提供者 1 的副手（Glosa_NonDH S1）、手語影像提供者 2 的副手（Glosa_NonDH S2）、手語影像提供者 1 的瑞典語翻譯（Översättning S1）、手語影像提供者 2 的瑞典語翻譯（Översättning S2）。在播放語料時，螢幕上的標記區塊會與影像同步顯示，使用者可按暫停或放慢播放速度（0.25 倍速、0.5 倍速及原速）。

在網頁介面上，以關鍵字檢索句（KWIC, keyword in context）的方式呈現，並附有上述的語料標記。此外，使用者亦可勾選查看更多資訊，例如：語料庫來源（Korpus）、該檢索句的時間長度（Längd）及起迄時間（Start, Slut）、語料來源檔名（Radnamn）、檔案敘述（Filbeskrivning），未來希望可以依照這些資訊的選項排序。在語料標記區塊的下方則標示了符合搜尋結果的時間，因此使用者可在特定檔案中縱覽時間軸上的搜尋結果，此一設計亦可看出該搜尋結果的分布狀況（dispersion）。

STS-korpus 的語料亦加入瑞典手語辭典（Svenskt teckenspråkslexikon, Swedish Sign Language Dictionary）的資訊。使用者可在主手與副手的標記欄位中點選欲了解的詞彙，便會跳出畫面顯示

該詞彙在瑞典手語辭典的影像檔、釋義、以及速記符號，在此畫面的最下方則有兩個連結，分別連結至語料庫中該詞彙的語料（Öppna i korpus）、或連結至該辭典的完整說明頁面（Öppna i lexikon）。另外，在辭典查詢設計上，也可連結回語料庫，是雙向設計的概念。

自建置 SSLC 語料庫時，便使用 ELAN 軟體作為語料處理的工具，進行語料標記及資料統計。在資料庫設計上，STS-korpus 的標記檔亦為 ELAN 軟體的 .eaf 檔案。網頁設計的前端語言為 Javascript，使用 Vue.js 及 Buefy 框架；後端語言為 Python，使用 Flask 框架；資料庫選擇為 MariaDB，並使用 JSON 檔作為前後端 API 串接的格式，至於.eaf 檔案的處理則借助 Python 中的 SQLAlchemy 與 pympi-ling 套件，但並無開放原始碼，可向原作者請求。

由於 STS-korpus 的建置時間較晚，擁有較多的資源與技術進行手語語料庫的建置，但在內容規劃方面，STS-korpus 亦將語料庫與辭典結合，並進一步提供雙向搜尋的功能。整體而言，STS-korpus 整合多個語料庫及提供語料上傳的設計具有未來性，可長遠地增添語料，但前提是共同的標記檔案格式，像是手語語料庫多採 ELAN 軟體的.eaf 副檔格式，但標示項目可隨不同語料庫來源而顯示不同的資訊。(Öqvist et al., 2020)

參、　國外群眾外包與語料收集機制的分析

在 2006 年《連線》雜誌的一篇文章中，豪伊（Jeff Howe）創造出了群眾外包（Crowdsourcing）此一新名詞，並將其定義為“. . . the act of taking a job traditionally performed by a designated employee and outsourcing it to an undefined, generally large group of people in the form of an open call.”，即「將過去由特定職員完成的工作，公開地交由不固定的一大群人完成。」此一概念其實最早在 1714 年時就已出現：英國政府當時為徵求判斷海上船隻位置的簡單方法，提供現金獎賞，公開向大眾求取不同想法。而在科技發達的今日，群眾外包借助網路無遠弗屆、跨越時空的力量，廣徵各方人士的貢獻能力或資料，亦已在商業、學術研究等各領域中成就了許多了不起的功業，最著名的群眾外包例子即為我們所熟知的維基百科(Wikipedia) —— 一本由全世界網民共創的百科全書。

過去十年來，群眾外包被大量地用在資料的收集、標記、合併，和其他人類智能作業(Human Intelligence Task, HIT)中。在這些作業當中，和語料庫最直接相關的則為資料收集(Data Collection)和資料標記(Data Annotation)。由過去的經驗看來，使用群眾外包方式完成這些工作的優點主要有以下三點：(1) 相較於由一小群員工完成一份大計畫，將其切割成若干小任務，分配給不特定的社會大眾完成，更能有效節省工作時間。(2) 群眾外包能使用相較於聘用專家或專業技術人員更低的成本，遠端獲取更多元、更詳盡的資料或產品。然而，另一方面，群眾外包亦有其缺點所在 —— 除了難以控管作業及作業成果的品質以外，甚或可能遇到有心人士混入作業人群中，蓄意妨礙作業進行。因此，使用

群眾外包進行語料庫作業時，須擬訂出一套完善的授權機制，以使成本和收益能盡可能對等。

將群眾外包應用至語言學領域， Munro et al. (2010) 透過亞馬遜 MTurk（Mechnical Turk）平台，蒐集下列主題的資料，包括：動詞短語的語意透明度（transparency of phrasal verbs）分析、語音檔案的切分，（segmentation of an audio speech stream）、語境預測（contextual predictability）、語法知識判斷（Judgment studies of fine-grained probabilistic grammatical knowledge）等，並將群眾外包的結果與實驗受試者相比，針對答題分佈是否有較多樣的回答、正確率與穩定性進行討論，認為群眾外包是值得推廣的方式，尤其是在已蒐集的資料呈現偏態（skew）之時。

此外，在亞馬遜 MTurk 平台上，潛在的語料提供者可以註冊登入該平台，平台有各式的外包樣板供外包方選擇，收費最低為 0.01 美元，且其中 20%為亞馬遜所有。外包方可以列出報酬、任務截止時間、所欲尋找的提供者背景，不同的任務可能會有不同或額外的支出。在 2010 年時平均每小時報酬為 5 美元。Ortega-Santos (2019) 透過 MTurk 平台蒐集西班牙語系國家的語料，該計畫利用兩週的時間蒐集了 269 位提供者的語料，並逐一分析提供者的所在國家別、第一語言背景、其他會說的語言、教育程度、提供者所在地區的人口數、每週工作時數、工作型態（學生、兼職、全職、退休、身障或無業）以及在 MTurk 獲得的酬勞對他們而言意義為何（此題與我無關 irrelevant to me、感覺不錯，但不一定能夠改善[經濟]狀況 nice, but doesn't necessarily change my circumstances、有時候能讓我有基本的收支平衡 sometimes necessary to make basic ends meet、總是能讓我有基本的收支平衡 always necessary to make basic ends meet），對群眾外包的語料提供者結構進行詳細的探討。

然而，與實驗室設計相較，群眾外包因為無法實際接觸到語料提供者，如果發生語料與大部分資料相差甚遠時，無從得知原因為何，無法知道語料提供者當時的精神狀況與是否瞭解題意，但群眾外包可以大幅縮減語料蒐集所需的時間，且語料來源可能更加廣泛。除了MTurk 平台的綜合介紹，以下僅列舉芬蘭、美國國家語料庫、Mozilla公司《同聲計劃》、英文方言 App、當代威爾斯國家語料庫作為群眾外包案例之參考。

3.1. 群眾外包—芬蘭

和芬蘭國家語料庫建設相關之計畫主要有 FIN-CLARIN 和 National Digital Library (NDL)這兩項，以下將分別介紹芬蘭利用群眾外包來收集資料的方法：

3.1.1. 芬蘭常用語言資料建設計畫（Finland-Common Language Resources and Technology, FIN-CLARIN）

CLARIN（Common Language Resources and Technology Infrastructure，常用語言資料建設）為歐盟的 ESFRI（European Strategy Forum on Infrastructures）底下 34 支子計畫的其中之一；而 FIN-CLARIN 則是芬蘭政府參照 CLARIN、並以和其整合為目標所開展的語言資料庫計畫。FIN-CLARIN 當中包含了 20 項子計畫，其主要目標如下：

(1) 設立關於資料的標準（格式等）和運用方法。

(2) 從其他的來源取得所需資料。

(3) 取得並運用各種文字、語音形式的資料，並尋找、研發使用的工具和方法。

(4) 處理授權議題，使語料庫之成果能夠進一步被應用在日後相關領域之活動或學術研究之上。

此外，FIN-CLARIN 也提出了一些在執行計畫時可能會遭遇的問題，還有其解決方法：

(1) 即便有數位化過後的資料，也很難得知資料之所在，因此需要 metadata 來進一步分析。FIN-CLARIN 底下有許多存放 metadata 的資料庫，如 Meta-Share 以及其他 metadata 資料庫等等：https://www.kielipankki.fi/tools/。

(2) 即使找到了資料，也很難得到使用之許可權，因此需要一個部門/單位專門處理、接洽授權等相關事項。FIN-CLARIN 有一系列協議都在處理授權之相關事項，其網站上以及協議書中亦有對於各種使用方法以及引用情況的詳細定義；使用者也須得先申請方得使用語料庫當中的資源。

(3) 即使得到使用許可，也可能有資料之間彼此格式不相容，或者無工具可處理資料的狀況，因此需要擬定一規定資料格式以及介面之標準、還有研發處理資料相關的工具。Tools 頁面中有提供一些外部連結，供使用者選擇需要的工具做使用，如 Sparv（標記語料用，且不限語言）。

3.1.2. 國家數位圖書館計畫（National Digital Library, NDL）

此計畫的重點在於透過數位化的方式，長久保存各種文化、科學類相關的資料、數據，並確保有需要者（研究人員等）能夠藉此更輕

易得取用這些資料。此計畫底下並有 Finna，作為一整合、保存、管理、維護芬蘭國家圖書館、博物館內館藏資料之服務。其主要作業如下：

(1) 藉由數位化的方式，整合各處、各領域、各類型的資料，收錄在一處，並記載資料本身、資料出處、資料所有者……等資訊。

(2) 招募相關領域專家成立維護系統、數位化資料的工作團隊。

(3) 資料系統之維護 (the maintenance of standard portfolio)。

(4) 訂立和以上工作相關之規定、指導原則。關於 NDL 的詳細說明：此連結包含 NDL 的使用方法、介面說明、資料類型以及結構等等說明。

3.1.3. 群眾外包語料庫成果分析：以 FIN-CLARIN 為例

在 FIN-CLARIN 的計畫底下，一共有 208 個語料庫，皆為主題較細的語料庫，如 Corpus of Age-related Voice Disguise (AVOID)、ArkiSyn Database of Finnish Conversational Discourse、 HelsinkiKorp Version 等等。以下僅就 FIN-CLARIN 取得語料授權之方式及其授權予終端使用者之規則、及其管理語料庫之分類方式作介紹。

授權方式而言，FIN-CLARIN 提出了三種不同的方式授權語料庫予使用者做後續使用：(1) PUB（公眾皆可申請使用）、 (2) ACA（僅限學術使用）及 (3) RES（須經特殊申請認證）。使用者在申請取得使用權時，須依照自身之需求及身份等，選擇不同的授權方式向語料庫方做申請使用，待語料庫方核可後，始取得語料之查詢及使用權。另一方面，關於語料庫方收集語料的部分，FIN-CLARIN 官方則採取開放公眾自由捐贈資料的方式，賦予資料所有者決定該資料適用 PUB、

ACA 或 RES 何者授權，惟捐贈資料者須先確認資料為其所有、或者確保資料捐贈不會產生任何法律問題，且依照 FIN-CLARIN 官方訂定的上傳格式及方法，方得完成授與資料的程序。

服務狀態則的則是語料庫中的語料是否為可取得之狀態，一共分為 A、B、C 三類：A 狀態為資料仍處於待激活狀態，須經語料庫方做進一步的處理（格式修訂、疑難排解等）方能提供給使用者使用；B 狀態為資料使用者須先要求存取，並待語料庫方核可後開放資料方得做後續使用；C 狀態則為資料已處於可供使用者取用之狀態。FIN-CLARIN 官方收錄之部分語料庫之頁面如下圖：

Abbreviation	Name and metadata	License	Apply	Location	Service level	Help	Cite
acquis-ftb3	The Finnish Sub-corpus of the JRC-Acquis Multilingual Parallel Corpus	PUB		Korp	B		
agricola-v1-1-korp	The Morpho-Syntactic Database of Mikael Agricola's Works version 1.1, Korp	PUB		Korp	B		
ai2d-rst	AI2D-RST: A multimodal corpus of 1000 primary school science diagrams	PUB		Download	B		
aku-egg	Speech and EGG (Electroglottography) Simultaneous Recordings	ACA		LAT	B		
amph	amph-Corpus	ACA		Download	B		
ArkiSyn-korp	ArkiSyn Database of Finnish Conversational Discourse, Helsinki Korp Version	PUB		Korp	B		
AVOID	Corpus of Age-related Voice Disguise (AVOID)	RES		Download	B		
BeserCorp	The Corpus of Beserman Udmurt	PUB		Korp	B		
ceal-dl	The Downloadable Version of Classics of English and American Literature in Finnish	RES		Download	A		
ceal-o	Classics of English and American Literature in Finnish, Sentences and Paragraphs in the Original Order	RES		Korp	A		
ceal-par-s-korp	Classics of English and American Literature as translated by Kersti Juva, English-Finnish parallel corpus, scrambled, Korp	ACA		Korp	A		

圖 10. FIN-CLARIN 收錄之語料庫列表頁面

在 FIN-CLARIN 計畫底下之各語料庫皆受 FIN-CLARIN 計畫之管轄，因此皆適用上述之授權許可方式及服務狀態等分類。然而，這些語料庫雖然適用相同的分類原則，但是因主題不同、各自規模不一，範圍從 484,010 tokens 至 13 sentences 不等，主要語言也有芬蘭語和英文

等不同種類。縱然如此，就其中規模較大的語料庫（如 ArkiSyn Database of Finnish Conversational Discourse 等）而言，FIN-CLARIN 仍可說是以群眾外包之方式，建造了為數不少的、具相當規模的語料庫；除此之外，一如先前專家會議中所得出的結論，透過群眾外包，語料庫方可蒐集到更多元、更平衡之語料，避免語料庫建成之後只適合少數特定族群（如學術研究者等）使用。

以此作一小結，本團隊認為，於建置語料庫而言，群眾外包雖仍有其需要注意之處，然其不失為一可有效之蒐集不同來源之語料的管道，因此建議日後執行語料庫建置團隊可採用此方法多方蒐集語料，藉此達成平衡語料庫之目的。

3.2. 群眾外包—美國

美國國家語料庫（American National Corpus）是個運用協作開發計畫（Collaborative Development Project）來收集資料的語料庫，語料庫的語料仰賴語言學學者和一般民眾等主動提供或進行加註整理，若學者或民眾想要提供語料或針對語料進行編註的話，可以遵從網站上的指示，對語料庫提供貢獻。該語料庫底下設有 OANC 和 MASC 兩個子語料庫，分別收錄了語料本身和其註釋資料，以下將分別介紹 OANC 和 MASC 利用群眾外包來收集資料的方法：

3.2.1. 美國開放國家語料庫（Open American National Corpus, OANC）

3.2.1.1. 貢獻語料的流程（Contribute Texts）

(1) 閱讀許可協議，並確認語料內容格式是否符合 ANC 所訂定的條件。許可協議的原文和中譯，還有貢獻語料的條件如下：

許可協議
Grant of license

本人在此同意將所貢獻之文件全球性、永久性、買斷式地授權與美國國家語料庫計畫，以電子形式或未來開發之任何媒體形式來使用、重組格式、再製與散佈。本人知悉所貢獻之文件將作為美國國家語料庫網站其美式英語語言資料的一部份，這些資料將提供給他人作為語言學之教學、研究、商業與非商業開發等目的之使用。
By contributing my document through the ANC web page, I hereby grant to the American National Corpus project a worldwide, perpetual, royalty-free license to use, reformat, reproduce, and distribute, in electronic form or any and all media hereinafter developed, my submission as part of a collection of American English-language material. I understand that the collection will be made available to others for the purposes of linguistic education, research, and development, including commercial development.

（請注意，本許可協議並不會將您的著作權轉讓給美國國家語料庫。）
(Note that this license does not assign copyright to the ANC.)

不過，關於 ANC 的許可協議，葉茂林委員也提醒道：「英美法系採取契約自由原則，與我國大陸法系多有國家公權力介入，如遇糾紛多採有利一般公眾或消費者之解釋不同，此節差異建請留意。」

(2) 登入 ANC 網站，並至 **the upload page**。

(3) 填寫使用者基本資料（如，年齡、性別、國籍和種族），還有關於欲貢獻語料的基本資料，以作為研究美式英語的參考。如果需要的話，貢獻者欄位可填寫匿名。

(4) 假如欲貢獻的語料之前曾經被出版發行過，請填寫相關資訊，若無，則可忽略此步驟。

(5) 按照網站說明上傳文件，如果要上傳多個檔案，可以將它們放在一個文件夾中並上傳。

3.2.1.2.語料的條件

欲貢獻的語料必須符合下列各項條件，才會被 ANC 採用。以下簡單說明 ANC 網站明訂的各項條件。

(1) 語料內容：包括所有類型的已出版和未出版的書面和口頭（轉寫）語料，如小說、非小說、詩歌、報紙、雜誌、期刊、小冊子、日記等，以及網路上的語料，像是部落格、網頁、tweet、電子郵件、饒舌歌詞等等。

(2) 資料類型：

a. 必須是 1990 年或之後的語料。

b. 作者/發言人必須是美國英語的母語使用者 (Who qualifies as a native speaker of American English?)。

c. 貢獻者必須擁有這些語料的著作權，或者這些語料必須是公共領域的（請參閱 著作權問題）。

d. 語料各別文件應不少於 1000 詞。

e. 文檔應主要由語言資料所組成，即檔案中盡量不要包含表格、公式、圖像等。

(3)檔案格式：由於 ANC 主要是採取自動處理文檔的形勢，因此 ANC 有可能會主動剔除自動處理非常困難的文檔。以下是一些網站建議的檔案格式：

a. 使用格式正確的 XML 標記，並使用“標準”詞彙，例如 XCES，TEI 或 DocBook。
b. 屬於 Word doc 或 docx 文件或 rtf 文件，而且不論是使用 Word 內建的或個人擁有的字體樣式，請盡量保持一致。
c. 使用格式正確的 XHTML 標記，並使用“嚴格的” XHTML DTD。
d. 這些文檔建議是「純文字檔」，各章節標題和段落之間需空行，另外也建議使用 UTF-8 或 UTF-16 格式。
e. 語料是用 HTML 手工標記的，意即不是由 Dreamweaver 或 FrontPage 之類的網頁生成程序生成的。

另外 ANC 也提出了一些網站不建議，但尚可處理的格式：

a. 由 FrontPage、DreamWeaver 等程式生成的 HTML 檔案。
b. PDF 檔。

最後是一些 ANC 網站完全無法處理的資料格式：

a. 採用 Quark、InDesign 或其他“出版”軟體格式（“publishing” software format ）
b. double-column 的 PDF 檔。
c. 使用了非常不標準的字體的檔案。

另外，有關 ANC 網站的一些常見問題可參考以下網址：http://www.anc.org/contribute/texts/faq/；而有關美國著作權相關議題可參考 Brad Templeton「著作權簡介」、「有關著作權的十大迷思」這兩篇部落格文章，網址分別如下：https://www.templetons.com/brad/copyright.html、https://www.templetons.com/brad/copymyths.html。

3.2.2. 美國人工標記子語料庫（Manually Annotated Sub-Corpus, MASC）

倘若任何人想對 ANC 網站上的語料貢獻註釋的話，可以先在網站上下載語料和標記工具，再按網站建議的格式進行加註。語料可在以下連結取得：https://www.anc.org/data/masc/downloads/data-download/；標記工具可在以下連結取得：https://www.anc.org/data/masc/downloads/tool-downloads/。ANC 網站並沒有特別明訂某一種資料或標記的格式為標準格式，而是盡可能提供各種不同的格式供使用者自行選擇，這些資料格式包括 CoNLL IOB、XML、UIMA 等；語言學標註格式包括 Penn Treebank 句法註釋、WordNet 詞義註釋、FrameNet 語義框架註釋等；詞性標註格式包括 Penn、CLAWS5、CLAWS7 等詞性集。豐富的資料或標註格式使得美國國家語料庫有別於其他英語相關語料庫。最後，如果加註過程中有

任何問題，或是已完成加註的話，可以寄信至 anc-contrib@anc.org 進行聯繫。

3.2.3. 群眾外包語料庫成果分析：以美國國家語料庫為例

美國國家語料庫的最大特色有兩點，首先它是個運用協作開發計畫（Collaborative Development Project）來收集資料的語料庫，另外該語料庫也提供各種不同的資料和標註格式讓使用者自行運用。運用群眾外包的方式不但有助於美國國家語料庫省下不少時間與經費成本，也能收集到更多元的語料。此外，美國國家語料庫所提供的多種資料和標註格式的模式增加了語料庫的多元度與自由度，而且也有助於各種統計語言模型（statistical language model）的開發。

3.3. 同聲計畫（Common Voice by Mozilla）

3.3.1. 計畫簡介

「同聲計畫」是 Mozilla 為開發語音轉文字和文字轉語音引擎及訓練模型、輔助其下另一語音辨識技術開發工作——「深度語音辨識」(Deep Speech) 專案——所推出的計畫。Deep Speech 為精確處理人類語音的開源語音辨識引擎模型，於 2017 年 11 月釋出；同聲計畫則自 2017 年 7 月開始啟動，與 Mycroft、Snips.AI 以及威爾斯的 Bangor 大學等新創企業或校園夥伴進行語音收集與技術合作，目標是建立一全球化之開源語音資料庫，以收集用於廣泛訓練語音辨識技術的聲音數據，至今已有共超過兩百位開發者參與計畫的軟體開發。

3.3.2. 計畫規格

同聲計畫的最終目的是大量收集可用於訓練人工智慧語音辨識之語言資料，故每一種語言約需蒐集來自不同人、共計 10000 小時左右的錄音檔，方能訓練出完備之語音辨識系統。而到目前為止，同聲計畫已經募集了來自超過四萬多人所貢獻的聲音、27 種語言音檔的收集計畫，另外還有高達 72 種語言的錄音計畫正在進行中，成為同種類語言資料庫中最大的開源語言資料庫。而在 Common Voice 資料集和其他資料源的輔助下，Deep Speech 在技術上已經能夠以人類的精確度即時（即在語音串流的當下）將語音轉譯為文字。

圖 11. Common Voice 語言計畫規格：以臺灣腔華語為例

圖 12. 同聲計畫中目前有 27 種語言的收集計畫已正式上線，另 72 種語言收集計畫正在準備中

3.3.3. 運作方式

在 Common Voice 現有之語音資料集中，每一筆語音資料係來自貢獻者（註：貢獻者還可選擇提供年齡、性別和腔調等後設資料，以便提供更多的語音片段標籤予訓練語音引擎使用。）自願讀出一系列由他人捐贈的語料庫文句，將其錄音後、使其進入所謂「聆聽佇列」，等待接受其他志願者的聆聽，確認說話者是否正確地讀出了該語句。若有兩位以上驗證者投下「正確」票，就會標示為有效資料、並且正式進入到 Common Voice 的「資料集」當中，幫助開發者打造語音識別工具；若大於兩位驗證者投下「不正確」票，該片段就必須回到佇列

重新來過，而若被退回第二次，則該片段就會進入「片段回收桶」。在此之上，「資料集」和「片段回收桶」當中所有的資料皆以 MP3 格式收錄於 Common Voice 網站當中，採用 CC0 的授權模式（”No Rights Reserved”，即使用資料時不必標記出處）開放予大眾下載和使用。

除此之外，Mozilla 亦根據社群的回饋進行可用性研究，持續改善 Common Voice 之網站藉由設法讓貢獻過程更加有趣，鼓勵更多的人持續貢獻他們的聲音。貢獻者現在可以在錄製和驗證的過程中，看到每種語言的進度，並改善了移動到剪輯片段的提示。貢獻介面增加了審查、重新錄製以及跳過剪輯等新功能，方便貢獻者操作語音錄製，另外，現在也可以創建儲存配置文件，跨多語言追蹤貢獻者自己的進度以及指標。

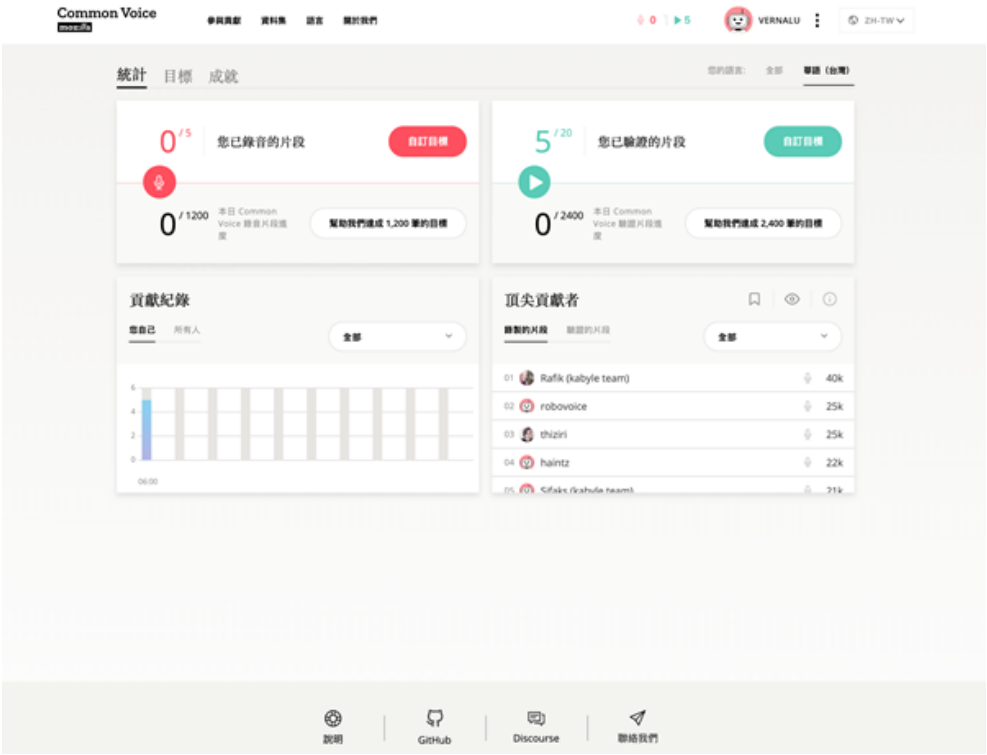


圖 13. 貢獻者在網站上創建帳號之後，就可以擁有自己錄音和驗證的所有記錄

圖 14. Common Voice 音檔資料收集流程

圖 15. 志願者可聆聽他人提供之音檔，協助判定該資料是否可用

圖 16. 志願者朗讀隨機跳出之例句，錄音之後等待他人驗證

3.4. 英文方言 App（The English Dialects App, EDA）& 英文方言 App 語料庫（The English Dialects App Corpus, EDAC）

The English Dialects App 發行於 2016 年 1 月，為一款用以收集英式英語當中不同方言的應用程式軟體，相容於 Android 和 iOS 兩種系統。其特色為完全開放使用者自己參與方言測驗以及錄下自己的音檔，並將此紀錄於其資料庫當中，最後放入 The English Dialects App Corpus，

故亦可視為使用「群眾外包」的方式蒐集語料。以下僅就其取得語料之機制作介紹及說明。

The English Dialects App 將收集語料的過程分為兩步驟：（一）方言檢測（Dialect Quiz）和（二）錄音（Recording）。在方言檢測的程序當中，應用程式首先會先跳出一視窗，讓使用者確認是否要參與檢測，待使用者點擊 "Ok, I'm in!" 按鈕後，方開啟檢測。方言檢測的題目主要聚焦於英語各地方言間的差異（包含語音、詞彙、句型等層面），使用者在檢測的過程中，讀取或聽取自己通常使用的選項，程式並透過最後分數的加總，判定使用者可能為使用哪一地區的方言。下圖為 The English Dialects App 之方言檢測中作為題目的變項：

Variables chosen for the dialect quiz, prompts, example variants, variant count, and variable type.

Variable	Prompt	Example variants	N	Type
Lexical variation in *autumn*	autumn	autumn, fall	3	Lexical
Lexical variation in *splinter*	splinter	spelk, speel	10	Lexical
Lexical variation in *snail*	snail	hodmedod, dod-man	3	Lexical
Pronunciation of ⟨room⟩	room	[ɹʊm], [ɹʏm]	3	Phonolexical
Masculine reflexive pronoun	himself	himself, hisself	2	Morphological
Feminine possessive determiner	hers	hers, hern	2	Morphological
3rd person habitual present	feed	do feed, feeds	3	Morphological
Velar Nasal Plus (cf. Wells 1982) – presence or absence	tongue	[tʌŋ], [tʌŋg]	2	Phonetic
Yod – presence or absence	new	[njʊː], [nʊː]	2	Phonetic
BATH vowel	last	[lɑːst], [last]	3	Phonetic
STRUT vowel	butter	[ˈbʊtə], [ˈbʌtə]	2	Phonetic
C/r/C realization (rhoticity)	arm	[ɑːm], [ɑɹm]	2	Phonetic
#/θ/C realization	three	[θɹiː], [fɹiː]	4	Phonetic
Intrusive /r/– presence or absence	thawing	[θɒːɪŋ], [θɒɹɪŋ]	2	Phonetic
V/l/C realization	shelf	[ʃɛɬf], [ʃɛʊf]	3	Phonetic
KIT/SCHWA in unstressed syllables	pocket	[ˈpʰɔkɪt], [ˈpʰɔkət]	2	Phonetic
/ai/ before voiceless consonants	night	[nɛɪt], [niːt]	8	Phonetic
/ai/ before voiced consonants	five	[fɛɪv], [fɔiv]	7	Phonetic
Presence or absence of /h/	hands	[handz] [andz]	2	Phonetic
CLOTH vowel	off	[ɔːf], [ɑf]	3	Phonetic
MOUTH vowel	house	[huːs], [hæus]	8	Phonetic
FACE vowel	bacon	[ˈbɨəkən], [ˈbeːkən]	4	Phonetic
V/t/V realization	bit of	[bɪd əɣ], [bɪʔ əɣ]	4	Phonetic
HAPPY vowel	happy	[ˈhæpi], [ˈhæpe]	4	Phonetic
Variation in *scone*	scone	[skəʊn], [skɒn]	2	Phonetic
Dative alternation	give it to me	give it me, give me it	2	Syntactic

圖 17. EDA 方言檢測用於出題的變項

另一方面，除了進行方言檢測外，The English Dialects App 也開放使用者錄下自己的音檔、以及聽他人錄下的音檔，以檢測自己是屬於哪一方言群。所有使用者錄音時，讀的皆為從童話故事「狼來了」（The Boy Who Cried Wolf）中所截取的、同一段文字；此段文字並已

先經過審核，確認其與其他可選的片段相比較少有重複的字詞、也因此更具有詞彙及聲韻多樣性。錄音完成後，使用者可回放並聆聽自己的錄音，若有不滿意的地方亦可重錄，一直到使用者滿意為止。接著，使用者會移往下一個頁面，選填如性別、年齡、所居住的城市……等相關個人資料，提供語料庫方作為後設資料（Metadata）使用。下圖分別為此階段使用者錄音、選擇居住地、以及完成一切程序後資料歸檔的程序圖：

圖 18. EDA 錄音階段中的三程序

（註：此階段對於「所居住城市」的預設值為「方言測驗」的結果，然若不符合，使用者可自己點擊地圖上其他地點做更改。）

至 2017 年五月以前，一共有超過 99,000 人下載 The English Dialects App，至 2018 年時資料庫裡則已有超過 50,000 份方言測試結果以及超過 4300 份語音檔案，以成果而言頗為豐碩。然而，其分析報告

指出，這些所收集而來的檔案中，族群的比例並不符合現實情況——也就是說，某些地區、或者某個階級有較高比例的民眾參與使用此應用程式，而造成語料的不平衡。另外，The English Dialects App 亦無提供吸引大眾自發參與群眾外包計畫的誘因（獎品、獎金、或者設計有趣的流程供使用者參與……等），因此在此一面向上無法為我們提供範例。然結論而言，其設計之外包流程仍值得我們作為參考，且另一方面，我們也可以其為借鏡，改善及加強其沒有處理到的面向。

3.5. 當代威爾斯語國家語料庫（Corpws Cenedlaethol Cymraeg Cyfoes, CorCenCC; National Corpus of Contemporary Welsh）群眾外包的方式與流程

威爾斯國家語料庫充分利用群眾外包來收集語料，非常值得參考。運作的流程首先透過 app 取得語料、貢獻者的授權、語料相關詮釋資料（metadata）（由貢獻者所填寫）等資訊，並儲存在檔案儲存系統中；接著會由專業人員進行語料審查、數據整理（data cleaning）、資料度用（data curation）、語料轉寫等工作；待完成後語料庫團隊就會利用由 Steven Neale (2018)等所開發的 CyTag 工具將語料自動標註，標註內容含詞性標註、詞形還原（Lemmatization），所使用的詞性集共有 145 個類別，包含性別、數量、人稱、時態等語法標註。威爾斯國家語料庫的運作流程圖如下：

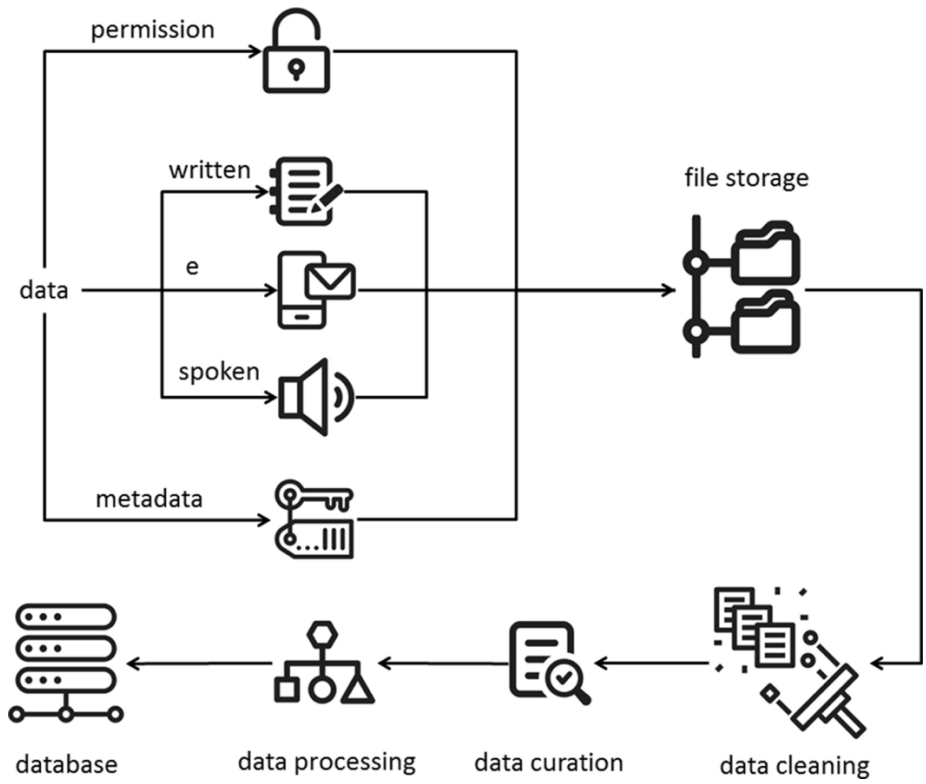


圖 19. 威爾斯國家語料庫的運作流程圖

為了要監測語料庫是否平衡，威爾斯國家語料庫在收集語料的當下就會紀錄每筆語料的詮釋資料（metadata）。威爾斯國家語料庫所收錄的詮釋資料（metadata）細項包括：語料類型（書面、口語或線上）、體裁、是否為自發性語料、是否為翻譯語料、來源、語料的交談對象、語境、對話參與者、地點（如商店、郵局）、地區（如某地地名）、貢獻者基本資料、貢獻者語言能力，詳見 https://link.springer.com/article/10.1007/s10579-020-09501-9/tables/2。所收錄的貢獻者基本資料細項包括：姓、名、性別、出身年、目前居住地、職業（Occupation）、就職情形（Employment status）、法人實體（Legal entity，如個人、公司或組織），詳見 https://link.springer.com/article/10.1007/s10579-020-09501-9/tables/3。所收錄的貢獻者語言能力細項包括：方言別（地區）、在哪裡習得威爾斯語的、語言能力自評、是否將威爾斯語作為第二語言來學習、學習

程度，詳見 https://link.springer.com/article/10.1007/s10579-020-09501-9/tables/4。透過收集詳細的詮釋資料（metadata）資訊，威爾斯國家語料庫便可以收集到各種類型、體裁、使用情境、語言變體等等的語料，這些資料可以反應威爾斯語的真實使用情況，並且作為教學、語料庫研究等等目的來使用！

3.6. 群眾外包之應用分析

從上述的例子中，可發現群眾外包的幾項優劣：(1) 在資料蒐集方面，相較於實驗室資料，發包方較無法掌握語料提供方之背景，亦較無從得知受試時的情形，因此群眾外包有品質不定的風險。(2) 從另一方面來看，群眾外包在資料偏態（skewed）時是解決此問題的重要管道，有些語料無法取得時，便需要尋求其他方式。(3) 在建置語言資源時，可視語料取得難易、專業知識要求等原則，將部份語料蒐集外包給有興趣且有能力的個人或團體，例如：原住民族語的田野調查、手語語料之建置須取得社群支持、語料標記可交由受過訓練的標記人員處理等，依專業與多管道的方式進行。

肆、　國外相關數位典藏計畫、資料格式、與工具的分析

在建置國家語料庫的過程中，資料的數位化（digitalization）是不可或缺的一環，本章選擇澳洲的太平洋區域瀕危文化數位典藏計畫（Pacific and Regional Archive for Digital Sources in Endangered Cultures, PARADISEC）作為介紹，因為該典藏計畫不以澳洲為範圍，更延伸到太平洋區域的語言文化保存，總計有 500 個子典藏計畫，對於資料本身的描述，也就是後設資料（metadata）的建立，值得借鏡。

為了尋求一致的後設資料標準，該典藏計畫採用語言典藏公開群體（OLAC）的所制定的準則。目前世界上兩大主流後設資料標準分別為 OLAC（Open Language Archives Community） 及 IMDI（ISLE Meta Data Initiative），前者奠基於都柏 林核心集（Dublin Core），其他領域亦遵循都柏林核心集的分類，主要由美國所使用，後者是馬克思普朗克學會（Max Planck Institute）所制定，使用者多為歐洲國家，注重多模態（multimodal）的後設資料描述，且比 OLAC 的分類更細。

在 OLAC 的典藏計畫列表上，臺灣是其中一員，中研院的漢語平衡語料庫、近代漢語標記語料庫、臺灣南島語數位典藏是採用其標準的典藏計畫，未來若使用 OLAC 的後設標準分類，在整合上會更有效率，因此以 PARADISEC 數位典藏計畫作為介紹，探討其架構及資料交換的方式。

4.1. 太平洋區域瀕危文化數位典藏計畫（Pacific and Regional Archive for Digital Sources in Endangered Cultures, PARADISEC）

4.1.1. PARADISEC 典藏計畫簡介

澳洲研究委員會下轄的語言活力卓越研究中心（ARC Centre of Excellence for the Dynamics of Language）致力於探索與保存語言的多樣性及演化，並以其數位典藏計劃聞名。太平洋區域瀕危文化數位典藏計畫（Pacific and Regional Archive for Digital Sources in Endangered Cultures, PARADISEC）是眾多典藏計畫的集合，每個典藏計畫亦有不同的使用條款，不過 PARADISEC 計畫特別點出其後設資料（metadata）皆適用創用 CC 授權條款的「姓名標示-相同方式分享 4.0 國際（ShareAlike 4.0 International License）」條款，並需要註冊會員始得進一步下載相關資料與檔案。（註：有關創用 CC 授權條款之介紹，請詳見「4.1.2 創用 CC（Creative Commons）授權條款簡介」小節。）

該數位典藏計畫自 2003 年起由三所大學負責，分別為雪梨大學、墨爾本大學及澳洲國立大學，值得注意的是，澳洲雖然已完成國家語料庫的設置，但原住民語言並非該語料庫的重心，因此大眾在使用該語料庫時，可能無法順利獲得想要搜尋的語言資料，而 PARADISEC 包含豐富的語料，總計 500 個子典藏計畫，超過 1 千 200 個語種。目前的語言已不限於澳洲或太平洋地區的語言，因此未特別將澳洲國內的原住民語言劃分出來，但可依條件進行相關檢索。

在其網站上，可進入目錄頁面，依照搜尋條件找到相關的典藏資料，網址為：https://catalog.paradisec.org.au/collections/search，其中澳洲的典藏資料共有 116 筆，每筆資料都有對應的典藏編號（Collection

ID），點擊連結會出現有關該典藏的詳細描述，包括資料收集者、維護組織與負責人、資料涵括的國家或語種、引用格式與授權使用範圍等。此外，其開源精神不僅展現在資料的豐富程度，該典藏計畫所建構的內容管理系統也是開源的。名為 Nabu，是埃法特語（Efate）「連結各地的道路（road）」之意，Nabu 是 PARADISEC 於 2012 年開發的多媒體的管理系統，主要是讓該典藏計畫擁有一致的後設資料呈現方式，原始碼託管於 GitHub，網址為：https://github.com/nabu-catalog/nabu，並提供 API 及 GraphQL。以下簡單介紹 Nabu 的內容：

Nabu 是太平洋區域瀕危文化數位典藏計畫（Pacific and Regional Archive for Digital Sources in Endangered Cultures, PARADISEC）於 2012 年所建構以用來管理資料的系統，以 Ruby 寫成。全系統除支持 PARADISEC 資料的上傳、下載與管理外，另也遵循公開檔案典藏後設資料協議（Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMI），方便使用者透過一套協議過的後設系統標記有效地管理與使用資料。圖 20 為 Nabu 系統的使用流程圖：

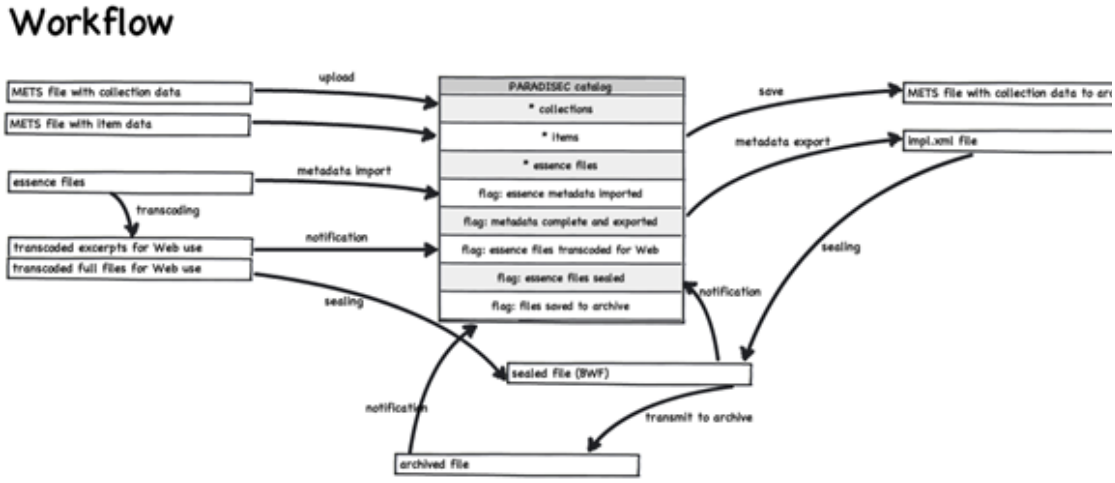


圖 20. Nabu 系統的使用流程圖

於 Nabu 上傳與下載之後設檔案皆為 METS（Metadata Encoding and Transmission Standard）或 CSV 檔，而上傳時資料本身亦需儲存於 PARADISEC 使用者可以取得之處所，方便資料的檢視與取得。語音與影音資料則無論上傳時的格式，一律將透過系統內部轉檔將格式統一。

於本計畫而言，Nabu 作為 PARADISEC 之核心管理系統，所揭示的重點有二：一是後設資料必須完備且符合相關協議，以方便使用者與其他典藏合併查詢、使用。如語言典藏公開群體（Open Language Archives Community, OLAC）（關於 OLAC 的介紹請參閱「4.2 語言典藏公開群體（Open Language Archives Community, OLAC）」一節）、人文網絡設施（Humanities Networked Infrastructure, HuNI）與人際溝通科學虛擬實驗室（Human Communication Science Virtual Lab, HCS vLab）即透過統一的後設資料，將 PARADISEC 融入查詢範圍中（Thieberger

2014）。二來，PARADISEC 與 Nabu 亦展現了統一的格式如何可以作為個人與研究群體之間的中介點，方便使用者在共享所蒐集之資料的同時，亦接軌於同語言的其他資料，透過開放的平台讓個別的田野調查結果得以自然群聚與累積（Thieberger, 2010）。

另外，PARADISEC 亦收藏了 5 個臺灣的典藏計畫[6]，以下分述之：

(1) 編號 AC2，Arthur Capell 的太平洋田野調查筆記，目前收藏於澳洲國家圖書館（National Library of Australia）。

(2) 編號 AIT1，Apay Tang 錄製之太魯閣語音檔，夏威夷大學提供，目前無法取得。

(3) 編號 CLV1，Bert Voorhoeve 錄製之語料（包括敘事、神話、詞表、訪談及對話），澳洲國立大學提供。

(4) 編號 RB1，Robert Blust 錄製之馬來西亞、臺灣、印尼、巴布亞紐幾內亞的音檔，夏威夷大學提供，目前無法取得。

(5) 編號 WL1，Wolfgang Laade 錄製之音樂檔案，大英圖書館音訊檔案室（British Library National Sound Archive）提供。

上述資料的授權方式多為「開放（須符合 PDSC 規範） Open (subject to agreeing to PDSC access conditions)」，該典藏的授權方式共有四種，分別為「尚未標示 as yet unspecified」、「未開放（須符合授權條款規範） Closed (subject to the access condition details)」、「開放（須符合 PDSC 規範） Open (subject to agreeing to PDSC access

[6] 在 PARADISEC 典藏計畫中，無臺灣手語的資料，但在 OLAC 的查詢頁面上共有 6 筆資料，2 筆為有關臺灣手語的描述（language descriptions），編號為 oai:glottolog.org:taiw1241 及 oai:wals.info:tzi，另外 4 筆列為其他資源（other resources），內容是臺灣手語的綜覽報告，編號為 oai:sil.org:9125、oai:sil.org:9053、oai:ethnologue.com:tss、以及 oai:linguistlist.org:lang_tss。

conditions)」、「混合（依據個別資料而有所不同）Mixed (check individual intems)」。資料收集人在一開始上傳其欲貢獻之資料時，能夠決定資料要依哪一種存取狀態呈現給其他典藏計畫的使用者。「尚未標示」即為資料所有者尚未決定此資料此資料的存取狀態；「開放」表示該資料庫底下所有的資料皆是以公開的形式呈現，使用者可自由存取資料，惟須遵從此平台所擬訂之使用規範（詳見 http://www.paradisec.org.au/deposit/access-conditions/）；「未開放」表示該資料庫底下所有的資料須經由有意使用之人申請，經過資料所有者核可之後，方能存許、使用該資料，且資料不受此平台所擬訂之使用條款規範，而是參照資料所有者額外自訂之規範；「混合」則表示該資料庫底下之每筆資料的可存取狀態皆不相同，即有些會是「開放」、有些會是「封閉」，此時則依照每筆資料所顯現之可存取狀態，決定套用「開放」或是「未開放」之使用條款。

整體而言，該典藏計畫的架構方式值得參考，因為其後設資料的欄位完整、詳細，且有說明文件指引使用者將語料納入該典藏計畫。除此之外，該典藏計畫亦提供開源碼與API串接，讓資料能夠更靈活地擷取，而不受限於網頁的頁面呈現，與此同時，卻仍顧及到各個資料的授權範圍，如此一來能夠將龐大的資料檔案相互交流，並確保個別的授權設定。

有關 PARADISEC 典藏計畫的 API 串接之說明文件，可在下列網址取得：https://catalog.paradisec.org.au/apidoc，除此之外，PARADISEC 亦是語言開放典藏社群的一員（Open Language Archive Community, OLAC），與後設資料相關的說明文件存放於下列網址：http://www.language-archives.org/documents.html，以下分別敘述之。

在 PARADISEC 的 API 架構下，可藉由「資料交換的蒐集與服務格式（Registry Interchange Format——Collections and Services, RIF-CS）」取得各個子數位典藏，而欲取得個別的語料，則可使用 OLAC 所訂定的方式擷取資料，從這樣的設計來看，可知後設資料的紀錄和語料內容的取得是分開的兩個流程。RIF-CS 的本質是 XML 檔案，在最上層的是 OAI-PHM 的標籤（tag），接著是 ListRecords，再下一層是一個個的 record，以 header 和 metadata 兩個標籤分別紀錄關於該 record，即各子典藏在 PARADISEC 的編號與子典藏的後設資料訊息，包括：名稱（name）、簡述（description）、授權方式（right）、電子資訊的典藏網址（location>address>electronic）、實體收藏地點（location>address>physical）、引用格式（citation）、原始典藏連結（relatedInfo>identifier）、語料涵蓋地理範圍（coverage>spatial）、語料涵蓋時間（coverage>temporal）等。這些資料不僅能夠以網頁瀏覽器的方式呈現，也可以透過機器可讀（machine-readable）的方式取得。

在取得個別語料的方面，也是透過相似架構的 XML 檔案處理，有 OAI-PHM、ListRecords 及 record 的標籤，不同的是 metadata 下多了 olac:olac 的標籤，如下圖：

圖 21. PARADISEC 使用 OAI-PMH 及 OLAC 架構下的 API 串接內容

圖 21 中的 olac:olac 中記錄了如何從 OLAC 系統中取得特定的語料資料，OLAC 標準是以都柏林後設資料標準（Dublin Core）為基礎，從 OLAC 的說明文件網頁中，也可看到如「篇章類型（discourse type）」、「語言編碼（language extensions）」、「語料類型（linguistic data type）、「語言學子主題（linguistic subject）」、「語料貢獻者蒐集語料時的角色（role）」以及如何宣示這些後設資料的名稱（recommended metadata extensions）的使用詞表（vocabulary），有了這些共同的名稱，就可以進行跨資料庫的檢索與資料的交換，尤其是語料庫涉及的後設資料龐大而複雜，這些後設資料的標準成了建置大型語料庫的重要基礎。

從澳洲國家語料庫與 PARADISEC 典藏計畫的架構來看，由於像英語、華語等優勢語言往往擁有較多的語料，而原住民語言的資料常常會涉及到後設資料的補充，因此在資料呈現上可能需要考量統合或是分開的問題，尤其是當遇到個人資料去識別化的部分，可以先以命名實體識別（named entity recognition, NER）的技術找出涉及當事人的

個人資訊，並由人工檢查，因此如何平衡後設資料的完整與語料提供者的隱私，是建置語料庫典藏的重要課題。

4.1.2. 創用 CC（Creative Commons）授權條款簡介

依據現行之著作權法，著作權人全權保留了一項著作的使用權利，即所謂「所有權利保留」（All Rights Reserved）。也就是說，超出法規中所規定之「合理使用」範圍內，任何人都必須事先取得著作權人的授權，才可更進一步地利用、使用某一資源。這對於資訊之流通、想法之交流、發現之前進等，某種程度上都造成了一定的不便。於是，著名法律學者 Lawrence Lessig 於 2001 年時發起在美國成立 Creative Commons 組織，並提出相對於「所有權利保留」之「保留部分權利」（Some Rights Reserved）的做法。Creative Commons 並透過 4 大授權要素的排列組合，提出了 6 種公眾授權條款，臺灣將此稱為「創用 CC 授權條款」。創作者（著作權人）可以挑選出最適合應用於自己作品的授權條款，標示於其作品上，將作品釋出給他人使用。以下就創用 CC 條款之四個授權要素及六種授權條款做介紹：

四個授權要素

(1) 姓名標示表示：使用者須按照著作人獲授權人所指定之方式表彰其姓名。

(2) 非商業性表示：使用者不得用該作品來獲取商業利益或金錢報酬。

(3) 禁止改作表示：不可變更、變形或修改該作品，僅可重製。

(4) 相同方式分享表示：若使用者變更、變形或修改該作品，則其僅能依相同之授權條款來散布該衍生作品。

六種授權條款

(1) 姓名標示：允許使用者重製、散布、傳輸以及修改該作品（包含商業性利用），惟使用者須在使用時依照著作人或授權人所指定之方式表彰其姓名。

(2) 姓名標示－非商業性：允許使用者重製、散布、傳輸以及修改著作，但不得為商業目的之使用；使用時並須按照著作人指定的方式表彰其姓名。

(3) 姓名標示－非商業性－相同方式分享：允許使用者重製、散布、傳輸以及修改著作，但不得為商業目的之使用；使用時並須按照著作人指定的方式表彰其姓名。此外，使用者僅能依本授權條款或與本授權條款類似者來散布修改該著作後之衍生作品。

(4) 姓名標示－禁止改作：允許使用者重製、散布、傳輸著作（包括商業性利用），惟不得修改該著作；使用時並須按照著作人指定的方式表彰其姓名。

(5) 姓名標示－非商業性－禁止改作：允許使用者重製、散布、傳輸著作，但不得為商業目的之使用，亦不得修改該著作。使用時並須按照著作人指定的方式表彰其姓名。

(6) 姓名標示－相同方式分享：允許使用者重製、散布、傳輸以及修改著作（包括商業性利用）；使用時並須按照著作人指定的方式表彰其姓名。此外，使用者僅能依本授權條款或與本授權條款類似者來散布修改該著作後之衍生作品。

4.2. 語言典藏公開群體（Open Language Archives Community, OLAC）

考量到語言資源的分散與後設資料的不一致（圖 22），語言典藏公開群體（Open Language Archives Community, OLAC）是一個致力於（1）統一語言資源的後設資料（2）將分散於各處的語言資源集中管理，讓使用者方便集體搜索的聯合機構。其資料管理主要建立於兩套標記系統：都柏林核心後設資料組（Dublin Core Metadata Set）與公開檔案典藏後設資料協議（Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMI）。

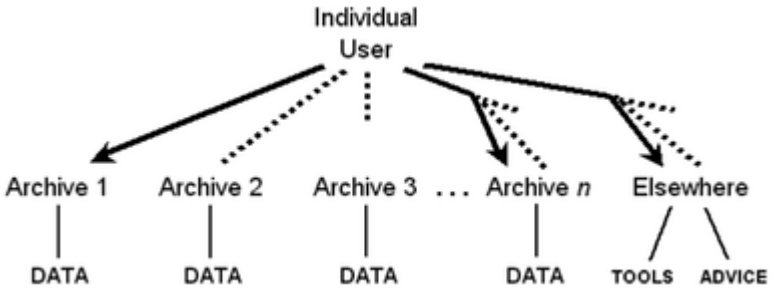


圖 22. 語言資源的分散讓使用者難以查詢 (Bird & Simons, 2003)

在 OLAC 的脈絡中，後設資料的存在很大一部分是為了因應網路搜索的不足。舉例而言：

(1) 雖文字資料即便散播各處亦能透過如 Google 等搜尋軟體取得，然語言資源除文字資料以外，尚有語音與影音等無法被直接搜索出的資料形式。

(2) 如欲搜索個別語言的資源，若拼法的羅馬化不一致（e.g. Fadicca, Fadicha, Fedija, Fadija, Fiadidja, Fiyadikkya, 與 Fedicca），或其名

稱與其他名詞重疊（e.g. Mango 與 Santa Cruz），會造成不有效的搜索。

(3) 許多語言資源也不會有文字敘述使搜索引擎能查得，而是直接存放於資料庫中，方便己身使用。（Bird & Simons，2003）

有鑑於以上問題，OLAC 運用都柏林核心後設資料組與都柏林後設資料協議（Dublin Core Metadata Initiative, DCMI）所進行的一些延伸，開發了以下 15 項語言資源專用標記（Bird & Simons，2003）：

a. 貢獻者（contributor）：貢獻此資源的人
b. 範圍（coverage）：地理與時間上的範圍
c. 創造者（creator）：實際創造此資源的人，如原住民語的發音人
d. 日期（date）：資源創造循環中的重要日期
e. 描述（description）：關於資源內容的描述
f. 格式（format）：數位格式
g. 指稱（identifier）：如網路連結或 ISBN 等明確的資源指稱
h. 語言（language）：資源內容的語言
i. 出版者（publisher）：使該內容公開的出版者
j. 相關資源（relation）：相關資源的連結
k. 相關權利（rights）：此資源的權限人
l. 來源（source）：此資源的來源
m. 主題（subject）：此資源的主題，以關鍵字描述
n. 名稱（title）：此資源的名稱
o. 類別（type）：此資源的類別

OLAC 所開發的此套後設標記系統，在資源描述與資源統合的意義上取得了本計畫可以參照的顯著成果。如 David Nathan 與 Peter K. Austin 所言：”Much of the activity of traditional language description can be understood as creating metadata, ‘data about data’, that can potentially provide indexing, access, annotation, and classification for all data types, including recordings “（Nathan & Austin，2004）。應當在多大程度上採納與仿效其他計畫所開發的標記，以能夠在配合語言典藏群體的同時，亦能夠對於臺灣語言的典藏進行最有效的標記，將會是本計畫重要的課題。

4.3. 都柏林核心集（Dublin Core）

因為有了多項技術的發展，數位典藏愈臻成熟，例如：XML檔案格式、Unicode 編碼，使得資料能夠靈活地被保存下來。都柏林核心集（Dublin Core, DC）是 1995 年時於美國俄亥俄州都柏林制定的跨領域後設資料描述標準，其 15 個類別可從三個面向理解：與資源相關的「內容」資訊、資源的「智財權」資訊以及「例式」，例如：日期（Date）、格式（Format）、識別符（identifier）及語言（Language）(Bird & Simons, 2003)。

值得注意的是，都柏林核心集有幾項在處理資料時會遇到的問題(語料庫建置入門工作流程指南, 2010)，例如：(1)「一對一原則」：同樣的文本可能會有不同的版本，因為創作者（creator）或貢獻者（contributor）不同，故視為不同的典藏項目。(2)「簡化原則」：可不需要修飾語。(3)「適當的資料值」：應審慎選擇相對應的元素讓後設資料能夠發揮最大用處。

4.4. ISO 639 語言代碼國際標準（International Organization for Standardization 639 Language Codes）

有關語言方面的後設資料，以 ISO 639 系列為語言代碼，由 3+2 個字母所組成，前面 3 個字母是 2002 年出版的第一部份，為該語言所屬的主要語言分類，後面 2 個字母是 1998 年出版的第二部分，標示大語言（macrolangauge），另外有 mis 表示「未被編碼」、mul 表示資料包含多種語言、und 標示尚未確定的語言，前後共有六個部分，但由於結構層次不同，可能會有不同的代碼，針對各個臺灣的國家語言，在 ISO 639-3 的分類下可以找到代碼，包括閩南語為 nan，客語為 hak，原住民部分的阿美語再細分了 ami 與 ais（苙蘭阿美語），臺灣手語為 tts 表 8 為蕭素英老師製作的對照表：

表 8. 蕭素英老師製作的語言國際標準代碼對照表 (語料庫建置入門工作流程指南, 2010)

英文名稱	中文名稱	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	備註
Amis	阿美語		ami			
Amis, Nataoran	苙蘭阿美語		ais			
Atayal	泰雅語		tay			
Austronesian	南島語系	map		map		語言集合

Languages						
Bunun	布農語		bnn			
Chinese	中文、漢語		zho	zho/chi	zh	大語言
Chinese, Hakka	客語		hak			
Chinese, Min Nan	閩南語		nan			
English	英語		eng	eng	en	
Formosan languages	臺灣南島語族	fox				語言集合；階層關係 map:fox
Kanakanabu	卡那卡那富語		xnb			
Kavalan	噶瑪蘭語		ckv			
Ketangalan	凱達格蘭語		kae			
Kulon-Pazen	巴宰語		uun			

Paiwan	排灣語		pwn			
Puyuma	卑南語		pyu			
Rukai	魯凱語		dru			
Saaroa	沙阿魯阿語		sxr			
Saisiyat	賽夏語		xsy			
Sign languages	手語	sgn		sgn		語言集合
Sino-Tibetan languages	漢藏語系	sit		sit		語言集合
Siraya	西拉雅語		fos			
Taiwan Sign Language	臺灣自然手語		tss			
Taroko	太魯閣語（賽德克語）		trv			
Thao	邵語		ssf			

Tsou	鄒語		tsu			
Yami	達悟語 （雅美語）		tao			

伍、 本國國家語言相關之語言資料庫

根據第一章所介紹的世界各國語料庫，可收錄作為語言資料庫材料的資料包括語料庫、資料庫、線上辭典、網路論壇、電子報等各種線上資料，還有報紙、期刊、雜誌、文學作品、教材、文宣、論文等各種紙本資料。本章搜尋各種國家語言相關的資料，並依華語、閩南語、客語、原住民語、閩東語、手語這些語言別來區分，接著再分別以線上資料、紙本資料小節來做呈現。以下所盤點的線上資料包括語料庫、資料庫、典藏網、電子報、線上辭典、資源網、電視台網站等；紙本資料則包括學習教材、童話集、故事集、詞典、文學作品、地方志語言相關介紹等。

5.1. 華語

5.1.1. 線上資料

以下為可收錄至國家語言資料庫之華語線上資料，(1)至(5)為語料庫，(6)及(7)為資料庫。(6)可以成為語料庫的材料。

(1) 中央研究院漢語平衡語料庫：為中央研究院資訊科學研究所、中央研究院語言學研究所所建置。總計含有一千萬詞左右。語料庫的語料都經過自動分詞及詞性標記且經過人工校對。網址為：http://asbc.iis.sinica.edu.tw/。

(2) 中央研究院中文詞彙特性速描系統：為中央研究院語言學研究所中文詞彙網路小組所開發、詞庫小組所維護。包含由臺灣、中國大陸、新加坡 14 億字中文新聞語料組成的大型語料庫。建議只收錄臺灣的語料。網址為：http://wordsketch.ling.sinica.edu.tw/。

(3) 國家教育研究語料庫索引典系統：為國家教育研究院語文教育及編譯研究中心所建置。目前國教院書面語語料庫已收錄四億五千八百萬詞以上，口語語料庫一千六百萬詞以及數百萬詞的華英雙語平行語料。語料庫的語料都經過自動分詞。另有華語中介語約一百一十二萬詞。建議收錄書面語和口語語料以及華英雙語平行語料。網址為：https://coct.naer.edu.tw/cqpweb/。

(4) 教育電台華語語音語料庫：由教育電台提供語料，臺北科技大學團隊提供語音辨識技術所建立的華語語音語料庫，已建置 2 千多個小時的語料庫且持續建構中。

(5) 中研院漢語對話語音語料庫（Sinica MCDC）：為中央研究院語言學研究所曾淑娟所創。包含 60 位三大年齡層（16-25 歲、26-35 歲及 36-45 歲）組成的發音人 30 個對話，共 25.6 個小時音檔與文字轉寫檔及標記。目前該資料庫之學術授權需經由中華民國計算語言學學會申請，相關說明可參考下列網站：http://www.ling.sinica.edu.tw/v3-3-1.asp-auserid=20.htm。

(6) 臺灣現當代作家研究資料彙編計畫：由國立臺灣文學館於 2010 年發起的彙編計畫，包括賴和、王拓、吳晟、席慕容、隱地、吳漫沙等百位文學家的年表、珍貴照片、手稿與評論文章，於 2017 年底時已出版 100 冊彙編書籍，回顧時代橫跨日治時期到當代。其實，該彙編計畫是奠基於 2004 年的「臺灣現當代作家評論資料目錄」編纂計畫而成，該計畫委託由財團法人臺灣文學發展基金會執行，當時已收集了 310 位文學家的評論資料條目，累積了十餘萬筆的資料，網址為 http://cw.nmtl.gov.tw。目前該資料以條目的方式收錄在該研究資料庫中，無法下載電子檔案。

(7) 各種華語學習資源：如（a）教育部重編國語辭典修訂本（網址：http://dict.revised.moe.edu.tw/）（b）教育部國語辭典簡編本（網址：http://dict.concised.moe.edu.tw/）（c）教育部國語小字典（網址：http://dict.mini.moe.edu.tw/）（d）教育部異體字字典（網址：https://dict.variants.moe.edu.tw/）（e）教育部成語典（網址：http://dict.idioms.moe.edu.tw/）...等等。

5.1.2. 小結

目前華語相關資源以國教院和中研院所收錄的相關語料庫為大宗。國教院書面語料約有四億五千八百萬詞以上，口語語料約有一千六百萬詞左右；中研院漢語平衡語料庫所收錄的書面語料則約有一千萬詞左右，中研院漢語對話語音語料庫共收錄約 25.6 個小時的口語語料[註 1]，而中央研究院中文詞彙特性速描系統的語料規模雖達 14 億字，但當中包括不少中國大陸和新加坡的語料。在標記方面，中研院採用了 46 個詞類標記，國教院的語料庫提供中研院所建立的 46 個詞類標記，另外也根據華語文教學需要，使用鄧守信教授所建立的 8 大詞類標記。而在應用層面，雖然兩者的應用層面皆不少，不過因為中研院的語料標記較為精細，比較偏向學術研究方面的應用，而國教院則比較偏向教育方面的應用。[7]

[7] 中華民國計算語言學學會網站（http://www.aclclp.org.tw/use_mat_c.php#mcdc）上所列的舊版中研院漢語對話語音語料庫(Sinica MCDC8)簡介提到，語料庫所收錄的聲檔長度約 8 小時，共約 12 萬字左右。雖然目前該語料庫網站沒有更新 25.6 個小時的口語語料約等於多少字數的資料，不過和國教院的一千六百萬詞相比，中研院漢語對話語音語料庫的口語語料應該還是比較少。

5.2. 閩南語

以下為可收錄至國家語言資料庫之閩南語線上資料，(1)至(5)為語料庫，(6)至(8)為資料庫，(9)、(10)為典藏網，(11)為電子報，(12)為線上詞典，(13)至(18)為拼音方案、選字原則、教學資源等各種閩南語相關資源網，(19)則為電視台網站。(6)到(19)的資源當中，(6) 、(7) 、(8) 、 (9) 、(10) 、(11) 、(19)可以成為語料庫的材料，其餘適合成建為資料庫。

5.2.1. 線上資料

(1) 國立中正大學臺灣閩南語口語語料庫：該語料庫為國立中正大學語言學研究所麥傑教授與蔡素娟教授共同主持的一系列國科會計畫之成果，約為 80 萬詞，其中已公開 28 小時錄音之轉寫，總共約 31 萬 5 千詞。網址為：http://lngproc.ccu.edu.tw/Corpus/。

(2) 教育閩南語語音語料庫：由教育部委辦、臺北科技大學廖元甫教授所主持的計畫，收集臺灣各區的閩南語，目標是發展閩南語語音辨識技術並開發應用軟體。為了達到此目標，將邀請發音人以閩南語讀出腳本。總經費是 999 萬 5966 元。該計畫內容為建置以閩南語語音辨識、分析為目的之語音語料庫，至少蒐集完成 200 小時的語音內容，並製作閩南語語音比對、識別等工具軟體。計畫成果未來可釋出供各界自由使用，包括商業性目的。該語料庫目前正在建構中，計畫執行時間為 108 年 12 月 1 日至 110 年 11 月 30 日。

(3) 中央研究院 iCorpus 臺華平行新聞語料庫：是目前唯一有華語與閩南語雙語語料文本的語料庫，該計畫語料的主要來源為網路上的新聞報導，原文為華語，翻譯成閩南語拼音。網址為：

http://icorpus.iis.sinica.edu.tw/，另將網站原始碼託管於 GitHub https://github.com/sih4sing5hong5/icorpus。

(4) 台語文語詞檢索：由台中教育大學臺灣語文學系楊允言副教授所設置。語詞查詢可以選擇由漢羅文本或由全白話字文本搜尋。網址為：http://ip194097.ntcu.edu.tw/TG/concordance/form.asp。

(5) 教育部閩南語詞彙分級計畫：由臺中教育大學程俊源教授建置之閩南語平衡語料庫，目前正在進行當中，完成後將有 100 萬詞，語料經過斷詞處理，並以中研院詞庫小組簡化版的詞性集為基礎進行詞性標記。

(6) 國立臺灣文學館-臺灣民間說唱文學歌仔冊資料庫：為國立臺灣文學館所建置的資料庫，裡面收錄了不少台語唸歌、歌仔冊相關資料，收錄檔案類型為文字檔。網址為：http://koaachheh.nmtl.gov.tw/bang-cham/thau-iah.php。

(7) 臺灣民間文學館：行政院文化建設委員會經費補助之計畫，由元智大學羅鳳珠教授擔任主持人，東海大學胡萬川教授（第一二三期）、中研院范毅軍博士（第三期）共同主持。本計畫將胡萬川教授所採集之臺灣民間故事謠諺數位化（含語料、語音資料），建置「臺灣民間文學館網站」。網站上目前收錄了宜蘭縣、桃園縣、苗栗縣、台中縣、南投縣、彰化縣、雲林縣、台南縣以及高雄縣民間文學集，涵括神話、傳說、民間故事、笑話、歌謠、諺語、謎語等內容，並提供聲音檔與文字檔，聲音檔的部分無法直接在網站上聆聽，需要透過「研究區」聯絡製作團隊並取得帳密後方能使用；而文字檔的部份達到 5,053,989 字的規模，沒有提供華語對照翻譯，用字也非依照教育部所頒布的用字規範。依語言別的字數統計如下表：

表 9. 不分縣市作品總數量與字數統計表(語言別)（資料來源：臺灣民間文學館）

語言類別	作品數量	字數(含註解)
客家話	4,288	1,145,584
客家話、華語	1	147
客家話、閩南話	11	2,113
華語	12	3,936
華語、閩南話	1	170
閩南話	8,797	3,895,727
閩南話、日語	10	3,513
閩南話、客家話	5	1,839
閩南話、華語	1	960
總計	13126	5,053,989

此外，該網站還同時設有「詞彙語意辭典」資料庫、「文學地理資訊」系統，前者針對方言中特殊的詞彙語料提供音標與解釋；後者則將民間文學集語料與 GIS（Geographic Information System）科技作結合，使用者只要輸入如行政區、文體等資訊，即可得知相關語料的確切採集地。「臺灣民間文學館網站」網址為：http://cls.lib.ntu.edu.tw/TFL2010/。

(8) 閩客語典藏：中央研究院在 2002 至 2012 年進行了 2 期 11 年的「拓展臺灣數位典藏計畫」，其中「閩客語典藏」為前述計畫底下之分項計畫「語言典藏」的成果之一。「閩客語典藏」相關計畫共分兩期來完成，第一期計畫（2002 至 2006 年）名稱為「閩南語典藏－歷史語言與分布變遷資料庫」，其內容主要是建置閩南語文獻標記語料庫與語言分布變遷的地理資訊系統，所整理的閩南語文獻包括嘉靖、萬曆、順治、光緒年四種版本的《荔鏡記》，戲曲《同窗琴

書記》、《金花女》、《蘇六娘》，以及清末至現代的閩南語歌仔冊。

第二期計畫（2007 至 2012 年）名稱為「閩客語典藏」，則接續第一期的計畫，大量拓展閩南語和客語的典藏。第二期計畫的內容除了將閩、客語相關文獻加以標記、系統化典藏，還有將閩、客語字典數位化、並建立辭典整合查詢外；另外計畫團隊也到閩客雜居的雲林縣崙背鄉、二崙鄉、新竹縣新埔鎮、苗栗縣後龍鎮、南庄鄉等地進行語言田野調查，並將調查結果結合地理資訊技術，製作成語言分佈調查資料庫。「閩客語典藏」所收錄的閩南語文獻包括《基督要理》（Doctrina Christiana en letra y lengua china）、《廈英大辭典》、《英廈辭典》、《廈門音新字典》、《台日大辭典》、早期羅馬字（白話字）醫學書籍、1885-1968 的《臺灣教會公報》、當代的文本與口語資料等。網站分有中文版和英文版，第一期計畫的網址為：http://cls.lib.ntu.edu.tw/southernmin/index.htm；第二期計畫的網址為：http://minhakka.ling.sinica.edu.tw/bkg/index.php。

(9) 臺灣白話字文獻館： 本計畫為臺灣師範大學臺灣文化及語言文學研究所李勤岸副教授所主持的國科會計畫之成果，以「臺灣教會公報」為焦點，揀選重要內容並將之數位化，於線上公開。網址為：http://pojbh.lib.ntnu.edu.tw/。

(10) 台語文記憶：由台中教育大學臺灣語文學系楊允言副教授所主持的國科會計畫，主要為台語教學與使用之教材文本，例如：台語辭典（含英、日語版）、讀本、及教科書等。另外有文學期刊、教會出版之各種文本、歌集等。大部分資料之主題語言為閩南語。網址為：http://ip194097.ntcu.edu.tw/memory/TGB/mowt.asp。

(11)教育部悅讀越懂閩客語電子報：教育部發行的閩南語和客語電子報、其中閩南語的部分約 50 萬詞，該資料庫的閩南語以漢字書寫，可作為拼音與漢字轉換的參考。網址為：https://epaper.edu.tw/learning.aspx?classify_sn=6。

(12)教育部臺灣閩南語常用詞辭典：約有 13,000 詞，加上附錄資料約 3,000 筆、非語詞的單音字約 3,000 筆、及改版增加約 4,000 筆，共約 24,000 筆。網址為：https://twblg.dict.edu.tw/holodict_new/default.jsp。

(13)臺灣閩南語羅馬字拼音方案：是教育部推廣的拼音系統（臺羅）為標準及提供簡單的介紹。網址為：https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=446。

(14)臺灣閩南語漢字之選用原則：為教育部所公布，內容為說明閩南語漢字使用的習慣、理論、及教育部的建議。網址為：https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=15。

(15)臺灣閩南語推薦用字 700 字詞：為教育部所公布，內容介紹閩南語最常使用的語詞及其推薦使用的漢字。網址為：https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=45。

(16)臺灣閩南語我嘛會每日一詞：為教育部委請國立教育廣播電臺所製作，內容為閩南語每日一詞教學。網址為：https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=494&content_sn=4。

(17)九年一貫台語教學資源網：九年一貫台語教學資源網主要對象為小學生，提供小朋友閩南語學習的補充資料。為了讓小朋友學習更容

易，該網站的特點以注音符號的方式教台語。網址為：http://www.taiwanwe.com.tw/。

(18)本土語言資源網：為教育部所整理製作的網站，裡面整合了閩、客、原等臺灣國家語言的各種學習資源、學習課程、學習活動、學習評量、社群連結等相關訊息。網址為：https://mhi.moe.edu.tw/sidemap.jsp。

(19)公視台語台：為公共電視文化事業基金會的電視頻道之一，前身為2004年7月1日開播的「Dimo TV」、和2012年10月1日更名的「公視2台」。2018年《國家語言發展法》通過後，由政府支持並於2019年7月1日改稱作「公視台語台」，是目前第一個以全台語播出的電視頻道。目前該頻道的節目字幕有全台文漢字、台華夾雜、華語字幕這幾種，若之後能將這些節目的台文、華文字幕皆整理出來，將能提供國家語言資料庫可觀的台語口語語料。網址為：https://taigi.pts.org.tw/。

5.2.2. 紙本資料

以下為可收錄至國家語言資料庫之閩南語紙本資料，(20)為教材，(21)為故事集，(22)至(24)為文學作品(25)為辭典。除(25)適合成為資料庫外，其餘都可以收錄成為語料庫的材料。

(20)教育部《咱來學臺灣閩南語》：共有7冊，包含拼音、語詞、語句、文章等，是教育部委託國立臺灣師範大學規劃研編的網路學習資源，包括音檔。可在教育部網站取得相關資料的 PDF 檔，網址為：http://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=506。

(21)各縣市文化局出版之閩南語故事集：如臺南縣閩南語故事集（胡萬川、林培雅，台南市政府文化局主編），鹿鎮閩南語故事集，台中縣民間文學集等。這些紙本資料需先進行數位化並校正。部分資料如桃園及台中鄉鎮閩客語故事、傳說、笑話、歌謠等之合集，已收錄於中央研究院語言學研究所語言典藏之「閩客語典藏」第二期網站中，網址為：http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa。

(22)教育部歷年來舉辦本土文學創作獎得獎作品集當中的閩南語作品，包含教育部 97 年用咱的母語寫咱的文學／用恩兜个母語寫恩兜个文學：97 年本土文學創作獎得獎作品集、98 年臺灣閩客語文學獎作品集 、100 年教育部臺灣閩客語文學獎作品集、102 年教育部閩客語文學獎作品集、104 年教育部閩客語文學獎作品集。可在教育部網站取得相關資料的 PDF 檔，網址為：https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=481。

(23)文化部本土語言創作及應用補助作業要點所補助的閩南語作品，例如：2020 年 3 月 18 日出版、由蔡雅菁翻譯之《小王子》，該書之出版為文化部「本土語言創作及應用補助出版」計畫之成果，採教育部閩南語常用詞辭典之用字，亦有朗讀音檔。(“說台語的《小王子》！法文譯者蔡雅菁的母語之旅”, 2020; “【藝術文化】大人囝仔上深的感動 經典文學小王子台語有聲版問世”, 2020)。

(24)具有代表性的臺灣閩南語文學作品。儘管此部分作品的漢字和羅馬字與目前教育部建議用字很多不一致，但是其代表性和文化及歷史的價值毋庸置疑，建議收錄。

(25)具有代表性的臺灣閩南語相關字典。如由吳守禮教授的《國臺對照活用辭典》；董忠司、城淑賢主編，陳惠玉、楊蕙菁、楊菁惠、徐曉萍協編的《簡明臺灣語字典》等。

5.2.3. 小結

閩南語因為沒有專責機構負責主導整理與建置語言資料庫，因此目前沒有一個具官方、代表性、較大宗的語料庫。撇除正在建置中的語料庫，若以國立中正大學臺灣閩南語口語語料庫和台語文語詞檢索作比較的話，前者收錄的語料以電台節目的口語語料為主，語料拼音採用教育部民國 87 年公布之「閩南語拼音系統」，語料分詞結果也都經過人工檢測；後者所收錄的語料以書面語料為主，內含不少漢羅夾雜的語料，分詞等主要是透過程式處理。

5.3. 客語

以下為可收錄至國家語言資料庫之客語線上資料，(1)至(3)為語料庫，(4)為資料庫，(5)為典藏網，(6)為電子報，(7)為線上詞典，(8)為資源網，(9)為電視台網站。(5)、(6) 、(9) 及部分（8）的資源可以收錄成為語料庫的材料。

5.3.1. 線上資料

(1) 國立政治大學賴惠玲教授建立的客語口語語料庫：至 2016 年則已收錄 198,877 個字，語料腔調涵蓋四縣、海陸、大埔與詔安等，邀請各年齡層及職業別擔任發音人，亦考慮到發音人的性別平衡。網址為：http://140.119.172.200/hakka/，不過目前網站似乎無法使用。

(2) 國立中央大學臺灣客家語語料庫：由中央大學江俊龍副教授所建立，除了東勢（大埔）腔調，四縣、海陸、饒平及詔安等腔調已各累積了最少 7 萬個字的資料。

(3) 「建置臺灣客語語料庫」：為客家委員會公告之巨額採購案，由政治大學團隊得標，政大英語系教授賴惠玲、資訊科學系教授劉吉軒及新聞系教授劉慧雯等主持。此計畫預計分為 3 個期程，最終於 2022 年底完成語料庫的建置。在語料蒐集上，以逐步擴增的方式，期望能夠在第三期程結束時達到 1,800 萬字書面語料以及 30 萬字口語語料的規模。口語語料將具有音檔、語音與文字對齊的時間訊息、以及一部份言談分析常用的標記，另外也包含各種腔調的平衡。

(4) 客家委員會客語認證詞彙資料庫：客家委員會針對客語能力認證所要求的詞彙製成資料庫，作為線上學習資源，提供給客語學習者，並依照認證難度分為三級（初級、中級、中高級），又按主題整理成 18 個分類，總計 26,925 條詞彙，可在網頁上的「詞彙列表」區塊瀏覽各級詞彙。網址為：https://wiki.hakka.gov.tw/。

(5) 閩客語典藏：中央研究院在 2002 至 2012 年進行了 2 期 11 年的「拓展臺灣數位典藏計畫」，其中「閩客語典藏」為前述計畫底下之分項計畫——「語言典藏」——的成果之一。「閩客語典藏」相關計畫共分兩期來完成，第一期計畫（2002 至 2006 年）名稱為「閩南語典藏－歷史語言與分布變遷資料庫」，其內容主要是建置閩南語文獻標記語料庫與語言分布變遷的地理資訊系統，其中計畫裡所整理標註的文獻亦包含少量的客語資料（如《渡台悲歌》）。

第二期計畫（2007 至 2012 年）名稱為「閩客語典藏」，則接續第一期的計畫，大量拓展閩南語和客語的典藏。第二期計畫的內

容除了將閩、客語相關文獻加以標記、系統化典藏，還有將閩、客語字典數位化、並建立辭典整合查詢外；另外計畫團隊也到閩客雜居的雲林縣崙背鄉、二崙鄉、新竹縣新埔鎮、苗栗縣後龍鎮、南庄鄉等地進行語言田野調查，並將調查結果結合地理資訊技術，製作成語言分佈調查資料庫。「閩客語典藏」所收錄的客語文獻包括《客英大辭典》、《客法大辭典》、傳教士文獻、台中縣與桃園縣等客語民間文學集等。網站分有中文版和英文版，第一期計畫的網址為：http://cls.lib.ntu.edu.tw/southernmin/index.htm；第二期計畫的網址為：http://minhakka.ling.sinica.edu.tw/bkg/index.php。

(6) 教育部悅讀越懂閩客語電子報：教育部發行的閩南語和客語電子報中客語語的部分已累積數十萬詞。網址為：https://epaper.edu.tw/learning.aspx?classify_sn=6。

(7) 教育部臺灣客家語常用辭典：目前該辭典共計有 15,464 筆詞目。該辭典亦提供各地讀音之音檔、詞條的「釋義」、「對應華語」、「近反義」等語意關係，更特別的是附加了其他辭典中相對應的詞目，對於交叉查詢是很方便的設計。該辭典的附錄，則有以四縣、海陸腔為主的主題式詞表以及常用虛詞表等。網址為：https://hakkadict.moe.edu.tw/。

(8) 本土語言資源網：為教育部所整理製作的網站，裡面整合了閩、客、原等臺灣國家語言的各種學習資源、學習課程、學習活動、學習評量、社群連結等相關訊息。網址為：https://mhi.moe.edu.tw/sidemap.jsp。

(9) 客家電視台：於 2003 年 7 月 1 日開播，目前由臺灣公共廣播電視集團所擁有的全客語發音電視頻道（包含四縣腔、海陸腔、大埔腔、詔安腔、饒平腔）。不過，該頻道雖為全客語發音，字幕卻是華語，

倘若未來能將頻道的客語版字幕整理出來，將能提供國家語言資料庫可觀的客語口語語料。網址為：http://www.hakkatv.org.tw/。

5.3.2. 紙本資料

以下為可收錄至國家語言資料庫之客語紙本資料，(10)為教材，(11)、(12)為童話，(13)為故事集，(14)至(16)為文學作品，(17)為辭典。其中（10）到（16）可以收錄成為語料庫的材料。（17）可成為資料庫。

(10)教育部《客家語部編版分級教材》：為教育部召集「部編版客語分級教材編輯委員會」所編輯的教材，該套教材共分為四縣腔、海陸腔、大埔腔、饒平腔、詔安腔、南四縣腔六個版本，每個版本皆有九冊教材。關於這些教材的電子書、有聲教材、教師手冊等資源可在國家教育研究院《客家語部編版教育資源》網站取得，網址為：http://hakka.naer.edu.tw/hakka/。

（11）《客家話小王子》是由徐兆泉老師擔任編譯，在2000年所出版的寓言式童話書籍，本書採用苗栗腔客家話，並以「通用拼音」來標注。目前國家圖書館存放有該套書籍。

(12)《安徒生童話全集》客語版由謝杰雄老師擔任總編輯，分「四縣腔、海陸腔」，並有華語對照。是目前少數有客語華語對照的文學作品。目前國家圖書館存放有該套書籍。

(13)各縣市文化局出版之客語故事集，如東勢鎮客語故事集（胡萬川主編台中縣文化局出版）。部分資料如桃園及台中鄉鎮閩客語故事、傳說、笑話、歌謠等之合集，已數位化收錄於中央研究院語言學研究所語言典藏之「閩客語典藏」第二期網站中，網址為：

http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa。其他紙本資料則需要經過數位化並校正後才可收錄。

(14)教育部歷年來舉辦本土文學創作獎得獎作品集當中的客語作品，包含教育部 97 年用咱的母語寫咱的文學／用恩兜个母語寫恩兜个文學：97 年本土文學創作獎得獎作品集、98 年臺灣閩客語文學獎作品集 、100 年教育部臺灣閩客語文學獎作品集、102 年教育部閩客語文學獎作品集、104 年教育部閩客語文學獎作品集。可在教育部網站取得相關資料的 PDF 檔，網址為：https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=481。

(15)文化部本土語言創作及應用補助作業要點所補助的客語作品。

(16)具有代表性的臺灣客語文學作品。儘管此部分作品的漢字與目前教育部建議用字未必一致，但是其代表性和文化及歷史的價值毋庸置疑，建議收錄。

(17)具有代表性的臺灣客語相關字典。如中原週刊社出版的《客話辭典》，徐兆泉老師編著的《臺灣四縣腔海陸腔客語辭典》，以及由曾彩金老師總編輯、2019 年 12 月才剛出版的《六堆辭典》等等。

5.3.3. 小結

目前客語相關語料庫以客家委員會公告之「建置臺灣客語語料庫」巨額採購案、國立政治大學的客語口語語料庫、與國立中央大學臺灣客家語語料庫為大宗。在語料庫規模上，前者預計收錄達到 1,800 萬字書面語料以及 30 萬字口語語料；另外前者亦重視客語各腔調間的語料平衡，倘若某腔調資料較為不足，也會考慮採用田野調查的方式來補足相關資料。

5.4. 原住民語

以下為可收錄至國家語言資料庫之原住民語線上資料，(1)為語料庫，(2)、(3)為典藏網，(4)為線上詞典，(5)至(7)為推薦新詞、教學資源、影音資源等各種原住民語相關資源網，(8)為電視台網站。(1)、(2)、(3)、(8)及一部份（7）的資源可以收錄成為語料庫的材料。其餘資源適合發展成資料庫。

5.4.1. 線上資料

(1) 台大臺灣南島語多媒體語料庫：原為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），由臺灣大學語言學研究所黃宣範、蘇以文及宋麗梅教授共同主持。後又得到國科會人文學研究中心（2006-2010）及行政院原住民族委員會臺灣原住民族圖書資訊中心（2012-present）經費補助，由宋麗梅教授負責語料蒐集及轉寫，原住民族圖書資訊中心同仁負責典藏技術，在既有的基礎上進行改版、修訂、轉檔與擴增工作。目前語料庫已建置賽夏語、噶瑪蘭語、鄒語、阿美語、薩奇萊雅語、賽德克語、布農語（卓群、郡群）、泰雅語、魯凱語、卡那卡那富語、卑南語等十一族語料，並且持續增加中。原先的網址為：http://203.66.168.190/。因受到駭客攻擊，該網站已遭破壞，但資料仍然存在。

(2) 蘭嶼達悟語口語資料典藏網：為達悟語線上學習平台，方便居住都市的達悟族年輕一代，還有其他想學習達悟語的人來學習。由原住民族委員會委託靜宜大學達悟語研究團隊執行，主持人為何德華、董瑪女，團隊成員包括楊孟蒨、張惠環、郭惠桐、戴印聲、曾佳

瑩、饒承恩、及蘭嶼顧問謝永泉、曾喜悅。網址為：http://yamiproject.cs.pu.edu.tw/。

(3) 臺灣南島語數位典藏（The Formosan Language Digital Archive）：是由中研院語言所齊莉莎所主持的「中央研究院國家典藏數位化計畫」底下的「語言典藏」子計畫之一，其最終目標為建立所有臺灣南島語的語音、詞彙、單句和長篇故事語料等，並加以中、英文翻譯。「臺灣南島語數位典藏」主要包括語料庫查詢、語言地理查詢及書目查詢等三大項目，其中語料庫的語言別有魯凱語（含萬山、茂林、多納、大南、霧台、大武等 6 方言）、雅美語、鄒語（含久美、特富野、達邦等 3 方言）、賽夏語（含東河、大隘等 2 方言）、泰雅語（含賽考利克、澤敖利等 2 方言）、排灣語（含東南部、西北部等 2 方言）、布農語（含郡社群、卓社群、卡社群、巒社群、丹社群等 5 方言）、阿美語（含中部 1 方言）、卑南語（含南王、知本 2 方言）、巴宰語（含巴宰、四庄 2 方言）、卡那卡那富語、沙阿魯阿語、賽德克語（含霧社、春陽、太魯閣 3 方言）及噶瑪蘭語等 14 語族 32 方言的查詢選項，部分語言的方言分類和原民會的版本不同（如，鄒語、賽夏語、排灣語等）。語言地理查詢部分採用地理資訊系統（GIS）技術，讓使用者可以依地圖查詢各語言在詞彙、語音、語法上的異同，進而比對臺灣南島語的同源詞與非同源詞的分佈情形，同時亦將建立有聲檔案（voice files）。最後，書目查詢則提供四類臺灣南島語相關書目查詢，即「臺灣南島語言學書目資料」、「臺灣原住民鄉土文化及母語教材」、「臺灣原住民文學相關書目資料」和「臺灣原住民音樂相關書目資料」。網址為：https://museum02.digitalarchives.tw/ndap/2001/AustronesianLang/formosan.sinica.edu.tw/m/index.html。此外，根據「人文社會

資料庫名錄檢索（HUSSCat）」網站的說明，2014 年底「臺灣南島語數位典藏」永續經營計畫結束後網站就暫時關閉了，因此目前網站上的部分功能無法正常使用。HUSSCat 相關說明網址為：http://husscat.hss.ntu.edu.tw/xmlui/handle/123456789/7709。

(4) 原住民族語言線上詞典：為原住民族委員會從 2007 年開始計畫進行編輯的 16 族線上字典，現在各族的字典都已經完成，只要從網站點進任何一個族語的字典，就可以使用。網址為：https://m-dictionary.apc.gov.tw/。

(5) 臺灣原住民語言推薦新詞：原民會與教育部於 2015 年起，每年公布一批原住民族語新詞，網址為：http://ilrdc.tw/research/newwords/newword106.php。因科技進步與時代變遷，族語亦隨之擴增新詞，以符合日常溝通需要。透過田野調查，列出新詞在 16 族族語的說法，從 105 年度至 108 年度，已有 4 批新詞，族語也打出「族語 開始時尚」的口號。

(6) 族語 e 樂園：臺北市立大學族語數位中心設計製作，原住民族委員會版權所有。裡面提供相當豐富的 16 族族語學習資源，部分族語的資源甚至還有細分成數個不同方言的版本，例如阿美語教材就分南勢、秀姑巒、海岸、馬蘭、恆春五個方言版本。網址為：http://klokah.tw/。

(7) 本土語言資源網：為教育部所整理製作的網站，裡面整合了閩、客、原等臺灣各國家語言的各種學習資源、學習課程、學習活動、學習評量、社群連結等相關訊息。網址為：https://mhi.moe.edu.tw/sidemap.jsp。

(8) 原住民族電視台：於 2005 年 7 月 1 日開播，由原住民族文化事業基金會所擁有。該頻道包含目前法定原住民族 16 族發音的節目，各

語族節目輪流播出，字幕皆採用華語翻譯。倘若未來能將頻道的各原住民語字幕整理出來，將能提供國家語言資料庫可觀的相關口語語料。詳細資訊可參考原民會網站的原視新聞與原視節目專區，網址為：http://www.ipcf.org.tw/。

5.4.2. 紙本資料

以下為可收錄至國家語言資料庫之原住民語紙本資料，(9)為教材，(10)為詞典，(11)、(12)為文學作品。

(9) 原民會《原住民族語言能力認證測驗》網站：為原民會所設之網站，裡面有不少和族語能力測驗相關資源，其中「資源下載」專區可取得原住民 16 族 42 方言的教材 PDF 檔，網址為：http://lokahsu.org.tw/resource/。

(10) 各項由中研院、教會團體、各出版社所出版的詞典。如，由中研院李壬癸院士所發表的《巴宰語詞典》、《噶瑪蘭語詞典》；或是由董瑪女、何德華、張惠環編輯，國立臺灣大學出版中心出版的《達悟語詞典》等。這些書籍可經由書店購得，或者從相關網站上下載取得（如《語言暨語言學》LANGUAGE AND LINGUISTICS）。

(11) 教育部歷年來舉辦本土文學創作獎得獎作品集，包含 96 年原住民族語文學創作獎作品集、98 年原住民族語文學創作獎作品集、100 年原住民族語文學創作獎作品集、102 年原住民族語文學創作獎作品集、104 年原住民族語文學創作獎作品集。可在教育部網站取得相關資料的 PDF 檔，網址為：https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=480。

(12) 文化部本土語言創作及應用補助作業要點所補助的原住民族語文作品。

5.4.3. 小結

目前原住民語相關語料庫以臺灣南島語數位典藏、台大臺灣南島語多媒體語料庫，與族語 e 樂園為大宗，這三個語料庫接收錄多種原住民語言的相關資料。由於缺乏前者的完整資料，我們比較最後兩者的差異，首先族語 e 樂園所收錄的語言別、方言別資料數量更多也更齊全，目前台大臺灣南島語多媒體語料庫收錄了約 11 種語言的資料，族語 e 樂園除了收錄官方承認的 16 族語言資料外，部分語言還額外區分成不同方言別的版本。在語料內容方面，台大臺灣南島語多媒體語料庫主要收錄透過田野調查的方法所採集口語語料，族語 e 樂園則收錄包括閱讀、書寫、歌謠、動畫等各種學習相關教材語資源。在語料書寫與標記方面，台大臺灣南島語多媒體語料庫是參考德國 Max Planck 語言所及 Leipzig 大學語言所共同建置的 Leipzig Glossing Rules 來做標記，而族語 e 樂園因為是比較教育導向的網站，因此網站上的語言資料採用的是原民會所公告的族語書寫系統（http://ilrdc.tw/research/rwview/rwsystem.php）。

5.5. 閩東語

5.5.1. 線上資料

以下為可收錄至國家語言資料庫之閩東語線上資料，目前所找到的資料主要是教育學習相關的資源網站。

(1) 連江縣本土教學資源網：為連江縣政府教育處所設置的網站，裡面收錄各種閩東語的學習教材相關資源，然而目前部分教材所採用的是福州話的標準音，並非馬祖當地的常用讀音。網站上的學習教材資源包括：12 冊國小閩東語課本、幼兒園閩東語教材、31 首童謠、

《齊講馬祖話 120 句口袋書》、日常生活常用詞彙、綜合活動馬祖話、《五語快易通》等，這些教材皆附有線上 pdf 檔和音檔。而其中的《五語快易通》為包含華語、英語、閩南語、金門閩南語、馬祖閩東語、四縣客語、海陸客語、大埔客語、饒平客語、詔安客語、長樂客語、阿美語、泰雅語、布農語、卑南語、排灣語共計 16 種版本的語言/腔調的對照課文，使用者只要點選不同版本的課文，就可以聆聽比對同一句句子的不同語種/腔調的念法。此外，網站也提供國小閩東語教師手冊 pdf 檔下載，還有馬祖閩東語本字檢索系統(試用版)來查詢常用詞彙。連江縣本土教學資源網網址為：http://www.matsudialect.org/。

(2) 馬祖方言天地：為台大中文系博士陳高志老師在 2011 年 8 月 6 號所架設的網站，網站上有老師所寫的各種關於馬祖閩東語的介紹、講解與糾錯，另外也有一系列「方言講古」的文章，專門介紹馬祖當地相關的歷史文化資訊，網址為：http://mypaper.pchome.com.tw/matsuren。

5.5.2. 紙本資料

以下為可收錄至國家語言資料庫之閩東語紙本資料，主要是地方志針對語言的介紹。

(3)《連江縣志--語言志》：《連江縣志》為連江縣政府文化局於 2013 年所出版的縣志，該套書籍記載了關於連江縣的歷史、大事記、地理、社會、觀光、政事、兵事、人物、教育、文化、人民、語言、經濟財稅等各面向的資訊，而其中的《語言志》篇章是由陳高志老師所主纂，對於馬祖閩東語的聲韻、聲調、變音、用字、辭義、造字等語言議題皆有詳細的介紹與說明。目前可在連江縣政府的網站

上取得《連江縣志--語言志》的 pdf 檔，網址為：http://gov.matsu.idv.tw/lienchiang/language.html。

5.5.3. 小結

目前所能找到的馬祖閩東語相關資源不多，主要是資源網和地方志相關資料，再加上受到國語運動、馬祖開放觀光、馬祖當地居民移居臺灣本島等因素影響，現在馬祖閩東語的傳承狀況極不樂觀。建議國家語言資料庫除了可先彙整上述資料外，未來還需要再針對馬祖當地居民所使用之閩東語重新收集語料，並建置語料庫。

5.6. 臺灣手語

手語不只是聾人社群的溝通方式，更是具有完整架構的語言。如同世界上的人們說著不同的（口語）語言，手語亦隨著歷史推進與地理互動，發展出各個手語語言，在臺灣使用的手語稱為「臺灣手語（Taiwanese Sign Language, TSL）」。臺灣手語因殖民歷史的關係，與韓國手語同時涵納日本手語的詞彙 (Smith, 2005)。在 1945 年後，受到政府推行手語標準化與教材化的影響，臺灣手語分裂成聾人社群長時間使用與發展而成之「自然手語（natural sign language，前述之 TSL）」與聽人主導之「臺灣文法手語（grammatical sign language，即 Signed Chinese, SC）」兩套系統。(李, 2016) 如今，在越來越多聾人參與的情況下，將手語的語言權還予聾人社群 (回應《身心障礙者權利公約》首度國家報告審查會議結論性意見，2017)

在語言特色方面，臺灣手語是一種視覺語言，擁有豐富的擬形詞，顯現出手語具有高度的「象似性（iconicity）」。(戴 & 蔡, 2009) 不過以母語學習的角度來說，文法手語的打法多為「一字一字依循中文的

詞序及造字原則而有了文法」的方式，與自然手語的發展原則不同，即使是大量接觸的兒童亦未必能夠有效吸收與理解。(李, 2016) 因為不了解手語亦可將手勢分解成更小的基本單位 (李, 2016)，亦忽視口語仰賴的是非序列性（non-linearity）及三度空間，而非口語的序列性與一度空間，而讓文法手語凌駕於自然手語之上。(戴 & 蔡, 2009)

有鑒於上述考量，國家語言語料庫與資料庫之規劃應以自然手語為主，且以多媒體的形式呈現，更能符合手語使用族群的需求。目前臺灣手語的研究團隊以國立中正大學語言學研究所最為活躍，長年針對臺灣手語進行系統性的研究，已有多篇碩博士論文產出，亦釋出手語資源，如「臺灣手語線上辭典」、「臺灣手語電子資料庫」、「臺灣手語參考語法」，「臺灣手語研究群」研究團隊網站網址為：http://tsl.ccu.edu.tw/web/。

以下為可收錄至國家語言資料庫之臺灣手語線上資料，(1)、(2)為線上辭典，(3)、(4)為資料庫，(5)為新聞節目，(6)、(7)、(8)為學習資源。(5)作為語料庫的材料，其餘適合發展資料庫。

5.6.1. 線上資料

(1) 臺灣手語線上辭典：由國立中正大學語言學研究所蔡素娟教授與戴浩一講座教授負責編纂，為 2001 年開始的國科會計畫，第三版辭典的詞項已達 3500 個，來源為《手能生橋》、《臺北市手語翻譯培訓教材》、《臺灣手語參考語法》中之詞彙。並且增加「位置」（可點選畫面中人形圖示之身體各個位置）及「手形」（使用手指數目及手形）的查詢功能，或以中文筆畫查詢，可選擇影像檔的播放速率，並附上位置與手形資訊。網址為：http://tsl.ccu.edu.tw/web/browser.htm，本資料庫亦有英文版網頁：http://lngproc.ccu.edu.tw/

TSL/indexEN.html 及 APP：http://taiwansign.ccu.edu.tw/?page_id=59。

(2) 萬手網：由臺灣手語翻譯協會與中華民國聾人協會建置，網址為：http://www.wekeysign.org/，該網站收錄了手語的 124 個新詞，由 11 名聾人及 10 名手譯員打出各新詞，主題聚焦在交通、醫療與法律方面，像是 Uber 一詞，讓手語更融入生活。在檢索設計上，可依「詞條」、「詞組」與「例句」分層檢索，及「主手」、「副手」及「位置」的分類篩選。除了現有的資料，該網站亦採合作編輯的模式，使用者可註冊並上傳自己的手語。

(3) 臺灣手語電子資料庫：由國立中正大學語言學研究所張榮興教授在 2011 年 8 月所建置的資料庫，包括兩個資料庫，分別是「臺灣手語地名電子資料庫」及「臺灣手語姓氏電子資料庫」，臺灣的地名詞彙承載了歷史脈絡，尤其是如「打狗」一地地名為日治時期前的古地名，而非替換成日本手語或中國手語的打法，可作為日後追溯語源的參考。目前資料組成為 1000 個地名、407 個姓氏，打法由聾人社群提供，該網站附有打法影音檔、打法及造詞原則之註解說明，網址為：http://signlanguage.ccu.edu.tw/index.php。

(4) 手語拾遺：臺灣手語語料紀錄網站，發起人林亞秀參加帝亞吉歐（Diageo）Keep Walking 夢想資助計畫的成果，語料來源為 2008 年起發起人與手語使用者長輩的訪談內容，於 2019 年上線，這些長輩們都年過七十，受過日本聾校教育，希望萃取、回溯出當中的舊詞彙及語源，且為呈現臺灣手語的語法等面向內容，語料可輸入關鍵字搜尋相關例句。此外，亦有詞彙檢索的功能，使用者在查詢時可篩選「手形」與「位置」，附有影像檔與打法的註解說明，亦

可放慢播放速度，最下方則可點擊例句，前往相關頁面，網址為 https://www.twsl.cc。

(5) 手語新聞：目前可於公共電視台、客家電視台收看手語新聞。公共電視台的手語新聞節目於民國 91 年 9 月 2 日開播，節目時段為週一至週五晚間 9:45 至 10:00，由王曉書主播、牛暄文主播主持。客家電視台的「當晝新聞」節目則於民國 97 年 8 月 1 日起全程提供手語翻譯，由向盛言主播、丁立芬手語主播主持，從客家族群出發，為華語字幕、客語、手語雙主播的型態。(“當晝新聞打手語 服務聽障"看"新聞”, 2008)

5.6.2. 紙本資料

(6) 手能生橋：由中華民國聾人協會於 1997 年出版之手語課本，共兩冊，收錄 752 個手語詞彙（該詞彙資料已收錄於「臺灣手語線上辭典」中），亦有複合詞、會話與文法練習，以自然手語為主，希望補足聽人領導的中文手語不足之處。

(7) 手語大師：1997 年出版三冊，2002 年出版第四冊，由趙玉平先生編纂之自然手語課本，還包括臺灣手語的發展歷史介紹。

(8) 臺北市手語翻譯培訓教材：由臺北市勞動力重建運用處發行，共兩冊，第一冊為入門教材，第二冊為手譯員訓練導向的主題內容，涵蓋勞政、社政、醫療、稅務、法律及電腦等領域，另有電子化 APP，包括 1900 個手語單詞、600 個複合詞、480 句實例演練。

5.6.3. 小結

臺灣手語的資源反映手語的語言特性及手語在臺灣的發展軌跡，相關線上資源的設計皆採多媒體的形式呈現，包括影像檔及註解說

明，並可放慢影像的播放速度。在檢索功能方面，「手形」及「位置」是不可或缺的設計，尤其是圖示的介面能讓使用者更方便地選擇欲搜尋的內容。檢索結果的呈現亦是圖示，並附上註解說明。

從資料組成來看，詞彙、例句的分層是手語資料庫的特色，因為在打手語時，詞彙之間的關係是三度空間，不須總是一字一字依序打出。然而，現有的手語資料庫多收錄詞彙，例句較少。在上述資源中，尚無臺灣手語語料庫的例子，相較於口語語言語料庫的規模、平衡、主題類型（genre）等建置原則而言，仍亟需資源與人力挹注。

陸、　語料的轉寫、標記與工具

上一章的內容為各國家語言的語言資源盤點，考量語料的轉寫與標記係語料庫建置的一環，於此章依語料處理順序、以主題式的小節探討在轉寫與標記階段時的原則、可開發的輔助工具，方便團隊人員進行相關語言處理，以及過往研究所提供的寶貴經驗。

6.1. 用字規範與對應華語

從歷史的角度來看，臺灣國家語言的書寫系統經歷一段演變，雖然至今仍存有些許分歧，但各語言已有教育部編撰的常用詞辭典、輸入法、不同書寫系統的轉換等，且考量到跨語言檢索時，對應華語使各語言之間有了對照，綜合前述原因，若為書面語料或已轉寫成文字的口語語料，將其與華語對照，並可能可以保留原有文字，而未有轉寫文字部分的口語語料則採用教育部用字進行轉寫，再與華語對照。

閩南語及客語以《教育部常用詞辭典》及《推薦用字》為主。《常用詞辭典》共有 1 萬 3 千餘詞之閩南語詞彙與 1 萬 5 千餘詞（15,454 詞）之客語詞彙。《推薦用字》則有《臺灣閩南語推薦用字 700 字表》及《臺灣客家語書寫推薦用字》，於民國 98 年推出第一批 305 個用字、民國 100 年推出第二批 209 個用字，共 514 字。此外，《閱讀越懂閩客語》週刊電子報的專欄文章自民國 102 年起已累積 50 萬字語料，因使用漢字書寫，可作為閩客語與華語之間的對照。值得注意的是，蔡素娟教授於 2011 年（民國 100 年）建置閩南語兒童語料庫時，亦參考了董忠司教授總編纂之《臺灣閩南語辭典》、陳修主編之《臺灣話大詞典》、李榮主編之《廈門方言詞典》、楊秀芳主編之《閩南語字彙》、吳守禮編撰之《國臺對照活用辭典》、許極燉編撰之《臺語辭典常用

漢字》、楊青矗主編之《國台雙語詞典》等，亦是珍貴的資源。(蔡 et al., 2009; Tsay, 2007)

閩東語的部分目前還尚未有相關政府單位公布的用字規範，因此現階段可先參考《連江縣本土教學資源網》網站上的現有辭典相關資源，如《馬祖閩東語本字檢索系統(試用版)》（網址：http://fc-matsu.com/ ）、《日常生活常用詞彙》（網址：http://www.matsudialect.org/1000/index.html ）、《綜合活動馬祖話》（網址：http://www.matsudialect.org/1000_2/index.htm ）等；另外也可參考陳高志老師的《連江縣志-- 語言志》（網址：http://gov.matsu.idv.tw/lienchiang/language.html ）相關介紹。

針對原住民族語，原住民族語言研究發展中心於民國 107 年完成 15 族書寫系統修訂共識確認會議，並與其中 12 族取得共識，可參考《原住民族語言書寫符號》，網址為：http://ilrdc.tw/research/rwview/rwsystem.php。與此同時，原住民委員會自 2007 年起（民國 96 年）編製之 16 族線上辭典亦已完成。然而，由於客語與原住民族語皆有專責機構籌劃語料庫的建置，建議依專責機構的規劃進行語料庫的整合，並將其與華語對應。此外，相比閩南語及客語，原住民族語使用羅馬字母書寫的比例最高，且有詞彙原形（lemma）與屈折變化（inflection）之分，應以詞素（morpheme）分析結果為主。

在處理用字不一致的過程中，建議須建置用字對應表，將教育部為準的閩南語用字與對應華語建置成有索引的資料表，有關索引之討論請參見「 8.5.1 維運與管理 」一節。楊允言教授 (楊 et al, 2008) 提到，用字對應是分詞、詞性標記的基礎，可幫助處理閩南語、漢字混用的

漢羅書寫方式，自動檢閱混用的各個字是否出現在同一詞條中，相關討論請見「6.3 詞性標記」一節。

在輸入法方面，閩南語書寫系統不一致，須經過轉換，轉為教育部用字以求一致，例如：白話字（Pe̍h-ōe-jī, POJ）為西元 1810 年代基督教傳教時所推行之用字，又稱為教會羅馬字（Kàu-hōe Lô-má-jī，簡稱教羅）、TLPA 為臺灣語言音標方案（Taiwan Language Phonetic Alphabet, TLPA）是由臺灣語文學會於 1991 年所制定的音標系統，及 2002 年至 2008 年政府機關採行之通用拼音。教育部閩南語常用詞辭典網站上的「資源下載」專區提供「教育部臺灣閩南語羅馬字拼音輸入法（簡稱台羅拼音輸入法）」，並可將其他拼音系統轉換為台羅拼音，如：白話字、TLPA 及通用拼音系統轉換為台羅拼音系統，此功能對於語料庫建置過程的語料轉寫及用字轉換很有幫助。若是使用者欲輸入閩南語查詢語料庫，意傳科技亦開發了網頁介面，其 GitHub 為 https://臺語羅馬字轉換器.意傳.臺灣。

從使用者的角度出發，台語信望愛釋出電腦版的閩南語輸入法，會顯示可能的幾個漢字供使用者選擇，並顯示輸入法狀態為「全羅、全漢、漢字優先、羅馬字優先」等模式。此外，Phah Tâi-gí（打台語）的發起人吳家銘（Ngô͘ Ka-bêng）蒐集閩南語的字詞資料、開發核心技術與候選字演算法、常用詞詞頻、使用者詞頻、使用者自訂字詞、字詞語句學習等，於 2017 年陸續釋出 Android 及 iOS 的閩南語輸入法應用程式（網址為：https://github.com/PhahTaigi/PhahTaigi_Android），可輸入白話字、教育部羅馬字、教育部公告漢字，並支援半自動完成功能（auto-complete），即輸入部分羅馬字，會出現所有聲調及推薦的詞。相關計畫 ChhoeTaigi（找台語）於 2018 年接受零時政府 g0v 的資金贊助，整理 9 個文獻來源（依詞彙數量多寡有台文華文線頂字典、

台日大辭典、Maryknoll 台英辭典、Embree 台語辭典、教育部台語辭典、甘字典、iTaigi 華台辭典、臺灣白話基礎語句、臺灣植物名彙）共 317,526 個詞彙，採創用 CC 授權，相關資料說明、授權方式、csv 字詞檔案可參考 https://github.com/ChhoeTaigi/ChhoeTaigiDatabase?fbclid=IwAR0qa0-FwtahpuTxK2yrLLpE0wpXvheo9DrlaDv8mLLZWahY_lDdMOMg2XA，目前持續在嘖嘖平台上募資，用於字詞資料的人工校對、辭典網站與程式的維護與更新、電腦版輸入法開發等。其他閩南語輸入法工具，可參考維基教科書列表：https://zh.wikibooks.org/zh-sg/臺灣話/書寫/輸入法。

1. 台文華文線頂辭典

字詞資料代號：

```
ChhoeTaigi_TaibunHoabunSoanntengSutian
```

資料內容說明：

欄位名稱	說明
id	編號
poj_unicode	白話字
poj_unicode_dialect	白話字（其他講法）
poj_input	白話字輸入
poj_input_dialect	白話字輸入（其他講法）
hanlo_taibun_poj	漢羅台文（白話字）
kiplmj_unicode	教育部羅馬字
kiplmj_unicode_dialect	教育部羅馬字（其他講法）
kiplmj_input	教育部羅馬字輸入
kiplmj_input_dialect	教育部羅馬字輸入（其他講法）
hanlo_taibun_kiplmj	漢羅台文（教育部羅馬字）
hoabun	華文

授權說明：

```
【台文華文線頂辭典】
基礎資料提供：Tē͘ Liông-úi（鄭良偉）教授
資料增加kap編修：Iû͘ Ún-giân（楊允言）教授、眾phah字kap校對ê義工
以 姓名標示-Sio-kâng方式分享 4.0 國際 (CC BY-SA 4.0) 授權
https://creativecommons.org/licenses/by-sa/4.0/deed.zh_TW
```

圖 23. ChhoeTaigi（找台語）收錄之閩南語字詞資料說明畫面（圖片來源：https://github.com/ChhoeTaigi/ChhoeTaigiDatabase）

臺灣手語雖無書寫系統，但視覺語言的三度空間表達方式、時間與空間的互動，讓臺灣手語詞彙的組成、分析更加複雜，基本要素如打法「位置」與「手形」等，亦有主副手、對稱與交替性、臉部表情等細部資訊，這些資訊都是將手語打法細分成更小單位（smaller units）

的方式，亦有其系統性的原形化（lemmatization）原則，與口語語言大不相同，若要與華語詞彙對應，有一定的難度。

若從影像轉成可機讀的格式來看臺灣手語的轉寫，亦是一個標記的過程。2011 年建置的臺灣手語線上辭典已於 2017 年釋出第三版，並完成音韻標記（phonologically annotated）(Tsay, 2019)，共有 30 個打法「位置（location）」及「手形（handshape）」，亦包括南北腔調的標記（_S 及_N），更多的標記項目如：動作（movement）、手形方向（hand orientation）以及非手部之特徵（non-manual features，如臉部表情）等。藉由各個項目的標記，逐步將臺灣手語的最小對立體（minimal pair）分辨出來。

6.2. 分詞

若是以電腦能夠處理的目的進行分詞，對於英文或是其他以羅馬字母書寫的語言來說，分詞的困難度相對較小，多半以空格作為分詞的依據（space-delimited），並加上大小寫判斷、複合詞（compound word）詞典等資訊，使 New York、White House 等字串不會被斷開成兩個字詞。韓文、日文、華語（CJK, Chinese-Japanese-Korean）等語言基於語言特性，並未使用空格將字詞斷開，因此在自然語言處理上較無既定規則可循，須仰賴程式內建的分詞辭典與演算法進行基本斷詞，亦常需要另外的使用者辭典（user dictionary）或人工檢查（post-editing），以提升斷詞的準確度，因此斷詞是前處理很重要的一環。分詞錯誤會造成「集外詞的比例」（out-of-vocabulary ratio, OOV ratio）攀高，意即分詞結果中，出現許多不在分詞程式詞典內的字詞，不利於往後的自然語言處理任務，雖然亦有研究結果發現 (Meng et al., 2019)，以字為單位的分詞（character-based segmentation）在高集外詞比例的情況下可能會表現得比經過斷詞的語料好，但並非代表毋需斷詞，因為

該研究針對的是語言模型等實驗結果的表現，若以語料庫檢索為考量，斷詞處理是必須且基本的，不同的分詞方式會使得檢索時的結果不同。華語的分詞程式已臻成熟，如受到廣泛使用的結巴斷詞，原為簡體中文的斷詞程式，亦推出繁體中文的字典、中研院於 2019 年最新釋出的 CKIP tagger 已有 Python 套件開放使用 (Li et al, 2020)，史丹佛大學的分詞程式 Chinese Word Segmentor 等。其中以中研院最新推出深度學習的分詞程式正確率最高。針對閩南語、閩東語、客語及原住民族語的斷詞處理，若初步以是否採用羅馬字母作為書寫系統來看的話，原住民族語較無困難，而閩客語則須搭配使用者詞典（user dictionary），讓程式得以基於此詞典進行斷詞。

另一方面來說，斷詞不僅僅是空格及羅馬字母使用與否的討論，更是何謂「成詞性」（wordhood）的課題。 (Magistry, 2013) 目前華語分詞多採用「中文資訊處理分詞規範調查研究計畫」之分詞準則，此計畫於民國 87 年由經濟部中央標準局委託中華民國計算語言學學會辦理，以下節錄分詞準則重點內容：

1. 以「字串基本語意單位」作為切分字詞的原則，並在語意及句法方面符合語言學理論。
2. 依自動化處理的難度分級：
   (1)「信級」為依照標準詞典將詞切分，其分詞結果可進行資料交換，在此層級可解決歧義（ambiguity）。
   (2)「達級」為符合簡單構詞（morphology）的字詞切分，其結果可應用於大部分的自然語言處理任務。
   (3)最理想的階級是「雅級」，即其結果可作為語言學理論的展現，但自動處理不易到達此階段。
3. 第二點提及之「標準詞典」可與時俱增，隨時更新。

4. 此準則列出兩條基本原則及六條輔助原則，後者可有變異性。
5. 根據各個詞類，主要是名詞及動詞，有不同的字詞結構，此準則詳列合併或分開的例子以供參考。

針對閩南語的分詞原則，楊允言教授於2019年在教育部本土語言資源網上轉錄曾金金教授於 1997 年臺灣文學出版物收集、目錄、選讀編輯計畫結案報告說明之內容 (曾, 1997)，認同經濟部中央標準局之分詞原則（以「搜文解字」稱之），亦以兩條基本原則與六條輔助原則說明閩南語的分詞，並附有詳細之閩南語例示。

針對臺灣手語的分詞處理，參考國內外的例子可發現手語資料庫「SignBank」或詞彙資料庫「lexical database」是了解各國手語詞彙很重要的語言資源，因為手語在單詞及成句的情境下是不同層面的呈現方式，除了從臺灣手語語料庫中挖掘詞彙，亦可另就臺灣手語詞彙建置資料庫，並由臺灣手語使用者及社群組成專業團隊，提出符合臺灣手語語言特色的處理原則。

另外，有了該分詞準則的建立，「只要定時更新基本詞庫或特殊領域的專門詞庫，便可維持分詞規範的不變性」(語料庫建置入門工作流程指南, 2010)，此亦反映出前述之使用者詞典的重要性，若能在程式完成自動分詞後進行校對，有利於提升其準確度。

6.3. 詞性標記

詞性標記與分詞的處理常是同時進行的，例如：中研院詞庫小組的 CKIP tagger 與史丹佛大學的 Chinese Word Segmentor 都能處理分詞及詞性標記，但其分詞標準與詞性集（tagset）不同。中研院的詞性標記集共分為三個階層，原始階層共有 122 個詞類，較精簡的版本為 47 個詞類，最精簡的版本為 19 個詞類，中研院平衡語料庫的檢索介面採

用的是 47 個詞類的階層。鄧守信教授 (2010) 提出八大詞類，是目前詞類數目較少且使用範圍廣泛的詞性集，多應用於華語教學。在諮詢會議中，曾淑娟教授亦提出共用詞性集（Universal POS）的建議。

表 10. 中研院平衡語料庫之詞性標記集 (詞庫小組, 1995)

精簡詞類	簡化標記	對應的 CKIP 詞類標記[8]	
A	A	A	/*非謂形容詞*/
C	Caa	Caa	/*對等連接詞，如：和、跟*/
POST	Cab	Cab	/*連接詞，如：等等*/
POST	Cba	Cbab	/*連接詞，如：的話*/
C	Cbb	Cbaa, Cbba, Cbbb, Cbca, Cbcb	/*關聯連接詞*/
ADV	Da	*Daa*	/*數量副詞*/
ADV	Dfa	Dfa	/*動詞前程度副詞*/
ADV	Dfb	Dfb	/*動詞後程度副詞*/
ASP	Di	Di	/*時態標記*/
ADV	Dk	Dk	/*句副詞*/
ADV	D	*Dab*, Dbaa, Dbab, Dbb, Dbc, Dc, Dd, Dg, Dh, Dj	/*副詞*/
N	Na	Naa, Nab, Nac, Nad, Naea, Naeb	/*普通名詞*/
N	Nb	Nba, Nbc	/*專有名稱*/
N	Nc	Nca, Ncb, Ncc, Nce	/*地方詞*/
N	Ncd	Ncda, Ncdb	/*位置詞*/
N	Nd	Ndaa, Ndab, Ndc, Ndd	/*時間詞*/
DET	Neu	*Neu*	/*數詞定詞*/.
DET	Nes	*Nes*	/*特指定詞*/
DET	Nep	*Nep*	/*指代定詞*/
DET	Neqa	*Neqa*	/*數量定詞*/
POST	Neqb	*Neqb*	/*後置數量定詞*/
M	Nf	Nfa, Nfb, Nfc, Nfd, Nfe, Nfg, Nfh, Nfi	/*量詞*/
POST	Ng	Ng	/*後置詞*/

[8] 斜體詞類，表示在技術報告#93-05中沒有定義，即後來增列的。

N	Nh	Nhaa, Nhab, Nhac, Nhb, Nhc	/*代名詞*/
Nv	Nv	Nv1,Nv2,Nv3,Nv4	/*名物化動詞*/
T	I	I	/*感嘆詞*/
P	P	P*	/*介詞*/
T	T	Ta, Tb, Tc, Td	/*語助詞*/
Vi	VA	VA11,12,13,VA3,VA4	/*動作不及物動詞*/
Vt	VAC	VA2	/*動作使動動詞*/
Vi	VB	VB11,12,VB2	/*動作類及物動詞*/
Vt	VC	VC2, VC31,32,33	/*動作及物動詞*/
Vt	VCL	VC1	/*動作接地方賓語動詞*/
Vt	VD	VD1, VD2	/*雙賓動詞*/
Vt	VE	VE11, VE12, VE2	/*動作句賓動詞*/
Vt	VF	VF1, VF2	/*動作謂賓動詞*/
Vt	VG	VG1, VG2	/*分類動詞*/
Vi	VH	VH11,12,13,14,15,17,VH21	/*狀態不及物動詞*/
Vt	VHC	VH16, VH22	/*狀態使動動詞/
Vi	VI	VI1,2,3	/*狀態類及物動詞*/
Vt	VJ	VJ1,2,3	/*狀態及物動詞*/
Vt	VK	VK1,2	/*狀態句賓動詞*/
Vt	VL	VL1,2,3,4	/*狀態謂賓動詞*/
Vt	V_2	V_2	/*有*/
T	DE	/*的, 之, 得, 地*/	
Vt	SHI	/*是*/	
FW	FW	/*外文標記*/	

COLONCATEGORY	/* 冒號 */
COMMACATEGORY	/* 逗號 */
DASHCATEGORY	/* 破折號 */
ETCCATEGORY	/* 刪節號 */
EXCLAMATIONCATEGORY	/* 驚嘆號 */
PARENTHESISCATEGORY	/* 括弧 */
PAUSECATEGORY	/* 頓號 */
PERIODCATEGORY	/* 句號 */
QUESTIONCATEGORY	/* 問號 */
SEMICOLONCATEGORY	/* 分號 */
SPCHANGECATEGORY	/* 雙直線 */

表 11. 鄧守信教授 (2010) 提出之八大詞類

編號	詞類標記	詞類名稱
1	V	動詞
2	N	名詞
3	ADV	副詞
4	Prep	介詞
5	M	量詞
6	Det	定詞
7	Ptc	語氣詞
8	Conj	連接詞

不同階層的詞性集適合不同的用途，例如：十類左右的詞性集常使用於自動詞性標記程式的開發，若是詞性過多，有些詞性的語料便會過少，所需的訓練語料更大，否則會造成稀釋的效果，詞性分布落差甚大，機器學習的正確率較差，及標記者面臨「主觀強制性的歸類」問題。(蔡 et al, 2009)

另外，華語已有諸多關於形容詞詞類的討論 (Huang et al., 2017)，上述中研院與鄧守信教授提出的詞類標記中，前者僅有「非謂形容詞（a）」之詞類，多數形容詞的標記為「狀態不及物動詞（VH）」，後者更是沒有形容詞的詞類，如此設計反映華語語言特性。在一致性原則與國家語料庫的架構下，建議以現有詞性集作為標準，並視閩南語、客語、原住民族語、閩東語的語言特性調整成為新的詞性集。蔡素娟教授建置之臺灣閩南語兒童語料庫（Taiwanese Child Corpus, TAICORP，網址為：http://www.ccunix.ccu.edu.tw/%7Elnglab/TAICORP.htm）基於中研院之詞性標記，加入了 Di/T、CIT 及 Comp 三個詞性，如表 12，Di/T（“marker following pseudo-transitive active verb”）針對閩南語「le0」一字，常出現於動詞之後及句尾，因此給予時態標記 Di 及語助詞 T 之詞類，其例句為「你 坐 le0」；CIT 針對閩南語「得 2」一字（“special

tag for the word "得 2"”），意思是「能夠」，可加在動詞之後，或與動詞分開而置於句尾，華語無此情形，故新增此詞類，例句為「你 未使 去 偷挽 別人 辛苦 所 種 e0 果子 得 2」；Comp（“complementizer”）針對「得、甲 1、了 2、予 3」等字，可將兩個動詞連接起來，例句為「穿 予 3 水水 2」；意傳科技亦將中研院詞性集（CKIP）、蔡素娟教授、楊允言教授及 Chinese PennTree 的比較上傳至 GitHub，網址為：https://github.com/i3thuan5/su5-sing3/blob/master/詞類比較.csv，其中 CKIP 及 Chinese PennTree 為華語詞性集，Chinese PennTree 的詞性集如表 13。

表 12. 蔡素娟教授 (Tsay, 2007) 閩南語兒童語料庫之詞性集

詞性標記	詞類	詞類（中文）
A	non-predicative adjective	非謂形容詞
Caa	coordinate conjunction	對等連接詞
Cab	listing conjunction	連接詞
Cba	conjunction occurring at the end of a sentence	連接詞
Cbb	following a subject	關聯接接詞
Da	possibly preceding a noun	數量副詞
Dfa	preceding VH through VL	動詞前程度副詞
Dfb	following adverb	動詞後程度副詞
Di	post-verbal	時態標記
Dk	sentence initial	句副詞
D	Adverbial	副詞
Na	common noun	普通名詞
Nb	proper noun	專有名稱
Nc	location noun	地方詞
Ncd	localizer	位置詞
Nd	time noun	時間詞

Neu	numeral determiner	數詞定詞
Nes	specific determiner	特指定詞
Nep	anaphoric determiner	指代定詞
Neqa	classifier determiner	數量定詞
Neqb	postposed classifier determiner	後置數量定詞
Nf	classifier	量詞
Ng	postposition	後置詞
Nh	pronoun	代名詞
I	Interjection	感嘆詞
P	Preposition	介詞
T	particle	語助詞
VA	active intransitive verb	動作不及物動詞
VAC		動作使動動詞
VB	active pseudo-transitive verb	動作類及物動詞
VC	active transitive verb	動作及物動詞
VCL	transitive verb taking a locative argument	動作接地方賓語動詞
VD	ditransitive verb	雙賓動詞
VE	active transitive verb with sentential object	動作句賓動詞
VF	active transitive verb with VP object	動作謂賓動詞
VG	classificatory verb	分類動詞
VH	stative intransitive verb	狀態不及物動詞
VHC	stative causative verb	狀態使動動詞
VI	stative pseudo-transitive verb	狀態類及物動詞
VJ	stative transitive verb	狀態及物動詞
VK	stative transitive verb with sentential object	狀態句賓動詞
VL	stative transitive verb with VP object	狀態謂賓動詞
V_2		有
DE	*special tag for the word "的"	的
SHI	special tag for the word "是"	是
FW	foreign words	外文標記

*Di/T	*marker following pseudo-transitive active verb	*le01
*CIT	*special tag for the word "得 2"	*得 2
*Comp	*complementizer	*補語連詞

表 13. Chinese PennTree 的詞性集 (Xia, 2000)

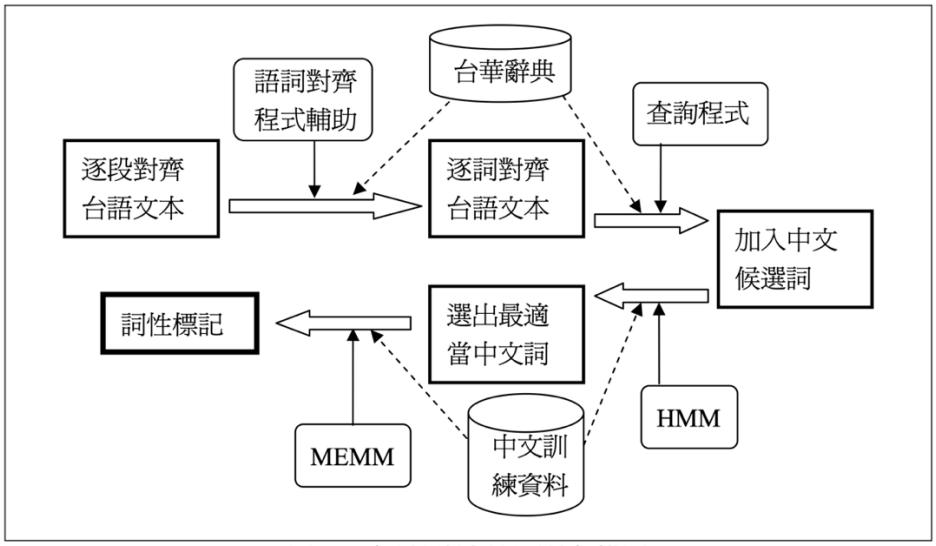
詞性標記 Tag	**說明 Description**	**詞例 Example**
AD	Dverb	還
AS	aspect marker	著
BA	把 in ba-construction	把、將
CC	coordinating conjunction	和
CD	cardinal number	一百
CS	subordinating conjunction	雖然
DEC	的 in a relative-clause	的
DEG	associative	的
DER	in V-de const. and V-de-R	得
DEV	地 before VP	地
DT	determiner	這
ETC	for words 等, 等等	等、等等
FW	foreign words	ISO
IJ	interjection	啊
JJ	other noun-modifer	男、共同
LB	被 in long bei-const	被、給
LC	localizer	裡
M	measure word	個
MSP	other particle	所
NN	common noun	書
NR	proper noun	美國
NT	temporal noun	今天
OD	ordinal number	第一
ON	onomatopoeia	哈哈、嘩嘩
P	Prepositions (excluding 把 and 被)	從
PN	pronoun	他
PU	punctuation	、？。
SB	被 in short bei-const	被、給

SP	sentence-final particle	嗎
VA	predicative adjective	紅
VC	copula	是
VE	有 as the main verb	有
VV	other verbs	走

表 14. 共用詞性集（Universal POS tags）（資料來源：https://universaldependencies.org/u/pos/）

開放詞類 **Open class words**	**封閉詞類** **Closed class words**	**其他** **Other**
形容詞 ADJ 副詞 ADV 感嘆詞 INTJ 名詞 NOUN 專有名詞 PROPN 動詞 VERB	介詞 ADP 助詞 AUX 對等連接詞 CCONJ 限定詞 DET 數詞 NUM 介副詞 PART 代名詞 PRON 從屬連接詞 SCONJ	標點符號 PUNCT 符號 SYM 其他 X

楊允言教授提到，閩南語詞性集確立的困難點，在於現有的字典多非以中研院或其他華語的詞性集為基礎，而是以稍嫌簡略的詞性標記帶過（“We did not have a Taiwanese dictionary that contained the Mandarin POS tagset.”），而詞性標記是耗費極大人力的工程，因此該計畫以統計分析的方法將詞性標記自動化，圖 24 為其標記流程設計 (楊 et al, 2008)：



圖一、台語詞性標記系統架構圖

圖 24. 楊允言教授閩南語詞性標記系統架構圖 (楊 et al, 2008)

在解決漢字與羅馬拼音混用（簡稱漢羅）的問題方面，此系統將該詞彙以臺華字典對照，核對是否出現在同一詞條。反之，如果該詞彙沒有出現在同一詞條，可能是漢字用字不一致或未知詞，程式須標示出該詞彙，以讓標記者檢查修正，或是直接提醒使用者相關情形。另外，由於詞性標記的訓練結果沒有正確答案可參循，以抽樣的方式進行人工校對與計算錯誤率，包括華語對應詞是否正確、詞性是否正確。楊允言教授提到，詞性標記錯誤可能歸因於華語對應錯誤、缺少華語對譯或受前一錯誤詞性標記影響而造成的傳播錯誤（propagation error），表 15 為其校對語料的詞性標記錯誤分析：

表 15. 楊允言教授閩南語詞性標記系統之錯誤分析 (楊 et al, 2008)

錯誤原因	次數	比例	說明
選錯中文詞	13	27.08%	
沒有正確的中文詞可選	2	4.17%	
未知詞	8	16.67%	

人名	4	8.33%	
傳播錯誤	4	8.33%	包含一未知詞
總計	30	62.50%	扣除重複算的

各語言的特性使得華語對應不易，也加重詞性標記的負擔，例如：不同語言有不同的語序，而有錯誤的華語對應；另外，未知詞的最終原因可能是無華語對應，若以此次研究與往後其他研究的訓練語料作為基礎，可進一步提升自動化詞性標記的正確率。若是因無華語對應而造成錯誤，可返回增補與華語的對應字典，但楊允言教授表示僅能提升 5%內的準確率。(楊 et al, 2008)

針對臺灣手語詞性，可見名詞、動詞、助詞（auxilaries）等詞性(Tai & Tsay, 2015)，儘管助詞在其他手語語言較少見，但臺灣手語有助詞，主要使用時點為句中動詞動作性不強（does not move in space）時，用以標示主、受詞間的關係。(Tai & Tsay, 2015) 此外，Smith (2005) 以自身研究所蒐集的三位手語使用者語料及《手能生橋》一書中的句子為基礎，列出以下的分類，可作為臺灣手語的詞性參考之一：

- 主詞 S – subject
- 動詞 V – verb
- 不及物動詞 Vi – intransitive verb
- 受詞 O – object
- 對應主受詞的曲折動詞 sVo – verb inflected for both subject and object
- 僅對應受詞的曲折動詞 Vo – verb inflected for object only
- 副詞 A – adverbial
- 時間副詞 At - Adverbial of time
- 方式副詞 Am – adverbial of manner

- 因果副詞 Ar – adverbial of reason
- 地方副詞 Ap – adverbial of place
- 助詞 X – auxiliary
- 否定詞 N – negative
- 疑問詞 Q – question word
- 感嘆詞 I – interjection
- 已詞彙化疑問詞標誌 Q – lexicalized interrogative marker

6.4. 口語語料轉寫與標記

蔡素娟教授開發閩南語兒童語料庫時，設計了一套轉寫系統 Segmentor，包含詞彙存取功能（lexical access）、轉寫工具（transcription tool）及分詞工具（segmentation tool），由於轉寫者多半較熟悉華語輸入法，轉寫工作常仰賴閩華對照，然而在轉寫時遇到許多常見閩南語字詞，其華語本字（cognate Chinese character）卻為低頻字（low-frequency character）的情形，甚至有些閩南語字詞根本沒有對應的華語，且為了將閩南語對照至華語，而以華語音韻系統思考，轉寫結果可能會較不貼近閩南語的音韻特徵。另一方面，儘管直接輸入閩南語能改善上述問題，卻仍存在其他根本問題，例如：轉寫者必須對閩華的音韻系統皆非常熟悉、須熟稔如何將閩南語口說等內容轉以至文字呈現等，因此該語料庫採用成人語料庫拼音輸入程式（Adult-Corpus Romanization Input Program, ACRIP），其架構系統如下圖：

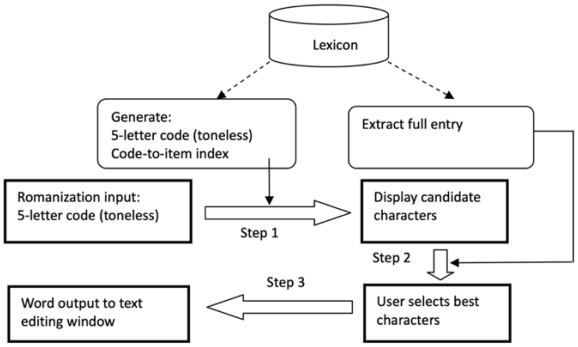


圖 25. 蔡素娟教授成人語料庫拼音輸入程式（Adult-Corpus Romanization Input Program, ACRIP）之架構系統圖示 (Ruan et al., 2012)

使用該系統時，標記者在程式輸入前五個羅馬字母（5-letter code），程式會將其與代號對照（code-to-item index），之後列出 3 個候選字詞予標記者選擇，因此能以半自動完成（auto-completion）的方式輔助標記者進行口語語料的轉寫。此外，考量到閩南語音調輸入不易，程式亦會一併顯示建議的音調選項。

在語料標記部分，英國國家口語語料庫 2014 年版本（Spoken BNC 2014）擴增了口語語料及釋出說明手冊，其中針對口語音檔轉寫成文字的步驟，列出應包含的口語標記（paralinguistic annotation），希望能讓轉寫文字更貼近原始內容，例如：

<pause>為停頓之標記，並細分較長的停頓<long>與較短的停頓標記<short>；同時發言的標記為<overlap>；不完整的字詞標記為<trunc>，常用於修復的語言行為（repair，意即說者在言談中意識到錯誤而修正的情形）。(McEnery et al, 2017)

在標記語料的過程中，可使用兒童語言交換系統（Child Language Data Exchange System, CHILDES）所開發之軟體 CLAN，下載網址為：https://childes.talkbank.org。(MacWhinney, 2000) 兒童語言交

換系統由卡內基美隆大學 Brian MacWhinney 教授與 Catherine Snow 教授所開發，提供各語料庫之語料與標記儲存，CLAN 為針對語料處理設計的軟體，其附檔名為*.cha，如下圖所示：

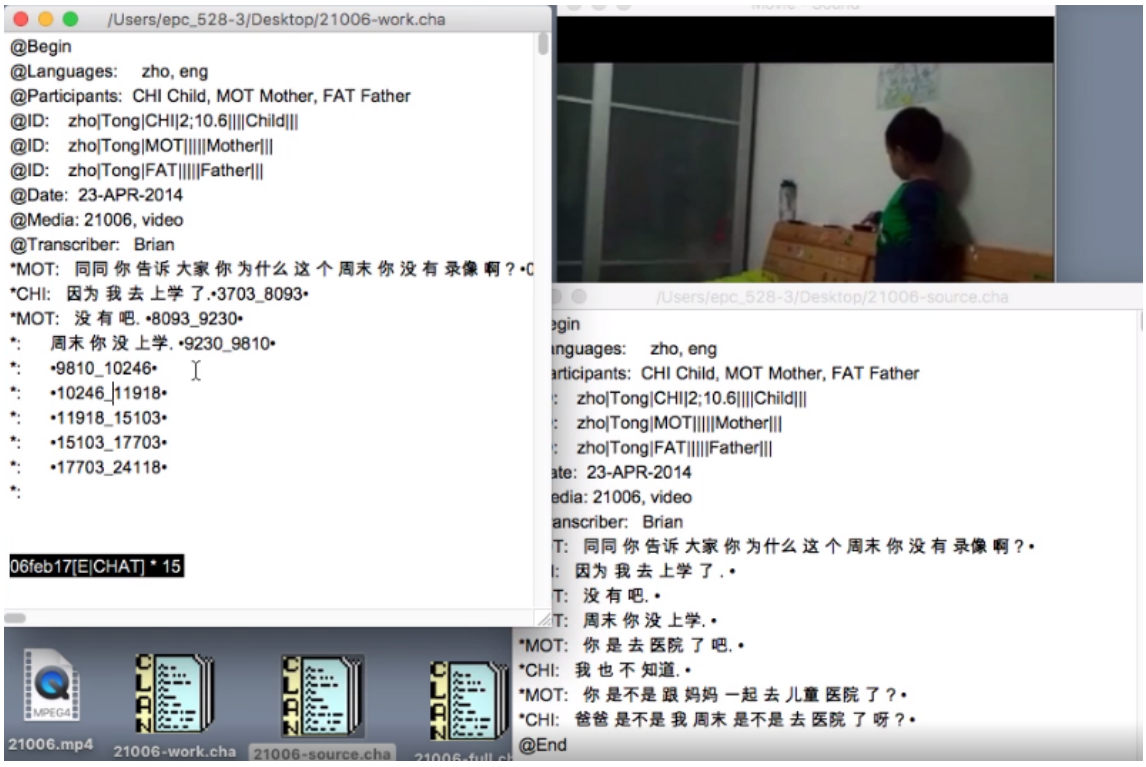


圖 26. CHAT 語料轉寫畫面 (MacWhinney, 2017)

以下是 CHAT 格式之說明，完整說明可參考 MacWhinney (2010) 的說明文件，網址為：https://talkbank.org/manuals/CHAT.pdf。此外，蔡素娟教授 (Tsay, 2007) 亦使用 CHILDES 系統進行語料的轉寫，並提供了相關說明。

(1) @Begin 和@End 宣告檔案的開始與結束。

(2) @Languages 採用的是國際語種代號標準 ISO 639-3，臺灣國家語言的個別代號（Individual language identifier）為華語 cmn、閩南語 nan、客語 hak，皆隸屬於 zho 這個大語言的代號（Macrolanguage identifier）之下。

(3) @Participants 的格式為「說話者 3 碼 ID 角色」（ID role），並以逗號（,）隔開，常見的 3 碼 ID 有 PAR（對話者，participant）、CHI（孩童，child）、MOT（母親，mother）。

(4) @ID 的格式為「@ID: 語言別|語料庫名稱|3 碼 ID|年齡|性別|分群名稱|社經地位|角色|教育程度|其他」（@ID: language|corpus|code|age|sex|group|SES|role|education|custom| ），圖 26 中年齡（age）為「2;10.6」代表「兩歲十個月六天」，而分群名稱（group）因該語料庫系統常用於非典型（atypical）兒童之語言研究，而有 ASD（autism spectrum disorder，自閉類群障礙）等資訊，教育程度則有 Elem, HS, UG, Grad, Doc（初等教育、中等教育、大學、碩士、博士）之分。

(5) @Media 的格式為「不包含副檔名的檔案名稱 音檔或影像檔」，如圖 26 中「21006, video」，若是音檔則為「sound」。

(6) 若為平行語料，有華語翻譯，可加入 %ort 的標記。(蔡 et al., 2009)

6.5. 手語語料標記

於「貳、國外手語語料庫、資料庫的現況分析」一章回顧之手語語料庫中，皆使用荷蘭馬克思普朗克研究中心（Max Planck Institute）所開發的 ELAN (EUDICO Linguistic Annotator) 軟體作為手語語料處理的工具，其副檔名為 .eaf (EUDICO annotation format)。在蒐集手語語料時，常會從不同角度拍攝，ELAN 軟體可以同時顯示不同角度的影像畫面，荷蘭 Corpus NGT 在 2008 年問世時，便拍攝了四個不同角度的影像。(Crasborn & Sloetjes, 2008) 在標記方面，荷蘭手語語料庫先建立好

一份標記範本，將標記項目的內容留空，之後每個語料切分檔（segments）會對應到一個標記檔，且進一步將左右手或不同發音人的標記分開成不同層列（tier）的標記，如下圖。

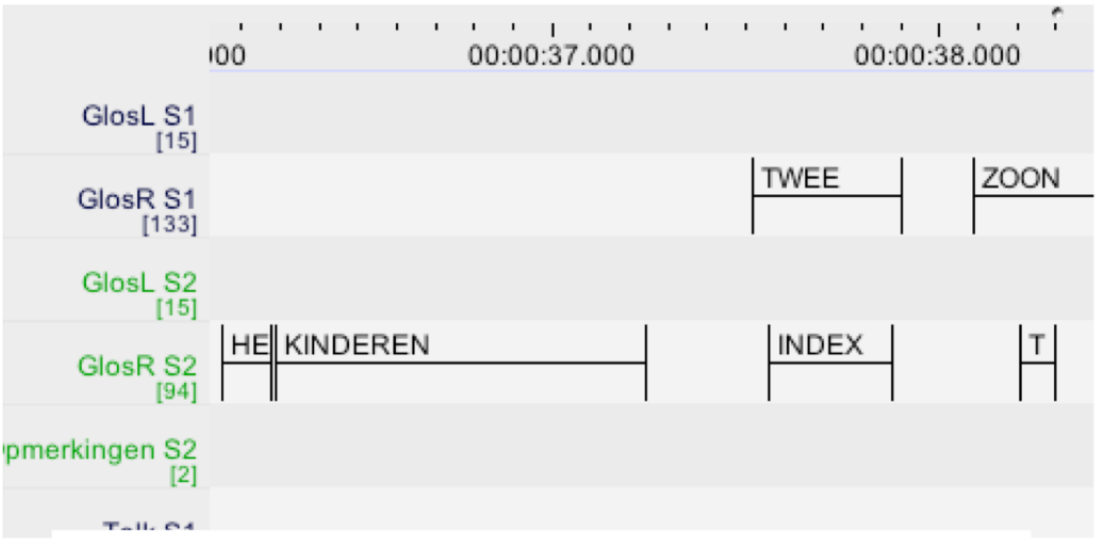


圖 27. ELAN 不同層列（tier）及時間對應（time-aligned）的標記畫面 (Crasborn & Sloetjes, 2008)

在軟體功能方面，荷蘭 Corpus NGT 在利用 ELAN 時，亦考量了 (1) 標記的複製貼上功能（duplicate annotation），並加上標記群組（dependent annotations）的選擇設定，因為有些手語打法是左右手對稱的，但標記層列的設計卻是分開的。(2) 針對搜尋功能，提供「全符合」、「部分符合」或正規表示（regular expression）的選擇，並可將搜尋結果輸出成 tsv 檔（tab separated values）。搜尋功能對於後續標記工作規劃、後設資料說明很有幫助。 (3) 標記者可就標記項目的編號進行標記，毋需完整標記。

在初步資料分析上，荷蘭 Corpus NGT 則提供些許描述性統計、標記分佈（density viewer）、標記內容列表（a list of unique annotation values）的功能，如下圖。

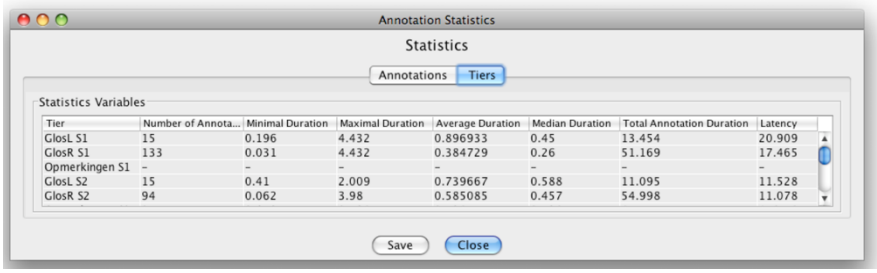


圖 28. ELAN 標記資料統計 (Crasborn & Sloetjes, 2008)

柒、　國家語言資料庫整體設計與規劃之建議

本章為國家語言資料庫整體設計與規劃之建議。本報告第一章回顧了北美洲、歐洲、大洋洲、亞洲各主要國家的國家語料庫現狀，第二章則聚焦於手語語料庫與資料庫，從這兩章的文獻分析可發現目前國外已有口語和書面語組成的國家語料庫（national corpus）的例子可參考，近期釋出的手語語言資源出現資料庫與語料庫兩種形式，因此首先需要區分語料庫（corpus）與語言資料庫（language database）的概念。一般而言，語料庫可視為一種特殊的資料庫。語料庫的特性可參考日本國立國語言研究所前川喜久雅教授（Dr. Kikuo Maekawa）於第二屆中研語言學論壇「國家語言語料庫：規劃與建置」的論述，而在前川喜久雅教授提到的七點特質中，語料庫的資料都是從真實的對話或文本而來，因此真實性（authenticity）是語料庫的一個特性。一般而言，語料庫是指根據某一個組成原則所收集具有前後文語境的語篇或對話的真實語言資料的取樣。具有較長的前後文語境可以視為語料庫的一個特性。

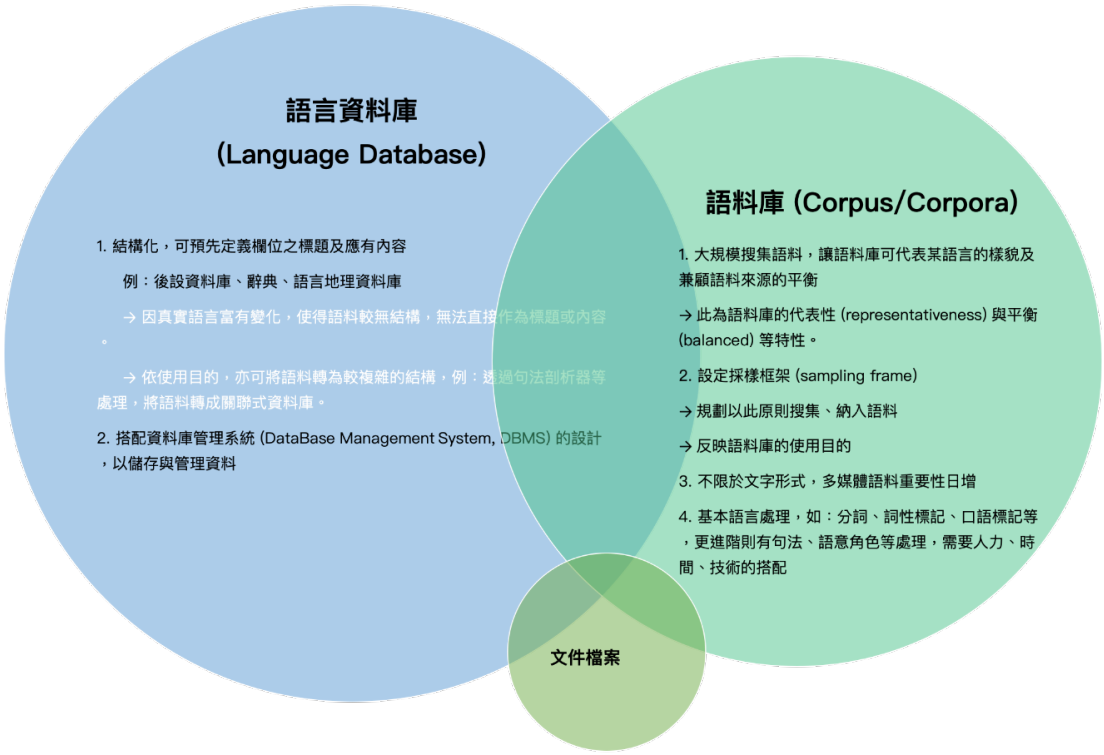


圖 29. 語言資料庫與語料庫之差異圖示

在第二屆中研語言學論壇「國家語言語料庫：規劃與建置」中，日本國立國語言研究所的前川喜久雅教授（Dr. Kikuo Maekawa）提到現代語料庫應包括以下七點特質：代表性（representativeness）、平衡（balance）、規模（size）、真實性（authenticity）、可機讀性（machine-readability）、公開取得（public availability）以及標記（annotation）。(Maekawa, 2019) **代表性**與平衡是相互關聯的概念，若語料庫的語料是具有代表性的，從小部分的語料分析即可看出該語言的趨勢，而**平衡**指的是語料庫須涵蓋各種語域（register），但每個文化有獨特的主題內容，有時候也讓這個指標變得相對主觀，這也是為什麼日本平衡語料庫採用隨機抽樣（random sampling）的方法建置而成，詳見「1.3.2 現代書面日語平衡語料庫（現代日本語書き言葉均衡

コーパス；Balanced Corpus of Contemporary Written Japanese, BCCWJ）」一節。**真實性**指的是語料庫內的語料皆來自實際使用的語言例子，若是新收語料則盡量不受到相關工作人員影響，而字典編纂的例句可能也不符合此一指標，至於口語語料則視其轉寫與標記等處理多少程度能夠反映原始語料，以及使用目的是否能被滿足而定，但作為語料庫建置者，不一定能夠決定所有的處理原則。語料庫**規模**並非純數學的問題，而是在研究成本、資訊技術及標記品質等多項因素考量下所做的決定，然而許多機器學習的訓練也是奠基在龐大的語料之上的。**可機讀性**在目前的時代來看已非字體顯示、儲存容量與速度的問題，更重要的是資料交換的標準化（standardization of the format for data exchange）及語料分析紀錄的方法。**標記**包含斷詞、詞性標記、句法依存關係、指代（anaphora）/後指（cataphora）、語意標記（semantic labeling）、音韻結構等，而日文語料遇到的問題是書寫的多樣化（multiplicity），在日本平衡語料庫中一個詞可能會有多達六種的書寫方式，若沒有處理這個問題，會大大降低語料庫檢索的意義。在後設資料方面，以不違反語料提供者的隱私為原則，提供作者姓名、性別、出生年份、出版年份、出版社、出版類型、書名或文章名、編輯者等。**公開取得**不一定要無償公開，但可免費提供部分語料，完整語料則可斟酌收費。(Maekawa, 2019)

至於臺灣的國家語言語料庫建置，前川喜久雅教授提到 (1) 許多國家語料庫都是單語（monolingual）的語料庫，而臺灣卻須顧及每個《國家語言發展法》的語言。(2) 語言使用的人口數無法決定各語言在國家語料庫的比例，而且口語語料對於原住民族語來說很重要。(3) 閩南語、客語、原住民族語等因為歷史因素，會有許多語言混用的情形，亦是國家語料庫建置能回饋的語言使用情形。(4) 語料庫建置是語言學

與資訊學的結合，但兩者常有不同的考量，尤其是對於語料的品質及規模，必須取得折衷的辦法（a reasonable point of compromise）。(5) 語料庫建置是不斷持續的過程，因為語言不會在建置完成便停止演變。(Maekawa, 2019)

另外，張榮興委員也建議道：「有關國家語料庫之建置，建議可從「人」（語料庫之使用者、建置者）、「物」（語料庫）及「機構」（整合平台）三大面向思考語料庫之架構規劃（如下圖所示），據此進一步研析相關重點。」

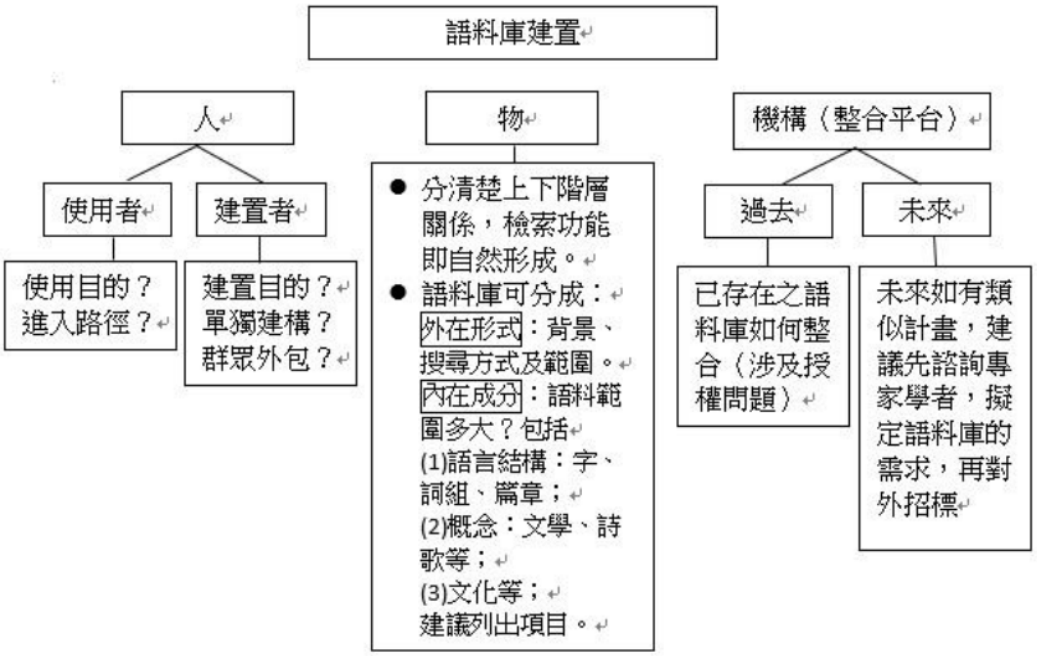


圖 30. 張榮興委員建議之語料庫架構規劃

以《國家語言發展法》為宗旨之下，將委員建議之目的、階層、各任務項目整理成之國家語言資料庫規劃內容與項目架構圖（請參考圖 31）。

圖 31. 國家語言資料庫規劃內容與項目架構圖

7.1. 國家語言現況

7.1.1. 語言瀕危、傳承危機

為了要讓國人瞭解並關注瀕危的國家語言，國家語言資料庫的入口介紹及背景說明可彙整母語流失相關主題的紀錄片、動畫或文字，例如：公共電視台、公視手語新聞、客家電視台及原住民族電視台。如果前述的資料不夠，建議可再製作其他短的紀錄片。

除紀錄片外，也可以放入相關國際網站連結，例如：《世界瀕危語言地圖冊》（UNESCO Atlas of the World's Languages in Danger，網址為：http://www.unesco.org/languages-atlas/）、《身心障礙者權利公約》（Convention on the Rights of Persons with Disabilities, CRPD，網址

為：https://www.un.org/development/desa/disabilities/convention-on-the-rights-of-persons-with-disabilities.html）手語語言權的相關內容、以及《國際手語日》（International Day of Sign Language，網址為：https://wfdeaf.org/iwdeaf2019/）等，讓民眾了解臺灣各語言的瀕危程度與傳承危機。

最後，國家語言資料庫可以設置專區，專門收錄我國政府機構所發佈，或者政府機構委託或補助學者所進行之相關語言調查報告，作為前述瀕危語言議題之相關影片和網站連結的進階補充。倘若國人在瀏覽過語言流失議題相關影片和網站連結後，還想得知進一步資訊的話，就可以到國家語言調查報告專區調閱相關報告。

7.1.2. 主題紀錄片、動畫或文字

針對口語語言，客家電視台和新媒體製作公司臺灣吧合作釋出《客客客棧-「參。拾母語」特輯》影片及《參。拾母語》網站。影片以閩南語、客語及原住民族語為題，深入淺出地介紹語言承載文化、國語運動、語言流失、語言復振等議題，華語版影片網址為：https://www.youtube.com/watch?v=3Td_jldDW44；另有客語版影片，網址為：https://www.youtube.com/watch?v=VZ3kkpuzHD4。《參。拾母語》網站則進一步介紹了母語斷層危機等相關議題，網址為：https://hakkafa.hakkatv.org.tw/。

與臺灣手語相關的媒材，可放入2019年臺灣國際聾人電影節的宣傳影片（臺灣國際聾人電影節於2015年首辦），此次電影節響應國際手語日（International Day of Sign Language）主題「人人都有使用手語的權利（Sign Rights for All）」，由中華民國聾人協會和臺灣文學館合作，影片網址為：

https://www.facebook.com/TWIDFF/videos/380324682639273/?vh=e&extid=ypHitj82QC1SivcK。

以上資料很適合放到國家語言資料庫上供國人參考，了解國家語言資料庫與語料庫的建置背景。如果前述的資料不夠，建議也可再製作其他短的紀錄片來補足。以下是《參。拾母語》網站所附的客語流失、不同語言的使用比例等相關圖表：

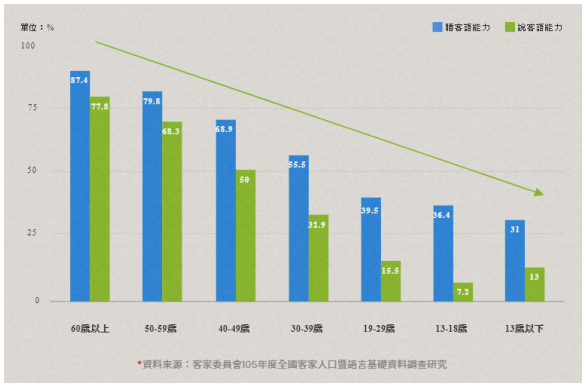


圖 32. 客家人的客語聽說能力也隨著世代越年輕而遞減（資料來源：用圖表帶你看母語斷層危機）

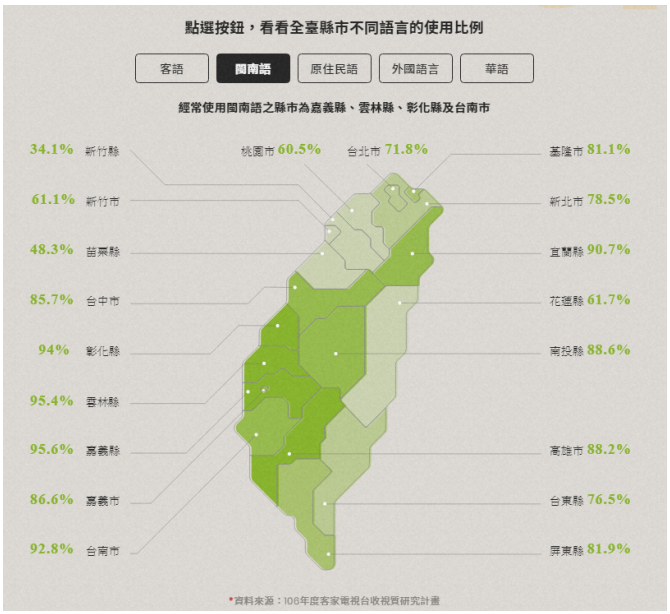


圖 33. 全台縣市閩南語使用比例（資料來源：用圖表帶你看母語斷層危機）

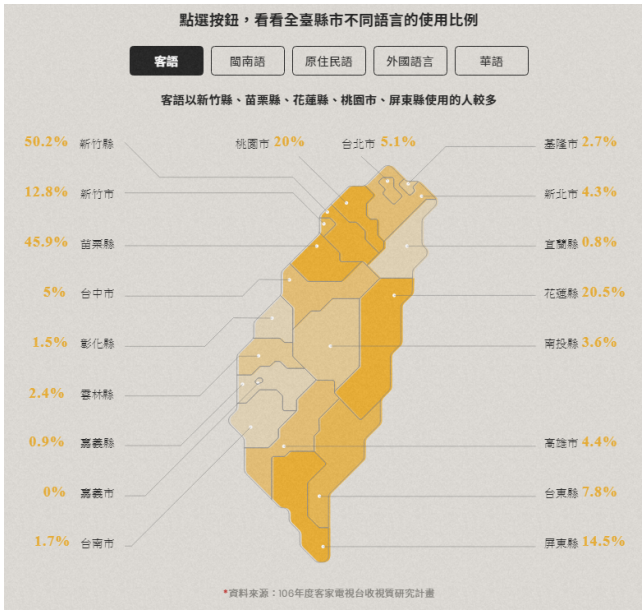


圖 34. 全台縣市客語使用比例（資料來源：用圖表帶你看母語斷層危機）

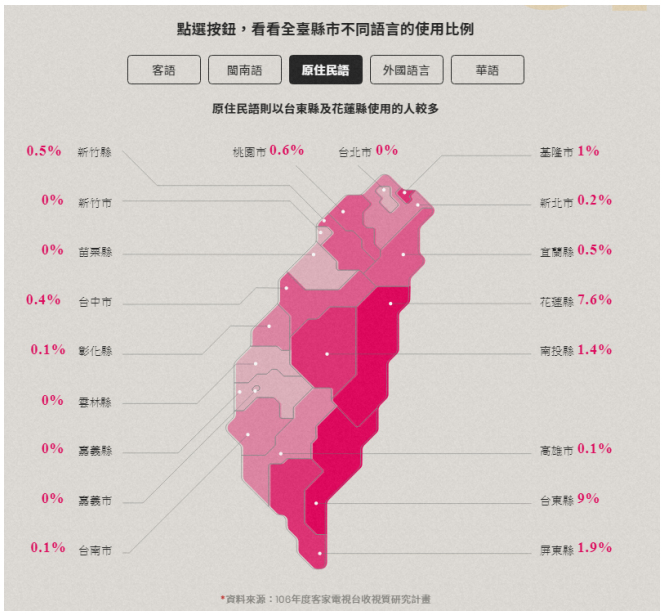


圖 35. 全台縣市原住民語使用比例（資料來源：用圖表帶你看母語斷層危機）

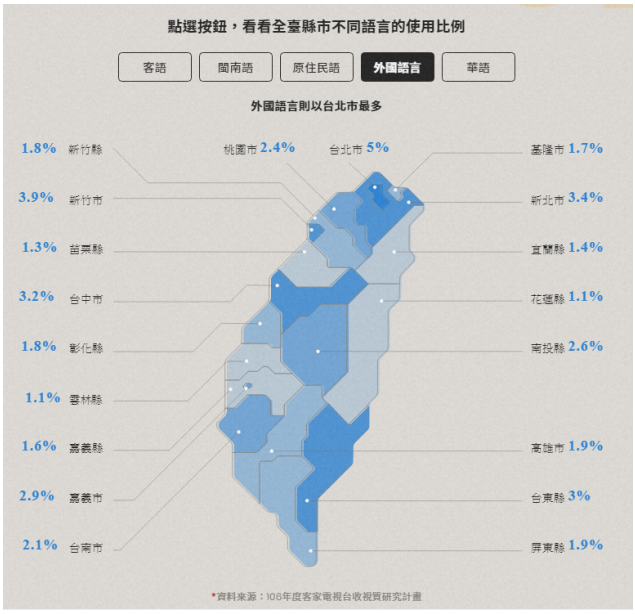


圖 36. 全台縣市外國語言使用比例（資料來源：用圖表帶你看母語斷層危機）

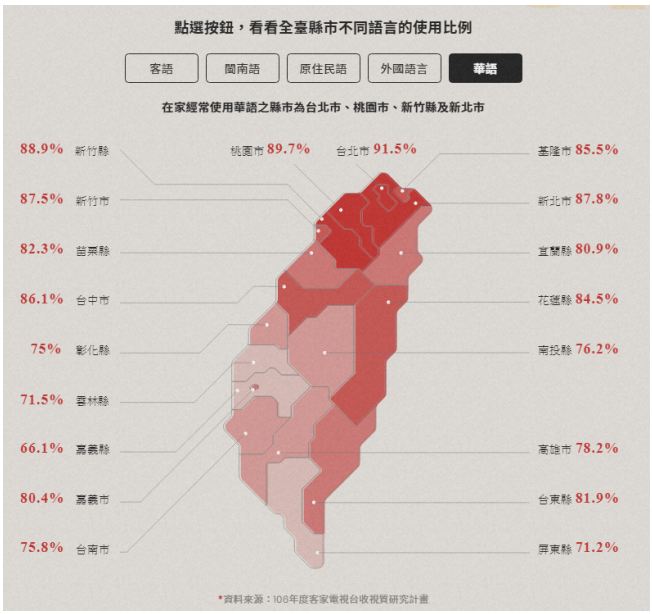


圖 37. 全台縣市華語使用比例（資料來源：用圖表帶你看母語斷層危機）

7.1.3. 整體分佈、個別語言介紹

此區關於國家語言的地理分佈，可邀請專家學者為各國家語言撰寫短篇介紹文。相較於「7.1.2 **主題紀錄片、動畫或文字**」的初步介紹，若想要參閱較學術的資料，可前往閱讀此區的短篇介紹文，例如：洪惟仁教授於附錄一的閩南語介紹文、張永利教授於附錄二的原住民族語介紹文、客語及手語部分亦希望邀文介紹個別語言，包括臺灣手語的歷史淵源、語言接觸與自然手語發展等面向。(Smith, 2015)

7.1.4. 相關議題連結

UNESCO(聯合國教科文組織)在 2009 年 2 月 19 日公布了「世界瀕危語言地圖冊」（UNESCO Atlas of the World's Languages in danger），並將語言依照情況輕重分為無危(Safe)、脆弱(Vulnerable)、危險(Definitely endangered)、重大危險(Severely endangered)、極度危險(Critically endangered)和滅絕(Extinct)六個等級。2016 年 UNESCO 總共

評估了臺灣 24 種原住民語言，其中 9 種平埔族語言巴賽語(Basay)、凱達格蘭語(Ketangalan)、龜崙語(Kulun)、道卡斯語(Taokas)、巴宰語(Pazeh)、拍瀑拉語(Papora)、巴布薩語(Babuza)、和安雅語(Hoanya)、西拉雅語(Siraya)已被列為「滅絕」；5 種語言噶瑪蘭語(Kavalan)、葛蘭語(Nataoran)、邵語(Thao)、卡那卡那富語(Kanakanavu)、拉阿魯哇語(Saaroa)為「極度危險」；賽夏語(Saisiyat)為「重大危險」；布農語(Bunun)為「危險」；還有 8 種語言泰雅語(Atayal)、太魯閣語(Taroko)、阿美語(Amis)、卑南語(Puyuma)、魯凱語(Rukai)、排灣語(Paiwan)、雅美語(Yami)和鄒語(Tsou)為「脆弱」。這項結果顯示臺灣的原住民語言正面臨嚴重的傳承與語言流失問題！而透過放入相關國際網站連結的方式也可以激發國人對於本國原住民語言流失議題的重視！

圖 38. 世界瀕危語言地圖冊之臺灣原住民語言瀕危情況

臺灣手語的使用人口則可參考衛福部每年的身心障礙統計資料，據衛福部 2020 年更新的資料統計顯示，聽覺機能障礙者（Hearing Mechanism Disability）計有 124,485 人，男性較女性多，分別為 70,311 人及 54,174 人。依縣市別區分，新北市 16,843 人、臺中市 14,434 人及臺北市 13,362 人居前三。年齡與聽障人口呈現正向成長，未滿 3 歲之人口有 362 人為聾人，而 65 歲以上則高達 87,048 人。中華民國聾人協會官方網站（網址：https://www.nad.org.tw/old-www/）上提供了百分比資料，臺灣約有 5%人口為身心障礙人士，其中 11%為聽障人士。依原因來看，15%為先天性失聰，而 65 歲以上的聽障人士超過一半，佔 66%。(「聽障的人口有多少？」, 2016) 不過，以上資料屬於聽障人口依性別、縣市、年齡及成因的統計資料，尚未有「會說手語」的人口統

計，因此僅能從間接資料推想手語使用人口的範圍值為何，而 Smith (2005) 對臺灣手語的歷史回顧中提及使用人數約 30,000 人，但未敘述數據來源或方法。

7.1.5. 國家語言調查報告

聯合國的「世界瀕危語言地圖」網站上雖沒有明確列出臺灣閩南語、臺灣客家語的語言活力現況，但根據行政院 99 年人口及住宅普查初步統計結果提要分析與中研院社會所 2013 年的臺灣社會變遷基本調查計畫結果顯示，在臺灣社會中閩南語與客家語的使用率隨著年齡層下降而遞減，代表著閩南語與客家語也正面臨世代傳承與語言流失的問題！關於這兩份報告的語言調查方式，首先 99 年人口及住宅普查初步統計結果提要分析針對年滿 6 歲者（ 2004 年 12 月 26 日及以後出生）詢問兩道題目：「您在家裡使用哪幾種語言？」「您的父母互相溝通使用哪幾種語言？」兩題的選項都是：「國語」、「閩南語」、「客家語」、「原住民族語」、「其他」。申報人可以複選，亦即只要是會講的語言都可以選（詳見普查表(pdf 檔)）。而臺灣社會變遷基本調查計畫則是詢問年滿 18 歲民眾：「請問您在家裡最常講國語、台語、客家話，還是哪一種語言呢？」為了捕捉人們最常用的一種語言，當受訪者堅持有兩種以上時，才開放複數選項（詳見臺灣社會變遷基本調查計畫報告中的《研究問卷II國家認同組》）。表 16 為這兩份報告的語言調查結果。

表 16. 2010 年人口普查與 2013 年臺灣社會變遷調查比較（資料來源：葉高華（2018））

	2010 年人口普查			2013 年臺灣社會變遷調查			
	會說就選			選出最常說的一種			
	華語	閩南語	客語	華語	閩南語	含客語	華語+閩南語

出生年	～1945	45.3	81.7	10.1	12.3	71.4	8.8	5.7
	1946～1955	71.9	86.8	8.6	15.2	59.1	4.4	21.4
	1956～1965	83.8	85.9	7.7	19.8	56.2	4.5	19.0
	1966～1975	90.4	84.1	6.4	32.6	38.1	5.2	23.2
	1976～1985	91.9	83.2	5.6	43.6	28.6	2.5	24.7
	1986～1995	94.9	78.6	4.8	57.3	22.3	1.0	19.1
	1996～2004	96.0	69.7	3.8				

註：單位為%。臺灣社會變遷調查的原住民受訪者太少，不宜比較原住民族語，故省略。

兩份報告的語言調查結果都顯示，閩南語與客家語的使用率隨著年齡層下降而遞減，而華語使用率則隨著年齡層下降而遞增。不過，因為這兩份報告的問法不同，造成兩者得出的數據有落差。在 99 年的報告中只要民眾會說該項語言就會列入統計，而 2013 年的報告則是要求民眾選出最常使用的語言；前者的問法較鬆，因此臺灣社會語言流失的問題在該份報告中顯得不明顯，後者的報告結果應該比較符合臺灣社會母語流失的實際現況。

國家語言資料庫可以設置專區，專門收錄教育部本土語言調查報告、客委會於 2002-2016 年所進行的臺灣客家民族所使用的客語使用狀況報告、原民會於 2012-2015 年所進行的原住民族語言調查研究三年實施計畫 16 族綜合比較報告、衛福部每年更新的身心障礙者統計資料、及未來國家語言調查的相關報告等。

報告主題內容為語言使用或語言田野調查。該專區會以連結或是檔案下載的方式呈現，語言調查報告依性質可以大略分成兩個類別，第一類是政府機構的報告或相關彙整資料，包括：

(1) 行政院主計處在 99 年人口及住宅普查初步統計結果提要分析中調查國家語言使用的結果：https://www.dgbas.gov.tw/public/Attachment/111171361171.pdf。

(2) 教育部本土語言調查報告：https://mhi.moe.edu.tw/newsList.jsp?ID=5。

(3) 客委會於 2002-2016 年所進行的臺灣客家民族所使用的客語使用狀況報告：https://www.hakka.gov.tw/Content/Content?NodeID=626&PageID=37585。

(4) 原民會 2012-2015 年所進行的原住民族語言調查研究三年實施計畫 16 族綜合比較報告：https://mhi.moe.edu.tw/file/files/1050601-1-原住民族語言調查研究三年實施計畫第 3 期實施計畫 1 至 3 期 16 族綜合比較報告摘要彙編_(公告).pdf。

(5) 衛福部身心障礙者統計資料：https://dep.mohw.gov.tw/DOS/lp-2976-113.html。

(6) 未來國家語言相關機構所進行的語言調查報告等。

第二類則是政府機構委託或補助學者之研究案，例如：

(1) 中研院社會所 2013 年的臺灣社會變遷基本調查計畫中的國家語言使用結果。

(2) 文化部 2017 年的《全國語言基礎資料研究計畫》（編號：PG10602-0036）：https://www.grb.gov.tw/search/planDetail?id=12068997。

(3) 科技部 2015 年委託的《族語保存現況調查研究計畫成果報告》（NSC 101–2410–H–001–094–MY3）：http://dx.doi.org/10.1080/01434632.2015.1022179。

(4) 客委會 2013 年補助的《臺灣南部客家語言使用態度與使用行為研究—屏東市》（編號：PG10205-0110）：https://www.grb.gov.tw/search/planDetail?id=2945780。

透過設置國家語言調查報告專區的方式，可以讓想進一步研究本國語言使用與流失議題的國人，得知進階的詳細資訊。

7.2. 國家語料庫

國家語料庫應包含我國所有國家語言的語料庫，包括華語、閩南語、客語、原住民族語、臺灣手語、及閩東語。在第一、二章我們參考了不少國外各國家語料庫的設計，這些設計優點包括，美國國家語料庫提供了豐富的語言處理工具；日本國立國語研究所收錄了歷史、將日語作為第二語言等各種不同類型的語言資料；韓國語意網絡研究中心展現了以人工智慧為導向的資料建置成果，其世宗國家語料庫則擴及韓語語言學習者為使用對象；俄羅斯國家語料庫對於語法和語義有精細的分析處理等等。在手語語料庫的例子中，各國的手語語料庫各有特色，英國手語語料庫以社會語言學為基礎，蒐集各個語言及社會背景手語使用者的語料；澳洲手語語料庫則希望蒐集對手語使用社群來說實用的手語語料，以醫療、教育等為主題蒐集特定的語料；瑞典手語語料庫則以回饋教學現場為建置目的，讓使用者上傳自己的語

料、以及釋出可線上檢索的網頁版語料庫。本章將會參考這些優點，並將其納入我國國家語料庫的設計當中。

另一方面，在可能遇到的難處部分，參考英美國家語料庫的建置過程，可發現口語語料比例仍舊較低，因為口語語料蒐集後須經過文字轉寫、發音人背景資料匯入後設資料庫等步驟，所需的時間與處理較書面語料繁複。此外，手語語料庫亦面臨類似的困難，須仰賴受過訓練的人員為原始影像加上各項標記。更長遠來說，若是將國家語言資料庫發展成平衡語料庫，例如：美國的 OANC 及英國的 BNC，則須大量蒐集新的語料，並將語料經過前處理與標記，尤其是需要轉成 XML 的格式方能進行檢索，因此採用如 OLAC、都柏林核心集等國際標準，能夠減少開發的困難度。

原則上，為了爭取時效，初期現有資料會採用連結方式，連結到各單位的語言資料庫為主；閩南語、臺灣手語語料庫、閩東語因為沒有專責機構負責，建議由文化部建置並維護這三個語料庫，其考量原因如下：首先，因為國家語言語料庫預計會收錄公視台語台的相關影音資料，而公視台語台又是由文化部編列預算所補助成立，因此由文化部專責主導閩南語語料庫的建置與維護最為適合。在臺灣手語方面，目前和臺灣手語較有相關的政府部門包括衛福部、教育部與文化部等，考慮到衛福部是有關公共衛生、醫療與社會福利事務的最高主管機關、教育部則側重教育相關的事務，因此，若要較全面地整合與建置手語相關語料庫，目前看來文化部是最適合做主導的機構。同理，閩東語並沒有中央主管機關負責，文化部為全國文化相關的中央主管機關，閩東語語料庫由文化部建置應是合理且最佳的安排。雖然客語和原住民族語都分別有客委會、原民會等專責機構主導建置並維護相關語言資料，然而目前建置國家語言資料庫的工作是由文化部所

主導，未來客語、原住民語相關資源也會納入國家語言資料庫，並且再視情況做一定程度的修改或增補；為了統合各語言資料的規格與內容設計，由文化部來主導建置並維護閩南語、臺灣手語、閩東語語料庫，是最為適切的做法。關於國家語言料的建置原則，在此提出六項建議：

(1) 依據重要性及時效性分階段逐步整合現有資源及建置新資料。

(2) 沒有專責機構負責或缺乏資源的國家語言資料優先建置。

(3) 規畫一部分開放資料提供下載。

(4) 訂定國家語言資料的通用格式，依據需求利用群眾外包向民間徵求各國家語言的相關語言資料。

(5) 與國際接軌，善用已開發的各種開發工具。

(6) 與學界及民間對國家語言資料庫開發有興趣的人士共同組成學術社群，開發相關應用軟體。

不過，考量到語料庫的內容項目基本上會因擁有者的背景知識而有所差異，加上國家語料庫的設置無法一次到位，必須分次逐步完成，因此，建議未來宜與相關的專家學者組成審查委員，指導專責機構去討論與執行語料庫的各項內容，以尋求最大共識。

在語料蒐集方面，可參考第五章所列各單位的語料庫，以著作權沒有疑義者優先納入國家語料庫並提供跨語言檢索功能。不同語料庫之間若有共同的部分，如華語解釋或共同的詞性標記就能夠進行交叉檢索，可以達到語言推廣、傳承之長遠目標。就第五章所收集的各語種資料來看，目前華語資料最充足；客語正由客委會主導建置書面語料達 1,800 萬字、口語語料達 30 萬字規模的客語語料庫；原住民語雖正由原民會主導整理相關語言資料的工作，但因語種和方言類別繁

多，加上部分族群人口數稀少（如，根據 2020 年 1 月現住原住民人口數按族別及年齡分統計表顯示，卡那卡那富族、拉阿魯哇族、撒奇萊雅族、邵族等族人口數不到千人），收集和整理相關語料的工作可能還需要更多的時間、與進一步的詳細規劃。綜合上述各點，建議未來建立國家語言資料庫時，無專責機構主導，或屬於整合不同資料，這類的語言資料庫宜及早規劃。這些包括 (1) 國家語言資料庫網站（含國家語言現況介紹）、(2) 閩南語語料庫、(3) 閩東語語料庫、(4) 臺灣手語語料庫、(5) 國家語言辭典資料庫及國家語言地理資訊系統等五項。建議國家語言資料庫建置原則採用由上而下及模組化設計，先進行國家語言資料庫的整體架構與網站的建置，內容部分則分階段於子語料庫及子資料庫陸續建置完成後再與主系統銜接。由於牽涉不同部會的業務，應舉行跨部會的協調會議，減少整合時可能會產生的問題。以下針對閩南語、臺灣手語、及閩東語這三個語料庫的建置提出建議。

7.2.1. 閩南語語料庫

針對閩南語語料庫的建置規劃，專家諮詢會議中獲得各委員的寶貴意見，以下是本團隊提出之閩南語語料庫規劃圖：

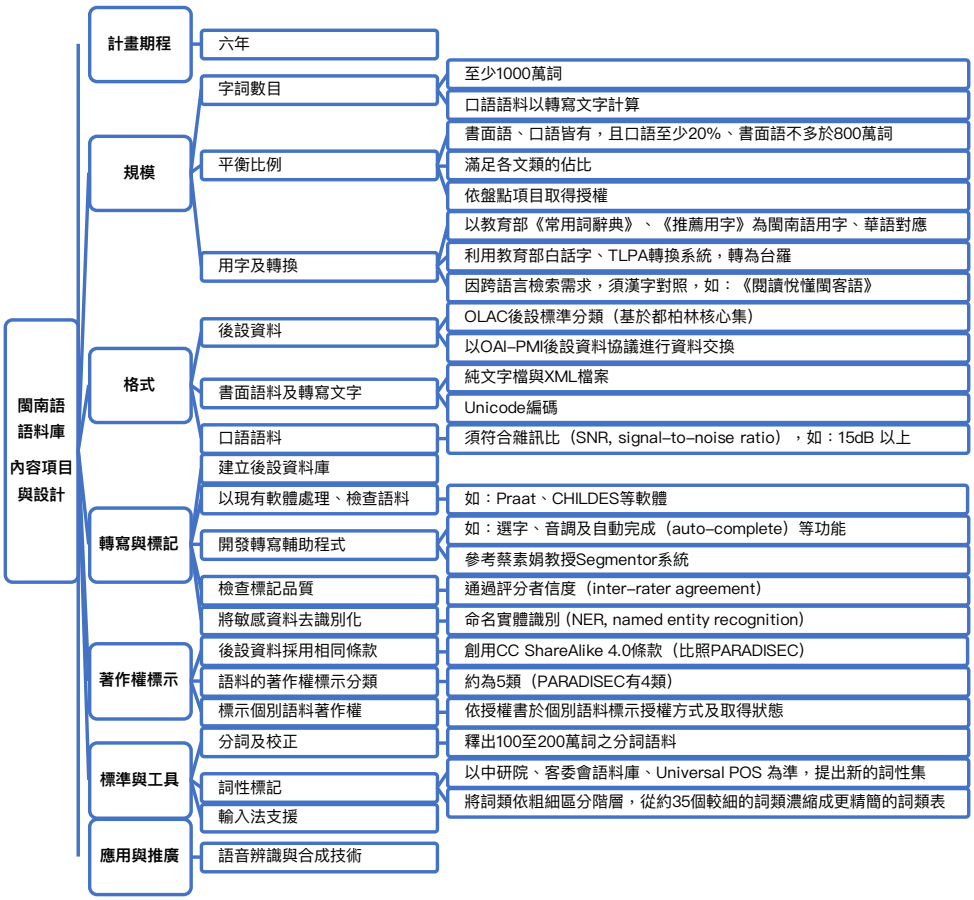


圖 39. 閩南語語料庫規劃圖

此外，江敏華委員在期中報告審查意見表中提到：「就第五章所列之既有語料庫，在閩南語部分顯然不足以建置一個類似平衡語料庫的檢索系統，一來資料量太少，二來其中有許多是教學資源，長篇語料及自發性語料不足」。關於閩南語語料，較近代的書面語語料來源建議可納入教育部歷年本土文學創作獎之得獎作品（客語、原住民族語亦有）。洪惟仁教授建議可以將閩南語老電影、老歌納入平衡語料庫（如國家電影中心的台語片 60 週年、國立臺灣歷史博物館的臺灣音聲一百年等資料），因為這些資料有不少都已經超過 50 年的著作權保護期，因此只要經過適當的轉寫與標記工作，便可作使用。另外，關於自發性語料不足的部分，可以先考慮將公視台語台各節目的閩南

語、華語字幕皆整理出來，應該就能作補足。關於字幕的部分，目前意傳科技開發了可以將閩南語轉換成華語的 app，另外該公司也提供影音字幕自動化等的服務；因此，未來或許可以諮詢或委託意傳科技利用機器自動辨識，來整理公視台語台的字幕資料，接著再由具閩南語語言學背景的專家來人工糾錯。意傳科技亦在進行泰雅語等原住民族語的語音辨識，可參考其 GitHub https://github.com/i3thuan5/tai5-uan5_gian5-gi2_kang1-ku7/pull/379。目前其他單位正在開發建置的語料庫部分，則先以超連結的方式引導使用者使用該語料庫，待其計畫完成後，再進入整合的階段，並檢視國家語料庫中何種語料需要添增，以求建立語料庫語料的平衡。

此外，目前閩南語語料還具有用字不統一的問題，關於這部分，江敏華委員認為建置閩南語語料庫無法一次到位，必須分階段執行，因此建議可以先典藏各閩南語相關資料，接著再逐步將各資料的用字統一，並做成平衡語料庫。而何信翰委員也建議，閩南語語料的羅馬字部分可以優先收錄較多人使用、且資料量也相對較多的教羅、台羅相關語料；至於漢字部分，可將教育部異體字字典的資料納入。

7.2.2. 臺灣手語語料庫

目前世界上的手語語言資源，資料庫是常見的資源，以「手語資料庫（SignBank）」或「詞彙資料庫（lexical database）」為名，如：美國手語資料庫（ASL SignBank，全名為 American Sign Language SignBank）、美國手語詞庫（ASL-LEX，全名為 A Lexical Database of American Sign Language）、荷蘭手語資料庫（NGT SignBank，全名為 Nederlandse Gebarentaal SignBank）等，亦有線上手語辭典。

目前有兩項由中正大學發展出來的重要資源，分別是臺灣手語線上辭典（網址：http://tsl.ccu.edu.tw/web/chinese/）以及臺灣手語電子資

料庫（網址：http://signlanguage.ccu.edu.tw/）。關於手語資料庫的授權，目前中正大學語言學研究所的蔡素娟教授表示，已詢問學校研發處有關語料授權的處理，希望能夠依據科技部的辦法，以學術研究、教育或公益用途的方式，無償提供語料，將來由執行單位與中正大學簽約，完成相關程序。

對於以視覺方式表達意義的語言來說，手語語料庫可提供不同於手語資料庫的語言資訊。透過臺灣手語語料庫的建置，可更加了解語言的多樣性，深化臺灣手語語言使用的社群基礎，並借助資訊技術的幫助，更細緻地處理影像與時間的資料，跳脫紙本資料對視覺語言的限制。

在新資料蒐集方面，可由臺灣手語使用者擔任發音人，錄製多角度、以素色為底的清晰影像，蒐集內容可參考英國手語資料庫設計詞彙列表、以圖像引導（elicit）發音人打出該手語詞彙，抑或是參考澳洲手語資料庫的作法，邀請二或多位發音人組成焦點團體（focus group），從訪談或對話中擷取主題式資料。在資料處理方面，由於臺灣手語無一定書寫系統，須仰賴標記處理及時間對應進行日後的檢審、利用，常見的標記項目包括「手形」、打法「位置」及「意義」說明等。另外，美國手語詞庫計畫（ASL-LEX）特別收錄語料庫較少出現的詞彙、以及心理語言學實驗的相關資料，增加資料量與標記量，亦是臺灣手語資料庫在資料擴充階段可參考的例子。

在資料呈現方面，應在檢索頁面提供影像及標記訊息，可讓使用者調整影像的播放速度、下載部分影像檔或標記檔等。此外，考量手語語料以影像檔為主，與口語語言的純文字檔不同，其影像檔應以一級/原始資料的形式存取，意即完整保存、多層級及異地備份，而標記

檔因有校正、更新之需求，應紀錄其版本，區分可上線及仍在處理之資料。

考量手語特色與口語語言迥異，建議讓相關研究團隊或組織參與，例如：瑞典手語語料庫（STS-korpus）在計畫初始即設定為多語料庫的系統；美國手語詞庫（ASL-LEX）及美國手語資料庫（ASL SignBank）合作，針對手語不同面向展開各項標記計畫；澳洲手語資料庫（Auslan SignBank）依醫療、教育等主題切分子計畫，便可集結各方人力投入，逐步充實臺灣手語語料庫的內容。

對手語語料庫來說，影像檔、時間訊息和語料標記是核心元素，應提供空間儲存畫質較高的原始影像，而語料庫檢索時可考慮讓使用者能夠調整畫素或播放速度，依「7.4.2 資源分享」的原則來看，原始影像應以原始資料的方式儲存，進行異地備份、多層級備份等。在語料蒐集方面，國外手語資源的發音人多為手語母語者、達母語程度的手語使用者，建議與相關單位、團隊合作，另外亦可採焦點團體（focus group）的方式，邀請手語使用者、手譯員、家屬與社群進行相關議題討論，從生活面了解手語社群的需求及奠定更深厚的社群基礎，而非直接帶入聽人的思維，造成自然手語的發展再次受挫。在語料處理方面，Johnston (2009) 提到手語語料庫需要的是語料標記，從各項標記逼近手語表達，例如：手語打法的「位置」及「手形」，而越多的標記越能將手語中的最小對立體（minimal pair）區別出來。(Crasborn et al., 2018)

有關於臺灣手語語料庫的規劃，我們的建議如下：

(1) 建立總共至少 10 小時長的臺灣手語影片及檢索系統。其中需至少有五分之一包含兩位臺灣手語者的手語對話影片。

(2) 領域及主題需儘量多元。

(3) 參考英國手語語料庫和澳洲手語資料庫的作法及設計，由臺灣手語使用者擔任發音人，錄製手語使用者的敘事、對話、訪談影片。蒐集內容可參考英國手語資料庫設計詞彙列表、以圖像引導（elicit）發音人打出該手語詞彙，亦或是參考澳洲手語資料庫的作法，邀請二或多位發音人組成焦點團體（focus group），訪談或對話中擷取主題式資料。一部份資料可採用公視手語新聞的影片。

(4) 須標記處理項目包括「手形」、打法「位置」及「意義」說明等項目，並須附上華語翻譯。

(5) 檢索頁面提供影像及標記訊息，可讓使用者調整影像的播放速度、下載部分影像檔或標記檔等。

(6) 手語語料庫應整合手語資料庫。建議使用開源的 NGT 資料庫並使用 ELAN 軟體以 EVC（external controlled vocabulary）檔來管理資料庫的詞彙，標記人員在標記語料時便從此 EVC 檔案尋找符合的詞條。可以 ID 查詢表確認該詞彙是否收錄在資料庫中，再確認該詞彙是否多義。在標記上，如果有多個選擇，以使用頻率（frequency）和象似性（iconicity）為原則，標記項目包括：複合詞的切分及對應詞彙、單手或雙手（handedness）及是否雙手動作對稱（symmetry）、手形變化（handshape changes）、動作方向（movement direction）及是否重複（repetition）、手勢打在身體的哪個部位上（location），此外亦提供實名（name entity）及語意欄（semantic field）的標記該資料庫的系統可即時更新最小對立體（minimal pair）的資料。

(7) 建議詞彙分析方面採用以下的原則標記 (Becker, 2020)，以便跨資源搜尋，例如：

(1) 手勢（handedness）：若手語打法為雙手時，通常會有對稱或交替的動作，標記為[SymmetricalOrAlternating]，若無則為[AsymmetricalSameHandshape]，單手打法則標記為[OneHanded]。另一種打法有主副手之分，標記為[AsymmetricalDifferentHandshape]。若不符合上述對稱原則（symmetry condition）或主副原則（dominance condition），則標記為[Other]。

(2) 主要身體部位（major location）：臂腕（arm, including wrist）、軀幹（body, signer's torso）、手（hand）、頭及臉部（head, including face）、無特定部位（neural, signing space in front of singer's body）等。

(3) 主手（dominant hand）、使用的手指（selected finger）、以及手指彎曲狀態（flexion）：使用的手指是依手指彎曲或伸直的狀態而定，而手指彎曲的標記分為 9 類（categorical），而非給予連續性的數值（numerical）。

(4) 尚未包含手語打法的動作方向（direction of movement），因此有些不同的手語詞彙在兩個資料庫的標記資料是相同的。

(5) 在網頁介面上，以關鍵字檢索句（KWIC, keyword in context）的方式呈現，並附有上述的語料標記。此外，使用者亦可勾選查看更多資訊，例如：語料庫來源（Korpus）、該檢索句的時間長度（Längd）及起迄時間（Start, Slut）、語料來源檔名（Radnamn）、檔案敘述（Filbeskrivning），未來希望可以依照這些資訊的選項排序。在語料標記區塊的下方則標示了符合

搜尋結果的時間，因此使用者可在特定檔案中縱覽時間軸上的搜尋結果，此一設計亦可看出該搜尋結果的分布狀況（dispersion）。

7.2.3. 閩東語語料庫

(1) 規模：以六年期計畫建置一百萬詞規模的閩東語語料庫。包含書面語及口語兩部份，共 100 萬詞。為兼顧平衡性之考量，書面語語料以不超過 80 萬詞為原則。口語部分以不少於 20 萬詞為原則。提供以句為單位的閩東語與華語對照。

(2) 資料格式與其它國家語料庫一致，採 XML 格式、Unicode 編碼。語料需提供分詞及詞性標記。

(3) 在語料蒐集方面，以現有閩東語語言資源為主，書面語料可以可以連江縣本土資源教育網裡面的資源為基礎，口語語料可納入兒歌、歌曲比賽得獎作品等。蒐集日常使用語詞、諺語、兒歌、歇後語，兒歌作為語料庫材料。馬祖社區電台的廣播節目內容，如一分鐘母語教學、60 秒母語形象、20 分鐘母語教學、母語線上廣播等可作為口語語料。

7.2.4. 語料庫整合、跨語言檢索

待各國家語言語料庫建置完成後，便可進行語料庫的整合與跨語言檢索。江敏華委員提到，考量到未來國家語料庫的使用者不一定皆具備用教羅或台羅做檢索的能力，因此建議可以設置一個「對照語料庫」，把常用的閩南語漢字和羅馬字列出來。而何信翰委員也建議可以在檢索系統建立像是 google 搜尋引擎的「模糊查詢」設計，如此一

來即使使用者無法打出正確的用字作檢索，系統也能自動糾正或者推薦相關的查詢結果。至於羅馬字聲調的輸入部分，何信翰委員表示可以考慮一次提供聲調符號和數字聲調兩種輸入法，另外也可在輸出結果部分提供圖檔或 pdf 檔下載，以解決部分使用者電腦無法輸入或者顯示閩南語聲調符號的問題。張榮興委員也提到手語語料蒐集方面，語料來源者的問題，若是從新聞蒐集手語語料，可能不會是原始的手語，而是基於華語等當時被翻譯者所使用的語言的翻譯，因此翻譯語料與單語語料也可以做出區別。

在圖書資訊學的領域範圍，跨語資訊檢索（Cross-Language Information Retrieval）的定義為「使用不同於書面語的查詢語言進行資訊的檢索（“select information in one language based on queries in another”）」。(陳, 1998) 應用於雙語或多語語料庫上，可採用詞彙對照（dictionary-based）或平行語料（corpus-based）兩種方式進行跨語言檢索 (陳, 1998; Hull, 1997)，以下分述之。葉茂林委員也建議，關於資料檢索（index）系統之建置，亦可諮詢圖書資訊領域之專家學者 。

7.2.5. 詞彙對照

以閩南語常用詞辭典、客語常用詞辭典及原住民族語線上辭典等為準，透過對應華語的詞彙檢索閩客語、原住民族語的詞彙，這部分的概念及工作與「用字規範與對應華語」一節類似，數位政委唐鳳的開源計畫「萌典」亦以閩客語常用詞辭典作為與華語的對照，以下是搜尋「我們」的結果，附有閩、客語與華語的對照：

圖 40. 於「萌典」網站搜尋「我們」之結果

陳光華教授表示 (1998)，詞彙對照會面臨歧義（ambiguity）的問題，可計算不同語意出現的頻率，選擇最佳 N 個意義（select best N），並融合模糊處理（Fuzzy processing）的設計，讓使用者能夠盡可能查詢到所需的資訊。

7.2.6. 平行語料

若成功蒐集到平行語料，可抽取不同單位的對列、對齊（alignment）以進行跨語言檢索。對齊技術指的是以句或詞為單位，將不同語言的語料對照，在口語語料方面則包含時間的對齊（time alignment），圖 41 是國教院華英雙語索引典系統的跨語檢索畫面。

圖 41. 國教院華英雙語索引典系統中，搜尋「爆發」一詞的結果

針對臺灣國家語言的平行語料蒐集，2000 年徐兆泉老師出版了客語版的《小王子》，2018 年年底客語教師謝杰雄合力百位客語教師，將《安徒生全集》翻譯成客語，分為四縣腔、海陸腔兩個版本，共有 166 篇故事。(“童話大師安徒生會說、也會寫客語了”, 2019) 2020 年 3 月 18 日甫有閩南語版本的《小王子》問世。(“說台語的《小王子》！法文譯者蔡雅菁的母語之旅”, 2020) 該書的出版源於文化部「本土語言創作及應用補助出版」計畫，採用閩南語常用詞辭典之用字，並附有 QR Code，讀者掃描後可同步聆聽朗讀音檔，(“【藝術文化】大人囝仔上深的感動 經典文學小王子台語有聲版問世”, 2020) 審定委員黃震南表示，閩華之轉換在詞彙上常無直接對應，例如：「午安」、「晚安」是華語詞彙，而「想了半天」不應翻成「想了半工」，而是「想歸

埔」，透過閩南語翻譯計畫「體現臺語文化特點」。如小王子、安徒生全集都是平行語料的例子，有了平行語料後，經過對齊的技術處理，讓使用者進行跨語言檢索。

另外，輸入法可參考第 7 章「用字規範與對應華語」一節。以閩南語或客語拼音方式輸入拼音，由輸入法列出對應漢字供使用者選字，「免除轉譯成華語音讀的困擾」，且由於部分漢字為 Unicode 缺字，使用者之電腦無法顯示該字，需要安裝造字集，或是以圖檔的方式顯示該字。另外，以臺灣的國家語言來說，華語無詞彙原形（lemma）與屈折變化（inflection）的差別，但原住民族語則有，因此在檢索條件的設定上，如能提供原形檢索的功能，將會更符合各語言的特性。

7.3. 國家語言辭典資料庫及地理資訊系統

有鑒於語料庫與資料庫性質不同，前者以在特定語境下自然產生的語料為內容，資料庫則可加入語料庫未涵蓋的資訊，尤其是詞彙資料庫（lexical database）、語言地圖（linguistic map、linguistic atlas）等，不論是獨立的查詢系統，或是進一步提供跨資料庫、資料庫-語料庫雙向搜尋的功能，都可讓使用者方便且快速地找到相關資訊。

7.3.1. 國家語言辭典資料庫

在「伍、本國國家語言相關之語言資料庫」一章中，已針對華語、閩南語、客語、原住民族語言、閩東語、臺灣手語列出語料庫及非語料庫資源，可知現有資源中，語言學習資源比例不少，像是辭典、教材、有聲書、教學網站等，江敏華委員建議可從建立跨語言辭典資料

庫開始，將目前市面上能夠找到且能夠取得授權的字辭典完整數位化，並提供單一辭典或跨辭典的查詢功能服務。

各國家語言辭典資源，建議可先整合「伍、本國國家語言相關之語言資料庫」一章中所提到的辭典相關資源，來逐步建置華語、閩南語（含閩南語各個主要的方言）、客語（含四縣腔、海陸腔、大埔腔、饒平腔、詔安腔）、原住民語（含各語族方言別）、閩東語、臺灣手語的辭典，附上釋義、例句及影音檔，從詞彙的層面呈現各國家語言的文化底蘊，並促進不同國家語言間的文化交流。各國家語言辭典相關資源如下：

(1) 華語辭典資源包括《教育部重編國語辭典修訂本》、《教育部國語辭典簡編本》、《教育部國語小字典》、《教育部異體字字典》、《教育部成語典》等。

(2) 閩南語辭典資源包括《教育部臺灣閩南語常用詞辭典》、《國臺對照活用辭典》、《簡明臺灣語字典》等，還有《閩客語典藏》網站所收錄的字典典藏（《廈英大辭典》、《英廈辭典》、《廈門音新字典》、《台日大辭典》）。另外像是《臺灣閩南語羅馬字拼音方案》、《臺灣閩南語漢字之選用原則》、《臺灣閩南語推薦用字 700 字詞》、《臺灣閩南語我嘛會每日一詞》等相關資源也可納入。

(3) 客語辭典資源包括《教育部臺灣客家語常用辭典》、《客話辭典》、《臺灣四縣腔海陸腔客語辭典》、《六堆辭典》等，還有《閩客語典藏》網站所收錄的字典典藏（《客英大辭典》、《客法大辭典》）。

(4) 原住民語辭典資源包括《原住民族語言線上詞典》、《巴宰語詞典》、《噶瑪蘭語詞典》、《達悟語詞典》等。另外像是《臺灣原住民語言推薦新詞》等相關資源也可納入。

(5) 閩東語辭典資源包括《連江縣本土教學資源網》的《馬祖閩東語本字檢索系統(試用版)》。不過該線上系統目前還是試用版階段，無論是收錄字詞數量或是功能方面都還有不足的地方，例如本團隊試著以「同學」一詞來做查詢，得到的結果有 3 筆，這 3 筆結果沒有一筆是「同學」一詞的解釋，如果再試著點入第一筆「豬頭瘴」的結果才會在底下釋義內文中發現「同學」一詞。另外像是《連江縣本土教學資源網》的《日常生活常用詞彙》、《綜合活動馬祖話》，還有《連江縣志--語言志》等相關資源也可納入。《馬祖閩東語本字檢索系統(試用版)》的相關頁面截圖如下：

圖 42.《馬祖閩東語本字檢索系統(試用版)》相關頁面截圖之一

圖 43.《馬祖閩東語本字檢索系統(試用版)》相關頁面截圖之二

辭典資料庫的概念在手語資源建置上更為常見與重要，從國外手語語料庫建置的歷程來看，語料庫與辭典計畫同時進行的例子很多，澳洲、美國、瑞典皆是，因為詞彙和篇章的表達對手語來說是不同層次的，且以手語為母語的人數稀少，較難有大型手語語料庫問世。不過，晚近的手語語料庫已見辭典與語料雙向搜尋的設計，如：瑞典手語語料庫可在時間軸上點擊個別詞彙，便會跳出該詞彙的辭典部分資訊，使用者可接著選擇查看辭典或查看語料庫中其他的語料。

最後，考量到活力較低的國家語言（如閩南語、客家語、原住民族語、閩東語、臺灣手語等）因為使用率及普及率不及華語，針對新觀念或新科技而產生的相關詞彙，常常會出現必須向華語「借詞」，或者使用混亂的情況。因此，建議國家語言資料庫也可以在網站上提供各國家語言的「公告新詞」。「公告新詞」可以參考並延用原民會的新創詞創制流程（網址：

http://ilrdc.tw/research/newwords/process.php），定期蒐集、討論、修訂並公告新詞，供民眾作參考。

7.3.2. 國家語言地圖及地理資訊系統

除了語料庫、辭典的形式，結合語言與地理的內容呈現亦是國家語言資料庫可納入的內容項目之一。語言資料庫不僅是展示「語言共時性（synchrony）」等語言文法系統，更可融合「語言蘊含的文化與地方知識（local knowledge）」(Geertz, 2002, in Hsieh, 2019) 與變異（variations），以建立「臺灣語言史觀」，相關例子有荷蘭馬克斯普朗克學會（Max Planck Institute）的語言典藏（Language Archive）及俄羅斯少數語言地圖等。(Hsieh, 2019) 此外，日本國立國語研究所的官方網站專闢一區塊為「語言地圖」（分頁網址為：https://www.ninjal.ac.jp/english/database/type/maps/），語言地圖可分為兩類：語言分佈地圖以及語言現象地圖，前者為綜觀的地理分布資訊展示，後者則如日本國立國語研究所之例，將用字及發音差異（word and pronunciation）、語法現象（grammatical phenomena）整理成各 300 張左右的語言地圖，每個詞彙或語言現象都是一張 pdf 檔案，下載後可看到地圖上各點標示著各地的方言差異，如下圖。

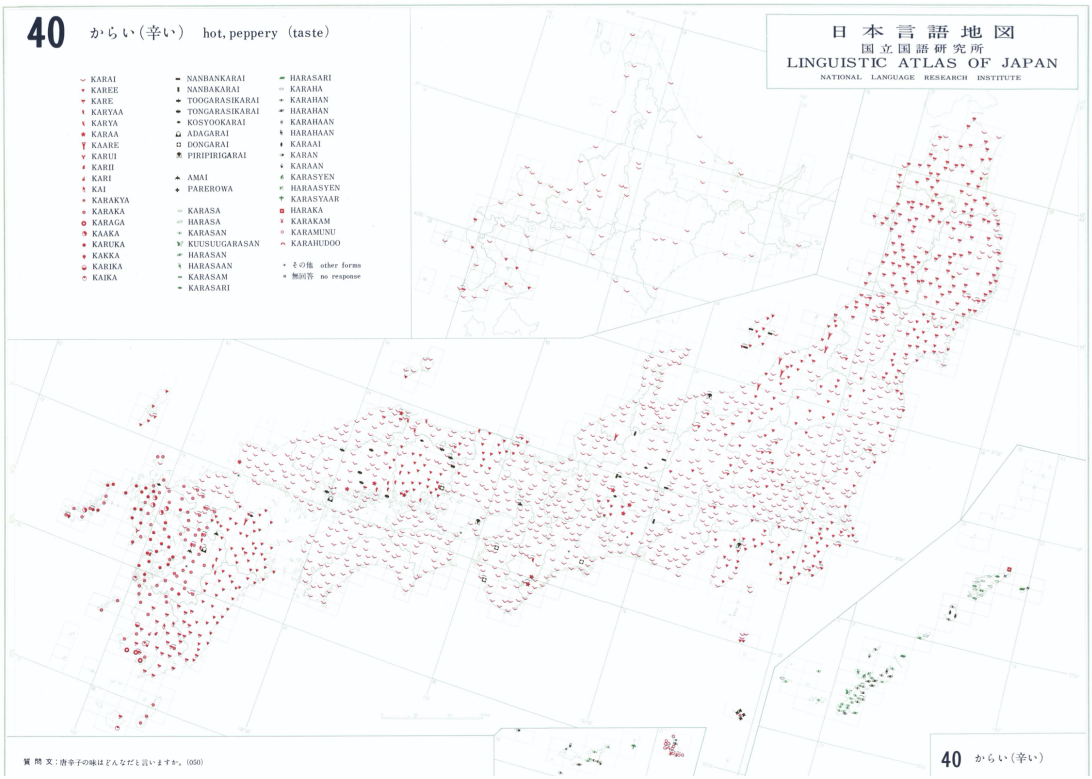


圖 44. 日本國立國語研究所收藏的方言語言地圖，以「辛い」一詞為例。

7.3.2.1. 國家語言地圖

在語言地圖的項目規劃上，建議先將過去學者們所繪製的語言地圖掃描、電子化，並整理過去田野調查所蒐集到的各種腔調音檔，製作成樣品音來展示，讓國人點選聆聽並比對各地方言的發音差異。當上述工作完成一定進度後，可將地理語言學研究成果與地理資訊系統（GIS）等技術結合，將樣品音以互動地圖的方式來呈現。

語言以及地理分布的相關資源如下：

(1) 洪惟仁教授於 2019 年出版的兩冊專書《臺灣社會語言地理學研究：臺灣語言的分類與分區 I》及《臺灣語言地圖集 II》，目前洪教授已經同意將這兩冊專書中的語言地圖授權給文化部。

(2) 卜溫仁（Warren A. Brewer）教授於 2008 年出版之《Mapping Taiwanese》。

(3) 張屏生教授之詞彙相關研究，詳見張教授之個人網頁：http://www.chinese.nsysu.edu.tw/zh_tw/Department_introduction/Teacher/%E5%BC%B5-%E5%B1%8F%E7%94%9F-2809422。

(4) 中研院鄭錦全院士建立的「歷史語言與分佈變遷資料庫」，結合語言分佈微觀計畫，研究閩客混合的雲林縣崙背鄉、二崙鄉、新竹縣新埔鎮、苗栗縣後龍鎮、南庄鄉的語言使用，勾勒出閩客語的地理互動，網址為：http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa。

7.3.2.2. 國家語言地理資訊系統

除語言地圖外，應以田野調查、群眾外包等模式來逐步擴增地理語言學相關發音語料，建立國家語言地理資訊系統。具體建議如下。

(1) 將語言地理學研究成果與地理資訊系統（GIS）等技術結合，將樣品音以互動地圖的方式來呈現。

(2) 建置結合地理資訊系統與方言田野調查的資料庫。

(3) 提供群眾外包模式來逐步擴增地理語言學相關發音語料。

設計 APP 以群眾外包的方式收集資料。

7.4. 語言資料徵求及各項資源分享

7.4.1. 擴增語料庫及資料庫

不論是語料庫或資料庫，在擴充內容階段時，可先了解現有資源為何，若符合該項目的收錄原則，則可納為新語料或資料。由於 (1) 國

家語言資料庫與語料庫定調為口語、多媒體的語言資源 (2) 各國家語言的文化各有特色，亦已有豐富多樣的各式資料，可評估作為語料或資料的可能，如：臺灣各國家語言具文化、歷史、風俗、藝術代表性的錄音檔、歌曲、重要儀式、典禮、電影、紀錄片等資料 (3) 考量授權難度及成本等情況，進一步篩選與評定收錄先後順序。以上述原則擴增語料庫或資料庫，可參考以下項目，其中不少資源皆為政府公部門的資料：

(1) 影片：

a. **國家電影中心-台語片 60 週年：**國家電影中心（Taiwan Film Institute，簡稱 TFI）原為行政院新聞局在 1978 年與民間集資成立的公設財團法人機構，2012 年 5 月 20 日改由文化部影視及流行音樂產業局管理。2016 年該中心規劃了台語片 60 週年影展，向民眾展示了目前已完成修復的台語片，包括《地獄新娘》、《王哥柳哥遊臺灣（上）》、《王哥柳哥遊臺灣（下）》、《三鳳震武林》、《危險的青春》等等。網址為：http://taiyupian60th.weebly.com/。

b. **教育部閩南語動畫：**為教育部自 2017 起所推動的計畫，該計畫目前已取得 5 部國內外動畫的授權，包括「貝貝生活日記」、「少年阿貝 GO！GO！小芝麻」、「櫻桃小丸子」、「九藏喵窩」及「卡滋幫」，並且已完成為動畫製作閩南語的配音與字幕。網址為：https://twbangga.moe.edu.tw/info。

c. **公共電視教育影音公播網：**為公視首度授權、針對教育單位建置的雲端串流影音平台，由尚儀數位學習公司製作發行。其中亦收錄有一部閩南語發音、華語字幕的連續劇

《祖師爺的女兒》，網址為：http://ptsvod.sunnystudy.com.tw/。

d. **看影片學客語(行動學習) - 哈客網路學院 - 客家委員會：**哈客網路學院是客委會所製作的客語學習網，其「看影片學客語」專區有 31 部客語發音的短片，每個短片皆附有客語字幕還有相關詞彙的解釋與例句，另外還提供只要點選特定一句字幕就能直接觀賞該段影片的功能，非常方便（詳見圖 45）。網址為：https://elearning.hakka.gov.tw/ver2015/allclass/default.aspx?group=g000000015。

圖 45. 哈客網路學院《南風六堆_我的樹媽媽》學習影片頁面截圖

e. **族語 e 樂園：**由臺北市立大學族語數位中心所建置，原民會版權所有，內容收錄原住民 16 語族各方言的教材、歌謠、動畫、影音等各種資源。網址為：http://web.klokah.tw/。

f. **影音平台-母語巢-臺北市原住民語言學習網：**為臺北市政府原住民族事務委員會所建立的網站平台，提供阿美族、泰雅族、賽夏族、布農族、排灣族、魯凱族、太魯閣族、鄒族、卑南族、雅美族等 10 族的短片、動畫、歌謠、教學相關資源。網址為：http://taipei.pqwasan.org.tw/video/。

g. **學習資源-本土語言資源網：**為教育部所建設的平台，網站彙整了相當豐富的各種國家語言的教學、影音等資源，不過其中亦有不少資源為非官方的資料，假如要放到國家語言多媒體檢索系統中的話需要額外授權。網址為：https://mhi.moe.edu.tw/infoList.jsp?ID=2。

(2) 歌謠：

a. **聽唱兒歌- 文化部-兒童文化館：**文化部曾於 2000 年至 2004 年策辦「臺灣兒歌一百」徵選活動，之後再從中挑選部分得獎作品，並為其譜曲和製作影音檔，並放置到文化部兒童文化館網站上。該網站收錄包括華語、閩南語、客語和原住民語的歌曲，不過有些閩南語歌曲的用字並不完全符合教育部用字規範，如《一禮拜》這首歌詞的「**歸**家出外來曝日」應寫作「**規**家出外來曝日」才對。網址為：https://children.moc.gov.tw/song_list?language=2。

b. **臺灣音聲一百年：**為國立臺灣歷史博物館所架設的網站，裡面收錄了從 1890 年代到 2010 年代的華語、閩南語、日語的歌謠、廣告的音檔與其簡介，其中以閩南語的資料最多。不過部分資料沒有附上歌謠歌詞或廣告台詞的文字檔，加

上有些音檔因為年代久遠有些部分可能聽不太清楚，因此轉寫工作可能較為不易。網址為 https://audio.nmth.gov.tw/。

c. **客家音樂 - 哈客網路學院 - 客家委員會：**哈客網路學院是客委會所製作的客語學習網，其「客家音樂」專區有不少客語歌謠相關教材，教材內附有歌謠講解、投影片、歌謠字幕等資料（詳見圖 46）。網址為：https://elearning.hakka.gov.tw/ver2015/allclass/default.aspx?group=20000004。

圖 46. 哈客網路學院《客家歌謠選集（二）》學習頁面截圖

d. **好客 ING-客家影音網路平台：**由行政院客委會所建置的影音網路平台，收錄包括客語各腔調的戲劇、電影、動畫、新聞、歌曲、童謠、廣播、教學影片等相當豐富的資源，字幕為華語。網址為：https://broadcasting.hakka.gov.tw/。

e. **客家歌謠- 臺北市客家文化主題公園：**為臺北市客家文化主題公園所設置的數位學習專區，裡面收錄有不少客家歌謠

的歌曲、歌詞、相關簡介等資料。網址為：https://ssl.thcp.org.tw/libraries/songs?page=9。

f. **族語e樂園：**由臺北市立大學族語數位中心所建制，原民會版權所有，內容收錄原住民 16 語族各方言的教材、歌謠、動畫、影音等各種資源。網址為：http://web.klokah.tw/。

g. **影音平台-母語巢- 臺北市原住民語言學習網：**為臺北市政府原住民族事務委員會所建立的網站平台，提供阿美族、泰雅族、賽夏族、布農族、排灣族、魯凱族、太魯閣族、鄒族、卑南族、雅美族等 10 族的短片、動畫、歌謠、教學相關資源。網址為：http://taipei.pqwasan.org.tw/video/。

h. **學習資源- 本土語言資源網：**為教育部所建設的平台，網站彙整了相當豐富的各種國家語言的教學、影音等資源，不過其中亦有不少資源為非官方的資料，假如要放到國家語言多媒體檢索系統中的話需要額外授權。網址為：https://mhi.moe.edu.tw/infoList.jsp?ID=2。

i. **《臺北褒歌之美》：**為洪惟仁教授在 2002 年所出版的互動式多媒體光碟，是「台北地區相褒歌保存計畫」之成果，光碟內收錄了「相褒歌」的錄影、錄音資料。

另外，以下是第二次專家諮詢會議中，針對閩南語的資料蒐集議題，洪惟仁教授推薦之閩南語教學相關網站：

a. 1999-（不斷更新中），《洪惟仁—臺灣語文工作者》個人網站，網址為：http://www.uijin.idv.tw。榮獲收入《臺灣研究網路資源選介》（國家圖書館，2006: 218）。

b. 2007，《臺灣羅馬字拼音方案及其發音學習網》網站，教育部補助製作，載於臺中教育大學臺灣語文學系網站之「教學資源」目錄下，網址為：http://www.ntcu.edu.tw/tailo/。榮獲國立臺中教育大學九十七年度「優良網路教學數位教材獎」。

c. 2005，《閩南語量詞資料庫》網站，教育部補助製作，載於臺中教育大學臺灣語文學系網站之「教學資源」目錄下，網址為：http://210.240.194.138/find/。

d. 2005，《閩南語發音與音標練習遊戲》網站，教育部補助製作，載於臺中教育大學臺灣語文學系網站之「教學資源」目錄下，網址為：http://www.ntcu.edu.tw/taiwanese/resource.html。（已下架，更新為(7) 2007 新版）。

e. 2005，《臺灣語文學系》網站，臺中教育大學及臺中教育大學補助製作，網址為：http://www.ntcu.edu.tw/taiwanese/mainpage.htm。

f. 2005，《臺語文學網：創作園地》網站，教育部補助製作，載於臺中教育大學臺灣語文學系網站之「教學資源」目錄下，網址為：http://192.83.167.52/TwlrWeb/。

g. 2002，《臺灣語文學會》網站，教育部資助教學計畫補助設立。網址為：http://www.tlls.org.tw/。今已改版，新址：http://www.twlls.org.tw/。

h. 2002，臺灣話會話教學網站《臺灣話e網情深》。教育部非同步遠距教學計畫補助製作，上載於洪惟仁教授個人網站上。網址為：http://www.uijin.idv.tw/class/index.htm。榮獲

僑委會九十一年度海華獎優良華語文教學軟體網頁組推薦獎。

i. 1999，《臺灣閩南語教材》網站，洪惟仁教授總編輯所編製成多媒體教材上載僑務委員會網站上。網址為：http://edu.ocac.gov.tw/language/taiwanese/。

j. 2005，《臺語文學網：臺語文庫》網站，教育部補助製作，載於臺中教育大學臺灣語文學系網站之「教學資源」目錄下，網址為：http://192.83.167.52/TwlrWeb/。

k. 2004，《民間歌謠網站》。數位典藏資料庫，元智大學補助製作，載於洪惟仁教授個人網站上，網址為：http://www.uijin.idv.tw/TAIWANSONG/index.htm。

在此階段，首要步驟會以整合現有資料為主，各種國家語言的多媒體資料在納入國家語言資料庫前需要先請求相關授權，如因為各種原因造成無法順利取得授權的情況的話，也應先留存相關資料的網址或出處，待之後取得授權後，再彙整至資料庫內。

7.4.2. 資源分享

關於國家語言資料庫中資料的儲存、存取與推廣，資料儲存考慮的是將資料存放的技術及規格，牽涉到資料保存、取出及查詢的難易、效率、成本等面向；資料存取則是政府之外的對象如學術機構、廠商、人民等取得該資料的可行性、方便性及成本等面向；資料推廣旨在增進國家語言資料庫之利用，以促進研究發展及加值應用。以下將先簡單介紹這三者之原則與目標，接著再提供語料庫後設資料（metadata）和國家語言資訊處理工具的建議。

首先，可以將要儲存的資料簡單區分成原始資料與二級資料這兩者。原始資料包括新聞的原始文字、田野調查的錄音檔、演出的錄影檔案等，或者有些像是因為隱私權的要求必須將原始資料去識別化的資料，也可歸在此類。這部分的資料希望能以完整保存為原則，取出的需求以完整檔案為主，因此查詢的需求較低或甚至無查詢需求；另外，也希望能達到永久保存的目標，因此可以考慮將這些資料異地備份、多層級備份（硬碟、光碟、磁帶等）等。倘若這些原始資料授權符合條件，甚至也可考慮加入國際級的計畫（如，AWS Public Dataset Program https://aws.amazon.com/opendata/public-datasets/）。二級資料則是將原始資料進一步作分句、分詞、標記、人工或自動文字辨識語音資訊等處理，因為查詢需求遠較原始資料高，因此在儲存上更需考慮各種查詢的彈性。二級資料若由演算法自動產生，該演算法及相關工具的保存非常重要；若由人工產生，則二級資料本身的保存非常重要，可參考原始資料的保存原則。以政府的立場，建立國家語言資料庫的原始資料應是最優先的；對於二級資料而言，不同領域的需求和範圍各有不同。因此，原始資料的存取建議應由政府來主導，二級資料則可交由民間作建置。

接著，在資料存取的部分，可以考慮存取者相對方便取得之「直接下載」方式、有利於查詢需求，卻可能不利於取得完整檔案的「API」、或者其他如「書信或 email 索取」、「光碟寄送等方式」。這些方式皆各有利弊，可再諮詢專家學者們的意見再考慮要採取哪些方式。另外，因為建置的語言資料庫可能有所變動（如，錯誤修正、增加新語料等），因此建議國家語言資料庫的所有「發行版本」都應有正式且明確之版號和相關說明，版號方面可參考「語意化版本 (Semantic Versioning, https://semver.org/)」之版號命名原則。

最後，為了推廣國家語言資料庫之利用，該資料庫必須易於使用（如，使用較通用的檔案格式和檔案編碼），也易於機器讀取，以增進學術或商業界使用的意願。資料說明也應完整且正確，並提供英語等外語版本說明。授權方式應儘量簡單而友善。然後，也可透過論文發表的方式讓學術界同儕認知該語言資料庫的存在，進而引起利用該資料庫進行研究的興趣及可能性。

考量到上述資料儲存、資料存取與資料推廣之原則，還有參考「肆、國外相關數位典藏計畫、資料格式、與工具的分析」，建議國家語言資料庫參考目前世界各主要語言資料庫的後設資料（metadata）和相關的自然語言處理和資訊處理工具。建議以都柏林核心集（Dublin Core）、語言典藏公開群體（Open Language Archives Community, OLAC）、以及公開檔案典藏後設資料協議（Open Archives Initiative Protocol for Metadata Harvesting, OAI-PMI）這三者為基礎。Dublin Core 是跨領域的工具，除了語言學界外，還有不少領域都會採用之。OLAC 是基於 Dublin Core 再開發的系統，目前很多政府都會採用之。而 OAI-MPH 是資料交換的方法，例如如果進到 PARADISEC 網站，會看到相關資料的描述（如，這個檔案是音檔、開放下載、貢獻者是誰等等）；這時，如果透過 OAI-MPH 的系統，就能下載資料。自然語言處理工具包括能自動標記段落、句子、詞的程式工具。相關資訊處理工具還包括字集和造字的程式。國家語言資料庫編碼建議採用 UTF 8 或 UTF 16。

針對跨語料庫的檢索功能，因每個語料庫原有的後設資料標記各不相同，在整合時，如有更細緻的標記應被保留，以附註的方式說明，或是在該語料庫的簡介部分特別提及該特色。另外，當語料庫採

用國際通用的標記原則時，更易整合，並可將其進一步列為確認標記正確與否的語料庫。

語料庫的授權狀態亦是後設資料的重點，像是澳洲的PARADISEC 數位典藏計畫中，授權狀態放在很明顯的地方，如下圖中的存取資訊（access information）附上「資料存取狀態（data access conditions）」以及對此狀態的「文字敘述（data access narrative）」，而在前一頁的資料庫資訊中，則有該資料庫的引用格式，如下圖。

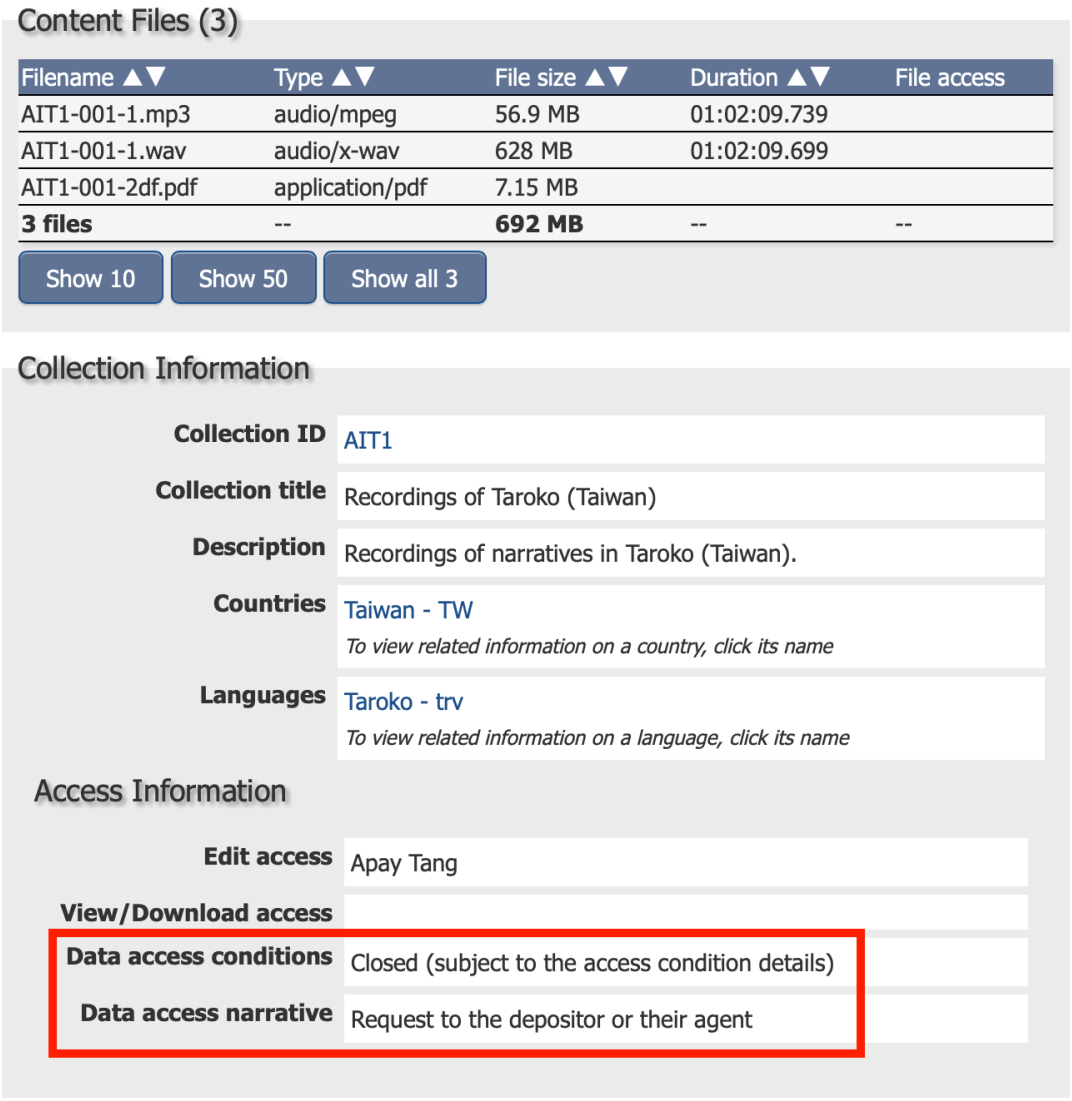


圖 47. 澳洲 PARADISEC 數位典藏計畫中的資料存取（access information）的資訊頁面

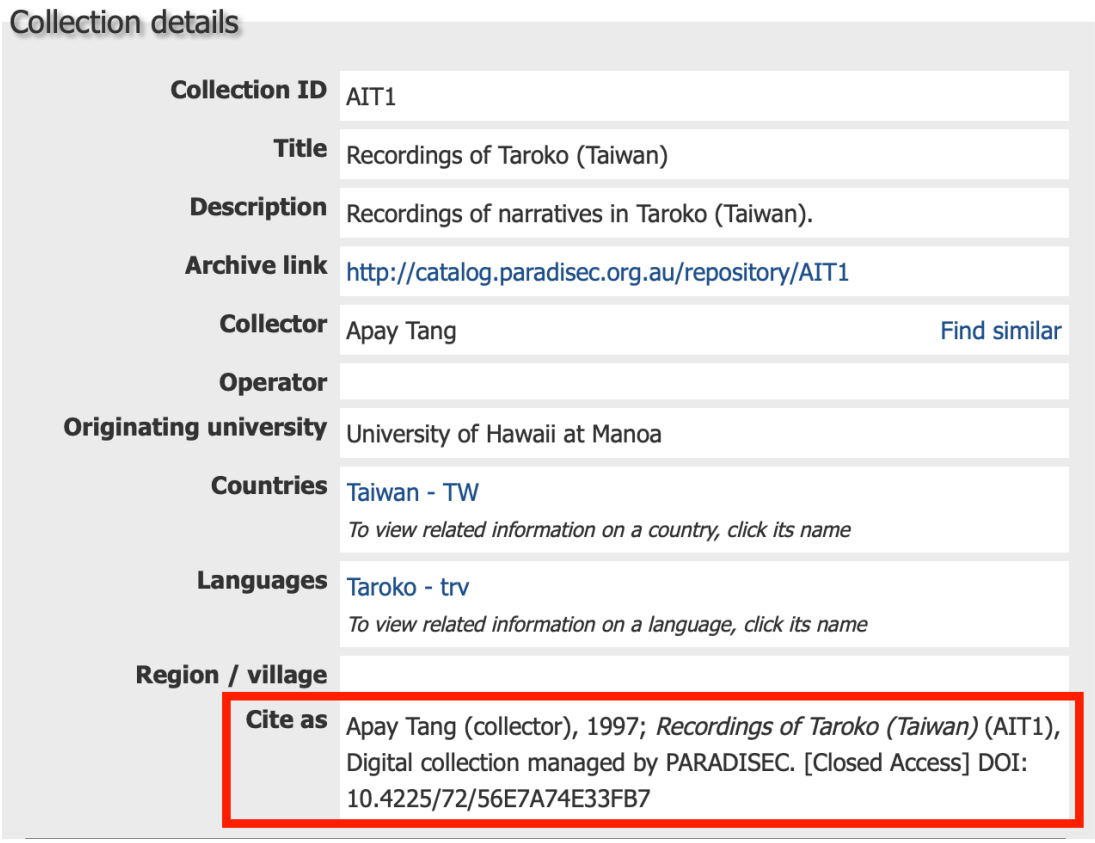


圖 48. 澳洲 PARADISEC 數位典藏計畫中資料庫的介紹頁面

此外，葉茂林委員也提到諸多法律層面的議題，包括甫於 2019 年通過的文化基本法、類似公共出借權（public lending right）的補償制度、對孤兒著作（orphan work）的補償金提存機制，或是為了教育與研究的目的，附上其超連結及簡短文字說明等作法，都是非常寶貴的意見。

捌、　國家語言資料庫招標項目與建置方式的建議

本章根據本計畫的需求說明書，針對國家語言資料庫的用途、使用對象、內容項目與建置方式、維運與管理、網站架構與功能、應用與推廣等各個面向提出現階段招標項目的規劃建議。第 1 至 3 節根據本計畫的需求說明書草擬國家語言資料庫的目標、用途、及使用對象。第 4 節提出現階段招標內容項目與建置方式的規劃建議。第 5 節至 7 節提出維運與管理、網站架構與功能、應用與推廣的規劃建議。

8.1. 國家語言資料庫的目標

(1) 激發國人對本國語言文化的興趣與熱愛，並關注瀕危的國家語言。具體作法可以採取像是製作相關主題的紀錄片、動畫、網站，還有提供相關議題的語言調查報告等，來讓國人理解各個國家語言的現況。

(2) 永續保存本國語言文化資產。具體作法像是盤點彙整各個國家語言現有相關語料和多媒體資料，建立國家語言資料庫，裡面除包含各個國家語言的語料庫外和還有國家語言辭典資料庫以及國家語言地圖及地理資訊系統。透過建置國家語言資料庫有助於各國家語言的正向發展，而資料的整合也能促進各國家語言的永續保存，並增進使用各國家語言的機會。另外，日本知名方言學家柴田武曾在其《語言地理學方法（言語地理学の方法）》一書中提到：「語言地理學是語言史的一種方法。」語言地圖的描繪可以不但讓國人了解當代臺灣的「語言共時性（synchrony）」，透過語言地圖資料的累

積，我們還可以從中推斷語言的演變情況，進而建立出「臺灣語言史觀」。

(3) 讓國家語言資料庫之建置更加完善，為國家語言資料庫建置奠定長遠發展的基礎。語言為文化的載體，隨著社會變遷，語言勢必也會有所改變，國家語言資料庫除了整合現有語言相關資源外，也應考慮永續發展的問題，以因應這些變化。因此，國家語言資料庫也應設置各種擴增機制，例如可以設置群眾外包機制，提供國人貢獻語料的管道，接著再由專家學者來審查這些語料並進行分詞、標記等工作，最後納入國家語言資料庫；或者也可以視需要定期推派由專家學者組成的語言調查小組，來進行各國家語言田野調查的工作，最後再將調查報告彙整至國家語言資料庫。透過上述做法，便可不斷更新國家語言資料庫內容，達到永續發展之目的。

(4) 促進未來研究發展及加值應用。國家語言資料庫成立後，可以定期舉辦和國家語言資料庫主題相關的研討會，並廣邀學界與業界人士探討國家語言資料庫在教育、翻譯、科技等領域之各種應用發展的可能性。此外，設置英文版的網站也有助於推廣本國語言研究，並促進國際研究或商業合作。

8.2. 國家語言資料庫的用途

國家語言資料庫除供學術之用外，並提供教育的功能，同時宣揚臺灣多元文化，凝聚文化共識，讓臺灣各族群以自己的語言文化為傲，並促進語言及文化平權。例如，利用影片、國內外相關網站連結、歷年國家語言相關調查報告等方式來簡介本國各國家語言現況，不但可以促進本國各族群彼此互相了解，也可以讓國人意識到臺灣是個具有豐富語言多樣性的寶島，進而讓國人對自身的語言文化感到驕傲。另外，透過建置多語的國家語言資料庫，也有助於翻譯相關領域的發展，

例如國家語言資料庫在建立跨語查詢、平行語料庫、華語對應等的過程時，勢必會觸及不同語言間的特殊用詞或用字的轉譯問題，一旦這些問題得到解決，未來翻譯人員在從事翻譯相關工作時便可參考，進而提升我國在國家語言的翻譯品質。最後，在建置國家語言資料庫時所採用或開發的各項工具，在未來也具備和各企業合作發展相關產品的商業潛力，可能的相關產品例如翻譯機、語音辨識、分詞程式、詞性標記程式等。

8.3. 使用對象

國家語言資料庫的使用對象包括學術界、教育界、社會大眾、以及對臺灣語言有興趣的國際人士。為鼓勵國際人士使用國家語言資料庫，未來國家語言資料庫應該充分國際化，需有以下配套措施，包括英文版的網站和檢索介面程式以及英文版的資料授權機制說明等。

8.4. 現階段國家語言資料庫內容項目與建置方式的建議

在前一章，我們建議分階段逐步整合現有資源及建置新資源。其中中央部會沒有專責機構負責或缺乏資源的國家語言應優先規劃建置。由於客委會正在建置客語語料庫，而原住民族語言語料庫的建置計畫有待其主管機關原民會通盤考慮後做決策，目前以下幾項或因為還沒有中央部會專責機關，或因為屬於整合不同的資料庫或語料庫，容易受忽略，應該及早規劃。這些包括 (1) 國家語言資料庫網站（含國家語言現況介紹）、(2) 閩南語語料庫、(3) 閩東語語料庫、(4) 臺灣手語語料庫、(5) 國家語言辭典資料庫及國家語言地理資訊系統等五項。建議國家語言資料庫採用由上而下建置原則及模組化設計，預先考慮國家語言資料庫的網站與資料庫與各子語料庫及子資料庫的銜接方式，先進行國家語言資料庫的整體架構與網站的建置，內容部分則分

階段於子語料庫及子資料庫陸續建置完成後再與主系統銜接。由於牽涉不同部會的業務，應舉行跨部會的協調會議，減少整合時可能會產生的問題。前面已經提出圖 31 作為國家語言資料庫整體架構與網站的建置的參考，建議以此架構圖，作為現階段國家語言資料庫招標有關內容項目與建置方式的參考。

8.4.1. 國家語言資料庫及網站（含國家語言現況介紹）

內容項目與建置方式	一、 國家語言資料庫網站包括圖 31所示的四個部分。(1)國家語言現況、(2) 國家語料庫（包括華語語料庫、閩南語語料庫、客語語料庫、原住民族語語料庫、閩東語語料庫、臺灣手語語料庫）、(3) 國家語言辭典資料庫及語言地理資訊系統、(4) 語言資料徵求及各項資源分享。除第一項國家語言現況之外，其餘都採取預留資料庫的方式處理。本案資料庫需保留後續擴充性，未來需能夠與各部會建置之語料庫或語言資料庫進行資料交換與介接作業。 二、 國家語言現況包含 (1) 網頁和影片介紹、(2) 國家語言調查報告之連結。以網頁和影片介紹臺灣各國家語言現狀並讓國人瞭解國家語言面臨的傳承危機以及保存國家語言的重要性。網頁及影片必須涵蓋所有臺灣國家語言的現況。影片必須透過剪輯現有的影片或重新拍攝，總長度至少必須達 30 分鐘，且需介紹臺灣各國家語言現況。第(2)項則連接與國家語言調查相關的報告。 **格式：**網頁設計採用 HTML 4 、CSS、 Javascript， 網頁內包含影片。網頁必須採取響應式設計(responsive web design)。影片畫質至少需達 1920x1080 且長寬比為 16:9 ，影片格式為 mp4。
期程	共 1 期為期 2 年。
人力	(1) 成立諮詢委員會：須聘請專精臺灣國家語言（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學專家學者至少 4 人。

	(2)計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上相關研究或實務經驗。 (3)專任助理 3 人。兼任助理 3 到 6 人。
經費	250 萬。

8.4.2. 閩南語語料庫

內容項目與建置方式	**語料規模**：包含書面語及口語兩個部分，共 1000 萬詞，其中書面語語料以不超過 800 萬詞為原則。 **文體**：含記敘、描寫、論說、說明等分類。 **語式**：語料庫應具備語式的分類。 **主題**：儘可能包含社會、生活、文學、科學、哲學宗教、教育、影視娛樂、體育、政治、藝文創作等各類主題，而不侷限於少數特定主題。 **書面語**：蒐集使用教育部建議之閩南語用字之圖書、雜誌、報紙、教材、文學作品、劇本、歌本、教科書……等資料。 **口語**：電視新聞、節目、廣播、電影等現有的影音資料。語音資料之品質須清晰可聽辨（包含聲音、畫面）。儘可能達到資料發音人男女比例均衡，且涵蓋不同年齡層。 **語料之授權**：需徵求相關資料所有者之同意並取得授權。 **「平衡」語料庫**：本語料庫包含書面語，口語語料，後者至少佔全部語料的 20%。 **開放資料**：(1) 提供 2 百萬有分詞及詞性標記的語料，附上閩南語例句及華語翻譯，開放民眾下載。(2) 閩南語華語平行語料及語音文字時間對齊語料：挑選 20 萬詞口語語料，以句為單位，將閩南語口語語音資料轉寫為羅馬拼音和漢字，且發音與文字時間必須對齊，

	並附上對應之華語翻譯。必須使用語音分析與標示工具 PRAAT 的格式，開放民眾下載。 **格式：** (1) 書面語：xml 格式文字檔及純文字檔。 (2) 口語：文字檔、影音檔（mp4）、聲音檔（wav 及 mp3）、PRAAT 的軟體格式。 (3) 需包含後設資料。 **分詞和詞性標記原則的擬定：**團隊成員必須包括語言學及計算語言學專家。需參考中研院詞庫小組中文詞性分析技術報告，由專家組成團隊並擬定一個符合閩南語特性的分詞和詞性標記原則。需提供如何實際應用分詞和詞性標記原則的操作手冊，包含明確的例子和判斷的方法。 詞性**標記集：**參考中研院詞庫小組中文詞性分析技術手冊及閩南語的語法特點與相關文獻，詞性標記集的數量以不多於中研院詞庫小組簡化標記，及不少於精簡標記為原則，建議標記總數介於 15 到 30 之間。 **分詞及詞性標記的正確性：**需由專家組成團隊提供包含 1 萬詞正確分詞和詞性標記的語料，作為助理正式標記前的訓練手冊。助理分詞和標記的正確率未達到八成五前，需不斷練習前述的訓練手冊，一直到分詞和標記的正確率達到標準後才可以正式進行語料分詞與標記。每份語料需經過至少兩位助理人工校對過，且不一致處需透過程式自動找出並交由專家組成的團對討論出這些例子中的正確的分詞與標記。助理常犯的錯誤應作為新進助理訓練用的資料。 **後設標記（meta data）：**應包含主題，出處，年代，文類，語式等訊息，並以 XML 格式呈現。 **系統功能：**建議參考 COCA 語料庫介面的功能，包含至少以下幾項 (1) 關鍵詞前後文檢索程式（KWIC），使用者輸入一個詞或字串後，可以選擇詞性、要檢索哪些語

	料庫、以及是否顯示詞性標記，輸出時會顯示這個字串在語料庫中出現的次數以及例句。 (2)輸入關鍵詞或字串顯示在不同語料庫出現的頻率。 (3)搭配詞檢索，輸入一個詞、詞性、搭配詞前後的範圍、搭配詞的詞性，程式自動顯示最常一起出現的搭配詞。 **介面程式：**除系統外，還需提供介面程式用以編輯錯別字、分詞、和詞性標記。每份編輯過的語料都記載標注的助理姓名及編輯時間。 **用字規範：**以教育部臺灣閩南語常用詞辭典為依據。 **閩南語口語語料庫語料來源：**中正大學蔡素娟教授所建立的臺灣閩南語語料庫內容為廣播節目。此外，已退休的清華大學胡萬川教授過去協助各縣市文化局所編輯的閩南語故事集，目前已數位化並收錄在中研院閩客語典藏計畫中。這些資料再加上已經不受著作權法保護 50 年以前的閩南語電影都適合作為閩南語口語語料庫的材料。 **閩南語書面語語料庫來源：**以教育部越讀越懂閩客電子報其中閩南語部分，教育部及文化部相關的閩南語創作比賽得獎者作品作為的材料。
期程	(1)以六年期計畫建置 1000 萬詞規模具有分詞和詞性標記的閩南語語料庫，其中口語多媒體資料不少於 200 萬詞。 第 1 期： (2)收集、校對、並整合現有口語及書面語語料至少 300 萬詞。 (3)擬定分詞和詞性標記原則與技術手冊。 (4)完成人工校正過 300 萬詞具有分詞及詞性標記的語料。 (5)利用前述語料和機器學習工具發展最初版本的閩南語分詞及詞性標記程式。必須有數據顯示不同機器學習工具和演算法訓練出來的結果。

	第 2 期： (1) 收集、校對、並整合現有口語及書面語語料至少 350 萬詞。 (2) 人工校正閩南語分詞性標記程式的結果。 (3) 每完成一百萬詞的語料就重新訓練一個新的閩南語分詞及詞性標記程式。 (4) 完成 350 萬詞經人工校正過分詞及詞性標記的閩南語語料。 第 3 期： (1) 收集、校對、並整合現有口語及書面語語料至少 350 萬詞。 (2) 完成 350 萬詞經人工校正過分詞及詞性標記的閩南語語料。 (3) 完成閩南語口語多媒體語料庫檢索系統。 (4) 利用 1 千萬詞人工校對過的閩南語分詞及詞性標記，重新訓練一個更準確的閩南語分詞及詞性標記程式，並開放下載，分詞及詞性標的正確率必須與當時機器學習相關技術文獻報告接近。 (5) 提供 200 萬詞有分詞及詞性標記的書面語語料，此資料同時還有閩南語句子的華語的翻譯，資料開放民眾下載。 (6) 挑選 20 萬詞口語語料，以句為單位，將閩南語口語語音資料轉寫為羅馬拼音和漢字，且發音與文字時間必須對齊，並附上對應之華語翻譯。必須使用語音分析與標示工具 PRAAT 的格式，開放民眾下載。
人力	(1) 成立諮詢委員會：依據期末審查會議委員的意見，閩南語語料庫的諮詢委員必須包括其它國家語言的專家以及著作權之專家，以便建置閩南語語料庫的經驗未來可以轉移到其它的國家語料庫。因此須聘請熟悉閩南語（至少 4 人）、閩東語、客語、原住民各族語、臺灣手語、著作權（各至少 1 人）之專家學者組成諮詢委員會。每年至少召開 4 次諮詢，協助閩南語語料庫之建置。 (2) 計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上語料庫相關實務經驗。

	(3) 專任助理 11 到 13 人，其中負責人工分詞和詞性標記的語言學助理至少應有 8 人。負責系統程式開發的助理不應少於 3 人。兼任助理 10 到 15 人。
經費	5 千萬元。

8.4.3. 閩東語語料庫

內容項目與建置方式	以六年期計畫建置一百萬詞規模的閩東語語料庫，其中口語部分以不少於 20 萬詞為原則，提供以句為單位的閩東語與華語對照資料。並從口語語料中挑選其中有音檔的 10 萬詞口語語料形成多媒體語料，以句為單位，將閩東語口語語音資料轉寫為羅馬拼音和漢字，且發音與文字時間必須對齊，並附上對應之華語翻譯。此口語多媒體語料必須使用語音分析與標示工具 PRAAT 的格式，開放民眾下載。在資料格式方面，與其它國家語料庫一致，採 XML 格式、Unicode 編碼。語料需提供分詞及詞性標記。 在語料蒐集方面，先整合現有閩東語語言資源，書面語可以以連江縣本土資源教育網裡面的資源為基礎，口語語料可納入兒歌、歌曲比賽得獎作品等。馬祖社區電台的廣播節目可作為口語語料，其內容涵蓋一分鐘母語教學、60 秒母語形象、20 分鐘母語教學、母語線上廣播成果側錄等。 分詞和詞性標記原則的擬定：團隊成員必須包括語言學及計算語言學專家。需參考中研院詞庫小組中文詞性分析技術報告，由專家組成團隊並擬定一個符合閩東語特性的分詞和詞性標記原則。需提供如何實際應用分詞和詞性標記原則的操作手冊，包含明確的例子和判斷的方法。 詞性標記集：參考中研院詞庫小組中文詞性分析技術手冊及閩東語的語法特點與相關文獻，詞性標記集的數量以不多於中研院詞庫小組簡化標記，及不少於精簡標記為原則，建議標記總數介於 15 到 30 之間。

	分詞及詞性標記的正確性：需由專家組成團隊提供包含 1 萬詞正確分詞和詞性標記的語料，作為助理正式標記前的訓練手冊。助理分詞和標記的正確率未達到八成五前，需不斷練習前述的訓練手冊，一直到分詞和標記的正確率達到標準後才可以正式進行語料分詞與標記。每份語料需經過至少兩位助理人工校對過，且不一致處需透過程式自動找出並交由專家組成的團對討論出這些例子中的正確的分詞與標記。助理常犯的錯誤應作為新進助理訓練用的資料。 後設標記（meta data）：應包含主題，出處，年代，文類，語式等訊息，並以 XML 格式呈現。 系統功能：建議參考COCA語料庫介面的功能，包含至少以下幾項 (1)關鍵詞前後文檢索程式（KWIC），使用者輸入一個詞或字串後，可以選擇詞性、要檢索哪些語料庫、以及是否顯示詞性標記，輸出時會顯示這個字串在語料庫中出現的次數以及例句。 (2)輸入關鍵詞或字串顯示在不同語料庫出現的頻率。 (3)搭配詞檢索，輸入一個詞、詞性、搭配詞前後的範圍、搭配詞的詞性，程式自動顯示最常一起出現的搭配詞。 介面程式：除系統外，還需提供介面程式用以編輯錯別字、分詞、和詞性標記。每份編輯過的語料都記載標注的助理姓名及編輯時間。 用字規範：以《連江縣本土教學資源網》的《馬祖閩東語本字檢索系統(試用版)》為依據。
期程	共 3 期 6 年。 第 1 期： (1)收集、校對、並整合現有口語及書面語語料至少 30 萬詞。 (2)擬定分詞和詞性標記原則與技術手冊。

	(3)完成 30 萬詞經人工校對具有分詞及詞性標記的語料。 (4)利用前述語料和機器學習工具發展最初版本的閩東語分詞及詞性標記程式。必須有數據顯示不同機器學習工具和演算法訓練出來的結果。 第 2 期： (1)收集、校對、並整合現有口語及書面語語料至少 35 萬詞。 (2)完成 35 萬詞經人工校正過分詞及詞性標記的閩東語語料。 (3)重新訓練一個新的閩東語分詞及詞性標記程式。 第 3 期： (1)收集、校對、並整合現有口語及書面語語料至少 35 萬詞。 (2)完成 35 萬詞經人工校正過分詞及詞性標記的閩東語語料。 (3)完成包含 20 萬詞的閩東語口語多媒體語料庫檢索系統。 (4)利用 100 萬詞校對過的閩東語分詞及詞性標記，重新訓練一個更準確的閩東語分詞及詞性標記程式，並開放下載，分詞及詞性標的正確率必須與當時機器學習相關技術文獻報告接近。 (5)提供 20 萬詞有分詞及詞性標記的書面語語料，此資料同時還有閩東語句子的華語的翻譯，資料開放民眾下載。 (6)挑選 10 萬詞有音檔的口語語料，以句為單位，將閩東語口語語音資料轉寫為羅馬拼音和漢字，且發音與文字時間必須對齊，並附上對應之華語翻譯。必須使用語音分析與標示工具 PRAAT 的格式，方便未來進一步的應用。
人力	(1) 成立諮詢委員會：須聘請熟悉閩東語語言學或計算語言學的專家學者至少 4 人，協助建置資料庫。 (2) 計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上相關研究或實務經驗。

	(3) 專任助理 8 人，其中語言學助理 4 人，負責系統程式開發的助理 4 人。兼任助理 5 到 10 人。
經費	共 1500 萬。

8.4.4. 臺灣手語語料庫

內容項目與建置方式	(1) 建立至少 10 小時長的臺灣手語多模態語料庫及檢索系統。 (2) 領域及主題需儘量多元。 (3) 需至少有五分之一包含兩位臺灣手語者的手語對話影片。 (4) 參考英國手語語料庫和澳洲手語資料庫的設計，由臺灣手語使用者擔任發音人，錄製手語使用者的敘事、對話、訪談影片，蒐集內容可參考英國手語資料庫設計詞彙列表、以圖像引導（elicit）發音人打出該手語詞彙，抑或是參考澳洲手語資料庫的作法，邀請二或多位發音人組成焦點團體（focus group），從訪談或對話中擷取主題式資料。一部份資料可採用公視手語新聞的影片。 (5) 須標記處理項目包括「手形」、打法「位置」及「意義」說明等項目。 (6) 檢索頁面提供影像及標記訊息，可讓使用者調整影像的播放速度、下載部分影像檔或標記檔等。需有華語的翻譯。 (7) 手語語料庫應整合手語資料庫。建議使用開源的 NGT 資料庫並使用 ELAN 軟體以 EVC（external controlled vocabulary）檔來管理資料庫的詞彙，標記人員在標記語料時便從此EVC檔案尋找符合的詞條。可以 ID 查詢表確認該詞彙是否收錄在資料庫中，再確認該詞彙是否多義。在標記上，如果有多個選擇，以使用頻率（frequency）和象似性（iconicity）為原則，標記項目包括：複合詞的切分及對應詞彙、單手或雙手（handedness）及是否雙手動作對稱（symmetry）、手形變化（handshape changes）、動作方向（movement direction）及是否重複（repetition）、手勢打在身體的哪個部位上（location），此外亦提供實名（name entity）及語

	意欄（semantic field）的標記該資料庫的系統可即時更新最小對立體（minimal pair）的資料。 (8) 建議詞彙分析方面採用以下的原則標記 (Becker, 2020)，以便跨資源搜尋，例如： (6) 手勢（handedness）：若手語打法為雙手時，通常會有對稱或交替的動作，標記為[SymmetricalOrAlternating]，若無則為[AsymmetricalSameHandshape]，單手打法則標記為[OneHanded]。另一種打法有主副手之分，標記為[AsymmetricalDifferentHandshape]。若不符合上述對稱原則（symmetry condition）或主副原則（dominance condition），則標記為[Other]。 (7) 主要身體部位（major location）：臂腕（arm, including wrist）、軀幹（body, signer's torso）、手（hand）、頭及臉部（head, including face）、無特定部位（neural, signing space in front of singer's body）等。 (8) 主手（dominant hand）、使用的手指（selected finger）、以及手指彎曲狀態（flexion）：使用的手指是依手指彎曲或伸直的狀態而定，而手指彎曲的標記分為 9 類（categorical），而非給予連續性的數值（numerical）。 (9) 尚未包含手語打法的動作方向（direction of movement），因此有些不同的手語詞彙在兩個資料庫的標記資料是相同的。 (10) 在網頁介面上，以關鍵字檢索句（KWIC, keyword in context）的方式呈現，並附有上述的語料標記。此外，使用者亦可勾選查看更多資訊，例如：語料庫來源（Korpus）、該檢索句的時間長度（Längd）及起迄時間（Start, Slut）、語料來源檔名（Radnamn）、檔案敘述（Filbeskrivning），未來希望可以依照這些資訊的選項排序。在語料標記區塊的下方則標示了符合搜尋結果

	的時間，因此使用者可在特定檔案中縱覽時間軸上的搜尋結果，此一設計亦可看出該搜尋結果的分布狀況（dispersion）。
期程	分 3 期每期 2 年，共 6 年。 第一期建立總共至少有 3 小時長的臺灣手語多模態語料庫及標記與華語翻譯。其中出現在臺灣手語語料庫並切割為概念的單位必須同時擴充至臺灣手語資料庫並與臺灣手語語料庫連結。 第二期建立總共至少有 3.5 小時長的臺灣手語多模態語料庫及標記與華語翻譯。其中出現在臺灣手語語料庫並切割為概念的單位必須同時擴充至臺灣手語資料庫並與臺灣手語語料庫連結。 第三期建立總共至少有 3.5 小時長的臺灣手語多模態語料庫及標記與華語翻譯並完成檢索系統。其中出現在臺灣手語語料庫並切割為概念的單位必須同時擴充至臺灣手語資料庫並與臺灣手語語料庫連結。完成臺灣手語多模態語料庫檢索系統。
人力	(1) 成立諮詢委員會：須聘請熟悉臺灣手語的語言學或計算語言學的專家學者至少 4 人，協助建置資料庫。 (2) 計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上相關研究或實務經驗。 (3) 專任助理 8 人，其中語言學助理 4 人，負責系統程式開發的助理 4 人。兼任助理 5 到 10 人。
經費	1500 萬。

8.4.5. 國家語言辭典資料庫

內容項目與建置方式	將目前市面上能夠找到且能夠取得授權的字辭典完整數位化，並提供單一辭典或跨辭典的查詢功能服務。介面的設計需參考萌典。需能支援以任何一種國家語言或華語檢索，且能檢索一個以上的國家語言。檢索模式需支援精確比對和模糊比對。

	優先整合「伍、本國國家語言相關之語言資料庫」一章中所提到的辭典相關資源，來逐步建置華語、閩南語（含閩南語各個主要的方言）、客語（含四縣腔、海陸腔、大埔腔、饒平腔、詔安腔）、原住民族語（含各語族方言別）、閩東語、臺灣手語的辭典，附上釋義、例句及影音檔，從詞彙的層面呈現各國家語言的文化底蘊，並促進不同國家語言間的文化交流。各國家語言辭典相關如下： (1) 華語辭典資源包括《教育部重編國語辭典修訂本》（網址：http://dict.revised.moe.edu.tw/）、《教育部國語辭典簡編本》（網址：http://dict.concised.moe.edu.tw/）、《教育部國語小字典》（網址：http://dict.mini.moe.edu.tw/）、《教育部異體字字典》（網址：https://dict.variants.moe.edu.tw/）、《教育部成語典》（網址：http://dict.idioms.moe.edu.tw/）等。 (2) 閩南語辭典資源包括《教育部臺灣閩南語常用詞辭典》（網址：https://twblg.dict.edu.tw/）、《國臺對照活用辭典》、《簡明臺灣語字典》等，還有《閩客語典藏》（網址：http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa）網站所收錄的字典典藏（《廈英大辭典》、《英廈辭典》、《廈門音新字典》、《台日大辭典》）。另外像是《臺灣閩南語羅馬字拼音方案》（網址：https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=446）、《臺灣閩南語漢字之選用原則》（網址：https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=15）、《臺灣閩南語推薦用字 700 字詞》（網址：https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=45）、《臺灣閩南語我嘛會每日一詞》（網址：https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=494&content_sn=4）等相關資源也可納入。

	(3) 客語辭典資源包括《教育部臺灣客家語常用辭典》（網址：https://hakkadict.moe.edu.tw/）、《客話辭典》、《臺灣四縣腔海陸腔客語辭典》、《六堆辭典》等，還有《閩客語典藏》（網址：http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa）網站所收錄的字典典藏（《客英大辭典》、《客法大辭典》）。 (4) 原住民語辭典資源包括《原住民族語言線上詞典》（網址：http://e-dictionary.apc.fishweb.com.tw/）、《巴宰語詞典》、《噶瑪蘭語詞典》、《達悟語詞典》等。另外像是《臺灣原住民語言推薦新詞》（網址：http://ilrdc.tw/research/newwords/newword106.php）等相關資源也可納入。 (5) 閩東語辭典資源包括《連江縣本土教學資源網》的《馬祖閩東語本字檢索系統(試用版)》（網址：http://fc-matsu.com/）。另外像是《連江縣本土教學資源網》的《日常生活常用詞彙》（網址：http://www.matsudialect.org/1000/index.html）、《綜合活動馬祖話》（網址：http://www.matsudialect.org/1000_2/index.htm），還有《連江縣志--語言志》（網址：http://gov.matsu.idv.tw/lienchiang/language.html）等相關資源也可納入。 (6) 其它第五章所列及審查委員建議的各項資料都包括在內。紙本資料需要先被轉成電子檔。 提供各國家語言的「公告新詞」，定期在國家語言資料庫網站上公佈各國家語言因新觀念或新科技而產生的新詞彙用法。「公告新詞」可以參考並延用原民會的新創詞創建流程（網址：http://ilrdc.tw/research/newwords/process.php），定期蒐集、討論、修訂並公告新詞，供民眾作參考。
期程	共 1 期 2 年。
人力	(1) 成立諮詢委員會：須聘請專精臺灣國家語言（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學專家學者至少 4 人。

	(2)計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上相關研究或實務經驗。資訊工程背景之專任助理 2 人，兼任助理 2 到 3 人。
經費	250 萬。

8.4.6. 國家語言地圖及語言地理資料庫

內容項目與建置方式	分二期共 4 年完成國家語言地圖及語言地理資料庫的建置。第一期整合過去的研究成果，第二期結合地理資訊系統、語言調查、田野調查、和群眾外包，建置以科技為導向的國家語言地理資料庫。 **第一期：**先將過去學者所繪製的語言地圖掃描、電子化，並整理過去田野調查所蒐集到的各種音檔，製作成樣品音來展示，讓國人可以在網頁地圖上點選聆聽並比對各地方言的發音差異。 **第一期依據的相關資源如下：** (1)洪惟仁教授於 2019 年出版的兩冊專書《臺灣社會語言地理學研究：臺灣語言的分類與分區 I》及《臺灣語言地圖集 II》，目前洪教授已經同意將這兩冊專書中的語言地圖授權給文化部。 (2)卜溫仁（Warren A. Brewer）教授於 2008 年出版之《Mapping Taiwanese》，據洪惟仁教授轉述，卜教授願意提供所收集的樣本音。 (3)張屏生教授之詞彙相關研究，詳見張教授之個人網頁：http://www.chinese.nsysu.edu.tw/zh_tw/Department_introduction/Teacher/%E5%BC%B5-%E5%B1%8F%E7%94%9F-2809422。 (4)中研院鄭錦全院士建立的「歷史語言與分佈變遷資料庫」，結合語言分佈微觀計畫，研究閩客混合的雲林縣崙背鄉、二崙鄉、新竹縣新埔鎮、苗栗縣後龍鎮、南庄鄉的語言使用，勾勒出閩客語的地理互動，網址為：http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa。

	第二期： (4) 將語言地理學研究成果與地理資訊系統（GIS）等技術結合，將樣品音以互動地圖的方式來呈現。 (5) 建置結合地理資訊系統與方言田野調查的資料庫。 (6) 提供群眾外包模式來逐步擴增地理語言學相關發音語料。 (7) 設計 APP 以群眾外包的方式收集資料。
期程	2 期 4 年。
人力	(1) 成立諮詢委員會：須聘請熟悉臺灣國家語言及語言地理學（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學的專家學者至少 4 人，協助建置資料庫。 (2) 計畫主持人及共同計畫主持人或協同計畫主持人必須分別具有語言學和計算語言學專長，且有 3 年以上相關研究或實務經驗。 (3) 專任助理 8 人，其中語言學助理 4 人，負責系統程式開發的助理 4 人。兼任助理 5 到 10 人。
經費	600 萬。

8.5. 網站架構與功能

考量國家語言資料庫包含語料庫及資料庫，以下整體網站的導覽地圖（sitemap）參考美國國家語料庫、英國國家語料庫與日本國語研究所網站之架構，包括：

(1) 簡介：語料庫概述、資料庫概述、更新狀況訊息、聯絡方式等。

(2) 語料庫檢索：以後設資料、語料特性提供相關「多條件」搜尋設定，並依特定方式排序。以中研院平衡語料庫為例，可提供文類、文體、媒體、主題等搜尋設定。常見的檢索功能有上下文關鍵詞句（concordances）、搭配詞（collocation）、n-grams 與詞頻表（frequency list）等。當使用者的搜尋設定不符合條件的時候，系統應告知使用者問題為何。

(3) 資料庫檢索：以資料庫欄位（field）為搜尋設定，並依特定方

式排列。

(4) 語料及資料下載：無授權疑慮的語料及資料可開放下載，語料部分提供 XML 與純文字格式的個別壓縮檔；資料則以 csv 或 json 檔下載。

(5) 相關工具：語料庫使用手冊、分詞標準、詞性集、後設資料說明文件等，英國國家語料庫的做法是與語料下載部分結合，會員註冊後下載的壓縮檔即包含語料及後設資料說明文件、使用手冊等。

(6) 語料、資料提供專區：提供平台上傳語料及標記等，資料部分則可提供聯絡窗口處理相關資料取得及授權。

8.5.1. 維運與管理

語料庫可視為一種特殊的資料庫。除非另有說明，以下有關國家語言資料庫的維運與管理同時適用於語料庫及資料庫。

8.5.2. 會員功能與管理

以英國國家語料庫為例，該網站提供會員制的功能。註冊會員後，可瀏覽登入活動紀錄，包括：搜尋歷史紀錄，並附上連結，以及可建立個人的字詞表，例如：在每個檢索句（concordance lines）前加上「收藏」的按鈕，方便會員建立屬於自己的列表。

從語料庫管理者的角度來看，此設計提供會員分類管理及會員資料統計，並可將相關資料匯出，例如：註冊會員人數、各語料庫的流量與點閱率等，以作為語料庫團隊調整、更新語料庫的參考。由於語料庫的建置雖是以共時（synchronic）的角度出發，從而窺見特定時代的語言樣貌，但長遠的語料庫規劃可以是歷時的（diachronic），英國國家語料庫擁有 1994 年及 2014 年兩個版本便是一例。

此外，語料庫在處理授權問題方面，從會員分類區分使用權限，基於會員需求和身份提供服務。會員分類也有助於設定API資料取得內容，不過澳洲PARADISEC典藏計畫表示，所有的後設資料都適用創用CC 授權條款的「姓名標示-相同方式分享-4.0 國際（ShareAlike 4.0 International License）」條款，使用者於註冊會員後方能取得相關資料。關於此創用 CC 授權條款的說明，請參見「4.1.2 創用 CC（Creative Commons）授權條款簡介」一節。

建議國家語言資料庫在會員功能與管理可以仿照國內外類似的平台針對不同類型的會員提供不同的功能。例如，可以將會員區分為一般社會大眾，有興趣專研的學生或社會人士，及各級學校教師或研究機構研究人員三類，分別給予不同的權限。一般社會大眾不需額外註冊可以檢索三次。超過三次就會被要求註冊帳號和密碼。最後一類會員被賦予權限可以檢視更多的細節和統計，可以將檢索的結果存在系統一週，並允許使用比較複雜的檢索功能。第三類會員在註冊時需透過任職機構的郵件地址認證。

會員功能與管理必須有忘記帳號和密碼時透過輸入註冊時提供的郵件地址取回帳號和密碼。另外為避免系統被人透過程式爬蟲濫用，每次登入必須輸入隨機出現的數字。

8.5.3. 後台內容管理系統

後台管理系統主要是內容更動的權限設定與系統的使用紀錄，資料庫和語料庫都適用。在權限設定方面，視更新頻率與需求提供更動權限給相關團隊人員，例如：語料庫須將網站上方橫幅（Banner）及最新消息專區的權限賦予相關業務承辦單位。

根據蘭卡斯特大學發表的一份論文 (Coole, 2016)，談論了語料庫管理系統的現況，傳統的內容管理系統（database management system,

DBMS）與資訊檢索系統（information retrieval, IR）不一定能夠滿足現今語料庫的功能需求，往往須仰賴電腦軟體，例如：WordSmith 和 AntConc，但單機軟體能夠處理的語料量較小；由伺服器端呼叫的工具（server-based tools），如 Wmatrix、CQPWeb、SketchEngine 等，通常是基於 Open Corpus Workbench（CWB）、MySQL 及 Lucene 之上，能夠處理較大量的語料。其中，Lucene 是以 Java 語言開發而成的全文搜尋引擎套件，可產出倒置索引檔（inverted file），省去逐字比對的檢索時間，提升檢索的效率。MongoDB 等系統則更能滿足語料庫的特定需求，並搭配平行計算（parallel computing）的方式批次處理詞性標記等任務，但仍有許多需求無法直接以現有的內容管理系統與平行計算方式實現。

有鑒於語料庫管理系統的相關需求，蘭卡斯特大學開發了 lexiDB 系統 (Coole, 2016)，亦以 Java 語言開發，針對語料儲存與搜尋設計而成，可以使用四種方式搜尋，分別是上下文關鍵詞句（concordances）、搭配詞（collocation）、n-grams 與詞頻表（frequency list）。其特點在於資料庫索引的設計（index scheme），符合齊夫定律的分佈特性（Zipfian nature），意即將詞彙依照出現頻率高至低排序，第二常見的頻率是最常見頻率的二分之一（1/2）。此外，相同字形、不同詞性的詞彙分成兩個詞條檢索。

由於語料庫規模龐大，上述提到的搜尋方式有一部份限制，例如：上下文關鍵詞句須提供結果數目上限，避免使用者在搜尋高頻率詞彙、常見詞彙時系統無法負荷；在 n-grams 功能上，lexiDB 與其他系統不同，多數系統的 n-grams 檢索是事先建立好的資料，因此只能檢索 n 為特定數字的資料，但 lexiDB 可快速計算出 n 為任意數字的資料。詞頻表也不僅是高低頻率的呈現，使用者可以進一步計算詞彙的關鍵值

（keyness），讓使用者選擇不同的統計方法，像是對數似然比（log-likelihood test）、卡方檢定（chi-squared test）等，再搭配效果量（effect size）的權重，更反映語言學的分析基礎。

無論何種設計都無法滿足所有的使用者，因此需允許使用者可以下載一定數量範圍內的檢索內容，自行使用他們熟悉的工具進一步運用他們下載的資料。這一個功能必須限制只提供給有興趣專研的學生或社會人士，以及各級學校教師或研究機構研究人員。

透過後台的管理工具，可以找出哪些詞被檢索，這些詞當中有哪些詞在資料庫中找不到，這些資訊可以協助管理者瞭解語言資料庫需要增加哪些詞彙。

無論是語言資料庫或語料庫的建置都需跨領域人才合作，應組成同時具有語言學背景及資訊技術之團隊執行語料庫建置計畫，在初始階段提出後台管理需求表及資料呈現形式，以讓資訊人員了解後台管理系統規劃內容。

8.5.4. 系統效能與上線測試

在網站正式上線之前，預留時間進行測試、模擬實際使用情形、修正，主要的任務為測試網站速度與解決資安問題。應說明以何種工具或方式確保資訊安全，並提供網頁檢測報告與資訊安全管理計畫，例如：IPv6 協定、OWASP 10 大漏洞等。

8.5.5. 語料及資料的處理與管理

資料庫/語料庫從初建、擴充到完善是一個長時間的生命週期（life cycle），許多資料庫/語料庫最初都是創建者田野調查的成果，亦有實驗室為了發展相關技術而蒐集之語料。另外，英國國家語料庫、開放美國國家語料庫等參考語料庫（reference corpus），特色是大量多樣的語料，需要自動化及譯後編輯（post-editing）協助轉寫與標記工作

的進行。(Bird et al., 2009) 以語料的轉寫和標記為例，格式的一致性是資料品質控管的其中一項，CHILDES 提供的 CLAN 軟體（Computerized Language ANalysis）和 SALT（Systematic Analysis of Language Transcript）軟體是兒童語言習得常見的工具，尤其適合以貫時性研究（longitudinal study）的語料維護，讓研究者得以長時間追蹤受試者的資料。(Behrens, 2008) CLAN 附有檢查功能，可列出不符格式的行號及錯誤訊息，如果想要一次檢查多個檔案，也有提供命令列介面給使用者。

此外，由於標記須耗費龐大人力與時間，通常會將標記拆分成許多細項，逐步完成，因此須建立標記手冊（annotation guidelines），定期更新與補充新的標記標準，並記錄哪些檔案已完成標記、已通過標記驗證或尚未完成等。同一份文件兩位標記者有不同的標記或分詞結果，也可以透過程式工具自動檢查出來。在資訊技術方面，須記錄更新或修改的時間（timestamp）與更動內容，並區分已可發布之內容或仍在檢審等階段訊息。

8.6. 應用與推廣

建議未來國家語言資料庫允許使用者下載一部份資料，最終朝向公開資料（open data）的方向發展，另也開發資料庫/語料庫工具及 API 等，讓使用者對資料和語料能有更多元深入的處理與應用。從語言資料庫/語料庫的建置過程中，將國家語言的資源進行整理，包括資料的檢查與修正、後設資料的建立、以及相關語言處理工具的開發等，將現有資源典藏，逐步加值應用，轉化為資料庫、語料庫的形式。在學術上集結各研究團隊的專業，共同充實各項語言資源。在應用方面，日前 Musgrave & Haugh (2020) 發表了一篇論文，提到國家語料庫的資料具有規模上的優勢，澳洲語言學期刊（Australian Journal of

Linguistics）也在第 34 卷第 1 期特刊廣徵以澳洲國家語料庫作為語料的研究論文。應用部分可從多語、多元文化、各層面的語言資料內容與工具的建構，扮演教育、翻譯、科技等領域發展的基石角色。除此之外，國家語料庫的長遠規劃還包含推廣的部分，英文版網站即是方法之一。其他推廣可嘗試開發手機版程式 app，增加其使用，也可以舉辦競賽，激發大眾對國家語料庫的想像與創意等，例如閩南語等國家語言與長照服務的結合也是一實用的發想主題。只要國家語言資料庫完成，以下幾種 app 程式都很容易可以設計出來，包括各國家語言詞彙的自動轉換、以華語跨語言檢索閩南語或其他國家語言詞彙，辭典資料庫與語料庫的整合檢索。其他像閩南語或其他國家語言的分詞和詞性標記程式，各國家語言的語音辨識和合成程式，只要訓練語料夠大，以目前相關的技術和工具成熟的程度，都是可行的。甚至各國家語言之間的機器翻譯系統，只要有足夠的平行語料，假以時日也可以達到一定的水準。

玖、　文化部「建置國家語言資料庫」勞務採購案需求規範說明書建議之草案

專案源起及目標：

語言為文化傳承之重要載體，為促進語言永續發展、並豐富國家之文化內涵，文化部特制定《國家語言發展法》，並於108年1月9日經總統公布施行；爰本部現依該法第八條規定：「政府應定期調查提出國家語言發展報告，建置國家語言資料庫」，辦理建置國家語言資料庫之計畫。國家語言資料庫除應含國家語言語料庫外，亦應納入各國家語言史料、調查統計等相關資料，以作為國家語言傳承、復振及發展之基石。然目前國內語言研究資料及語料之蒐集整理工作，除民間有零星的計畫與成果外，亦分散在各相關政府機關，如教育部、科技部、客家委員會、原住民族委員會、中央研究院、國家教育研究院等，本部因此於108年辦理「建置國家語言資料庫先期規劃研究案」，盤點、搜羅及整理國內外近代具代表性之語料庫之建置型態、維運管理、應用推廣、著作權議題等面向進行研究。本部現以此研究成果為基礎，著手落實建置國家語言資料庫之工作，盼能透過此國家語言資料庫之建置，促進不同族群間之互相理解，並能裨益未來學術研究發展，以期永續保存國家語言文化資產。

一、 **執行期程：**自簽約日起至　年　月　日止，分成3期程，共計6年期。第1期程為自簽約日起至　年　月　日；第2期程自完成第1次

後續擴充起至　年　月　日；第 3 期程自完成第 2 次後續擴充起至　年　月　日止。

二、 **預算金額：**本案總採購金額共計為新臺幣 9100 萬元整（含稅）。

(一)　國家語言資料庫網站（含國家語言現況的介紹）：250 萬元。

(二)　閩南語語料庫：5000 萬元。

(三)　閩東語語料庫：1500 萬元。

(四)　臺灣手語語料庫：1500 萬元。

(五)　國家語言辭典資料庫：250 萬元。

(六)　國家語言地圖及語言地理資料庫：600 萬元。

本案　年度至　年度預算如未獲立法院審議通過或經部分刪減，雙方得另行協商契約內容；如本部無法依本契約金額履約或無法達成契約目的時，得依政府採購法第 64 條規定辦理。**另廠商所投計畫書報價超過預算者為不合格標，不予減價機會。**

三、 **採購方式**

本案依採購法第 22 條第 1 項第 9 款辦理限制性招標公開評選。

四、 委辦需求事項：

(一) 成立顧問諮詢委員會

(1) 為利國家語言資料庫建置，須聘請熟悉閩南語、閩東語、客語、原住民各族語、閩東語、臺灣手語、計算語言學、著作權之專家學者（含各語言文化專家）組成諮詢委員會。

(2) 諮詢委員會人數需求如下：

A. 國家語言資料庫網站（含國家語言現況之介紹）：須聘請專精臺灣國家語言（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學專家學者至少 5 人。每年至少召開 4 次諮詢，協助國家語言資料庫網站之建置。

B. 閩南語語料庫：依據期末審查會議委員的意見，閩南語語料庫的諮詢委員必須包括其它國家語言的專家以及著作權之專家，以便建置閩南語語料庫的經驗未來可以轉移到其它的國家語料庫。須聘請熟悉閩南語語料庫（至少 4 人）、客語、原住民各族語、臺灣手語、閩東語、著作權、語料庫計算語言學（各至少 1 人）之專家學者（含各語言文化專家）組成諮詢委員會。每年至少召開 4 次諮詢，協助閩南語語料庫之建置。

C. 閩東語語料庫：成立諮詢委員會：須聘請熟悉閩東語（至少4人）、著作權（至少1人）、語料庫計算語言學之專家學者（至少2人）組成諮詢委員會。每年至少召開4次諮詢，協助閩東語語料庫之建置。

D. 臺灣手語語料庫：須聘請熟悉臺灣手語（至少 4 人）、著作權（至少 1 人）、語料庫計算語言學（至少2人）之專家學者組成諮詢委員會。每年至少召開4次諮詢，協助臺灣手語語料庫之建置。

E. 國家語言辭典資料庫：須聘請專精臺灣國家語言（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學專家學者至少4人組成諮詢委員會。每年至少召開4次諮詢，協助國家語言辭典資料庫之建置。

F. 國家語言地圖及語言地理資料庫：熟悉臺灣國家語言及語言地理學（包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語）的語言學或計算語言學的專家學者至少4人，組成諮詢委員會，每年至少召開4次諮詢，協助國家語言地圖及語言地理資料庫的建置。

(3) 各項目諮詢委員會於計畫執行期程間每季須至少召開 一次座談會，協助建置語料庫及資料庫等相關事宜，相關出席費及交通費等費用由本案支付。

(4) 廠商須於服務建議書中提出對於語料資料內容規範、語料蒐集及處理規範、用字規範，及斷詞、詞性標記初步構想，並於　年　月　日前取得諮詢委員會共識，經本部同意後實施。

(二)　國家語言資料庫之建置：

(1) 國家語言資料庫當中包含：

A. 國家語言資料庫網站（含國家語言現況之介紹）（共兩年）

i. 參考國家語言資料庫先期規劃研究計畫結案報告及各種相關文獻，以網頁和影片介紹臺灣各國家語言之現況，使國人瞭解國家語言面臨的傳承危機以及保存國家語言之重要性。

ii. 網頁及影片必須涵蓋所有臺灣國家語言的現況，且長度須至少達 30 分鐘，用以介紹其語言現狀。

iii. 格式：網頁設計採用 HTML 4、CSS、Javascript，網頁內包含影片超連結。網頁必須採取響應式設計（responsive web design）。影片解析度至少需達到 FULL HD 畫質，影片格式包含 mp4 等常見影片格式。資料庫需保留未來各國家語言完成後之介接。

B. 閩南語語料庫（橫跨第 1、2、3 期程，共六年）

i. 規模：包含書面語及口語兩部份，共 1000 萬詞。為兼顧平衡性之考量，書面語語料以不超過 800 萬詞為原則。

ii. 語料內容規範：語料分類屬性須包含以下內容，並應兼顧平衡性之考量：

(a) 至少包含書面語及口語兩個部分。

(b) 文體：含記敘、描寫、論說、說明等分類。

(c) 語式：廠商須規劃語料庫應具備相關語式分類。

(d) 主題：如社會、生活、文學、科學、哲學宗教、教育、影視娛樂、體育、政治、藝文創作……等各類主題，而不侷限於某些特定主題。

iii. 語料輸出與儲存格式

廠商須依據本案目標用途及未來語料庫應用模式，蒐集下列語料庫輸出與儲存格式：

(a) 書面語：xml 格式文字檔及純文字檔。

(b) 口語：文字檔、影音檔、聲音檔。

iv. 建立後設資料（Metadata）：以 xml 格式呈現

語言調查的資料應至少包含以下後設資料項目：

(a) 性別（若無可免）

(b) 地區（若無可免）

(c) 語言

(d) 形式（如：書面、口語、口語書面……等等）

(e) 來源

(f) 年份

(g) 文類

新收集的資料則應至少包含以下後設資料項目：

(a) 性別

(b) 地區

(c) 語言

(d) 形式（如：書面、口語、口語書面……等等）

(e) 來源

(f) 年份

(g) 文類

v. 語料蒐集及處理規範

(a) 書面語

(i) 使用教育部閩南語常用詞辭典規定之圖書、雜誌、報紙、教材、文學作品、劇本、歌本、教科書……等資料。

(ii) 徵求相關資料所有者之同意並取得上述資料之授權。

(iii) 遵照閩南語用字規範（請參照「用字規範」一節）將文本數位化並完成建檔。

(iv) 數位化文本斷詞及詞性標記。

(v) 至少兩次以上之交叉校訂用字及格式。

(b) 口語

(i) 語料來源：電視新聞、節目、廣播、電影等現有的影音資料。

(ii) 徵求相關資料所有者之同意並取得上述資料之授權。

(iii) 語音資料之品質須清晰可聽辨（包含聲音、畫面）。

(iv) 儘可能達到資料發音人男女比例均衡，且涵蓋不同年齡層。

(v) 各語言之資料數量須盡可能達成平衡。

(vi) 資料轉寫成文字、數位化及建檔，並且提供以語句為單位之文字與口語參照時間點標記。

(vii) 語料須遵照用字規範。

(viii) 語料須斷詞及詞性標記。

(ix) 錄音檔案須規劃編號（檔案命名）處理原則，做成光碟母片（至少包含原始檔 WAV、WMA 檔及 MP3 檔等 3 種聲音檔格式）；且影音檔需經加工，含隱私處理、音檔切割、其他後製（字幕或浮水印）等作業。

(x) 另，廠商應規劃考量上線後使用媒介之需求（如網站、APP 軟體等媒介）擴增聲音檔格式，如 RA、RM 系列等聲音檔格式。

(xi) 廠商須至少完成 2 次以上之交叉校訂用字及格式。

vi. 用字規範：參照教育部閩南語常用詞辭典書寫。

vii. 分詞、詞性標記

(a) 原則擬定：分詞和詞性標記須由專家組成團隊制定。團隊成員必須包括語言學及計算語言學專家。有關於分詞和詞性標記操作性的原則必

須先擬定一個類似中研院詞庫小組中文詞性分析技術手冊的檔案，提供明確的例子和判斷的方法。

(b) 詞性標記集：參考中研院詞庫小組中文詞性分析技術手冊及閩南語的語法特點與相關文獻，語法標記集的數量以不多於中研院詞庫小組簡化標記，及不少於精簡標記為原則。

(c) 每一份語料的分詞及詞性標記皆需經由兩位熟悉分詞和詞性標記原則的助理處理，再由程式自動找出兩者不一致的地方，交由專家團隊開會決定。且這些資料當中常犯的錯誤應作為新進助理訓練用的資料。

viii. 系統功能規劃：參考 COCA 語料庫介面的功能，應包含至少以下幾項

(a) 關鍵詞前後文檢索程式(KWIC)，使用者輸入一個詞或字串後，可以選擇詞性、要檢索哪些語料庫，以及是否顯示詞性標記，輸出時會顯示這個字串在語料庫中出現的次數，以及例句。該句如有華語的翻譯，也會一併顯示。

(b) 字串在不同語料庫出現的頻率。

(c) 搭配詞檢索，輸入一個詞，及前後的範圍，顯示最常一起出現的搭配詞。

ix. 介面工程：除系統外，還需提供介面程式用以編輯錯別字、分詞、和詞性標記。每份編輯過的語料都記載標注的助理姓名及編輯時間。

(a) 廠商需依本部需求，規劃、設計與開發本案使用者介面，包含建置語料庫使用檢索介面、語料庫後端上傳介面、書面及口語語料入庫系統等。

(b) 廠商應於服務建議書說明設計之資料庫系統功能，並提出對於使用者管理存取機制控管之建議。廠商並應就本案期程，說明各期規劃之工作項目、內容及進度。

(c) 除規劃開發建置檔案外，廠商尚須設計線上登入或整批資料檔匯入功能（即除開發單筆線上建檔外，開發整批上傳的功能）。

(d) 批次上傳（下載）作業：開發提供整批建檔資料或審查資料，提交「各臺灣語言專家學者」審查。並提供判斷，將重複性資料一併標示及呈現，俾利本部及「各臺灣語言專家學者」逐一檢視與審查流程。

(e) 資訊網站

(i) 會員管理功能

1、 開放大眾申請為會員，並提供會員分類管理及會員資料統計之功能，包含匯出自網站流量與註冊會員人數，及各類型語料庫之點閱率、點閱記錄等資料。

2、 將管理權限依據會員需求和會員身份（如學術、教學、學習……等等）做不同的授權。

3、 規劃個人收藏功能以供會員作詞彙收藏與下載之使用。

4、 聘雇人員專責回應、處理、回報各類型會員留言或建議事項。

(ii) 搜尋功能

1、 提供網站導覽地圖功能。

2、 提供可選定多條件（如：語料庫屬性、關鍵字等常用條件）的搜尋模式。

3、 提供語言設定、特定欄位資訊以及關鍵字搜尋法。

4、 依選定之條件出現相對應之結果，並可依指定方式排序。

5、 依選定資料特性檢核輸入欄位中之資料，當使用者輸入不符合檢核條件之內容時，應提示錯誤訊息，告知產生錯誤之原因，並給予使用者排解錯誤之指示。

(iii) 後臺管理功能：廠商應依本部需求，規劃、設計與開發國家語言資料庫之後臺管理系統，包括各項權限控管、參數設定等。對於國家語言資料庫之各項系統，包括使用者介面等重要之操作，並應保存紀錄備查。

廠商規劃之後臺管理系統至少須包含下列功能：

1、 系統日誌：登入(出)紀錄、查詢紀錄、下載與輸出操作等紀錄，以確保系統能被正確、合法使用。

2、 Banner 管理：

- 增設播放時間設定，控制前臺輪播時間。
- 開放本部網站之管理單位有增、修及刪之權限。

3、 最新消息分類：

- 消息區分以動態方式增設，保留擴充彈性。

x. 系統效能與上線測試

(a) 廠商應於上線至少 14 個日曆天前，提出網頁應用程式之弱點掃描、滲透測試、及程式原始碼安全檢測報告、壓力測試報告，以及提交「資訊安全管理計畫」，另需配合機關資安檢測作業修正相關弱點。網頁需符合 IPv6 規範（支援 IPv6 協定）及提供 IPv6 服務。並提供至少 3 組測試帳號、密碼供機關進行線上試營運。

(b) 廠商應提出具體檢查方式或工具，確保於相關開發及維護之軟體系統中，無植入木馬程式、後門程式或任何有違害機關資訊安全之程式碼，並應經原始碼檢測工具進行檢測有無最新 OWASP TOP 10 大漏洞及其他缺失，經審核無重大缺失後，再行上線，廠商應預留本項工作檢測及修正期間，另檢測報告並納入結案驗收項目。

(c) 廠商完成本案語料庫累計至少書面語語料數位化達 800 萬詞、口語語料轉寫及數位化達 200 萬詞時，應模擬 1,000 人同時上線作業為驗證基準，在 1 分鐘內從機關內部網站，瀏覽本案開發之網站，在未引發任何例外狀況下，須滿足

每項作業仍可持續提供服務，且語料庫使用檢索查詢平均反應時間不超過 5 秒，最長反應時間不超過 10 秒，並作為驗收檢測條件之一。

xi. 本案語料庫需保留後續擴充性，未來需能夠與本會建置之資料庫進行資料交換與介接作業。

xii. 其他需求

(a) 前述需求之細部需求規格，仍應以需求訪談正式確認之結果為準。

(b) 廠商於設計本案各項資訊系統權限控管方式時，應依本部實際需求賦予不同的操作範圍或查詢範圍。

C. 閩東語語料庫（橫跨第 1、2、3 期程，共六年）

i. 規模：以三期共六年建置一百萬詞規模的閩東語語料庫。包含書面語及口語兩部份，共 100 萬詞。為兼顧平衡性之考量，書面語語料以不超過 80 萬詞為原則。口語部分以不少於 20 萬詞為原則。提供以句為單位的閩東語與華語對照。挑選 10 萬詞有音檔的口語語料，以句為單位，將閩東語口語語音資料轉寫為羅馬拼音和漢字，且發音與文字時間必須對齊，並附上對應之華語翻譯。必須使用語音分析與標示工具 PRAAT 的格式，方便未來進一步的應用。

ii. 資料格式與其它國家語料庫一致，採 XML 格式、Unicode 編碼。語料需提供分詞及詞性標記。

iii. 在語料蒐集方面，以現有閩東語語言資源為主，書面語以可以連江縣本土資源教育網裡面的資源為基礎，口語語料可蒐集日常使用語詞、諺語、兒歌、歇後語，兒歌作為語料庫材料。馬祖社區電台的廣播節目可作為口語語料，其內容涵蓋一分鐘母語教學、60 秒母語形象、20 分鐘母語教學、母語線上廣播等。

iv. 語料內容規範：語料分類屬性須包含以下內容，並應兼顧平衡性之考量：

(a) 至少包含書面語及口語兩個部分。

(b) 文體：含記敘、描寫、論說、說明等分類。

(c) 語式：廠商須規劃語料庫應具備相關語式分類。

(d) 主題：如社會、生活、文學、科學、哲學宗教、教育、影視娛樂、體育、政治、藝文創作……等各類主題，而不侷限於某些特定主題。

v. 語料輸出與儲存格式

廠商須依據本案目標用途及未來語料庫應用模式，蒐集下列語料庫輸出與儲存格式：

(a) 書面語：xml 格式文字檔及純文字檔。

(b) 口語：文字檔、影音檔、聲音檔。

vi. 建立後設資料（Metadata）：以 xml 格式呈現

語言調查的資料應至少包含以下後設資料項目：

(a) 性別（若無可免）

(b) 地區（若無可免）

(c) 語言

(d) 形式（如：書面、口語、口語書面……等等）

(e) 來源

(f) 年份

(g) 文類

新收集的資料則應至少包含以下後設資料項目：

(a) 性別

(b) 地區

(c) 語言

(d) 形式（如：書面、口語、口語書面……等等）

(e) 來源

(f) 年份

(g) 文類

vii. 語料蒐集及處理規範

(a) 書面語

(i) 圖書、雜誌、報紙、教材、文學作品、劇本、歌本、教科書……等資料。

(ii) 徵求相關資料所有者之同意並取得上述資料之授權。

(iii) 遵照用字規範（請參照「用字規範」一節）將文本數位化並完成建檔。

(iv) 數位化文本斷詞及詞性標記。

(v) 至少兩次以上之交叉校訂用字及格式。

(b) 口語

(i) 語料來源：電視新聞、節目、廣播、電影等現有的影音資料。

(ii) 徵求相關資料所有者之同意並取得上述資料之授權。

(iii) 語音資料之品質須清晰可聽辨（包含聲音、畫面）。

(iv) 儘可能達到資料發音人男女比例均衡，且涵蓋不同年齡層。

(v) 各語言之資料數量須盡可能達成平衡。

(vi) 資料轉寫成文字、數位化及建檔，並且提供以語句為單位之文字與口語參照時間點標記。

(vii) 語料須遵照用字規範。

(viii) 語料須斷詞及詞性標記。

(ix) 錄音檔案須規劃編號（檔案命名）處理原則，做成光碟母片（至少包含原始檔 WAV、WMA 檔及 MP3 檔等 3 種聲音檔格式）；且影音檔需經加工，含隱私處理、音檔切割、其他後製（字幕或浮水印）等作業。

(x) 另，廠商應規劃考量上線後使用媒介之需求（如網站、APP 軟體等媒介）擴增聲音檔格式，如 RA、RM 系列等聲音檔格式。

(xi) 廠商須至少完成 2 次以上之交叉校訂用字及格式。

viii. 用字規範：參照《馬祖閩東語本字檢索系統》。

ix. 分詞、詞性標記

(a) 原則擬定：分詞和詞性標記須由專家組成團隊制定。團隊成員必須包括語言學及計算語言學專家。有關於分詞和詞性標記操作性的原則必

須先擬定一個類似中研院詞庫小組中文詞性分析技術手冊的檔案，提供明確的例子和判斷的方法。

(b) 詞性標記集：參考中研院詞庫小組中文詞性分析技術手冊及閩東語的語法特點與相關文獻，語法標記集的數量以不多於中研院詞庫小組簡化標記，及不少於精簡標記為原則。

(c) 每一份語料的分詞及詞性標記皆需經由兩位熟悉分詞和詞性標記原則的助理處理，再由程式自動找出兩者不一致的地方，交由專家團隊開會決定。且這些資料當中常犯的錯誤應作為新進訓練助理訓練用的資料。

x. 系統功能規劃：參考 COCA 語料庫介面的功能，應包含至少以下幾項

(a) 關鍵詞前後文檢索程式(KWIC)，使用者輸入一個詞或字串後，可以選擇詞性、要檢索哪些語料庫，以及是否顯示詞性標記，輸出時會顯示這個字串在語料庫中出現的次數，以及例句。該句如有華語的翻譯，也會一併顯示。

(b) 字串在不同語料庫出現的頻率。

(c) 搭配詞檢索，輸入一個詞，及前後的範圍，顯示最常一起出現的搭配詞。

xi. 介面工程：除系統外，還需提供介面程式用以編輯錯別字、分詞、和詞性標記。每份編輯過的語料都記載標注的助理姓名及編輯時間。

(a) 廠商需依本部需求，規劃、設計與開發本案使用者介面，包含建置語料庫使用檢索介面、語料庫後端上傳介面、書面及口語語料入庫系統等。

(b) 廠商應於服務建議書說明設計之資料庫系統功能，並提出對於使用者管理存取機制控管之建議。廠商並應就本案期程，說明各期規劃之工作項目、內容及進度。

(c) 除規劃開發建置檔案外，廠商尚須設計線上登入或整批資料檔匯入功能（即除開發單筆線上建檔外，開發整批上傳的功能）。

(d) 批次上傳（下載）作業：開發提供整批建檔資料或審查資料，提交「各臺灣語言專家學者」審查。並提供智慧判斷，將重複性資料一併標示及呈現，俾利本部及「各臺灣語言專家學者」逐一檢視與審查流程。

(e) 資訊網站

(i) 會員管理功能

1、 開放大眾申請為會員，並提供會員分類管理及會員資料統計之功能，包含匯出自網站流量與註冊會員人數，及各類型語料庫之點閱率、點閱記錄等資料。

2、 將管理權限依據會員需求和會員身份（如學術、教學、學習……等等）做不同的授權。

3、 規劃個人收藏功能以供會員作詞彙收藏與下載之使用。

4、 聘雇人員專責回應、處理、回報各類型會員留言或建議事項。

(ii) 搜尋功能

1、 提供網站導覽地圖功能。

2、 提供可選定多條件（如：語料庫屬性、關鍵字等常用條件）的搜尋模式。

3、 提供語言設定、特定欄位資訊以及關鍵字搜尋法。

4、 依選定之條件出現相對應之結果，並可依指定方式排序。

5、 依選定資料特性檢核輸入欄位中之資料，當使用者輸入不符合檢核條件之內容時，應提示錯誤訊息，告知產生錯誤之原因，並給予使用者排解錯誤之指示。

(iii) 後臺管理功能：廠商應依本部需求，規劃、設計與開發國家語言資料庫之後臺管理系統，包括各項權限控管、參數設定等。對於國家語言資料庫之各項系統，包括使用者介面等重要之操作，並應保存紀錄備查。

廠商規劃之後臺管理系統至少須包含下列功能：

1、 系統日誌：登入(出)紀錄、查詢紀錄、下載與輸出操作等紀錄，以確保系統能被正確、合法使用。

2、 Banner 管理：

- 增設播放時間設定，控制前臺輪播時間。
- 開放本部網站之管理單位有增、修及刪之權限。

3、 最新消息分類：

- 消息區分為一般消息或各國家語言能力認證考試相關資訊等數種分類資料，並以動態方式增設，保留擴充彈性。

xii. 系統效能與上線測試

(a) 廠商應於上線至少 14 個日曆天前，提出網頁應用程式之弱點掃描、滲透測試、及程式原始碼安全檢測報告、壓力測試報告，以及提交「資訊安全管理計畫」，另需配合機關資安檢測作業修正相關弱點。網頁需符合 IPv6 規範（支援 IPv6 協定）及提供 IPv6 服務。並提供至少 3 組測試帳號、密碼供機關進行線上試營運。

(b) 廠商應提出具體檢查方式或工具，確保於相關開發及維護之軟體系統中，無植入木馬程式、後門程式或任何有違害機關資訊安全之程式碼，並應經原始碼檢測工具進行檢測有無最新 OWASP TOP 10 大漏洞及其他缺失，經審核無重大缺失後，再行上線，廠商應預留本項工作檢測及修正期間，另檢測報告並納入結案驗收項目。

(c) 廠商完成本案語料庫累計至少書面語語料數位化達 80 萬詞、口語語料轉寫及數位化達 20 萬詞時，應模擬 1,000 人同時上線作業為驗證基

準，在 1 分鐘內從機關內部網站，瀏覽本案開發之網站，在未引發任何例外狀況下，須滿足每項作業仍可持續提供服務，且語料庫使用檢索查詢平均反應時間不超過 5 秒，最長反應時間不超過 10 秒，並作為驗收檢測條件之一。

xiii. 本案語料庫需保留後續擴充性，未來需能夠與本會建置之資料庫進行資料交換與介接作業。

xiv. 其他需求

(a) 前述需求之細部需求規格，仍應以需求訪談正式確認之結果為準。

(b) 廠商於設計本案各項資訊系統權限控管方式時，應依本部實際需求賦予不同的操作範圍或查詢範圍。

D. 臺灣手語語料庫（分三期共六年）

i. 建立總共至少有 10 小時長的臺灣手語多模態語料庫及檢索系統。

ii. 領域及主題需儘量多元。

iii. 需至少有五分之一包含兩位臺灣手語者的手語對話影片。

iv. 參考英國手語語料庫和澳洲手語資料庫的設計，由臺灣手語使用者擔任發音人，錄製手語使用

者的敘事、對話、訪談影片，蒐集內容可參考英國手語資料庫設計詞彙列表、以圖像引導（elicit）發音人打出該手語詞彙，抑或是參考澳洲手語資料庫的作法，邀請二或多位發音人組成焦點團體（focus group），從訪談或對話中擷取主題式資料。一部份資料可採用公視手語新聞的影片。

v. 標記處理項目包括「手形」、打法「位置」及「意義」說明等項目。

vi. 檢索頁面提供影像及標記訊息，可讓使用者調整影像的播放速度、下載部分影像檔或標記檔等。

vii. 需有華語的翻譯。

viii. 手語語料庫應整合手語資料庫。建議使用開源的 NGT 資料庫並使用 ELAN 軟體以 EVC（external controlled vocabulary）檔來管理資料庫的詞彙，標記人員在標記語料時便從此 EVC 檔案尋找符合的詞條。可以 ID 查詢表確認該詞彙是否收錄在資料庫中，再確認該詞彙是否多義，是否可能已被翻譯成其他意思。在標記上，如果有多個選擇，以使用頻率（frequency）和象似性（iconicity）為原則，標記項目包括：複合詞的切分及對應詞彙、單手或雙手

（handedness）及是否雙手動作對稱（symmetry）、手形變化（handshape changes）、動作方向（movement direction）及是否重複（repetition）、手勢打在身體的哪個部位上（location），此外亦提供實名（name entity）及語意欄（semantic field）的標記該資料庫的系統可即時更新最小對立體（minimal pair）的資料。

E. 國家語言地圖及地理資料庫（分兩期共四年）

i. 分兩期共四年完成國家語言地圖及地理資料庫的建置。第一期整合過去的研究成果；第二期結合地理資訊系統、語言調查、田野調查、和群眾外包等，建立以科技為導向的國家語言地理資料庫。

ii. 第一期：將過去學者所繪製的語言地圖掃描、電子化，並整理過去田野調查所蒐集到的各種腔調音檔，製作成樣品音展示，讓國人點選聆聽並比對各地方言的發音差異。第一期依據的相關資源如下：

a. 洪惟仁教授於2019年出版的兩冊專書《臺灣社會語言地理學研究：臺灣語言的分類與分區I》及《臺灣語言地圖集II》，目前洪教授

已經同意將這兩冊專書中的語言地圖授權給文化部。

b. 卜溫仁（Warren A. Brewer）教授於 2008 年出版之《 Mapping Taiwanese 》。

c. 張屏生教授之詞彙相關研究，詳見張教授之個人網頁：
http://www.chinese.nsysu.edu.tw/zh_tw/Department_introduction/Teacher/%E5%BC%B5-%E5%B1%8F%E7%94%9F-2809422

d. 中研院鄭錦全院士建立的「歷史語言與分佈變遷資料庫」，結合語言分佈微觀計畫，研究閩客混合的雲林縣崙背鄉、二崙鄉、新竹縣新埔鎮、苗栗縣後龍鎮、南庄鄉的語言使用，勾勒出閩客語的地理互動，網址為：
http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa

iii. 第二期：

a. 將語言地理學研究成果與地理資訊系統（GIS）等技術結合，將樣品音以互動地圖的方式來呈現。

b. 將田野調查與上述系統結合。

c. 提供群眾外包模式來逐步擴增地理語言學相關發音語料。

d. 設計 APP 以群眾外包的方式收集資料。

iv. 廠商須負責並確保取得相關資料所有人之授權同意，方得將資料收錄進國家語言資料庫中。

F. 國家語言辭典資料庫（一期二年）

i. 將目前市面上能夠找到且能夠取得授權的字辭典完整數位化，並提供單一辭典或跨辭典的查詢功能服務。

ii. 介面的設計須參考萌典，並須能支援以任何一種國家語言或華語檢索，且能檢索一個以上的國家語言。檢索模式需支援精確比對和模糊比對。

iii. 優先整合《文化部建置國家語言資料庫先行研究計畫》之第五章中所提到的辭典相關資源，來逐步建置華語、閩南語（含閩南語各個主要的方言）、客語（含四縣腔、海陸腔、大埔腔、饒平腔、詔安腔）、原住民語（含各語族方言別）、閩東語、臺灣手語的辭典，附上釋義、例句及影音檔，從詞彙的層面呈現各國家語言的文化底蘊，並促進不同國家語言間的文化交流。

(三) 系統平臺設備需求、網路環境及資訊安全規劃

(1) 廠商需提出下列相關系統平臺、施測設備、網路環境及資訊安全等之軟硬體環境、相關軟體程式版本訊息等整體開發規劃，於取得本案顧問諮詢委員會共識且經本部同意後，方得據以執行。主機伺服器軟硬體及網路頻寬：

A. 主機伺服器中央處理器之等級及伺服器數量。

B. 主機伺服器主記憶體容量及擴充性。

C. 主機伺服器網路卡性能等級。

D. 主機伺服器資料庫 RAID 性能及支援。

E. 主機伺服器硬碟容量及數量。

F. 主機伺服器作業系統。

G. 主機伺服器資料庫軟體。

H. 主機伺服器防毒軟體。

I. 網路承載、頻寬及不斷電系統之評估與規劃。

J. 資訊安全、個人資料防護措施及規劃。

K. 網路設備規劃。

L. 異地備援設備規劃。

(2) 系統平臺營運，系統架構所建置之各項伺服器或運作環境，必須具備下列功能或條件：

A. 標準性：中文化操作介面，不論是前端使用者或後臺管理者，且本案網站 環境需具備高度相容性，能於各種瀏覽器正常顯示【例如 IE10(含) 以上、Edge、Firefox、Google Chrome、Safari 等】且功能操作正常，並符合 W3C 規格。另外，亦須考慮網路傳輸因素，整體製作以瀏覽平順為原則；若需使用外掛軟體，需取得授權，以供使用者免費安裝使用。

B. 擴展性：系統架構規劃時，需考量未來需求及應用增加時，系統擴充及整合之彈性與延續性，當新的功能需求出現時，只需在現有機制上增加新的應用與服務模組，而不需更換整體系統；如因應新技術開發，必須調整整體系統時，原有資料庫資訊必須完整移轉至新系統，並維持正常營運，且須經本會同意後方能執行，本會無須負擔額外費用。

C. 先進性：應採用市場領先且成熟之技術【如 Web2.0(含)以上、HTML5、CSS3、AJAX 等】，使本計畫不僅滿足目前需求，而且能夠符合數年內資訊科技主流趨勢，例如配合國家發展委員會規劃之雲端發展趨勢。

D. 安全性：確保國家語言資料庫內容與資料不被第三者竊取利用，主機存取安全，資料庫資料可定期備份。

(3)若廠商使用之作業系統及資料庫，需自行提供及安裝有版權之軟體，安裝之軟體需為最新版本，並需定期更新維護。

(4)建置完成之網站需可使用 Android、iOS 各類行動裝置，及電腦系統閱讀，並符合響應式網頁(RWD)設計規範，當使用不同載具時，系統必須自動調整畫面大小與解析度，且不得有照片變形情形出現。

(5)廠商架設網站應依國家發展委員會訂頒之無障礙網頁開發規範辦理。並依據行政院 106 年 8 月 29 日院授發資字第 1061502362 號函訂定之「政府資料開放優質標章暨深化應用獎勵措施」中資料開放可取得性、易於被處理、易於理解等原則進行規劃（如附件一）。

(6)依本會要求維護、新增、修改、更新本案語料庫網站版面、功能及相關資料內容，並定期檢視及更新各語言版本內容，及維護系統及功能正常運作。

(7)相關資料庫及網站素材需於本案結束後交予本部，廠商並需確保資料庫相關之素材無侵害著作權之虞，若有侵權之情事，概由廠商負責。

(8)資訊安全管理需求

A. 一般資通安全規範：

(a) 密碼規則之定義及參數設定（密碼之長度、內容、重複次數、錯誤次數、有效期限、鎖定及解除機制）。

(b) 登入後在一定時間內若未執行任何動作，系統需強制登出。強制登出的時間長短，可由系統管理員統一設定。

(c) 所開發之應用系統需與作業系統及資料庫等最高權限使用者及群組區隔。

(d) 密碼長度應至少為 6 個字元（含）以上（內容至少包含大小寫英文字及數字）。

(e) 設定密碼輸入錯誤三次即須鎖住，帳號或密碼錯誤不直接明示，只顯示「帳號或密碼錯誤」。

(f) 系統要能夠記錄使用者登入登出的時間，存取、查詢及列印的資料表及資料欄位，系統操作行為，包括編輯、儲存、刪除、查詢及列印活動等紀錄。

(g) 密碼輸入時皆以暗碼顯示。

B. 網站安全監控與維護：

(a) 須使用檢查程式全面檢查應用程式，未含病毒、後門及間諜程式。

(b) 須使用弱點掃描工具，檢測本案開發之網站系統，至少能檢測：

i. 跨網站指令碼攻擊(Cross site scripting)。

ii. SQL 程式碼注入攻擊(SQL injection)。

iii. 程式碼執行攻擊(Code execution)。

iv. 目錄遊走(Directory traversal)。

v. 檔案引入攻擊弱點(File inclusion)。

vi. 網站程式原始碼暴露(Script source code disclosure)。

vii. 偵測 CRLF 注入攻擊。

viii. 偵測 OWASP Top 10 弱點。

ix. 跨頁框指令碼攻擊(Cross frame scripting)。

(a) 須檢查網頁程式已設有防 URL 攻擊功能。

(b) 檢查程式應由委外廠商自備，並提供檢測報告。

(c) 隨時監控網站，當偵測到任何未經授權的網頁檔案內容變更或異常流量時，立即通知本部。

(d) 網站更新查核機制功能。

(e) 廠商應提供檔案上傳檢核機制，檢查是否含有病毒或木馬等惡意程式，以確保網站資訊安全。

(f) 提供維護者維護紀錄列管追蹤，包含維護紀錄及網頁內容更新查核機制並製作統計表。

(g) 廠商應定期進行資安演練（每年至少一次），並配合本部資安演練相關作業。

C. 本案作業系統、資料庫及應用程式層級，除系統作業架構特殊需求外，所有密碼資料皆不得以明文型態存放。

D. 廠商於執行本案相關工作時，需確實遵守本部資訊安全相關規定（見）並於辦理此案需相關技術時，廠商並須提供本部協助

E. 保密條款要求：

(a) 廠商與其工作人員應遵守「個人資料保護法」及「國家機密保護法」等相關法律規定。

(b) 廠商對於本部所提供相關業務內容及規劃工作資料，負有保密責任，如有洩密情形發生導致損害時，應負完全賠償及法律責任。

(c) 廠商應簽具「資訊委外廠商安全同意書」（附件二）。

(d) 工作小組人員應簽具保密切結書（附件三）。

(e) 配合本部資訊內部稽核檢查（附件四）。

五、 國家語言資料庫建置完成上線後，本案廠商須配合本部執行成果發表會之廠商辦理行銷宣傳作業，免費提供國家語言資料庫成果紀錄、照片等相關資料。

六、 人力組合：廠商應成立工作小組，成員名單、經歷及分工須列入本案服務企劃書及工作計畫書項目內。

(一) 本案計畫主持人與共同主持人至少分別需具備語言學、計算語言學之專業能力，且有至少三年以上相關研究或實務經驗，工作小組至少包括：

(1) 專案主持人：各項目至少 1 人。專案主持人須全程參與本案，並擬訂計畫方針及綜理計畫相關事宜。

(2) 共同主持人：各項目至少 1 人。

A. 須共同參與本案，負責控管、規劃、分析及設計等統合本案全面之事宜，參與本案協調會議。

B. 資歷：須具相關之專案工作經驗不得少於 3 年，協助本案各國家語言資料之語音、文字及影像內容、各國家語言電子報及其他書面文本、口語語料內容等作業。

(二) 專任助理：

(1) 國家語言資料庫網站（含國家語言現況）：4-5 人。

(2) 國家語言地圖及語言地理資料庫：3-4 人。

(3) 閩南語語料庫：12 人，其中負責分詞和詞性標記的語言學助理 8-9 人。負責系統程式開發的助理 3-4 人。

(4) 國家語言辭典資料庫： 3-4 人。

(5) 閩東語語料庫：6-7 人，其中負責分詞和詞性標記的語言學助理 4-5 人。負責系統程式開發的助理 2 人。

(6) 臺灣手語語料庫：6-7 人，其中負責分詞和詞性標記的語言學助理 4-5 人。負責系統程式開發的助理 2 人。

(三) 兼任助理：

(1) 國家語言資料庫網站（含國家語言現況）：1 至 3 人。

(2) 國家語言地圖及語言地理資料庫：1 至 3 人。

(3) 閩南語語料庫：2 至 4 人。

(4) 國家語言辭典資料庫：1 至 2 人。

(5) 閩東語語料庫： 2 至 3 人。

(6) 臺灣手語語料庫： 4 至 5 人。

(四) 得標廠商需提供每位駐點人員工作使用之電腦相關設備，派駐人員電腦軟硬體規格如附件五。

(五) 廠商工作小組成員依廠商「工作計畫書」所載為準，如有變動，應以書面說明原因及替換人員學經歷條件，經本會同意後方可變更。

七、 專案工作及查核時程

(一) 本案期程自決標日翌日起依下表所定進度執行，工作期程以「日曆天」計，星期例假日、國定假日或其他休息日一併計入。廠商應依規定之工作時程及期限，完成各期工作項目，各期進度規劃如下表，如必要時，本部得視實際情形調整時限及共作項目：

年度	內容	工作項目
	1. 提送工作計畫書	決標日翌日起 20 日曆天內，提送書面工作計畫書（含工作進度）10 份及電子檔。（備註：完成後方可依據契約書第 5 條規定請領第 1 期款項）
第 1 期	2. 國家語言資料庫內容相關規範初步構想共識	廠商應於　年　月　日前，取得本案之顧問委員會針對國家語言資料庫內容及相關規範等初步構想之共識。
	3. 國家語言資料庫網站（含國家語言現況）	工作項目至少須包含： (1) 成立諮詢委員會 (2) 製作（含剪輯）各國家語言之介紹影片 (3) 完成資料庫架構及網頁 (4) 架設後台管理系統 (5) 完成上傳各國家語言介紹影片 (6) 後台管理系統擴充與維護 (7) 其他相關功能之設計與開發
	4. 閩南語語料庫	(1) 成立諮詢委員會 (2) 收集、校對、並整合現有口語及書面語語料至少 300 萬詞 (3) 訂定閩南語分詞原則及詞性標記集 (4) 訂定用字、詞性標記 (5) 訂定後設資料格式 (6) 完成 300 萬詞具有分詞及詞性標記的語料 (7) 利用前述語料和機器學習工具發展最初版本的閩南語分詞及詞性標記程式
	5. 閩東語語料庫	(1) 成立諮詢委員會 (2) 收集、校對、並整合現有口語及書面語語料至少 30 萬詞 (3) 訂定閩東語分詞原則及詞性標記集 (4) 訂定用字、詞性標記集

		(5) 訂定後設資料格式 (6) 完成 30 萬詞具有分詞及詞性標記的語料 (7) 利用前述語料和機器學習工具發展最初版本的閩東語分詞及詞性標記程式
	6. 臺灣手語語料庫	(1) 成立諮詢委員會 (2) 收集、校對、並整合至少 3 小時的臺灣手語多模態語料庫 (3) 領域及主題需儘量多元 (4) 需至少有五分之一包含兩位臺灣手語者的手語對話影片 (5) 訂定臺灣手語標記集 (6) 訂定後設資料格式 (7) 完成至少 3 小時具有標記的臺灣手語多模態語料庫
	7. 國家語言辭典資料庫	(1) 成立諮詢委員會 (2) 包含華語、閩南語（含閩南語各主要的方言）、客語（含四縣腔、海陸腔、大埔腔、饒平腔、詔安腔）、原住民族語（含各族語和方言別）、閩東語、臺灣手語的辭典 (3) 取得資料的授權 (4) 提供單一辭典或跨辭典的查詢功能服務。介面參考萌典的設計。需能支援以任何一種國家語言或華語檢索，且能檢索一個以上的國家語言 (5) 包含釋義、例句及影音檔 (6) 檢索模式支援精確比對和模糊比對 (7) 架設網站（含網頁設計）及後臺管理系統 (8) 其他相關項目
	8. 國家語言地圖及語言地理資料庫	蒐集、掃描並電子化過去學者所繪製的語言地圖，包括洪惟仁教授於 2019 年出版的兩冊專書《臺灣社會語言地理學研究：臺灣語言的分類與分區 I》及《臺灣語言地圖集 II》，卜溫仁（Warren A. Brewer）教授於 2008 年出版之《Mapping Taiwanese》及張屏生教授之詞彙相關研究 (1) 整理過去田野調查所蒐集到的各種腔調音檔，並製作成樣品音展示 (2) 確立後設資料格式 (3) 建立後設資料 (4) 取得語料授權

年度	內容	工作項目

第 2 期	1. 閩南語語料庫	(1) 收集、校對、並整合現有口語及書面語語料至少 350 萬詞 (2) 完成 350 萬詞具有分詞及詞性標記的語料 (3) 利用第一期 300 萬詞加上前述 350 萬詞語料共 650 萬詞語料以機器學習工具發展第二版本的閩南語分詞及詞性標記程式
	2. 閩東語語料庫	(1) 收集、校對、並整合現有口語及書面語語料至少 35 萬詞 (2) 完成 35 萬詞具有分詞及詞性標記的語料 (3) 利用第一期完成的 30 萬詞語料和第二期完成的 35 萬詞以機器學習工具發展第二版的閩東語分詞及詞性標記程式
	3. 臺灣手語語料庫	(1) 收集、校對、並整合至少 3.5 小時的臺灣手語多模態語料庫 (2) 領域及主題需儘量多元 (3) 需至少有五分之一包含兩位臺灣手語者的手語對話影片 (4) 完成 3.5 小時具有標記的臺灣手語多模態語料庫
	4. 國家語言地圖及語言地理資料庫	(1) 學研究成果與地理資訊系統（GIS）等技術結合，將樣品音以互動地圖的方式來呈現 (2) 地理資訊系統與方言田野調查的資料庫 (3) 逐步擴增地理語言學相關發音語料 (4) APP 以群眾外包的方式收集資料

年度	內容	工作項目
	1. 閩南語語料庫	(1) 收集、校對、並整合現有口語及書面語語料至少 350 萬詞 (2) 完成 350 萬詞具有分詞及詞性標記的語料 (3) 利用第一二三期共 1000 萬詞語料以機器學習工具發展第二版本的閩南語分詞及詞性標記程式 (4) 完成閩南語閩南語語料庫及檢索介面 (5) 開放一部份有分詞和詞性標記的閩南語語料提供民眾下載 (6) 開放閩南語分詞和詞性標記應用程式和程式碼供民眾下載及使用
	2. 閩東語語料庫	(1) 收集、校對、並整合現有口語及書面語語料至少 35 萬詞 (2) 完成 35 萬詞具有分詞及詞性標記的語料 (3) 利用第一二三期共 100 萬詞以機器學習工具發展第二版的閩東語分詞及詞性標記程式

第 3 期		(4) 完成閩東語語料庫及檢索介面 (5) 開放一部份有分詞和詞性標記的閩東語語料提供民眾下載 (6) 開放閩東語分詞和詞性標記應用程式和程式碼供民眾下載及使用
	3. 臺灣手語語料庫	(1) 收集、校對、並整合至少 3.5 小時的臺灣手語多模態語料庫 (2) 領域及主題需儘量多元 (3) 需至少有五分之一包含兩位臺灣手語者的手語對話影片 (4) 完成至少 3.5 小時具有標記的臺灣手語多模態語料庫 (5) 完成臺灣手語多模態語料庫檢索系統的介面與資料庫，並可以利用語意標記檢索影片內容

(二) 本案執行期間，廠商本案主持人或各工作團隊負責人須不定期至本部就工作方式、工作進度、網站架構、作業流程、程式設計、界面設計及資料結構等項目進行報告與溝通。

(三) 進度管理：

(1) 廠商須規劃設計工作進度報告表單，每月提出執行進度、成果內容與相關績效數據等資料，並依提送本部備查之進度辦理。另外，亦須負責辦理本部之各項臨時交辦事項。

(2) 工作進行中如發生可能影響工作進度之事故時，廠商應主動回報本部。

(3) 任一工作項目如發生落後預計進度之情況，廠商應主動向本部報告，並提出因應對策。

(四) 其他注意事項：

(1) 本部得不定期召開工作會議，以了解工作進度及處理相關需協調事宜。若因本案整體性規劃業務調整之需，廠商需配合本部修正工作之需求，並於約定時間內提出修正方案。

(2) 廠商設計應符合本部網路資訊系統之規劃。

八、 系統保固與維護

(一)保固期：本案自完成履約且經驗收合格日翌日起，由廠商提供一年免費保固服務、維護及技術諮詢，且提供保固維護計畫書。

(二)保固期內發現之瑕疵（包括故障、損壞、功能或效益不符合契約規定），由本部通知廠商改正。

(三)凡在保固期內發現瑕疵，廠商至少應於每日（含星期例假日）8 時至 18 時接受本部維修通知（書面或電子郵件方式）後，應由廠商於本部指定之期限內負責免費無條件改正。屆期不為改正者，本部得逕為處理，所需費用由廠商負擔，或動用保固保證金逕為處理，不足時向廠商追償。

(四)保固期內，採購標的因瑕疵致無法使用時，該無法使用之期間得不計入保固期。

(五)保固期滿，廠商得出具保固完成確認單通知本部已完成保固工作。

(六)保固維護期間內，得標廠商需提供下列服務：

(1) 維持系統正常操作之必要維護、正常操作中所發生缺點事項作必要之改善，進行維護須以不妨礙正常作業為原則。

(2) 當系統不能正常運作時，得標廠商於接獲通知後，須於 4 小時內進提出處理方案，並於本部之時程內完成修正。

(3) 保固期間若發現數位化瑕疵、影像掃瞄及資料登錄錯誤情形，得標廠商應於本部指定之期限內負責免費更正，逾期不為更正者，本會得逕為處理，所需費用由廠商負擔

(4) 保固期間廠商須指派專責技術工程人員負責維護諮詢之工作，提供必要之系統技術擴充諮詢支援。

(5) 廠商應依保固維護計畫書進行本案例行性維護作業（如資料庫維護或系統修補程式安裝）或其他異常處理時，應就該次維護之範圍及維護方式提供完整書面文件或電子紀錄，並由本會指定人員簽收確認。

(七)於本案執行及保固期間，如與其他廠商負責之部份相關，得標廠商應明確提供其他廠商或本會應配合之作業項目。如有爭議，本會有責任確認之解釋權。

(八)得標廠商於本案執行及保固期間內應注意系統之網路資訊安全，若因廠商疏失而導致發生影響本會資安之事件，本會得要求廠商賠償相關損失，得標廠商並應負相關之法律責任。

(九)經本會驗收合格後，廠商可將履約保證金之一部分計新臺幣 50 萬元整，轉為保固保證金，其餘履約保證金新臺幣 150 萬元整，於履約驗收合格且無待解決事項後 30 日內發還；至保固保證金，於保固期滿且無待解決事項後 30 日內發還。

九、 智慧財產權相關規定

(一)廠商同意本案之成果著作財產權，及依本採購標的計畫完成之一切著作財產權，於著作完成時，無償讓與本部。廠商應保證對於其職員、受僱人及受聘人職務上完成之著作，應依著作權法第 11 條第 1 項但書及第 12 條第 1 項但書規定，與其職員、受僱人及受聘人約定以廠商為著作人，享有著作人格權及著作財產權（如附件六）。

(二)廠商交付本部之本案相關文件、圖文影像與電子媒體等，其著作與智慧財產權均歸屬本部所有，如含有第三者開發之產品（或無法判斷是否為第三者之產品時），應保證（提供授權證明文件等）其內容（文字、圖、表、照片等）確屬可供合法使用之、無違背現行法規（包含符合中華民國著作權法規範）或侵害他人著作權及出版權等情事，若有侵害他人智慧財產權及第三人合法權益，致使本部遭致任何損失，或涉入其他權利爭議糾紛時，概由廠商負責處理，並承擔一切法律責任（含訴訟、律師費用及一切損害賠償）。

(三)本案設計相關之平面造型、立體造型、相關圖面、著作、資訊、成果、專門技術及營業秘密等（以下統稱為「相關資料」）之智慧財產權均歸本部所有。廠商除不得申請相關資料之任何專利權或其他智慧財產權之註冊登記外，本部要求為註冊登記時，廠商並須提供一切必要之協助，然註冊登記費用及因此產生之費用均由本部負擔。

(四)本案之新聞發布權歸屬本部。非經本部同意，廠商不得對外發布，否則本部得終止契約關係，得標廠商並應賠償本部得標價款總金額之 10 %。

十、 服務建議書撰寫規範

(一)提送格式：

(1) 以 A4 之紙張裝訂，由左至右中文直式橫寫（佐證資料可為英文），但相關之圖說得以 A3 之紙張製作。

(2) 封面請註明投標廠商名稱、標案名稱及提出日期，內頁須編製目錄，並於各頁下方中央加註頁碼。

(3) 不含封面、目錄及附件，以雙面印製不超過 50 頁為原則（A4 及 A3 雙面印製一張計二頁）。

(4) 印製 12 份。

(二)內容須包括下列各項：

(1) 專案概述：專案名稱、目標、內容、範圍等說明。

(2) 專案工作團隊：此部分內容包括本專案組織架構、人力及職掌、以及團隊合作模式、資源及未來配合方式說明。

(3) 專案管理規劃與分析說明：內容包括對本專案需求之規劃、執行與建議（依本需求說明書之各項需求逐項說明），以及對國家語言資料庫系統功能的規劃及後續經營管理計畫：

A. 摘要（就廠商對本案之整體瞭解之摘要說明）。

B. 國家語言資料庫語料蒐集方式與流程。

C. 國家語言資料庫內容與處理規範，包含下列預計採行之方法與內容：

i. 國家語言資料庫多重分類原則分類：屬性特徵及階層訂定（文類、文體、語式、主題等）。

ii. 斷詞原則訂定。

iii. 符合語言語法之特徵訂定。

iv. 用字、詞性標記訂定。

v. 後設資料格式訂定。

D. 系統開發與建置、工具、方法。

E. 作業需求。

F. 資訊安全。

G. 系統維護。

H. 保固與後續經營管理計畫。

I. 工作時程、完成期限、交付項目、績效管制說明等（請以甘特圖表示，詳細說明各項工作預定時程之規劃）。

J. 預想方案之可行性及預期效果評估。

(4) 經費分析表：詳列執行本案所需各項費用、成本等，分項詳細列表估算，且所有價格均應含稅。（參照經費明細表如附件七）

(5) 廠商經驗與能力

A. 廠商簡介。

B. 人力配置與工作職掌。

C. 工作小組人員專長及學經歷。

D. 詳述廠商之營運現況（需包含年承辦公私部門案件數與簡介）。

E. 承接類似專案經驗（詳述與本案類似之經驗，包括建置語料庫、系統軟硬體、應用系統、網際網路、網站建置及經營等之經驗及工作成果說明）及相關本部經驗（須於附錄檢附證明文件）等。

(6) 廠商得就有助於提升本專案效益之作為，但未列為本專案需求，主動額外提出補充或建議。此部分可另闢章節描述。

(7) 附錄

A. 預定採購之軟、硬體設備清單。

B. 工作小組人員履歷，及相關專業證照。

C. 技術支援廠商承諾書（無者免附）。

D. 廠商服務建議書與評分項目對照表（詳見下節）。

E. 其他證明文件。

(三)服務建議書之格式與內容不符規定者，評選委員得斟酌較低之評分，此部分請投標廠商注意。

十一、 評選決標事宜：

(一)本案依據行政院公共工程委員會發布「採購評選委員會組織準則」，成立評選委員會，並依「採購評選委員會審議規則」及「機關委託專業服務廠商評選及計費辦法」規定辦理，評選優勝廠商之作業，準用最有利標決標之評選規定。

(二)評選辦法：

(1) 本案先進行投標廠商資格審查，再邀請符合資格廠商就所提服務建議書進行簡報說明，由採購評選委員會進行評選。

(2) 符合資格者，由本計畫之採購評選委員會，擇期召開評選會議。評選時間、地點由本部以書面另行通知。

(3) 依投標廠商投標文件到達先後順序決定簡報次序。

(4) 簡報時間原則上以 15 分鐘為限，倘投標廠商達 3 家（含）以上，簡報時間以 12 分鐘為限（時間結束前 2 分鐘按鈴一短聲提醒，時間結束按鈴一長聲，即停止簡報。）。詢答時間以 10 分鐘（採統問統答方式，委員提問時間不計）為限（時間結束前 2 分鐘按鈴一短聲提醒，時間結束按鈴一長聲，即停止答復。）。

(5) 簡報時由本部提供 1 台單槍投影機與基本電源，廠商請自備筆電。

(6) 簡報人員必須包含本案之計畫主持人在內，計畫主持人未出席簡報者，評選委員得酌扣廠商簡報項目之得分。每一廠商至多得派 3 人進入會場簡報，若經 5 分鐘內唱名 3 次未到場簡報者，簡報部分予以零分計算，其他部分以書面審查。

(7) 簡報不得更改廠商投標文件內容，廠商另外提出變更或補充資料者，該資料不納入評選。

(三)評選項目及配分：

評選項目	內容	權重
1. 履約能力	專業資歷背景、人力架構、過往相關經驗	30%
2. 服務建議書內容之完整性及可行性	計畫內容及研究架構、方法、流程、步驟規劃	25%
3. 經費概算之合理性	經費運用情形與價格之合理性	20%
4. 計畫管理	計畫進度控管、計畫預期成果、計畫之周延性、可行性	15%
5. 簡報及答詢	簡報內容是否具體詳實、答覆有無中肯切題、掌握重點	10%
總分		100%

(四)優勝廠商評定方式：

(1)評選時，將就各評選項目分別評分後予以加總，依加總分數高低換算為序位，並彙整合計各廠商之序位，以合計值最低者為序位第 1 名，並經出席評選委員過半數同意後，評定各廠商序位名次，再簽報本部部長或其授權人員核定各廠商序位名次。

(2) 合格門檻：投標廠商平均分數達 75 分者為合格廠商，未達 75 分者不得列為優勝廠商。若無合格廠商時，主席宣布廢標，本案另行辦理。

(3) 優勝廠商僅為一家者，以議價方式辦理。優勝廠商為二家以上者，依序位第一者取得優先議價權，其次取得第二順位議價權，餘類推。但有二家以上廠商為同一序位者，以標價低者優先議價，若標價相同，即擇權重最高之評選項目之得分合計值較高者優先，得分仍相同者，抽籤決定之。

(4) 優勝廠商於議價完成後訂定委託契約。優勝廠商如因故無法完成議價程序或棄權者，本部得依序遞補。

(5) 其他評選注意事項：本部得因故終止評選事宜，通知投標廠商領回服務建議書。

(6) 採購評選委員自接獲評選有關資料之時起，不得就該採購案參加投標、作為投標廠商之分包廠商或擔任工作成員。其違反者，機關應不決標予該廠商。

(7) 投標廠商之服務建議書中所呈現之工作成員（依據行政院公共工程委員會 96 年 8 月 7 日工程企字第 09600302640 號函，工作成員範圍，包含投標廠商之投標文件所述人力組織及參與或協助該採購案之相關人員均屬之），如有不屬投標廠商之人員，須取得其同意書，投標時並一併附具同意書之影本，未檢附時，則視為該人員未同意擔任　　工作成員，機關得於評選會議要求

廠商說明，並請評選委員酌予扣減　　相關評選項目分數。

十二、　其他

本次招標由廠商所提供之相關內容如有任何侵犯他人智慧財產權之情事者，概由廠商負一切法律責任。

十三、　驗收

(一)廠商應依系統規格展示平臺功能，並提供系統測試（包含效能與壓力測試等）及與語料標記之正確性結果文件。

(1)測試環境需要在本部現有測試環境執行測試，廠商不得要求本部提供額外之軟硬體設備，以滿足本案效能測試或網站運行之需要，如有額外軟硬體設備之需要，廠商應自行採購，並納入本計畫預算內。

(2)廠商應自備效能及壓力測試工具，並事前提請本部審查同意後，以該工具執行效能測試。

(3)效能及壓力能測試結果，應能呈現本案開發之網站的系統處理能力，包括每分鐘最大可承受之使用者數、系統同時可處理資料筆數或交易數，系統使用者數、處理數或交易數超過系統處理能力時，將產生資料錯誤、處理異常或資訊安全漏洞。廠商應提供本部效能及壓力測試腳本(Test Scenario 或 Test Script)、測試個案、測試資料、測試紀錄與測試報告等。

十四、　結案報告

本計畫期程結束日之前提交完整的執行報告、報告項目、內容，及格式須經本部同意後撰寫。

拾、參考資料

(一)「建置國家語言資料庫先期規劃研究」勞務採購案需求說明書（附件九）。

拾、研擬各種授權書及授權機制草案

以下為參考國外相關授權書並諮詢協同計畫主持人翁聖賢律師後所提出的草案。

10.1. 臺灣國家語言資料庫之使用者條款草案

一、臺灣國家語言資料庫之使用者條款草案

(一)對臺灣國家語言資料庫的存取或使用，應受本使用條款與相關條件約束，且本使用條款應納入終端用戶許可協議之一部。

(二)當您開始使用臺灣國家語言資料庫所提供的任何服務，即視為您已接受本使用條款以及受其相關用戶許可協議。

二、定義

(一)本使用條款中使用的所有術語均應參照以下定義：

「用戶身份」：係指由臺灣國家語言資料庫所指定或授予的身份。

「授權使用」：係指由國家或政府單位所提供、或由臺灣國家語言資料庫指定或授予之可使用臺灣國家語言資料庫的身份。

「學術使用」：係指使用者符合國家或政府單位對學術用戶、學術人員所設置之標準或定義。

「非商業用途」：係指，藉由使用臺灣國家語言資料庫，並不會直接產生任何收入或商業利益，或非用於促進創造收入或商業利益之任何用途。

三、存取及使用臺灣國家語言資料庫之服務

(一)臺灣國家語言資料庫特此授予您一附有條件的、不可轉讓之許可，根據本條款及所擇用之用戶身份類別，您即享有授予存取及使用臺灣國家語言資料庫資料之權利。

(二)臺灣國家語言資料庫可依其自行認定，以限制您存取及使用資料庫中某些功能、及/或資料庫中之資訊、或子資料庫。

(三)您必須為對臺灣國家語言資料庫所進行之所有存取和使用、以及其所產生之後果負責。 臺灣國家語言資料庫保留取消或撤銷任何用戶身份之權利，恕不另行通知。

(四)用戶身份僅供其您本人或您其所屬團體使用，本條款禁止您允許(無論明示或默示允許)任何第三方藉其身份以存取及使用臺灣國家語言資料庫之服務。

(五)資料內容類別：

臺灣國家語言資料庫主要將內容分為三個類別：

公共內容（PUB）

學術內容（ACA）

受限內容（RES）

基於此分類和其相對應之用戶身份，您對資料庫中某些內容之存取和使用可能受到限制。

臺灣國家語言資料庫亦可能要求您接受學術內容和受限內容中其他許可或使用條款、或第三方要求之各種授權或限制條件(包含且不限於揭露用戶身份、研究目的、最終受益單位、贊助學術研究之單位等資訊)，您須同意上述條件後，方可進行使用內容。

(六)子類別

臺灣國家語言資料庫中提供的內容可能屬於以下類別標籤指示的某些子類別：

- 識別和訪問條件
  - ID：需要對用戶進行身份驗證或標識。
  - AFFIL = x：用戶需要隸屬於某個社區，例如，學術研究人員社區（x = EDU）或更廣泛的語言研究和技術研究人員社區（x = META）。
  - PERM：僅根據具體情況（例如強制性費用或研究計劃）授予用戶使用資料的權限。
  - FF：存取、使用該資料需要付費。
  - 計畫：用戶需有一項以上研究計畫以得到使用權限。

- 一般使用條件
  - BY：必須註明出處，即作者身份。
  - NC：內容僅可用於非商業目的。
  - INF：必須告知臺灣國家語言資料庫及/或授權人資料的使用目的和使用情況。
  - LOC：內容僅在單個位置、中心或站點上可用，亦即資料不得在雙方約定範圍以外之場所被重製或重現。
  - LRT：內容僅能適用或應用於語言研究及/或技術開發。
  - PRIV：該資料包括個人資料保護法所涵蓋之個人資料。

- 散佈限制

- NORED：不允許您重新散佈資料。
- DEP：不允許用戶重新散佈資料，但作為此規則的例外，您仍可以通過臺灣國家語言資料庫散佈修改後的版本。
- SA：允許在類似條件下重新散佈資料。
- ND：不允許您製作衍生任何著作或衍生作品，其包含原著作權的創作、作品之全部或一部。

- 其他條件
- *：授權許可中尚包含其他非標準約定與條件，須請您注意。

您須同意遵守這些要求。

(七)特定資訊授權約款

除上述類別外，某些內容亦有自身的許可條件（如知識共享許可），並可能會設置其他限制或要求。 您亦須同意遵守這些限制和要求，方得使用。

四、研究倫理

您同意遵守有關實務上各項研究倫理之典範，包括以尊重和專業對待共事者、涉及各方利益相關者、以及一般公眾，並應在有適用必要時將保密性與隱私性列入考量，亦應尊重各項文化差異(包含且不限因種族、族群、社經地位所造成之差異)，並與政府、公眾、私部門和其他出資者、贊助研究者建立開放且明確的關係。

五、所提供之服務其保證與責任

對於臺灣國家語言資料庫所提供之服務、軟體、程式或其他內容之可用性、及時性、安全性或可靠性，臺灣國家語言資料庫不承擔任何責任，並且保留隨時修改、暫停或終止服務的權利，恕不另行通知。

六、適用法律和完整協議

(一)本使用條款之準據法為中華民國法律，不考慮可能導致與其他司法管轄區適用之法律衝突之情形。若產生與臺灣國家語言資料庫服務相關或其引起的任何類型的爭議，則您同意由臺灣臺北地方法院為專屬管轄法院，惟臺灣國家語言資料庫有權得在任何管轄領域採用該另一司法管轄領域之法律，對任何不當使用行為採取強制執行措施(包含各項禁制令、保全措施)。

(二)本使用條款即為雙方之間就有關使用資料庫所達成之全部協議，並且取代之前所有書面或口頭之合意或協議。 倘若因任何原因，具管轄權之法院認為本條款之任何規定之一部或全分係不可執行，該規定之其餘部分仍具全部效力。

七、資料保護與隱私

您須同意遵守臺灣國家語言資料庫服務有關資訊安全、隱私安全、資料安全暨相關之保護措施與政策。

八、資料庫使用情況統計與限制自動化查詢影響統計

臺灣國家語言資料庫保有統計使用情況之權利，以衡量研究人員或其他終端使用者使用臺灣國家語言資料庫服務之情況。一方直接或間接地、或鼓勵他人使用臺灣國家語言資料庫，以致影響下載統計訊息和

其他使用統計訊息，皆視為違反臺灣國家語言資料庫之使用條款。臺灣國家語言資料庫保留限制使用、刪除內容與調整使用情況統計之權利，以因應任何可能出現之違反使用條款之情形發生。

九、使用條款之修正

(一)若因法律，行政命令或其他原因而須修改本使用條款，則臺灣國家語言資料庫將會在其網站上發布相關訊息等管道以告知您修正後之改變。

(二)倘若於收到通知或知悉有關使用條款修訂後，您仍繼續使用臺灣國家語言資料庫所提供之服務，則視為您同意更新版本之使用條款。

十、終止服務

如果您違反本使用條款之規定或其精神，或對臺灣國家語言資料庫帶來任何可能造成損害之風險，則臺灣國家語言資料庫保有停止提供全部或部份服務之權利，臺灣國家語言資料庫並將在您下一次使用臺灣國家語言資料庫服務時通知您相關終止事宜。

10.2. 臺灣國家語言資料庫之授權協議書草案

臺灣國家語言資料庫

提供及授權利用資料協議書

一、 讓與標的

_____（以下簡稱甲方）願將其所享有之_____（以下簡稱本資料）之著作財產權讓與給臺灣國家語言資料庫（以下簡稱乙

方），並依據本協議書之規定一併授與乙方本資料相關之使用權、利用權、及/或散布權（Rights of Distribution）等之相關權利。

二、資料之交付與核可

甲方應以規範及規格中定義之電子、或其他電磁紀錄形式將其所有 之資料交付給乙方。乙方於收到甲方所交付之資料後，須在合理之時間範圍內進行驗證；若是資料不符合規範及規格，則乙方得自行修正其所檢測到之錯誤，並得再次向甲方求取符合規範、規格之資源。

三、資料之維護與更新

甲方保有更新和維護資料之最終權利，惟若甲乙雙方未能就資料之維護達成共識時，乙方有權出於技術目的自行、或是僱用第三方維護及更新資料，甲方不得異議。乙方於本協議書終止後仍保有一切本協議書授權之使用、利用、維護及更新資料之相關權利。

四、報酬之交付

乙方為取得本協議書中有關資料之許可授權，茲此同意：

[　　]因甲方同意無償提供，無需支付授權金 。

[　　]向甲方支付_____新台幣(稅含)以作為一次性、非經常性授權金。

[　　]向甲方支付_____新台幣(稅含)以作為其他性質授權金。

五、權利瑕疵擔保

甲方擔保對本資料享有或擁有著作權、再授權許可權或其他任何得履行本協議書約定之授權許可範圍內進行授權許可之權利。甲方並擔

保，在符合本協議規範之條件下，對資料進行之任何使用方式（包含重製、散佈），無論任何形式均不至侵害任何第三方之著作權、或其他無形財產權利。倘若有第三方提出主張或通知，指稱本資料違反前述甲方之擔保，則乙方得自行認定前述主張或通知合理與否，將本資料自臺灣國家語言資料庫中刪除、更改或更新公開散佈形式，且所有因甲方違反其前述擔保所造成乙方之損害皆由甲方承擔責任。

六、資料之使用、利用及散佈權

甲方確認甲方擔保本資料之所有權、著作權、或其他形式之知識財產權，除本協議書已另有明文規定移轉或授權外，均屬於甲方或資料原所有人（如適用再授權之情形），並不因簽訂本協議書而有所影響。甲方並同意於著作權（或其他無體財產權）存續期間內，授與乙方非獨家、不可撤銷、得以重製或使他人重製、於不逾成為衍生著作、編輯著作之範圍內得以修改、得以提供予乙方之終端客戶以行使散佈權之權利，惟應限於學術、教育或研究等目的。此外，乙方若為知識共享，在不修改原作者姓名之情況下提供本資料，則其終端用戶得修改資料以供終端用戶個人或其所屬之研究小組使用，然其不得隨意散步修改過後之資料。

七、違約賠償責任

甲乙雙方各自對其因違反本協議書之規定所造成之損害承擔責任，惟此責任僅限於對他方所造成之直接損害，不包含間接損害。因故意侵權或重大過失所造成之損害責任或人身損害不適用於本款之賠償責任限制。

八、終端用戶之權利及義務

乙方應告知其終端用戶有關本協議書資料授權許可之規範條款，以及許可協議中與其相關之權利和義務。

九、聯絡窗口、通知和報告

雙方以書面或電子郵件形式發送至以下地址之關於本協議的通知或報告均應視為已有效送達：

甲方：

聯絡資訊：

地址：

電子信箱：

乙方：

聯絡資訊：

地址：

電子信箱：

雙方並均得在通知另一方後更改本協議中所定義之聯絡窗口或聯絡資訊。

十、協議之終止

(一)當甲乙其中一方嚴重違約(Material Breach)，且於收到他方書面通知改善後的三十日內未採取改正、糾正或補正措施，並完成去除違約狀況時，則另一方有權於發出終止書面通知後，立即終止本協議。

(二)若因甲方嚴重違約而導致本協議終止，則乙方有權在本協議終止後繼續依本協議書之約定使用本資料；若因乙方嚴重違約而導致本協議終止，則乙方必須終止所有對資源之任何形式之使用，並返還或刪除其所擁有之資料副本，或銷毀其電磁紀錄。

十一、協議之效力、終止與終止後之法律效果

本協議經雙方簽署後生效。但本協議書之以下條款，於本協議終止後仍然有效：

第三條《資料之維護與更新》

第五條《權利瑕疵擔保》

第六條《資料之使用、利用及散佈權》

第十四條《法律適用與爭端解決》

十二、協議之作成與修改

(一)本協議一式兩份，由雙方各執乙份為憑。

(二)本協議取代所有雙方先前曾就本協議所欲達成之協議目的，及所有口頭及/或書面合意、協議和理解。

(三)本協議若有任何修改，應由雙方協議另以書面為之，且任何修訂或修正均須經有代表權之雙方簽署或用印後方生效力。

(四)若本協議之任何條款在有關司法管轄領域中為非法、無效或不可執行，則並不影響該協議或本協議任何其他條款於該司法管轄領域之有效性或可執行性。

十三、法律適用與爭端解決

本協議書應依中華民國法律為準據法。

有關本協議之一切爭議或糾紛應統經由雙方之相互友好協商解決之。若雙方未能通過協商達成解決方案，則爭議應提交給台北地方法院進一步處理。

立契約人

甲方：

乙方：

中華民國　　年　　月　　日

10.3. 鼓勵授權措施

關於鼓勵措施，在此擬將鼓勵措施分為兩種情形：(1) 資料完全符合語料庫需求 (2) 資料不完全符合語料庫需求。於第一種情形時，可以協議書中第四條《報酬之交付》一部，約定雙方可接受之報酬金額，藉此取得資料所有人之授權。第二種情形，則建議文化部能夠另外擬定計畫，藉由補助計畫經費等，招募有意願的專家、學者等，協助官方額外蒐集或修改（其原所有之）語言資料後，再另簽訂協議以取得授權。

10.4. 國外語料庫授權方法參考

關於授權議題，目前在網路上可取得之資料，多為語料庫方授權予終端使用者此一方向，處理語料庫方如何向著作所有權人取得授權之相關議題的則很少。舉例而言，在美國著名的兩大語料庫 COHA、COCA，以及歐盟 CLARIN 計畫底下多數語料庫，皆無對外公開其取得

資料所有人的授權的細節，而僅提及其授權機制是受該國何項法律管轄等訊息。此外，學者 Mark Davies 在其關於建置語料庫的數個著作中，亦未對這些語料庫是如何處理授權議題的部分加以說明，僅提及了「需要簽訂契約」等訊息。

前述之授權說明書草案為參考 FIN-CLARIN 之授權協議書後擬訂而來。然而，即使可能仍有少數像這樣的可參考資料存在，考量到各國適用著作權/授權法律各不相同，能從他國的資料當中取得的可用資訊恐十分有限，還是需要我國法律專業人士一同參與擬定語料庫授權契約書之過程，方能得到最合適使用的授權書。因此，團隊在此建議，未來在執行建置語料庫的計畫時，應將擬定授權契約書之一環另立為一獨立計畫，招募了解著作權議題的法律人士，處理相關的授權議題、並用以擬定適合此計畫的授權契約書。

拾壹、參考文獻

Australian National Corpus Incorporated. (2012). Australian National Corpus. http://www.ausnc.org.au/

AWS Public Dataset Program. https://aws.amazon.com/opendata/public-datasets/

Becker, A., Catt, D., & Hochgesang, J. A. (2020). Back and forth between theory and application: Shared phonological coding Between ASL Signbank and ASL-LEX. In Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives (pp. 1-6).

Behrens, H. (2008). *Data maintenance.* In Behrens, H. (Ed.), *Corpora in language acquisition research: History, methods, perspectives*. John Benjamins Publishing.

Bird, S., Klein, E., & Loper, E. (2009). Managing Linguistic Data. In *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly.

Bird, S., & Simons, G. (2003). Extending Dublin Core metadata to support the description and discovery of language resources. *Computers and the Humanities*, *37*(4), 375-388.

Blust, R. (1999). Subgrouping, circularity and extinction: some issues in Austronesian comparative linguistics. *Selected papers from the eighth international conference on Austronesian linguistics*, *1*, 31-94.

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). http://www.natcorp.ox.ac.uk/

Brewer, W. A. (2008). *Mapping Taiwanese*. Institute of Linguistics, Academia Sinica.

Caselli, N. K., Sehyr, Z. S., Cohen-Goldberg, A. M., & Emmorey, K. (2017). ASL-LEX: A lexical database of American Sign Language. *Behavior research methods*, *49*(2), 784-801.

Cassidy, S. (2013, March). Interoperable Annotation in the Australian National Corpus. In *Proceedings of the 9th Joint ISO-ACL SIGSEM Workshop on Interoperable Semantic Annotation* (pp. 35-50).

Cassidy, S., Haugh, M., Peters, P., & Fallu, M. (2012, January). The Australian National Corpus: National Infrastructure for Language Resources. In *LREC* (pp. 3295-3299).

Coole, M., Rayson, P., & Mariani, J. (2016). lexiDB: A scalable corpus database management system. 2016 IEEE International Conference on Big Data, pp. 3880-3884). IEEE.

Crasborn, O., & Sloetjes, H. (2008). Enhanced ELAN functionality for sign language corpora. In *6th International Conference on Language Resources and Evaluation (LREC 2008)/3rd Workshop on the Representation and Processing of Sign Languages: Construction and Exploitation of Sign Language Corpora* (pp. 39-43).

Crasborn, O., Zwitserlood, I., van der Kooij, E., & Schüller, A. (2018). Global SignBank manual.

Fenlon, J., Cormier, K., & Schembri, A. (2015). Building BSL SignBank: The lemma dilemma revisited. *International Journal of Lexicography*, *28*(2), 169-206.

FIN-CLARIN | Kielipankki. https://www.kielipankki.fi/organisaatio/fin-clarin/

Geertz, C. (2002). An inconstant profession: The anthropological life in interesting times. *Annual review of anthropology*, *31*(1), 1-19.

Gippert, J., Meurer, P., & Tandashvili, M. (2012). Georgian National Corpus. http://gnc.gov.ge/gnc/page

Global SignBank. https://signbank.science.ru.nl

Hochgesang, J., Crasborn, O. A., & Lillo-Martin, D. (2018). Building the ASL Signbank. Lemmatization Principles for ASL.

Hormia-Poutanen, K., Kautonen, H., & Lassila, A. (2013). The Finnish National Digital Library: a national service is developed in collaboration with a network of libraries, archives and museums.

Howe, J. (2006). The rise of crowdsourcing. *Wired*, *14*(6), 1-4.

Hsieh, Shu-Kai. (2019). Corpus as Cultural Infrastructure. *The 2nd ILAS Annual Linguistics Forum—National Language Corpora: Design and Construction.* pp111-123.

Huang, C. R., Hsieh, S. K. & Chen, K. J. (2017). Mandarin Chinese words and parts of speech: A corpus-based study. Taylor & Francis.

Hull, D. A. (1997, March). Using structured queries for disambiguation in cross-language information retrieval. In *AAAI Symposium on Cross-Language Text and Speech Retrieval* (pp. 84-98).

Hungarian National Corpus. http://corpus.nytud.hu/mnsz/index_eng.html

Ide, N., & Suderman, K. (2002). Open American National Corpus | Open Data for Language Research and Education. http://www.anc.org/

Infocomm Media Development Authority (IMDA). (2018). National Speech Corpus (NSC). Retrieved from https://www2.imda.gov.sg/programme-listing/digital-services-lab/national-speech-corpus

Institute of Linguistics of the Faculty of Humanities and Social Sciences, University of Zagreb. (1998). Croatian National Corpus. https://web.archive.org/web/20060424031437/http://hnk.ffzg.hr/

Institute of the Czech National Corpus (ICNC) in the Faculty of Arts,

Charles University. (1994). Czech National Corpus. https://ucnk.ff.cuni.cz/cs/

Johnston, T. (2009). Creating a corpus of Auslan within an Australian national corpus. In *Selected Proceedings of the 2008 HCSNet Workshop on Designing the Australian National Corpus: Mustering Languages.*

Johnston, T. & De Beuzeville, L. (2014) Auslan corpus annotation guidelines.

Kim, H. (2006). Korean national corpus in the 21st century Sejong project. In *Proceedings of the 13th NIJL International Symposium* (pp. 49-54). National Institute for Japanese Language Tokyo.

Lampert, A. (2009). Email in the Australian National Corpus. Haugh et al.(eds). *National Center for Sign Language and Gesture Resources (NCSLGR) Corpus*. Retrieved from https://www.bu.edu/asllrp/ncslgr-for-download/download-info.html

Li, Peng-Hsuan., Fu, Tsu-Jui., & Ma, Wei-Yun. (2020). Why Attention? Analyze BiLSTM Deficiency and Its Remedies in the Case of NER. AAAI 2020.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. (2017). The Spoken British National Corpus 2014.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. transcription format and programs. Psychology Press.

MacWhinney, B. (2017). Tools for analyzing talk part 1: The chat transcription format. https://talkbank.org/manuals/CHAT.pdf.

Maekawa, Kikuo. (2019). Some personal reflections on (national) corpora. *The 2nd ILAS Annual Linguistics Forum—National Language Corpora: Design and Construction.* pp2-21.

Magistry, P. (2013). Unsupervised Word Segmentation and Wordhood

Assessment: The case of Mandarin Chinese (Doctoral dissertation). Retrieved from https://hal.archives-ouvertes.fr/tel-01573561/document

McEnery, T., Love, R., & Brezina, V. (2017). Introduction: Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics, 22*(3), 311-318.

McEnery, T., & Ostler, N. (2000). A new agenda for corpus linguistics-working with all of the world's languages. *Literary and linguistic computing*, *15*(4), 403-420.

Meng, Y., Li, X., Sun, X., Han, Q., Yuan, A., & Li, J. (2019). Is word segmentation necessary for deep learning of Chinese representations?. arXiv preprint arXiv:1905.05526.

Ministry of Education and Culture (Finland). (2013). FINNA. https://www.kiwi.fi/display/Finna/In+English

Morozova, M., Rusakov, A., & Arkhangelskiy, T. (2012). Albanian National Corpus. albanian.web-corpora.net

Mozilla Foundation. (2017). Common Voice by Mozilla. https://voice.mozilla.org/

Munro, R., Bethard, S., Kuperman, V., Lai, V. T., Melnick, R., Potts, C., Schnoebelen, T. & Tily, H. (2010). Crowdsourcing and language studies: The new generation of linguistic data. In *NAACL Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk* (pp. 122-130). Association for Computational Linguistics.

Musgrave, S., & Haugh, M. (2020). The Australian National Corpus (and beyond). Retrieved from http://www.academia.edu/download/61655949/MusgraveHaugh_The_Australian_National_Corpus_-_pre-print_version20200101-105470-vp4esj.pdf

Nathan, D., & Austin, P. K. (2004). Reconceiving metadata: language

documentation through thick and thin. In Peter K. Austin (ed.) *Language Documentation and Description, Vol 2*, 179-187. London: SOAS.

National Corpus of Polish. http://nkjp.pl/index.php?page=0&lang=1

National Institute of the Korean Language Republic of Korea. https://www.korean.go.kr/front_eng/main.do

Ossetic National Corpus. http://corpus.ossetic-studies.org/search/index.php?interface_language=en

Öqvist, Z., Kankkonen, N. R., & Mesch, J. (2020). STS-korpus: A Sign Language Web Corpus Tool for Teaching and Public Use. In *Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives* (pp. 177-180).

Pacific and Regional Archive for Digital Sources in Endangered Cultures. http://www.paradisec.org.au/home.html

Peters, P. (2009). The architecture of a multipurpose Australian National Corpus. *Haugh et al.(eds).*

Ruan, J. C., Hsu, C. W., Myers, J., & Tsay, J. S. (2012, March). Development and Testing of Transcription Software for a Southern Min Spoken Corpus. *International Journal of Computational Linguistics & Chinese Language Processing*, *17*(1), 1-26.

Semantic Versioning. Retrieved from https://semver.org/

Semantic Web Research Center (SWRC). http://semanticweb.kaist.ac.kr/home/index.php/Home

Shibata, Takeshi (柴田武). 1969. *Gengo Chirigaku no Hoho 言語地理学の*

方法 [Methodology of Geolinguistics]. Tokyo: Chikuma Shobo.

Slovak National Corpus. https://korpus.sk/index_en.html

Smith, Wayne H. (2005). Taiwan Sign Language research: An historical overview. *Language and Linguistics*, *6*(2), 187-215.

"Swedish Sign Language Corpus Project" (Apr. 25, 2018). Retrieved from https://www.ling.su.se/english/research/research-projects/sign-language/swedish-sign-language-corpus-project-1.59270

Tai, J. H. Y., & Tsay, J. S. (2015). Taiwan sign language: history, structure, and adaptation. In Wang and Sun (Eds.), *The Oxford hanbook of Chinese Linguistics* (pp. 729-750).

Tatar National Corpus. http://tugantel.tatar/?lang=en

Thai National Corpus. http://www.arts.chula.ac.th/ling/tnc/

The Abkhaz National Corpus. http://clarino.uib.no/abnc/page

The Auslan Corpus Project. (2011). Retrieved from https://www.latrobe.edu.au/news/articles/2011/article/the-auslan-corpus-project

The Balanced Corpus of Contemporary Written Japanese (BCCWJ). https://pj.ninjal.ac.jp/corpus_center/bccwj/en/

The CorCenCC project team. (2016). CorCenCC – National Corpus of Contemporary Welsh. http://www.corcencc.org/

The Institute for Bulgarian Language. (2001). *Bulgarian National Corpus*. https://dcl.bas.bg/bulnc/en/

The Institute for Language and Speech Processing (ILSP / "Athena" R.C.). (2019). Hellenic National Corpus. Retrieved from http://hnc.ilsp.gr/

The Institute of Russian language, Russian Academy of Sciences. (2004). Russian National Corpus. http://ruscorpora.ru/en/

Thieberger, N. (2010). Anxious Respect for Linguistic Data: The Pacific and

Regional Archive for Digital Sources in Endangered Cultures (PARADISEC) and the Resource Network for Linguistic Diversity (RNLD). In Florey, M (ed). (2010). *Endangered Languages of Austronesia* (pp. 141-158). Oxford University Press.

Thieberger, N. (2014). Digital humanities and language documentation. In Gawne, L. and Vaughan, J. (eds), *Selected Papers from the 44th Conference of the Australian Linguistic Society, 2013* (pp. 144–159). Melbourne: University of Melbourne. Retrieved from http://hdl.handle.net/11343/40961

Tsay, J. S. (2007). Construction and automatization of a Minnan child speech corpus with some research findings. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 12, Number 4, December 2007: Special Issue on Speech and Language Processing for Taiwanese Minnan, Hakka, and Mandarin* (pp. 411-442).

Tsay, J. S. (2014). A phonological corpus of L1 acquisition of Taiwan Southern Min. In Durand, J., Gut, U. & Kristoffersen, G. (eds), *The Oxford handbook of corpus phonology* (pp. 576-587). Oxford University Press.

Turkish National Corpus (TNC). https://www.tnc.org.tr/

UNESCO Atlas of the World's Languages in danger. http://www.unesco.org/languages-atlas/

Wallin, L. & Mesch, J. (2015). Swedish sign language corpus. In *Proceedings of digging into signs workshop: Developing annotation standards for sign language corpora.*

Xia, Fei. (2000). The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). *IRCS Technical Reports Series*, 38.

10901 現住原住民人口數按族別及年齡統計表.htm。取自

https://www.apc.gov.tw/portal/getfile?source=2D838540F5D6F659FAFB9859EF31AC3B381A272F479D65D98D902DFAAFC2E1543725230652686A55FD98C7F142FDF378533605ECE7154E6F32B33F5ADCFBD6EA&filename=0DA4D4B5AFC8D6DDD683E9859EEB6509C7ECAD8DC33A153B3AFF45884251450BA7E7358D0DAF914F9F57F98A9CE66E09

99 年人口及住宅普查初步統計結果提要分析 - 行政院主計總處。取自 https://www.dgbas.gov.tw/ct.asp?xItem=30077&ctNode=3272

九年一貫台語教學資源網。http://www.taiwanwe.com.tw/

公共電視教育影音公播網。http://ptsvod.sunnystudy.com.tw/

中央研究院 iCorpus 臺華平行新聞語料庫。http://icorpus.iis.sinica.edu.tw/

中央研究院中文詞彙特性速描系統。http://wordsketch.ling.sinica.edu.tw/

中央研究院漢語平衡語料庫。http://asbc.iis.sinica.edu.tw/

中央標準局（1998）。中文資訊處理分詞規範調查研究計畫。取自 http://sighan.cs.uchicago.edu/bakeoff2005/data/as_spec.pdf

台大臺灣南島語多媒體語料庫。http://203.66.168.190/

台語文記憶。http://ip194097.ntcu.edu.tw/memory/TGB/mowt.asp

台語文語詞檢索。http://ip194097.ntcu.edu.tw/TG/concordance/form.asp

臺灣民間文學館。http://cls.lib.ntu.edu.tw/TFL2010/

臺灣南島語數位典藏。https://museum02.digitalarchives.tw/ndap/2001/AustronesianLang/formosan.sinica.edu.tw/m/index.html

臺灣南島語數位典藏- 人文社會資料庫名錄檢索。http://husscat.hss.ntu.edu.tw/xmlui/handle/123456789/7709

臺灣現當代作家研究資料庫。http://cw.nmtl.gov.tw/

本土語言調查報告- 本土語言資源網。
https://mhi.moe.edu.tw/newsList.jsp?ID=5
用圖表帶你看母語斷層危機- 參拾母語。
https://hakkafa.hakkatv.org.tw/hakkafa-infographic
好客 ING-客家影音網路平台。https://broadcasting.hakka.gov.tw/
全國客家人口暨語言基礎資料| 客家委員會全球資訊網。
https://www.hakka.gov.tw/Content/Content?NodeID=626&PageID=37585
回應《身心障礙者權利公約》首度國家報告審查會議結論性意見。（2017 年 11 月 9 日）。取自 http://20.enable.org.tw/news/detail.php?id=14504
李壬癸、章英華、林季平、劉彩秀（2015）。**族語保存現況調查研究計畫成果報告**。科技部委託研究計畫（編號：NSC 101－2410－H－001－094－MY3），中央研究院語言研究所執行。取自 http://dx.doi.org/10.1080/01434632.2015.1022179
李信賢（2016）。**比較臺灣手語與中文手語的音韻系統**（博士論文）。國立中正大學語言學研究所，嘉義。
吳中杰（2013）。**臺灣南部客家語言使用態度與使用行為研究—屏東市**。客委會補助研究計畫（編號：PG10205-0110）。取自 https://www.grb.gov.tw/search/planDetail?id=2945780
林宜靜（2018 年 1 月 12 日）。百冊看文學！「臺灣現當代作家研究資料彙編計畫」成果發表！。 **中時電子報**。取自 https://www.chinatimes.com/realtimenews/20180112000991-260405?chdtv&fbclid=IwAR1QSWcg56eN5i1cwYe6T86k0YMs71gk8LSp8Zb287Jk5Nax5VGk_M4LqUs
国立国語研究所。https://www.ninjal.ac.jp/

客客客棧-「參。拾母語」特輯(華語版)。
https://www.youtube.com/watch?v=3Td_jldDW44
客英大辭典查詢。http://minhakka.ling.sinica.edu.tw/bkg/hakyin/
客語認證詞彙資料庫 - 客家委員會。https://wiki.hakka.gov.tw/
客家委員會客語認證詞彙資料庫。https://wiki.hakka.gov.tw/
客家音樂 - 哈客網路學院 - 客家委員會。
https://elearning.hakka.gov.tw/ver2015/allclass/default.aspx?group=20000004
客家歌謠- 臺北市客家文化主題公園。
https://ssl.thcp.org.tw/libraries/songs?page=9
客家語部編版教育資源 - 國家教育研究院。
http://hakka.naer.edu.tw/hakka/
语料库在线。http://corpus.zhonghuayuwen.org/
「咱來學臺灣閩南語」學習手冊 - 教育部語文成果網。
https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=506
洪惟仁（2002）。**台北褒歌之美**。宜蘭縣：國立傳統藝術中心。
洪惟仁（2019）。**臺灣社會語言地理學研究：臺灣語言的分類與分區Ⅰ＋臺灣語言地圖集**II（**二冊套書**）。台北市：前衛。
看影片學客語(行動學習) - 哈客網路學院 - 客家委員會。
https://elearning.hakka.gov.tw/ver2015/allclass/default.aspx?group=g000000015
原住民族語文學創作作品集 - 教育部語文成果網。
https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=480
原住民族語言能力認證測驗-資源下載。http://lokahsu.org.tw/resource/

原住民族語言調查研究三年實施計畫 16 族綜合比較報告。取自 http://hanzi.cmex.ericjoung.idv.tw/file/files/1050601-1-%E5%8E%9F%E4%BD%8F%E6%B0%91%E6%97%8F%E8%AA%9E%E8%A8%80%E8%AA%BF%E6%9F%A5%E7%A0%94%E7%A9%B6%E4%B8%89%E5%B9%B4%E5%AF%A6%E6%96%BD%E8%A8%88%E7%95%AB%E7%AC%AC3%E6%9C%9F%E5%AF%A6%E6%96%BD%E8%A8%88%E7%95%AB_1%E8%87%B33%E6%9C%9F16%E6%97%8F%E7%B6%9C%E5%90%88%E6%AF%94%E8%BC%83%E5%A0%B1%E5%91%8A%E6%91%98%E8%A6%81%E5%BD%99%E7%B7%A8_(%E5%85%AC%E5%91%8A).pdf

原住民族語言線上詞典。https://m-dictionary.apc.gov.tw/

陳光華（1998）。**超越資訊檢索的語言藩籬**。臺北：大學圖書館。取自 https://www.lis.ntu.edu.tw/~khchen/writtings/pdf/ulq1998.pdf

族語 e 樂園。http://web.klokah.tw/

族語千詞表-16 族族語千詞表 - 原住民族語言研究發展中心。http://ilrdc.tw/research/athousand/area16.php

張屏生教授 - 國立中山大學中國文學系。http://www.chinese.nsysu.edu.tw/zh_tw/Department_introduction/Teacher/%E5%BC%B5-%E5%B1%8F%E7%94%9F-2809422

張榮興手語。http://signlanguage.ccu.edu.tw/

教育部悅讀越懂閩客語電子報。https://epaper.edu.tw/learning.aspx?classify_sn=6

教育部臺灣客家語常用辭典。https://hakkadict.moe.edu.tw/

教育部臺灣閩南語常用詞辭典。https://twblg.dict.edu.tw/holodict_new/default.jsp

教育部閩南語動畫。https://twbangga.moe.edu.tw/info

國立中正大學臺灣閩南語口語語料庫。http://lngproc.ccu.edu.tw/Corpus/

國立臺灣文學館-臺灣民間說唱文學歌仔冊資料庫。http://koaachheh.nmtl.gov.tw/bang-cham/thau-iah.php

國家教育研究院（2016）。**韓國國家編譯暨教科書發展考察報告**。取自 https://report.nat.gov.tw/ReportFront/PageSystem/reportFileDownload/C10404200/001

國家教育研究語料庫索引典系統。https://coct.naer.edu.tw/cqpweb/

國家電影中心-台語片 60 週年。http://taiyupian60th.weebly.com/

國家語言 111 年列國高中部定課程 含手語閩東語（2020 年 3 月 18 日）。**中央通訊社**。取自 https://www.cna.com.tw/news/ahel/202003180450.aspx

國家語言發展法相關法令研究（2017 年 5 月 3 日）。**文化部「國家語言發展法之研究與規劃」公聽會會議紀錄**。取自 https://www.facebook.com/NationalLanguageDevelopmentAct/posts/1909814259299522

國家語言發展法摘要說明（n.d.）。**文化部**。取自 https://www.moc.gov.tw/content_275.html

童話大師安徒生會說、也會寫客語了（2019 年 1 月 1 日）。**國立中央大學客家學院電子報**。取自 http://hakka.ncu.edu.tw/hakka/modules/tinycontent/content/paper/paper316/316(13).html

傅仰止、章英華、杜素豪、廖培珊（2014）。**臺灣社會變遷基本調查計畫第六期第四次調查計畫執行報告**。臺北：中央研究院社會學研究所。

程俊源、方耀乾（2017）。**全國語言基礎資料研究計畫**。文化部委託研究計畫（編號：PG10602-0036）。取自 https://www.grb.gov.tw/search/planDetail?id=12068997

曾金金（1997）。**臺灣文學出版物收集、目錄、選讀編輯計畫結案報**

告說明（45-72 頁）。行政院文化建設委員會。

曾淑娟 Shu-Chuan TSENG - 中央研究院語言學研究所。http://www.ling.sinica.edu.tw/v3-3-1.asp-auserid=20.htm

詞庫小組（1995）。**中央研究院平衡語料庫的內容與說明技術報告**（95-102 頁）。台北：中央研究院資訊科學研究所。取自 http://asbc.iis.sinica.edu.tw/.

楊允言、戴嘉宏、劉杰岳、陳克健、高成炎（2008）。利用統計方法及中文訓練資料處理台語文詞性標記。**第二十屆自然語言與語音處理研討會論文集**（166-179 頁）。台北：台師大資訊系。

葉高華（2018）。臺灣歷次語言普查回顧。**臺灣語文研究，13(2)**，247-273。

當畫新聞打手語 服務聽障"看"新聞（2008 年 8 月 3 日）。**公視聽聽看**。取自 http://fruit.pts.org.tw/~seehear/news/801-850/news-805.htm

說台語的《小王子》！法文譯者蔡雅菁的母語之旅（2020 年 3 月 6 日）。**島語見聞**。取自 https://islandnewstw.com/archives/6432

閩客語典藏。http://minhakka.ling.sinica.edu.tw/bkg/bkg.php?gi_gian=hoa

語料庫建置入門工作流程指南（2010 年）。數位典藏與數位學習國家型科技計畫。取自 https://books.google.com.tw/books/about/語料庫建置入門數位化工作流.html?id=qshH2fT41vsC&redir_esc=y

臺灣手語線上辭典。http://tsl.ccu.edu.tw/web/browser.htm

臺灣手語電子資料庫。http://signlanguage.ccu.edu.tw/index.php

臺灣白話字文獻館。http://pojbh.lib.ntnu.edu.tw/

臺灣本土語言文學獎作品集 - 教育部語文成果網。https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=481

臺灣音聲一百年。https://audio.nmth.gov.tw/

臺灣閩南語我嘛會每日一詞。

https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=494&content_sn=4

臺灣閩南語推薦用字 700 字詞。

https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=45

臺灣閩南語漢字之選用原則。

https://language.moe.gov.tw/result.aspx?classify_sn=23&subclassify_sn=439&content_sn=15

臺灣閩南語羅馬字拼音方案。

https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=446

蔡素娟、麥傑、戴浩一（2009）。**語料庫為本的兒童閩南語詞彙研究—音韻與詞類**。行政院國家科學委員會學術補助計畫（編號：NSC98-2410-H194-086-MY3）。取自 https://www.grb.gov.tw/search/planDetail?id=1875197

影音平台-母語巢- 臺北市原住民語言學習網。

http://taipei.pqwasan.org.tw/video/

噶瑪蘭語詞典 -《語言暨語言學》LANGUAGE AND LINGUISTICS。

取自 http://www.ling.sinica.edu.tw/LL/zh/monographs.Contect/%E5%99%B6%E7%91%AA%E8%98%AD%E8%AA%9E%E8%A9%9E%E5%85%B8_

學習資源- 本土語言資源網。https://mhi.moe.edu.tw/infoList.jsp?ID=2

蘭嶼達悟語口語資料典藏網。http://yamiproject.cs.pu.edu.tw/

鄧守信（2010）。**對外漢語教學語法**。臺北：文鶴出版有限公司。

戴浩一 & 蔡素娟. (2009). 手語的本質: 以臺灣手語為例. 語言與認知 (126-176 頁). 臺北: 國立臺灣大學出版中心. [Tai, Hao-Yi, & Tsay, Su-Jane. (2009). The essence of sign language: In case of Taiwan sign language. In I-Wen Su & Yung-O Biq (Eds.), Language and Cognition, 126-176.]

【藝術文化】大人囝仔上深的感動 經典文學小王子台語有聲版問世（2020 年 3 月 30 日）。**自由時報**。取自 https://ent.ltn.com.tw/news/paper/1362198

聽唱兒歌- 文化部-兒童文化館。https://children.moc.gov.tw/song_list?language=2

聽障的人口有多少？（2016）。**中華民國聾人協會**。https://www.nad.org.tw/old-www/

附錄一、　洪惟仁教授專文撰稿 – 臺灣的語種分布與分區

臺灣的語種分佈與分區

洪惟仁

國立台中教育大學退休教授

1. 緣由與本文焦點

語言地圖有兩種，一種是變體分佈圖，一種是語言分區圖。我把前者的地圖繪製歸入「地理語言學」的研究範疇，而把後者歸入「語言地理學」的範疇。世界上的語言地圖大部分是變體分佈圖，但中國、臺灣都是先繪製語言分區圖，再繪製變體分佈圖。其中原因頗耐人尋味。

臺灣第一張語言地圖，應該算是小川尚義所繪製的〈臺灣言語分布圖〉(【附圖/圖 49】)，附錄於《日臺大辭典》(1907)。這張地圖是語言分區圖，為臺灣所有的語言進行分類與分區。從這張地圖算起，臺灣的語言地理學至今 2020 年已經有 113 年的歷史了。

第二張語言地圖是小川尚義與淺井惠倫合著的《原語による臺灣高砂族傳說集》(1935)中所附〈臺灣高砂族言語分布圖〉，詳細標示著平埔社及高砂族部落的位置。

這兩張地圖可以標示著日治時代臺灣語言地理學開拓性的成就。接著，經過皇民化及蔣家大中華民族主義嚴峻的統治，臺灣的語言地理學沈寂了半個世紀以上。

蔣家時期根據語言調查繪製地圖唯一的成果是恩師鍾露昇教授在

《閩南語在臺灣的分佈》(1967)所附27張變體分佈圖。這是臺灣方言變體分佈圖的濫觴。

筆者承襲恩師的志業，自1985年起進行了遍及全臺灣的閩南語方言掃瞄調查。1988年參加龔煌城教授主持的「臺灣地區漢語方言調查研究計畫」，正式開始臺灣全局的閩南語方言調查，但是至今未出版地圖集，唯一出版的一本《臺灣方言之旅》(洪惟仁1992)是作者有關臺灣閩南語方言分佈的一個簡要報告，書末附錄的〈臺灣的漢語方言分佈圖〉是戰後第一張根據實際的調查資料繪製的臺灣漢語方言分區圖。

2005年起本人開始進行全臺灣所有語言的分佈調查。14年後的2019年出版了《臺灣語言地圖集》(*Language Atlas of Taiwan*，簡稱*LAT*，前衛出版社)。這是臺灣有史以來第一部語言地圖集。一般全國性的語言地圖集通常是傾全國語言學家組成團隊繪製而成，但*LAT*從頭到尾由語言調查、地圖繪製等全部過程花了34年的時間，由筆者一人領導的團隊一貫完成。

這部地圖集共收105張語種分佈地圖，分全圖、語言區圖、行政區圖，巨細靡遺。其精細度達村里以下的自然村，所有語言小片都交代了該語言區的語言特色，是世界上最精細的語言地圖。但這部著作是站在語言地理學分類的立場立論，沒有考慮政治問題。比如金門講閩南話、烏坵講莆田話、馬祖講福州話，雖然政治上她們屬於中華民國的語言，但語言學上只能歸入福建各個閩語方言區內，將來我們繪製福建地區的方言地圖會把她們畫進去，但在本地圖集只能從略，因此文末【附圖】中找不到這些語言。

本文因篇幅所限，擬聚焦於臺灣語言全圖，鳥瞰臺灣所有語種的分類、分區與分佈。至於分類、分區與分佈的細節只好請讀者查閱*LAT*了。如果要進一步了解分類、分區的理論請看其孿生篇《臺灣語言的

分類與分區：理論與方法》（前衛出版社，2019）。

2. 臺灣的語種及分類

眾所週知，臺灣的語言有漢語及南島語兩大語系。漢語包括閩南語、客語、華語共3種；南島語是臺灣原住民的語言。原住民族委員會認定的原住民族有 16 族，不過其中有些是同一語族的方言群，如太魯閣語其實是賽德克語的方言，被當成兩個語族；有的是瀕危的南島語，但因為被當成是平埔族，也不認定是原住民。所以原民會的「族」不等同於「語」。我們從語言學觀點，認定臺灣的南島語共 15 種。

臺灣還有一種特殊的語種「日語客里諣」（Japanese-based Creole）。其民族原屬於賽德克陶賽族（Tausa，即都達 Toda 的訛音），因逃避太魯閣的壓迫，遷居宜蘭南澳鄉，由於和混雜的泰雅語 Squliq、Ts'oli'等方言溝通困難，居民採用日語為共通語，因而創造了一種混合語，以日語為基礎，融合了底層的泰雅語或賽德克語，也吸收了一些閩南語和華語詞彙，成為臺灣特有種的「客里諣（克里歐; creole）」，留傳下來成為居民的母語。

LAT 根據全盤性、系統性、一致性的分類原則對全臺灣的語種進行階層性分類。漢語、南島語各分為三個階層，最底層的方言或次方言往往只根據一兩個方言特徵加以分類、分區。但以下只針對漢語及南島語的大分類，進行宏觀的鳥瞰。

3. 臺灣各語種的分佈大勢

按照臺灣各語種的分佈狀態進行語言區劃，可以發現不同語言區有不同的語言分佈特色。茲以《臺灣語言地圖集》圖 A1（參見【附圖】）為例，鳥瞰臺灣所有語種的分佈狀態：

(1) 南島語

南島語北由新北市烏來區綿延至臺灣南端屏東縣牡丹鄉的雪山山脈和中央山脈等高山地帶成連續性分佈，達悟語分佈在蘭嶼，做為其「飛地」，形成 *LAT* 所謂的「南島語州」；另外卑南語、阿美語分佈在花東海岸山脈兩側與漢語混雜分佈，劃入「花東州」的一部分。南島語有 15 種語言，但有些語言人口極少，散步在其他語言區之中，佔有村里以上特定分佈區的語言只有 10 種。

(2) 客家話

客語五大方言中四縣腔、海陸腔、大埔腔、饒平腔集中分佈在北自桃園縣觀音鄉綿延至南投縣的丘陵地帶，包括桃竹苗丘陵、台中東部丘陵，延伸至南投國姓鄉，成連續性分佈，形成所謂的「客語州」；另外在他州內呈語言島或群島形態分佈。中閩州內的雲林縣二崙、崙背是一個詔安客語島；南閩州高屏地區的「六堆」四縣腔客語呈群島分佈。混雜的客語零星散佈在花東縱谷，與閩南語、阿美語雜處。

(3) 閩南語

閩南語是臺灣最重要的語言，人口佔七成以上，三大方言（漳腔、泉腔、混合腔）分佈在所有的平地：西部由桃園新屋的蚵殼港往南綿延至臺灣南端的屏東平原、恆春平原以及宜蘭平原、桃園台地以北；所有的海岸：東北及北部海岸、桃竹苗海岸、花蓮海岸平原、台東海岸平原；大部分的盆地：台北盆地、南投埔里盆地、花東縱谷；大部分的島嶼：澎湖群島、小琉球、綠島.....等。幾乎所有海拔最低，最容易開發，交通最發達的平原、海岸、盆地、島嶼都是閩南語的分佈區。因分佈區廣大，*LAT* 區劃為北閩州、中閩州、南閩州、澎湖州四個閩南語州，而花東州亦以閩南語為主。

(4) 華語語言島

華語雖然是臺灣最強勢的語言，散佈在整個臺灣，但不定在而無所不在。具高度聚集性的最大範圍通常不及村里，沒有一個華語語言島達到市鄉鎮的規模，且所有的華語語言島都與當地語言形成華語優勢的雙語區。進入二十一世紀以後，隨著眷村改建為國宅，所謂「軍宅」一個一個失去了獨立的封閉性空間，即將走向消失的命運。

(5) 日語客里謳

散佈在宜蘭縣南澳鄉的花澳村、金洋村博愛巷、東岳村及其相鄰的大同鄉寒溪村。隨著年輕人逐漸改說華語或泰雅語，日語客里謳也瀕臨消失。

4. 結論

以上五個主要語種，能夠形成大面積連續性分佈的只有南島語、客語和閩南語。其分佈格局大約和地形與海拔相關。

閩南語分佈在低海拔的平原、海岸、島嶼，分佈最為廣闊。主要分佈在四個閩南語州，但也深入各語言州。客語除六堆客語群島包孕在南閩州內、詔安客語島包孕在中閩州內之外，主要分佈在低、中海拔的丘陵地帶；南島語主要分佈在島中央高海拔的山地，唯達悟語孤懸蘭嶼島，阿美語、卑南語與閩南語、客語混雜於東部海岸山脈兩側的濱海地帶及花東縱谷。

本文宏觀地鳥瞰臺灣的語言。金門講閩南話、烏坵講莆田話、馬祖講福州話，但【附圖】未能呈現。

圖 49.【附圖】臺灣語言分佈全圖（資料來源：引自《臺灣語言地圖集》圖 A1，由洪惟仁教授提供）

作者：洪惟仁

國立台中教育大學退休教授

附錄二、　張永利教授專文撰稿 – 臺灣原住民族語言簡介

臺灣原住民語言簡介

張永利

中央研究院語言學研究所研究員

Updated 19 May 2020

臺灣原住民語言，早期有二十幾種，現今僅存十幾種，官方認定的有十六個，分別是泰雅語、賽夏語、賽德克語、邵語、布農語、鄒語、卡那卡那富語、拉阿魯阿語、魯凱語、排灣語、卑南語、阿美語、撒奇萊雅語、太魯閣語、噶瑪蘭語和雅美語（達悟語）；其中太魯閣語和賽德克語、撒奇萊雅語和阿美語關係密切，骨肉相連：前一對共同擁有特殊的融合依附代詞 saku / maku，後一對則共享特殊的第一人稱主格代詞 kako / kaku。除孤懸於海外的達悟語之外，其餘十五個原住民語言主要通行於西部山區或山腳以及花東縱谷或海岸平原，其具體分布請參見圖 50。曾經通行於西部平原的原住民語言，即平埔族語，多半都已消失，巴宰語或其近親噶哈巫語仍有少數耆老對其隻字片語保有記憶，西拉雅語則恐已被歷史洪流吞沒，獨留部分後代子孫力圖復振，積極爭取正名，對凱達格蘭語，人們大概也只能從總統府前的凱達格蘭大道的名稱來追思憑弔。

圖 50. 原住民族分佈圖（取自中華民國原住民知識經濟發展協會網頁：http://www.twedance.org/aboriginal00.aspx）

官方認定的十六族，人口總數約為五十七萬人，佔臺灣人口約百分之二點四左右，是不折不扣的少數民族。為了升學和就業，許多原住民離開原鄉，日常多已不使用族語，族語正如水管漏水一般，一點一點流失。

臺灣原住民語言，學界通稱為臺灣南島語言，與新南向的菲律賓語言和印尼、馬來語同屬南島語系。臺灣南島語言彼此關係親近，這一點也可從其族名略窺一二：邵語 (Thao)、鄒語 (Tsou) 和達悟語 (Tau) 的族名，其實是系出同門，皆源自原始南島語言的 tsau，意為「人」；臺灣南島語言和其他南島語言的親屬關係也是相當明顯，例如「眼睛」一詞，排灣語為 matsa、賽夏語為 masa'、塔加洛語和印尼語/馬來語皆為 mata，具體而微地反映出這些語言之間親屬同源關係，同時也清楚顯示臺灣南島語言保留祖語的優良特質—排灣語的 matsa 保留原始南島語的語音形式，而塔加洛語和印尼語/馬來語的 mata 則經歷了某種語音

演變 (/ts/ > /t/) ；類似的現象也可以從數字「四」一詞的體現可以看得出來，排灣語為比較存古的形式 sepatj、但塔加洛語和印尼馬來語都是比較簡化的形式 pat—臺灣南島語言保存了原始南島語的/S/，而塔加洛語和印尼馬來語都已經把這個祖語的特徵丟失了。

臺灣南島語言的存古特性對於了解南島民族的起源、分化、遷徙等歷史問題提供了不可取代的窗口，因此臺灣南島語言雖在上千的南島語言裡僅佔非常小的一部分，但其重要性卻遠大於其他南島語言—在整個南島語族十個主要分支中，臺灣南島語言就佔了九支，參見圖51。

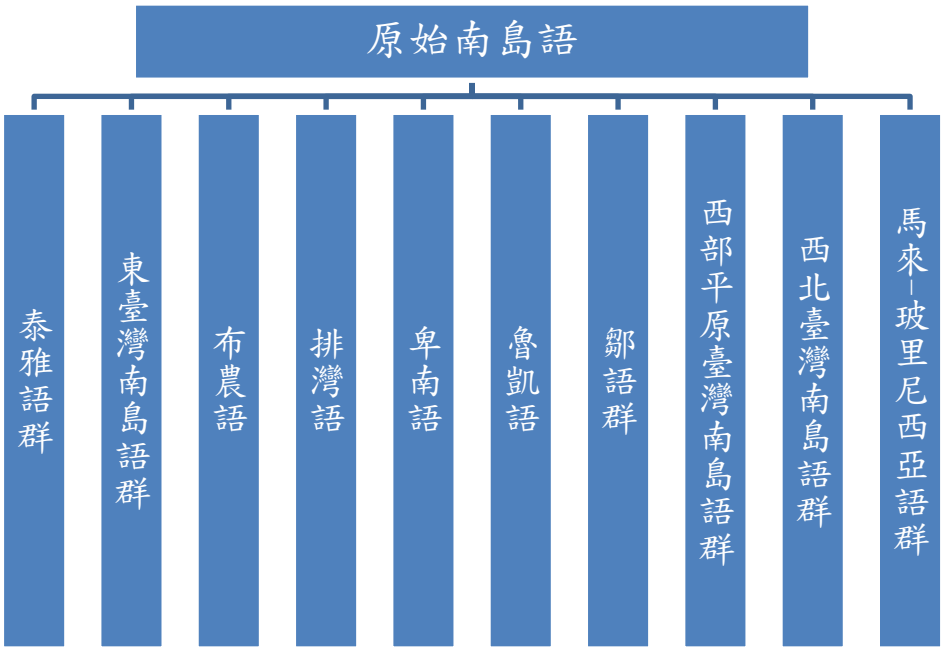


圖 51. 南島語系分群 (Blust, 1999: 45)

另外，臺灣南島語言在語言類型上也具有舉足輕重的地位。首先，臺灣南島語言擁有世界語言少見的語態系統：除了一般的主事者和受事者之外，處所、工具、受惠等事件邊緣成分都可成為主語，並且在動詞上進行相應的「語態」標記，其中工具和受惠「語態」通常使用

同樣的標記；這套語態系統通常也具有名物化功能，形成動詞語態和名詞語意角色標記同形的特殊現象，例如汶水泰雅語的 naniq-an 可以表達處所語態的「要吃」，也可以指「餐廳」，naniq-un 可以表達受事語態的「要吃」，也可以指「食物」；除了一般的陳述句區分語態外，祈使句也區分語態，例如噶瑪蘭語祈使句中，「吃」的主事語態為 qan-ka，受事語態為 qan-ika。

同時，臺灣南島語言也具有獨特的副動詞—英語、法語、日語和漢語的副詞在臺灣南島語言常常是以動詞的方式出現，因此會和動詞一樣，出現在句首、可以有語態標記，這包括一般副動詞和疑問副動詞，例如鄒語的一般副動詞 aupopoha'o「慢」可以帶有受事語態標記-a，形成 aupopoha'va；賽德克語的疑問副動詞組 huwa mesa「如何」可以帶有受事語態標記-un，形成 huwa kesun。。

臺灣南島語言的領屬語法也是一絕：「我有兩個小孩」往往是說成「我的小孩有兩個」，「我有很多錢」常常說成「我的錢很多」，其中主要數量詞是擔任謂語，而不是一般語言的修飾語。

臺灣南島語言的疑問句語法也相當特殊：「他喜歡誰」常常說成是「他喜歡的人是誰」，「誰把門打開」往往說成「打開門的人是誰」，疑問名詞充當謂語而非主語或賓語。

臺灣南島語言的重疊也值得一提。臺灣南島語言有非常豐富的重疊，除了可以表達一般的複數外，臺灣南島語言的重疊還可以表達非常特殊的意義或功能，例如未來時間（例如泰雅語 na-niqun「要吃」之詞根為 qaniq「吃」）、處所（例如鄒語 la-lauya「楓樹林；樂野村」之詞根為 lauya「楓樹」）、工具（例如賽夏語 sa-sapoeh「掃把」之詞根為 sapoeh「掃」）與縮小意義（例如排灣語 qa-tjuvi-tjuvi「蟲」之詞根為 qatjuvi「蛇」）等。

在文化層面上，臺灣南島語言也呈現和漢語截然不同的面貌。例如臺灣南島語言的「山豬」和飼養的「家豬」通常使用不同的詞彙，例如「山豬」在卑南語、魯凱語、排灣語、卡那卡那富語與鄒語分別為 babuy、babwi、vavui、vavuru 與 fuzu，「家豬」則分別為 lriyung、beeke、dridri / qacang、tutui 與 fex'x。

總之，臺灣原住民語言是上帝送給臺灣最好的禮物，是臺灣無價的瑰寶，值得大家珍惜保存流傳。

後記：本文應台大高照明教授之邀撰寫而成，收錄於高教授彙編之文化部國家語言資料庫規劃案結案報告。

附錄三、　宋麗梅教授提供之台大南島語語料庫說明資料

以下為台大南島語語料庫說明資料，目前因駭客入侵，網站被台大停權中，雖需經費和時間修復系統，然系統內的資料仍完好，可以使用，是多年來累積的寶貴語言資源。

簡單說明：

(1) 台大南島語語料庫建置於 2005 年，在沒有太多經費及沒有計算語言學專家協助下，多年來，以最低系統人力維運。

(2) 語料轉寫，都是靠訓練少許對南島語感興趣的碩班研究生來協助。每個語言所採集之類型的故事不一，有 pear/frog/日常生活/回憶/傳說/對話等。每個故事內，除了 metadata 外，還有族語、以 IU 為本的語法分析、中英 glossing 及中英翻譯、田野註解、IU 音檔、整句音檔，部份故事同時有影檔。

(3) 語料庫第一代是 linux 系統(語言學碩生協助)，第二代是 window(外包給 soho 族的工程師)。2019 年遭無聊駭客入侵，網站已被台大停權。

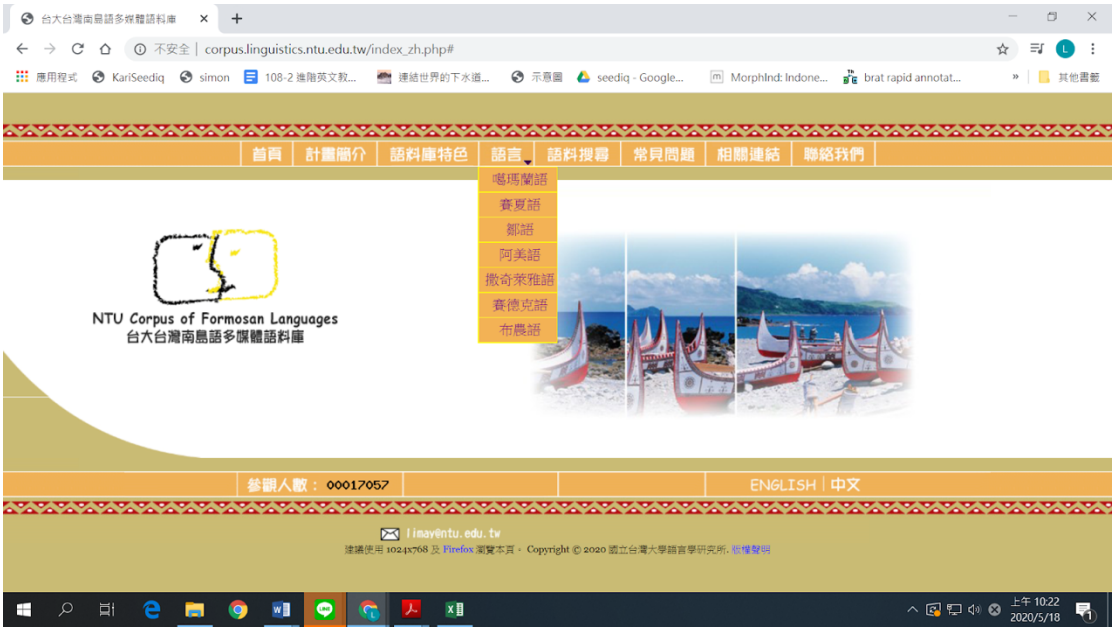
(4) 所有故事語料均在，我個人這半年已私下募集到足夠經費，目前需要找個懂 linux + 略懂南島語的工程師，重建或重做系統。

表 17. 台大臺灣南島語多媒體語料庫之轉寫統計資料（由宋麗梅教授提供）

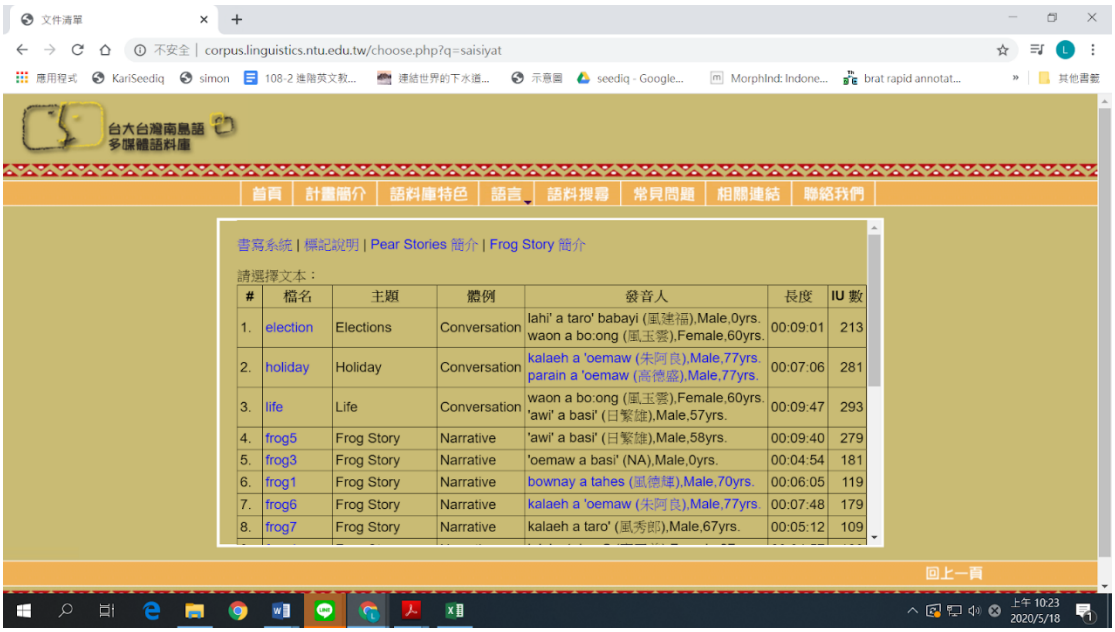
	故事數	時、分、秒數	已轉寫數	未轉寫數	備註
Kavalan	28	02:52:22	28		
Tsou	12	01:25:27	12		
Amis	19	01:02:28	19		
Sakizaya	68	06:05:27	59	9	
Seediq	27	02:41:00	27		
Saisiyat	26	02:33:49	22	4	
Bunun	17	01:14:39	7	10	
Rukai	21	02:00:23	21		
Atayal	11	01:15:36	11		
Kanakanavu	37	05:01:40	22	15	另有數千田野例句
Puyuma	17			17	
Total	**283**	**26:12:51**	**228**	**55**	

(停權中的舊網站)以舊網站之截圖，略微說明台大南島語語料庫內容

1. 首頁入口

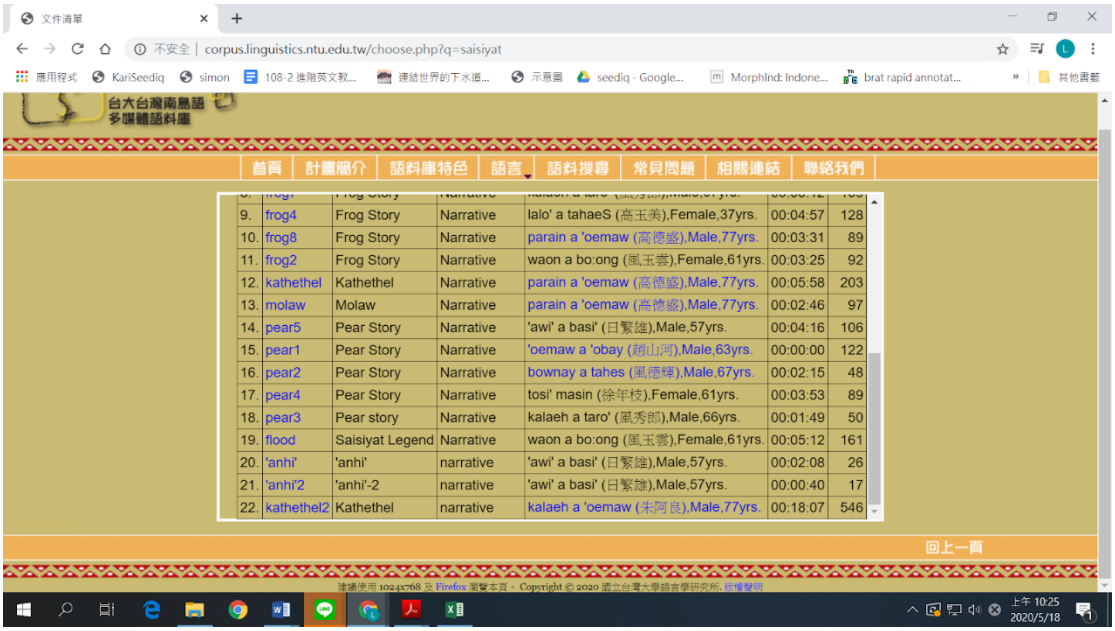


2. 以賽夏語為例，用有 22 個故事

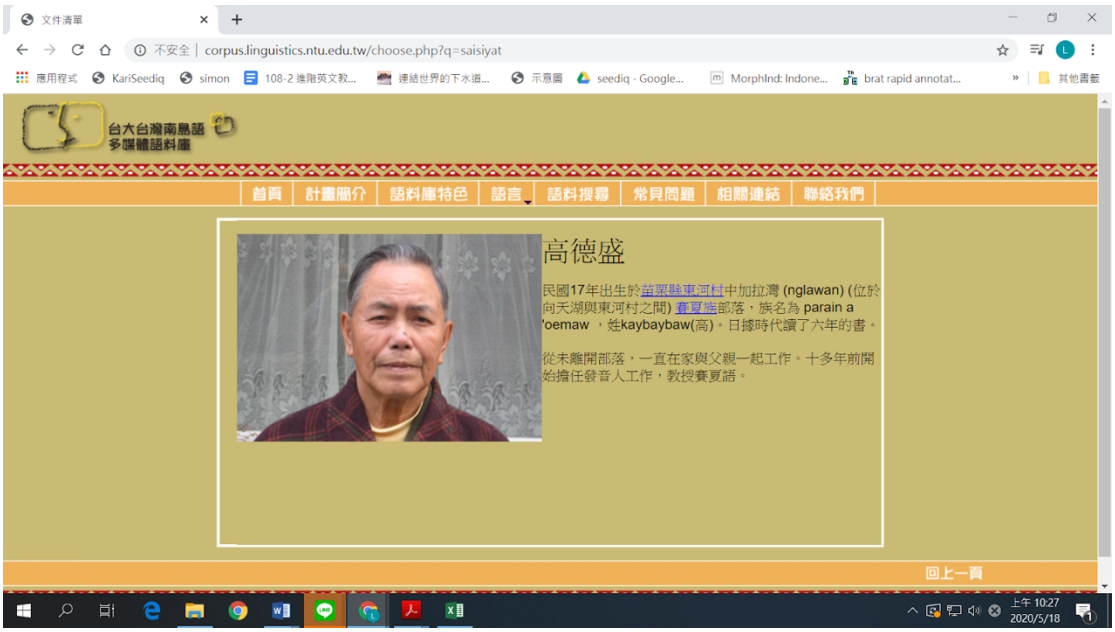


#	檔名	主題	體例	發音人	長度	IU 數
1.	election	Elections	Conversation	lahi' a taro' babayi (風建福),Male,0yrs. waon a bo:ong (風玉雰),Female,60yrs.	00:09:01	213
2.	holiday	Holiday	Conversation	kalaeh a 'oemaw (朱阿良),Male,77yrs. parain a 'oemaw (高德盛),Male,77yrs.	00:07:06	281
3.	life	Life	Conversation	waon a bo:ong (風玉雰),Female,60yrs. 'awi' a basi' (日繁雄),Male,57yrs.	00:09:47	293
4.	frog5	Frog Story	Narrative	'awi' a basi' (日繁雄),Male,58yrs.	00:09:40	279
5.	frog3	Frog Story	Narrative	'oemaw a basi' (NA),Male,0yrs.	00:04:54	181
6.	frog1	Frog Story	Narrative	bownay a tahes (風德建),Male,70yrs.	00:06:05	119
7.	frog6	Frog Story	Narrative	kalaeh a 'oemaw (朱阿良),Male,77yrs.	00:07:48	179
8.	frog7	Frog Story	Narrative	kalaeh a taro' (風秀郎),Male,67yrs.	00:05:12	109

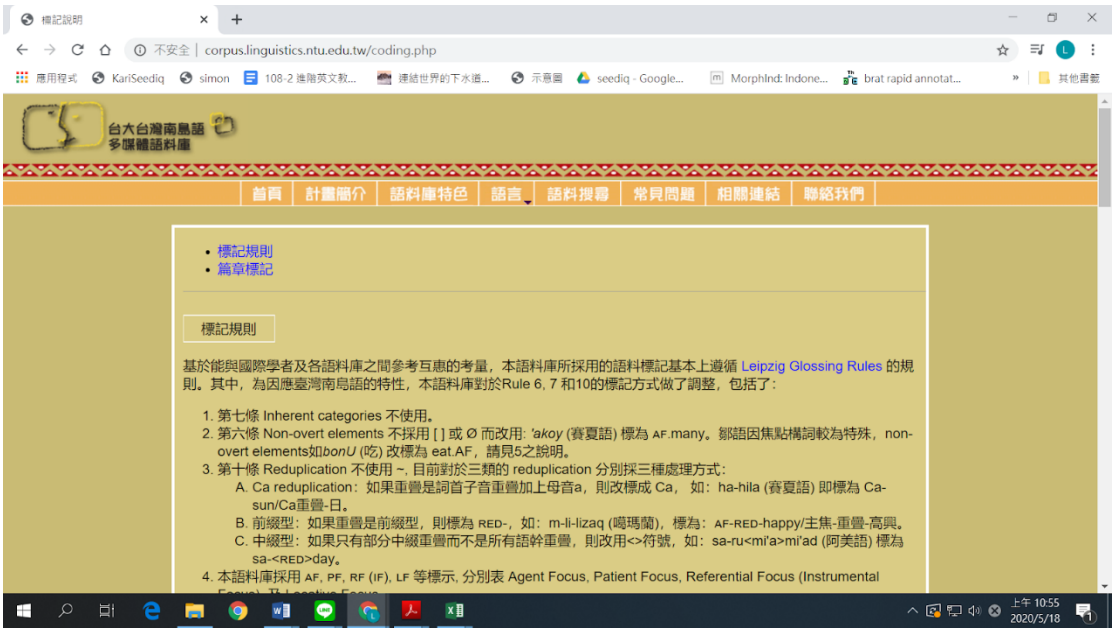
3. 賽夏語


corpus.linguistics.ntu.edu.tw/choose.php?q=saisiyat
台大台灣南島語多媒體語料庫
首頁 計畫簡介 語料庫特色 語言 語料搜尋 常見問題 相關連結 聯絡我們
9. frog4 Frog Story Narrative lalo' a tahaeS (高玉美),Female,37yrs. 00:04:57 128
10. frog8 Frog Story Narrative parain a 'oemaw (高德盛),Male,77yrs. 00:03:31 89
11. frog2 Frog Story Narrative waon a bo:ong (風玉雲),Female,61yrs. 00:03:25 92
12. kathethel Kathethel Narrative parain a 'oemaw (高德盛),Male,77yrs. 00:05:58 203
13. molaw Molaw Narrative parain a 'oemaw (高德盛),Male,77yrs. 00:02:46 97
14. pear5 Pear Story Narrative 'awi' a basi' (日繁雄),Male,57yrs. 00:04:16 106
15. pear1 Pear Story Narrative 'oemaw a 'obay (趙山河),Male,63yrs. 00:00:00 122
16. pear2 Pear Story Narrative bownay a tahes (風德輝),Male,67yrs. 00:02:15 48
17. pear4 Pear Story Narrative tosi' masin (徐年枝),Female,61yrs. 00:03:53 89
18. pear3 Pear story Narrative kalaeh a taro' (風秀郎),Male,66yrs. 00:01:49 50
19. flood Saisiyat Legend Narrative waon a bo:ong (風玉雲),Female,61yrs. 00:05:12 161
20. 'anhi' 'anhi' narrative 'awi' a basi' (日繁雄),Male,57yrs. 00:02:08 26
21. 'anhi'2 'anhi'-2 narrative 'awi' a basi' (日繁雄),Male,57yrs. 00:00:40 17
22. kathethel2 Kathethel narrative kalaeh a 'oemaw (朱阿良),Male,77yrs. 00:18:07 546
回上一頁

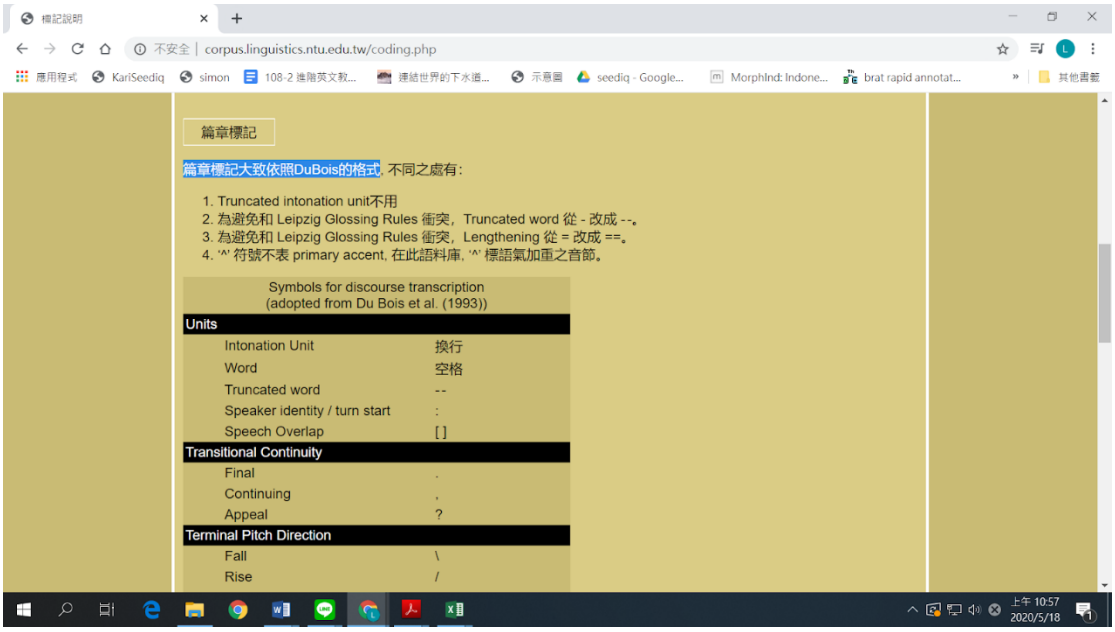
4. 發音人照片


corpus.linguistics.ntu.edu.tw/choose.php?q=saisiyat
台大台灣南島語多媒體語料庫
首頁 計畫簡介 語料庫特色 語言 語料搜尋 常見問題 相關連結 聯絡我們
高德盛
民國17年出生於苗栗縣東河村中加拉灣 (nglawan) (位於向天湖與東河村之間) 賽夏族部落，族名為 parain a 'oemaw ，姓kaybaybaw(高)。日據時代讀了六年的書。
從未離開部落，一直在家與父親一起工作。十多年前開始擔任發音人工作，教授賽夏語。
回上一頁

5. 基於能與國際學者及各語料庫之間參考互惠的考量，本語料庫所採用的語料標記基本上遵循 Leipzig Glossing Rules 的規則。



6. 篇章標記大致依照 DuBois 的格式



7. 每個語言所採集之類型的故事不一，有 pear/frog/日常生活/回憶/傳說/對話等。每個故事內，除了 metadata 外，還有族語、以 IU 為本的語法分析、中英 glossing 及中英翻譯、田野註解、IU 音檔、整句音檔，部份故事同時有影檔

附錄四、　程俊源教授書面諮詢建議

以下為第三次專家諮詢會議中，針對現有或建置中的閩南語語言資源議題，所列出之其一資源：「教育部閩南語詞彙分級計畫」。惟此計畫仍在進行中，網路上查無相關資料，僅於此附錄收錄詢問計畫主持人臺中教育大學程俊源教授之回覆說明。

此計畫案「教育部閩南語詞彙分級計畫」，目前執行期程猶尚未完竟，惟該語料庫所收語料蓋循平衡語料庫方式建置，語料皆已進行斷詞及詞類標記，俾利日後進行語言學分析或語言教學應用，並分口語語料庫暨書面語語料庫分別建置，務使語用風格判別釐然，詞庫建置詞項數擬定為 100 萬。

附錄五、　文化部「建置國家語言資料庫」勞務採購案需求規範說明書建議草案之附錄

附件一

政府資料開放優質標章暨深化應用獎勵措施

中華民國 106 年 8 月 29 日 院授發資字第 1061502362 號函

壹、　前言

有鑒於各機關對推動資料開放觀念逐步成熟，為強化與永續發展政府資料開放，並提升政府資料品質及其加值應用效益，爰規劃藉由標章認證及民眾參與機制，鼓勵各機關優化資料開放作業，期能促使各機關提供高品質、便於民眾利用之資料集，並善用資料輔助施政，以促進整體正向激勵作用，強化公共事務推動成效。

貳、　獎勵措施說明

一、　參獎對象

為鼓勵機關將資料集中列示於政府資料開放平臺，參獎對象為於政府資料開放平臺上架資料集之中央二級機關及地方政府。

二、　評獎類別及評獎標準

(一)　建立政府開放資料集品質標章機制

為鼓勵各機關提升資料品質，提供正確、易用、結構化之資料，針對政府資料開放平臺所有資料集進行機器檢測，依據各資料集完整性，分別授予金標章、銀標章或銅標章。

(二) 辦理資料開放金質獎評獎作業中央二級機關與地方政府分別依據資料量體分組評獎；另為鼓勵機關逐步提升其資料品質，本評獎另設「品質進步獎」，評核方式如下表：

評核構面	評核指標	評核重點	銅標章 (0.3 分)	銀標章 (0.6 分)	金標章 (1 分)
可取得性	資料資源連結有效性	資料資源連結可回傳連結成功狀態。	V	V	V
	資料資源可直接下載	使用者能透過連結直接獲取資料，無需透過登入或任何額外的操作形式。	V	V	V
易於被處理	屬結構化資料	➢ 固定欄位結構化資料：單一列標題的表格式資料，每筆資料的欄位數均相同，且無合併儲存格、無公		V	V

		式、無空行、無小計等。 ➢ 非固定欄位結構化資料：符合 W3C 之 XML、JSON 等結構化資料。 ➢ 其餘均為非結構化資料。			
易於理解	須依「資料集詮釋資料標準規範」提供詮釋資料	資料集詮釋資料之「編碼格式」、「主要欄位說明」與所提供之資料資源欄位相符。			V
	資料即時性*	資料集須依所填之「更新頻率」即時更新。			V
金質獎總分	[(銅標章資料集個數*0.3+銀標章資料集個數*0.6+金標章資料集個數*1)/該部會機關所屬之資料集總數*100]+加分項目				
加分項目*	資料集 API 若符合 Open API Specification(OAS)之驗證，則於總分加 0.1 分，加分項目至多 5 分。				

	加分項目由機關主動提報，並由國家發展委員會確認後，始得加分。
分組方式	中央二級機關及地方政府分別依據資料集量體採分組評獎： ➢ 第一組：資料集數量為一定數量以上。 ➢ 第二組：資料集數量一定數量以下。 ➢ 前開一定數量，由國家發展委員會依據每年政府資料開放推動情形定之。
品質進步獎	金質獎總分比前次進步 5 分者，可獲品質進步獎，惟排除金質獎各分組得獎之機關。

(三)　辦理資料開放應用獎評獎作業

鼓勵各機關提升資料應用及分析之能力，進而善用資料強化政府決策品質，並型塑公私協力應用示範案例。

由機關自主推薦優質活化應用案例，需說明該項應用案例可解決的問題、使用的資料集名稱、推薦原因、質化或量化之效益等，並開放民眾網路票選最佳之活化應用，以激勵各機關發想運用資料輔助施政之可能性，評核方式如下表：

民眾票選

評核指標	計算方式	配分
民眾網路票選最佳 之活化應用(20 分)	1.分數=(報名數量-名次+1) * 級距 2.級距=20/報名數量 舉例： 報名數量 5 組，每組分數級距 4 分，各名得分如下 第 1 名得分(5-1+1)*4=20 第 2 名得分(5-2+1)*4=16 第 3 名得分(5-3+1)*4=12 第 4 名得分(5-4+1)*4=8 第 5 名得分(5-5+1)*4=4	20

委員評獎

評核構面	評核指標	評核重點	配分

服務整合 (30 分)	資料之應用深度	說明資料分析及應用情形，並說明開放資料、內部或外部資料混搭情形。	10
	公私協力程度	說明與民間合作情形，並說明公私協力的合作模式。	10
	民間回饋	民間採用平臺上之原始資料集進行重整後回饋至平臺之民間資料集。	10
應用效益 (30 分)	創新程度	說明資料應用創新內容、步驟及方法。	15
	預期效益達成度	說明應用資料所解決之機關或民眾問題、或改善機關內部流程、提升機關服務品質等。	15
未來潛力 (20 分)	服務延續性	說明將資料應用納為機關常態運作的機制規劃。	10
	擴充應用之潛力	說明未來可再混搭其他資料的可能性與應用情境，及擴充應用服務之規劃。	10

總分		100

(四)　辦理資料開放人氣獎評獎作業

鼓勵各機關踴躍開放高價值、符合民間所需之資料，以提升政府透明治理，並驅動資料經濟發展。凡開放達 1 年且經品質檢測取得金標章之資料集，始能參與此評獎，以鼓勵機關開放及維持提供高品質、高應用價值之資料。

評核指標	配分	計算方式
資料集年度瀏覽量	30	30*(該資料集瀏覽量/同期間於本平臺瀏覽次數最多之資料集瀏覽量)
資料集年度下載量	40	40*(該資料集下載量/同期間於本平臺下載次數最多之資料集下載量)
資料集評分	30	30*(該資料集平均得分/5)
總分	100	

三、　獎勵額度及措施

	資料開放金質獎	資料開放應用獎	資料開放人氣獎
額度	• 第一組:中央二級機關、地方政府各取前 3 名 • 第二組：中央二級機關、地方政府各取前 2 名	前 3 名	前 10 名
獎勵方式	• 第一組 ➢ 第 1 名：主要專責人員及其主管各記小功 2 次 ➢ 第 2 名：主要專責人員及其主管各記小功 1 次 ➢ 第 3 名：主要專責人員及其主管各記嘉獎 2 次 • 第二組 ➢ 第 1 名：主要專責人員及其主管各記小功 1 次 ➢ 第 2 名：主要專責人員及其主管各記嘉獎 2 次 • 進步獎 主要專責人員及其主管各記嘉獎 2 次	• 第 1 名：主要資料應用人員及其主管各記小功 2 次 • 第 2 名：主要資料應用人員及其主管各記小功 1 次 • 第 3 名：主要資料應用人員及其主管各記嘉獎 2 次	• 資料集之業務單位專責人員及其主管各記小功 1 次 • 重複獲獎者，最高以小功 2 次為限

四、 作業時程

作業項目	時程
函請各機關提報參獎申請書	每年 4~5 月
機關提報優質資料應用案例	每年 6~7 月
民眾票選	每年 8~9 月
資料品質機器檢測	每年 8~9 月
評審委員評獎	每年 9 月
評審結果報院核定	每年 10 月
函請各機關依評獎結果辦理敘獎	每年 11 月

備註：以上作業時程得視實際狀況予以調整。

五、　　評獎方式

(一) 資料開放金質獎

於政府資料開放平臺上架之中央二級機關及地方政府，由國家發展委員會逕予進行品質檢核作業，並公開公布各資料集取得之標章。

(二) 資料開放應用獎

由機關填寫「參獎申請書」，於指定時間內函送國家發展委員會，逾期不受理。

本獎項民眾票選分數占 20%，將於政府資料開放平臺提供民眾票選。委員評獎分數占 80%，由「資料開放應用獎評審小組」負責本項評審工作，並由參獎機關透過簡報、示範展示等方式，展現資料應用成果及效益。

「資料開放應用獎評審小組」由國家發展委員會遴聘學者專家、資料社群、民間企業等代表組成。

(三)　資料開放人氣獎

於政府資料開放平臺上架之資料集，開放達 1 年且經品質檢測取得金標章之資料集，始能參與評獎，並排除前三屆曾獲獎之資料集。

附件二

保密切結書（廠商）

_____公司（以下簡稱廠商）受文化部（以下簡稱本部）委託辦理「建置國家語言資料庫」勞務採購案（以下簡稱本案），於本案執行期間有知悉或可得知悉或持有政府公務秘密及業務秘密（包含個人資料），為保持其機密性，同意遵守本同意書下列各項規定：

第一條廠商承諾於本契約有效期間內及本契約期滿或終止後，對於所知或持有一切本部未標示得對外公開之公務秘密，以及本部依契約或法令對第三人負有保密義務之業務秘密，均應以善良管理人之注意妥為保管及確保其機密性，並限於本契約目的範圍內，於本部指定之處所內使用之。非經本部事前書面同意，不得為本人或任何第三人之需要而複製、保有、利用該等秘密或將之洩漏、告知、交付第三人或以其他任何方式使第三人知悉或利用該等秘密，或對外發表或出版，亦不得攜至本部或本部處指定處所以外之處所。

第二條廠商知悉或取得本部公務秘密、業務秘密及任何個人資料，應限於其執行本契約所必須之業務範圍內、且僅限於本契約有效期間內，提供或告知有需要知悉該秘密之廠商團隊成員或其他相關人員。

第三條廠商在下述情況下解除其所應負之保密義務：

原負保密義務之資訊，由本部提供以前，已合法持有或已知且無保密必要者。

原負保密義務之資訊，依法令業已解密、依契約本部業已不負保密責任、或已為公眾所知之資訊。

原負保密義務之資訊，係自第三人處得知或取得，該第三人就該等資訊並無保密義務。

第四條廠商因違反本保密切結書之規定，致造成本部或其他相關第三者之損害或賠償時，廠商同意無條件負擔全部責任，包括因此所致本部或其他相關第三者涉訟時所須支付之一切費用及賠償。於其他相關第三者對本部提出請求、訴訟，經本部以書面通知廠商提供相關資料，廠商應合作提供，不得異議。

第五條廠商對工作中所持有、知悉之資訊系統作業機密或敏感性業務檔案資料、 個人資料等，均負有保密義務與責任，並遵循「營業秘密法」、「著作權法」、「商標法」、「專利法」、「個人資料保護法」及「個人資料保護法施行細則」等相關規定，非經本部相關權責人員之書面核准，不得擷取、持有、傳遞或以任何方式提供給無業務關係之第三人，如有違反，願賠償一切因此所生之損害，並擔負相關民、刑事責任，不得異議。此外，廠商處理個人資料檔案部分應於委託關係解除或終止時刪除或銷毀履行契約而持有之個人資料，及返還個人資料之載體；並提供刪除、銷毀或返還個人資料之時間、方式、地點等紀錄以茲證明，本部並保有查核之權利。

第六條廠商若違反本保密切結書之規定，本部保有請求廠商賠償本會因此所受之損害及追究廠商洩密之刑責之權利，如因而致第三人受有損害者，廠商亦應負所有賠償責任。

文化部

立切結書人

廠商名稱及蓋章：

廠商負責人姓名及簽章：

廠商地址：

廠商聯絡電話：

統一編號：

中華民國　　年　　月　　日

附件三

保密切結書（人員）

＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿公司（以下簡稱廠商）＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿（公司人員，以下簡稱甲方）受文化部（以下簡稱本部）委託辦理「建置國家語言資料庫」勞務採購案（以下簡稱本案），於本案執行期間有知悉或可得知悉或持有政府公務秘密及業務秘密（包含個人資料），為保持其機密性，同意遵守本同意書下列各項規定：

第一條甲方承諾於本契約有效期間內及本契約期滿或終止後，對於所知或持有一切本部未標示得對外公開之公務秘密，以及本部依契約或法令對第三人負有保密義務之業務秘密，均應以善良管理人之注意妥為保管及確保其機密性，並限於本契約目的範圍內，於本部指定之處所內使用之。非經本部事前書面同意，不得為本人或任何第三人之需要而複製、保有、利用該等秘密或將之洩漏、告知、交付第三人或以其他任何方式使第三人知悉或利用該等秘密，或對外發表或出版，亦不得攜至本部或本部處指定處所以外之處所。

第二條甲方知悉或取得本部公務秘密、業務秘密及任何個人資料，應限於其執行本契約所必須之業務範圍內、且僅限於本契約有效期間內，提供或告知有需要知悉該秘密之廠商團隊成員或其他相關人員。

第三條甲方在下述情況下解除其所應負之保密義務：

原負保密義務之資訊，由本部提供以前，已合法持有或已知且無保密必要者。

原負保密義務之資訊，依法令業已解密、依契約本部業已不負保密責任、或已為公眾所知之資訊。

原負保密義務之資訊，係自第三人處得知或取得，該第三人就該等資訊並無保密義務。

第四條甲方因違反本保密切結書之規定，致造成本部或其他相關第三者之損害或賠償時，甲方及其所屬之廠商同意無條件負擔全部責任，包括因此所致本部或其他相關第三者涉訟時所須支付之一切費用及賠償。於其他相關第三者對本部提出請求、訴訟，經本部以書面通知廠商提供相關資料，廠商應合作提供，不得異議。

第五條甲方對工作中所持有、知悉之資訊系統作業機密或敏感性業務檔案資料、 個人資料等，均負有保密義務與責任，並遵循「營業秘密法」、「著作權法」、「商標法」、「專利法」、「個人資料保護法」及「個人資料保護法施行細則」等相關規定，非經本部相關權責人員之書面核准，不得擷取、持有、傳遞或以任何方式提供給無業務關係之第三人，如有違反，願賠償一切因此所生之損害，並擔負相關民、刑事責任，不得異議。此外，甲方處理個人資料檔案部分應於委託關係解除或終止時刪除或銷毀履行契約而持有之個人資料，及返還個人資料之載體；並提供刪除、銷毀或返還個人資料之時間、方式、地點等紀錄以茲證明，本部並保有查核之權利。

第六條甲方若違反本保密切結書之規定，本部保有請求廠商賠償本會因此所受之損害及追究廠商洩密之刑責之權利，如因而致

第三人受有損害者，甲方及其所屬之廠商亦應負所有賠償責任。

文化部

立切結書人

姓名：

身分證字號：

電話：

地址：

中華民國　　年　　月　　日

保密切結書（駐點人員）

立切結書人_____（公司人員姓名），受_____（廠商名稱）委派至文化部（以下簡稱本部）處理業務，謹聲明同意遵守本同意書下列各項規定，對工作中所持有、知悉之資訊系統作業機密或敏感性業務檔案資料，均保證善盡保密義務與責任，非經本部權責人員之書面核准，不得擷取、持有、傳遞或以任何方式提供給無業務關係之第三人，如有違反願賠償一切因此所生之損害，並擔負相關民、刑事責任，不得異議。

第一條未經申請核准，不得私自將本會之資訊設備、媒體檔案及公務文書攜出。

第二條未經本會業務相關人員之確認並代為申請核准，不得任意將攜入之資訊設備連接本會網路。若經申請獲准連接本會網路，嚴禁使用數據機或無線傳輸等網路設備連接外部網路。

第三條經核准攜入之資訊設備欲連接本會網路或其他資訊設備時，須經電腦主機房掃毒專責人員進行病毒、漏洞或後門程式檢測，通過後發給合格標籤，並將其黏貼在設備外觀醒目處以備稽查。

第四條廠商駐點服務及專責維護人員原則應使用本部所配發之個人電腦與週邊設備，並僅開放連結本會內部網路。若因業務需要使用本部電子郵件、目錄服務、或另欲連接外部網路等，皆應經本部業務相關人員之確認並代為申請核准，方得執行。

第五條本部得定期或不定期派員稽查立切結書人是否遵守上列所有工作規定。

第六條本保密切結書不因立切結書人離職而失效。

第七條立切結書人因違反本保密切結書應盡之保密義務與責任致生之一切損害，立切結書人及其所屬公司或廠商應負連帶賠償責任。

文化部

立切結書人

姓名：

身分證字號：

電話：

地址：

立切結書人所屬廠商：

廠商名稱及蓋章：

廠商負責人姓名及簽章：

廠商地址：

廠商聯絡電話：

統一編號：

填表說明：

一、 廠商駐點服務人員、專責維護人員，或逗留時間超過三天以上之突發性維護增援、臨時性系統測試或教育訓練人員（以授課時需連結本部網路者為限）及經常到本部洽公之業務人員皆須簽署本切結書。

二、 廠商駐點服務人員、專責維護人員及經常到本部洽公之其他相關業務人員每年簽署本切結書一次。

中華民國　　年　　月　　日

附件五

文化部____年度資訊內部稽核（系統端）檢查表

受稽核單位名稱：_____ 系統主機 IP/名稱：_____

受稽核人員姓名：_____ 檢查日期：____ 年 ____月 ____ 日

受稽核單位名稱：					檢查日期： 年 月 日
項次	檢查重點		檢查情形		檢查說明
			合格	不合格	
1	系統存取政策及授權規定辦理情形	1-1 資訊單位是否訂定系統存取政策及使用管理規定。			
		1-2 系統存取政策及使用管理規定，是否以書面、電子或其他方式告知員工及使用者相關權限及責任。			
2	系統存取權限（帳號）管理情形	2-1 各機關是否建立系統使用者註冊管理等制度，加強使用者通行密碼管理，並要求使用者定期更新。			
		2-2 機關員工離（休）職時，資訊單位是否即時取消各項資訊資源及使用權限。			
		2-3 機關員工職務異動時，資訊單位是否依系統存取授權規定，調整其權限。			
		2-4 帳號刪除日期與員工離職日期是否有不一致者。			
		2-5 各機關開放外界連線作業，是否事前簽訂契約或協定，並明定其應遵守之資訊			

		安全規定、標準、程續及應負之責任。			
		2-6 各機關對系統服務廠商以遠端登入方式進行系統維修者，是否建立人員名冊及相關安全保密責任。			
		2-7 重要資料委外建檔者，不論在機關內外執行，是否採取適當及足夠之安全管制措施，以防止資料被竊取、竄改、販售、洩漏及 不當備份等情形發生。			
3	電腦資料庫查詢軌跡紀錄檔(Log)	3-1 資訊單位是否建立及啟動電腦資料庫查詢軌跡紀錄檔（Log），並保存至少 1 年，以作為日後調查及監督之用。			
		3-2 資訊單位系統紀錄檔是否定期備份轉出檔案後保存。			
		3-3 資訊單位是否有專人隨時（經常）檢視。			
4	系統存取異常狀況情形	4-1 登入「系統使用」紀錄之「登入次數」是否異常頻繁。			
		4-2 登入「系統使用」紀錄之「使用時間」是否有異常。			
		4-3 登入「系統使用」紀錄之「登入失敗次數」是否異常頻繁。			
		4-4 使用者「查詢內容」紀錄之「查詢成功次數（筆數）」是否有異常。			
		4-5 登入系統查詢時段是否有異常。			

		4-6 使用者「查詢內容」紀錄所登載之「案件編號」（如收文號）是否有異常。			
		4-7 查詢之資料與承辦案件（業務）有無不一致者。			
5	機關資訊系統或網頁資料安全控管情形	5-1 機關於網頁公開資訊是否符合「個人資料保護法」、「政府資訊公開法」等規定。			
		5-2 是否定期搜尋網站不當庫存資料並修正或改進設定。			
		5-3 發現資訊安全漏洞狀況時是否通報政風單位。			
6	系統存取異常狀況通報情形	6-1 異常狀況之界定是否符合現況。			
		6-2 有無建置資安異常通報機制。			
		6-3 發現異常存取狀況時是否通報反映予政風單位。			
稽核人員					受稽核單位/人員

附件六

派駐人員電腦軟硬體規格

序號	品項名稱	規格
1	____桌上型電腦 (　W~　W)	1. _____（設備型號） 2. CPU： 3. 晶片組： 4. RAM： 5. 硬碟： 6. 顯示介面： 7. 擴充槽： 8. 內建 I/O 介面： 9. 網路介面： 10. USB： 11. 光碟機： 12. 提供雙獨立顯示數位影像輸出（HDMI 或 DVI 或 Display Port）與 VGA。 13. 內接或外接 IC 卡讀卡機（可讀取自然人憑證 IC 卡）及記憶卡讀卡機。 14. 電源供應器： 15. 電源延長線： 16. 鍵盤/滑鼠： 17. 安全：
2	____吋桌上型寬螢幕液晶顯示器	1. 須與桌上型電腦主機同一品牌。 2. 可視區域： 3. 面板： 4. 解析度： 5. 點距： 6. 可視角度： 7. 對比率： 8. 亮度顯示： 9. 總反應時間： 10. 訊號輸入： 11. 喇叭：提供內建或外接____個（含）以上喇叭有效輸出率(RMS)為____W（含）以上。

		12. 內建多層膜防刮玻璃。 13. 安全：符合_____等規範。
3	軟體	_____

附件七

著作人約定書

受雇（聘）人＿＿＿＿＿＿＿＿於雇（聘）用人＿＿＿＿＿＿＿＿雇（聘）用期間內，在雇（聘）用人執行文化部「建置國家語言資料庫」勞務採購案契約由受雇（聘）人所完成之著作，茲約定均以雇（聘）用人為著作人，此證。

立約定書人　雇（聘）用人：

代表人：

地址：

立約定書人　受雇（聘）人：

身分證字號：

地址：

中華民國　　年　　月　　日

附件八

文化部

「建置國家語言資料庫」勞務採購案

經費分析表

項次	品項及規格	單位	數量	單價	複價	備註
1	**人事費**					
1-1						
1-2						
	小計					
2	**業務費**					
2-1	專家諮詢費					
2-2	國家語言資料庫 系統工程					
2-3	授權金					
2-4	訪談費					
2-5	資料彙整費					
2-6	交通費					
2-7	錄製語料					
2-8	撰稿費					
2-9						
	小計					
3	**差旅費**					
	小計					
4	**設備租賃及使用費**					
4-1						
4-2						
4-3						
4-4						
	小計					
5	**駐會人員**					

5-1	駐會人員薪資					
5-2	駐會人員之單位負擔勞保、健保（含補充保費）、勞退金					
5-3	駐會人員工作獎金					
5-4	駐會人員加班費					
5-5	駐會人員出差交通及住宿費					
6	**保險**					
6-1	公共意外險					
6-2	旅遊平安險及意外醫療					
7	**管理費**					
8	**雜支**					
	小計					
	合計（含稅）					

備註：

1、 請明列本案各項費用成本分析。

2、 請投標廠商自行填寫，品項內容應包含執行本案所需之細項費用。

附錄六、　文化部「建置國家語言資料庫先期規劃研究」勞務採購案需求說明書

附件九

壹、　計畫名稱

文化部「建置國家語言資料庫先期規劃研究」勞務採購案（以下簡稱「本案」）。

貳、　計畫目的

語言為文化傳承之重要載體，為促進語言永續發展，豐富國家之文化內涵，文化部特制定國家語言發展法，並於108年1月9日經總統公布施行；爰現依該法第八條規定:「政府應定期調查提出國家語言發展報告，建置國家語言資料庫」，辦理建置國家語言資料庫之先期規劃研究。

國家語言資料庫除應含國家語言語料庫外，亦應納入各國家語言史料、調查統計等相關資料，以作為國家語言傳承、復振及發展之基石，惟目前國內語言研究資料及語料之蒐集整理工作，除民間有零星的計畫與成果外，亦分散在各相關政府機關，如教育部、科技部、客家委員會、原住民族委員會、中央研究院、國家教育研究院等，且國家語言資料庫之建置涉及語言學、資訊工程、著作權等不同領域之高度專業性工作，應需事先詳細規劃研究。

另有鑑於教育部業於 106 年辦理「本土語言語料庫建置規劃研究案」並盤點本國近代具代表性之語料庫，爰本部以此研究成果為基礎，擬就國 家語言資料庫之建置型態(是否能整合現有之資料庫與語料庫)，以及後續 之維運管理、應用推廣等面向進行先期規劃研究，同時探討相關之著作權議題，讓國家語言資料庫之建置更加完善。期透過本先期規劃研究案，為國家語言資料庫建置奠定長遠發展的基礎，並得永續保存本國語言文化資產，促進未來研究發展及加值應用。

參、　經費預算

本案經費以新臺幣（下同）150 萬元整（含稅）為上限。本案係採跨年度撥款，108 年度撥付第 1 期款 30%，另第 2 期款 40%及第 3 期款 30%之款項辦理保留於 109 年度撥付。

肆、　採購方式

本案依採購法第 22 條第 1 項第 9 款辦理限制性招標公開評選。

伍、　廠商資格

請參閱投標須知。

陸、　履約期限

自決標日起至 109 年 5 月 29 日止。

柒、 計畫執行項目及工作內容

一、 研究項目：

(一) 國家語言資料庫：

1. 擇至少 3 個已建立國家語言資料庫之國家，並就其建置方式、內容項目、維運管理、使用對象、應用推廣等面向進行比較、分析及研討。

2. 盤點國內現有之語言資料庫，並就其建置方式、內容項目、維運管理、使用對象、應用推廣等面向進行說明。

3. 提出國家語言資料庫之建置方式、內容項目、維運管理、使用對象及應用推廣之規劃建議。

4. 盤點現可收錄於國家語言資料庫之資料清冊。

(二) 國家語言語料庫：

1. 研析目前本國具代表性之語料庫之規格，並提出國家語言語料庫之建議規格。

2. 國家語言語料庫之整合機制與方式，並請提出統一之語料格式或標準之建議方案，及研析跨語別語料共同搜索之可能性。

3. 依語料庫之建置目的，規劃各類語料（書面、口語及影像等）蒐集之方式。

(三)　研擬國家語言資料庫及國家語言語料庫對外授權應用機制，以及研析可能會涉及之著作權議題。

(四) 綜合上述分析及專家學者意見，提出「建置國家語言資料庫（含國家語言語料庫）」之需求說明書，需含評估過後之執行項目、執行期程、人力配置以及經費編列等。

(五) 視需求召開至少 2 次諮詢會議，邀請相關學者專家出席，進行研議。

二、　研究方法：

本案由廠商組成研究團隊，採蒐集學術專論、論文研究、調查資料、統計數據等文獻資料進行現況研究分析，經分類、綜整、研析，並配合未來勢，彙整後提出規劃及建議。另得就實際研究需求，適切規劃其他研究資料蒐集或分析方法，由廠商自行研議研究方案提報本部同意後執行之。

三、　研究計畫應交付項目：

廠商須交付含本案所有研究項目研析結果、至少 3 萬字之研究報告（圖表部分以彩色印製）。應印製期中報告 10 本、期末報告 10 本、審查修正後之定稿版期末報告 10 本，並連同印製底稿、報告內容、研究成果摘要、參考文獻原始資料及各式統計資料等，以電子檔格式儲存於光碟片中，一併依限送交本部。

期中報告及期末報告之撰寫內容、方式及印製格式，原則依「文化部及所屬機關（構）委託研究計劃作業規定」第六條（如附件 1）辦理。

四、　其他注意事項：

(一) 履約期間廠商需配合本部業務單位召開工作會議，就階段性工作成果或相關議題研討進行工作進度報告。

(二) 計畫執行期間，廠商如需更換研究團隊人員，須經本部書面同意後始得為之。

五、　計畫期程：

(一)　第一期：決標日起 10 個工作天內，修正服務建議書或提交細部工作計畫書（含工作期程表）。

(二)　第二期：109 年 1 月 31 日前完成期中報告事項。

(三)　第三期：109 年 5 月 29 日前完成期末報告事項。

捌、　組織及人力配置

本案研究人員人數由廠商依計畫執行需求規劃配置。其中研究主持人（含協同主持人）應至少兩位，需具備語言學、資訊工程及法學（著作權法）專業之學者或專業人員，或曾任教大專院校相關系所或從事相關專業工作。

玖、　經費編列

研究經費編列標準請參考「行政院所屬各機關行政及政策類委託研究計畫經費編列原則及基準」）附件 2）及「文化部及所屬機關（構）

委託研究 計畫作業規定」規定，預估本案所需費用，詳列經費分配情形（具體填寫數 量、規格、單價及總價），並參照本部經費明細表格式（如附件 3）填列。

壹拾、 服務建議書規範：

一、 提送格式：

(一) 以 A4 之紙張裝訂，由左至右中文直式橫寫（佐證資料可為英文），但相關之圖說得以 A3 之紙張製作。

(二) 封面請註明投標廠商名稱、標案名稱及提出日期，內頁須編製目錄，並於各頁下方中央加註頁碼。

(三) 不含封面、目錄及附件，以雙面印製不超過 50 頁為原則（A4 及 A3 雙面印製一張計二頁）。

(四) 印製 12 份。

二、 內容須包括下列各項：

(一) 專案概述：專案名稱、目標、內容、範圍等說明。

(二) 專案工作團隊：此部分內容包括本專案組織架構、人力及職掌、以及團隊合作模式、資源及未來配合方式說明。

(三) 專案管理規劃與分析說明：內容包括對本計畫之執行敘述，含研究架構、方法、流程、步驟、研究進度（附甘特條型圖）、預計完成之項目（需含期中、期末報告內容之規劃）。

(四) 總價及各單項價格成本分析。

(五) 廠商經驗與能力

1. 公司/財團法人團體/學術研究機構簡介。廠商須註明負責本案之聯絡人及電話，以便聯繫事宜之用。

2. 本案之研究團隊辦理類似專案經驗說明，並明列各人員之簡歷。

(六) 廠商得就有助於提升本專案效益之作為，但未列為本專案需求，額外補充或建議之部分，可另闢章節描述，但未列為本專案需求。

三、 服務建議書之格式與內容不符規定者，評選委員得斟酌較低之評分，此部分請投標廠商注意。

壹拾壹、 評選決標事宜：

一、 本案依據行政院公共工程委員會發布「採購評選委員會組織準則」，成立評選委員會，並依「採購評選委員會審議規則」及「機關委託專業服務廠商評選及計費辦法」規定辦理，評選優勝廠商之作業，準用最有利標決標之評選規定。

二、 評選辦法：

(一) 本案先進行投標廠商資格審查，再邀請符合資格廠商就所提服務建議書進行簡報說明，由採購評選委員會進行評選。

(二) 符合資格者，由本計畫之採購評選委員會，擇期召開評選會議。評選時間、地點由本部以書面另行通知。

(三) 依投標廠商投標文件到達先後順序決定簡報次序。

(四) 簡報時間原則以 15 分鐘為限，倘投標廠商達 3 家(含)以上，簡報時間以 12 分鐘為限（時間結束前 2 分鐘按鈴一短聲提醒，時間結束按鈴一長聲，即停止簡報。）。詢答時間以 10 分鐘（採統問統答方式，委員提問時間不計）為限（時間結束前 2 分鐘按鈴一短聲提醒，時間結束按鈴一長聲，即停止答復。）。

(五) 簡報時由本部提供 1 台單槍投影機與基本電源，廠商請自備筆電。

(六) 簡報人員必須包含本案之計畫主持人在內，計畫主持人未出席簡報者，評選委員得酌扣廠商簡報項目之得分。每一廠商至多得派 3 人進入會場簡報，若經 5 分鐘內唱名 3 次未到場簡報者，簡報部分予以零分計算，其他部分以書面審查。

(七) 簡報不得更改廠商投標文件內容，廠商另外提出變更或補充資料者，該資料不納入評選。

三、 評選項目及配分：

評選項目	內容	權重
1. 履約能力	專業資歷背景、人力架構、過往相關經驗	30%
2. 服務建議書內容之完整性及可行性	計畫內容及研究架構、方法、流程、步驟規劃	25%

3. 經費概算之合理性	經費運用情形與價格之合理性	20%
4. 計畫管理	計畫進度控管、計畫預期成果、計畫之周延性、可行性	15%
5. 簡報及答詢	簡報內容是否具體詳實、答覆有無中肯切題、掌握重點	10%
總分		100%

四、 優勝廠商評定方式：

(一) 評選時，將就各評選項目分別評分後予以加總，依加總分數高低換算為序位，並彙整合計各廠商之序位，以合計值最低者為序位第 1 名，並經出席評選委員過半數同意後，評定各廠商序位名次，再簽報本部部長或其授權人員核定各廠商序位名次。

(二) 合格門檻：投標廠商平均分數達 75 分者為合格廠商，未達 75 分者不得列為優勝廠商。若無合格廠商時，主席宣布廢標，本案另行辦理。

(三) 優勝廠商僅為一家者，以議價方式辦理。優勝廠商為二家以上者，依序位第一者取得優先議價權，其次取得第二順位議價權，餘類推。但有二家以上廠商為同一序位者，以標價低者優先議價，若標價相同，即擇權重最高之評選項目之得分合計值較高者優先，得分仍相同者，抽籤決定之。

(四) 優勝廠商於議價完成後訂定委託契約。優勝廠商如因故無法完成議價程序或棄權者，本部得依序遞補。

(五)　其他評選注意事項：本部得因故終止評選事宜，通知投標廠商領回服務建議書。

(六)　採購評選委員自接獲評選有關資料之時起，不得就該採購案參加投標、作為投標廠商之分包廠商或擔任工作成員。其違反者，機關應不決標予該廠商。

(七)　投標廠商之服務建議書中所呈現之工作成員（依據行政院公共工程委員會 96 年 8 月 7 日工程企字第 09600302640 號函，工作成員範　圍，包含投標廠商之投標文件所述人力組織及參與或協助該採購案之　相關人員均屬之），如有不屬投標廠商之人員，須取得其同意書，投　標時並一併附具同意書之影本，未檢附時，則視為該人員未同意擔任　工作成員，機關得於評選會議要求廠商說明，並請評選委員酌予扣減　相關評選項目分數。

五、 其他

本次招標由廠商所提供之相關內容如有任何侵犯他人智慧財產權之情事者，概由廠商負一切法律責任。

附錄七、　專家諮詢會議重要結論

本章節是三次專家諮詢會議及五次個別學者諮詢（分別為洪惟仁教授、張永利教授、蔡素娟教授、張學謙教授、程俊源教授）過程中，各位專家學者所提出的寶貴意見。以下將針對會議上所討論的各項議題分門別類作整理，這些議題包括：資料收集、著作權、如何保持逐漸流失的國家語言、語言地圖、閩南語語料處理與語音技術、語料庫跨語言檢索與數位加值應用。最後，在「8 專家諮詢會議與文獻整理總結」小節將針對各專家的意見們作統合與補充建議。

1.　國家語言資料庫的內容規劃[9]

高照明教授

我們希望國家語言資料庫的內容規劃，第一部分是「我們的母語」，網頁可以放上影片，如：客家電視台的動畫，並介紹語言流失的議題。

第二部分是「臺灣的國家語言與地理分佈」，邀請洪惟仁教授、張永利教授撰稿，分為整體介紹及以鄉鎮為單位的各個國家語言的介紹。如果有音檔的話，在網頁式地圖上可以標記發音，點選聆聽各地的樣本音，希望未來的執行單位能夠去做。

第三部分是「國家語言調查報告」，國家語言調查報告及國家語言資料庫是《國家語言發展法》通過後，三年內要提出的項目，希望是六年的計畫，分為前三年、後三年，包括各部會歷年的報告，如教

[9] 此為專家諮詢會議之紀錄，從諮詢會議的討論中獲得各諮詢委員之寶貴意見，逐步調整、修改「柒、國家語言資料庫整體設計與規劃之建議」之架構內容。

育部、客委會、原民會，以及內政部人口普查中與語言相關的內容，最近一次的普查是民國 101 年，約有三題題目，雖然題目數量較少，但調查人口較廣，且未來的報告也規劃於此區更新。此外，學者的調查報告也希望可以納入，提供下載。

針對國家語料庫，多次專家諮詢會議中都建議提出「典藏」與「新建語料」的區分，其中典藏內容為已經建置好的網站，可設計共同檢索的介面，例如：華語資源已有不少、客委會正在進行客語語料庫的建置、原住民族語語料庫雖須克服使用人數少且語種多的問題，但初步規劃上，齊莉莎教授建議可納入族語 e 樂園的資料，並重新檢審，以上部分以典藏的方式整合至文化部。國家語言資料庫分階段建置，第一階段為整合現有資料，第二階段是擴增語料，除了閩南語，客語在第二階段也可納入山歌、文化儀式、田野調查等語料。

閩南語語料庫為新建語料，比照華語及客語語料庫，建置一千萬詞的語料庫，口語語料為多媒體形式，提供影像檔、音檔及轉寫等內容。

2. 關於資料的收集

2.1. 資料收集的原則

張永利教授

首先，張教授提議之後或許可以開一個公開的 workshop（小型會議）作為平台，然後把建置國家語料庫時可能會碰到的各種議題區分成不同的 session，例如語料收集、語料庫架構內容、如何營運、著作權等等。接著就可以廣邀各界人士來參與討論、分享、交流各式相關資訊，如此才能取得更全面的看法與見解。

另外張教授也提及，workshop 最主要的目的是在收集非學術界的、未出版的、流落在民間的資料，例如一些小型語料庫、字典，以及前人已經蒐集到的關於各種臺灣國家語言的資料等。

最後，關於原住民語的的語料庫平衡，張教授也建議可以參考宋麗梅老師的原住民語料。

章忠信教授

章教授也認為第一步應該要先開設一個平台，讓大家分享「誰做過什麼」之類的不涉及著作權的資訊（即 metadata，上面標註各語言資料出處）；先了解大致狀況，然後再來討論決定要將哪些國家語言資料納入國家語言資料庫，這樣的做法會比較恰當。

章教授整理 metadata 這一步驟其實並不困難，困難之處在於之後要如何將那些語料進行數位化，因為通常要將語料數位化之前必須徵求資料擁有者的同意，如此就可能會有各種關於著作權的議題要考慮和處理。因此，章教授在此進一步建議，關於「伍、本國國家語言相關之語言資料庫」章節的議題應該要分階段進行，在現階段大家應先盡可能地蒐集 metadata，之後再去思考是否需要數位化蒐集到的資料，並且參照經濟利益、法律問題等去做相關的處理。

郭志忠博士

除了一般的書面語口語資料外，郭博士認為應該也可以納入像是歌謠、民間戲曲等語料，因為這些資料通常有押韻，可以用來判別字詞的原有讀音等資訊。而在原住民語語料的部分，郭博士認為原住民語言資料原本就相對不多，而且不少語言有瀕危的危機，因此原住民語的資料應該要是有多少就儘量蒐集多少。

2.2. 原住民族語的資料收集

湯愛玉教授

湯教授建議，原住民語言部分的田野調查應設置專門的推廣人員來進行，例如可以設成語言調查組底下的一個小組，這是因為原住民語言較多，再加上收集語料相對比其他國家語言不容易，因此特地設置一個專門的小組來進行工作會比較好。

齊莉莎教授

齊教授提到自己手上現有不少原住民語言相關資料（庫），或者未來可提供給國家語言資料庫使用。不過齊教授也提醒，除了想辦法蒐集新的資料以外，也要找相關人員來檢查原有語料庫裡不清楚的標記，並且重新檢視過之後才能公開。這是因為在校對詞表時，常常連原住民母語人士之間也會出現歧異看法，因此這點必須想辦法解決才能將語料對外公開。

齊教授進一步解釋，收集原住民語的困難點在於，一個族當中，不同地區的人可能會各自發展出次方言，而且隨著時間演進，這些次方言間的差異也會隨之擴大。所以，在收集原住民語料的過程中，必須標記地區，還有確認該方音的來源才可以。齊教授認為，因為次方言的不斷發展，再加上語言流失的因素，因此要完整了解所有原住民族語之（次）方言是件幾乎不可能的事。

2.3. 客語的資料收集

賴惠玲教授

目前，賴教授正在執行客委會語料庫建置的計畫，針對收集內容，賴教授提到胡萬川教授蒐集了許多閩南語、客語的故事集，客語部分如台中東勢的故事集。此外，客語語料庫建置時需要處理授權相

關的程序，在執行客委會語料庫建置計畫前，賴教授也已有多年蒐集的客語語料，但收進國家語料庫的語料，授權的程序都不能少，比語料產出本身要複雜很多。

2.4. 閩南語的資料收集

洪惟仁教授

針對閩南語的資料收集，以沒有經過分析的原始資料來說，洪教授表示已有許多資料，例如：臺語電影、胡萬川教授的閩南語故事集、楊允言教授的研究音檔（多以教會羅馬字書寫）。此外，洪教授也願意提供，在民俗歌謠部分，有傳統藝術中心出版之相褒歌 CD（資料範圍涵蓋出版當時的臺北縣 40 多處，文字檔及原像皆有），也提到如楊秀卿老師於廣播電台的作品等唸歌仔、亞洲唱片的出品等。

閩南語文字部分，因書寫系統關係，使得資料較少，但有閩南語雜誌、文學可作為書面語料，例如：《台語文摘》。

江俊龍教授

當代閩南語語料可納入綜藝節目、連續劇及訪談等。

劉昭麟教授

臺灣文學館收藏的全台詩、全台詞知識庫可考慮納入，其時間範圍大，資料也豐富。

高照明教授

語料庫的時間設定，希望能夠取得共識，但考量到資料取得、著作權問題及盡可能先羅列現有資料，想要詢問專家的建議。例如：教會公報最早的資料可追溯到清朝、閩客語典藏甚至有明朝嘉靖年間的語料。

劉昭麟教授

若以資料來源作為參考，劉教授舉例全台詩的時間橫跨明代到西元 1999 年。

洪惟仁教授

洪教授建議從近代開始收錄語料，在時代劃分上，日據時期的資料量不少，反而是戰後時期不若前期。此外，許多近代的資料已經過數位化（digitalized），像是高教授提到的教會公報，台日大字典亦有架設網站。

楊允言教授

羅鳳珠老師的網站有把胡萬川老師做的一系列的臺灣民間文學資料放進去，網址為 http://cls.lib.ntu.edu.tw/TFL2010/cht/cht_Article.aspx。

蔡素娟教授

雖然閩南語兒童語料庫的名稱是兒童，但其實是大人與兒童的對話，所以也可以納入閩南語語料庫。

程俊源教授之書面建議

程教授目前正在執行教育部「閩南語詞彙分級計畫」，期望建置一個平衡的閩南語語料庫，已完成斷詞與詞性標記，採中研院 CKIP 簡化版的詞性集，詞庫建置詞項數擬定為 100 萬。

以下是第二次專家諮詢會議中，專家們列舉之閩南語資料，以紀錄出現順序編排：

1. 1.《台語文摘》（洪惟仁教授）
2. 相褒歌，由傳統藝術中心出版（洪惟仁教授）
3. 閩南語故事集（胡萬川教授）
4. 唸歌仔（楊秀卿老師）
5. 亞洲唱片出品

6. 全台詩、全台詞知識庫，收藏於臺灣文學館
7. 教會公報
8. 台日大字典
9. 廖元甫教授語音辨識計劃所蒐集之音檔

另外，針對方言與腔調差異的語料，洪惟仁教授亦提供以下資源參考：

1. 洪惟仁教授於 2020 年出版之《臺灣語言地圖集》
2. 卜溫仁教授出版之《Mapping Taiwanese》
3. 張屏生教授的詞彙調查語料

2.5. 臺灣手語的資料收集

蔡素娟教授

在談論核心詞彙的主題時，回應張永利教授的發言，蔡教授表示不同的文化蘊涵不同的詞彙，而若以概念定義詞彙，臺灣手語已有很多資料。

3. 關於著作權問題

張永利教授

張教授建議，關於語料的著作權問題，或許可以成立類似中研院「智財組」的機構，專門來處理智慧財產權相關之業務。（蔡素娟教授其後也提到應該成立一個專門處理著作權的組。）另外，張教授也

認為，只要不侵害作者的著作人格權，通常就不會造成太大的問題，所以只要找握有著作權的出版社洽談即可。

章忠信教授

章教授認為，國家語言資料庫應該廣收資料，若買得到著作權的就先用買的，買不到的話可以再問問著作權握有方願不願意做提供。另外，如果將語料轉寫成文字後，有再額外加上註釋（annotation）的話，這個成果就算是利用別人的作品而成的著作，因為在加上註釋這個過程有智慧的投入。

章教授接著進一步補充，著作指的是，只要有包含創作的成分就算著作，因此口語也是著作的一種；沒有創作成分的成品只能算是「重製物」，所以有給提示稿的錄音也是重製而不是著作。

翁聖賢律師

針對著作權議題，翁律師提及應該考慮下列幾點：

(1) 是否會侵害到他人之權利？

(2) 我方權利受侵害時應如何處理？

(3) 屬於公眾領域（public domain）的資料，一但要「重製」，情況就不一樣，需要釐清著作權到底在誰手上。

因此，若想使用某些語料的話，建議還是要先找握有著作權的受讓人（assignee）洽談會比較好。

因為語料庫將會納入多媒體語料，翁律師提到可從屬於 public domain 的資料進行，視聽著作於公開發表 50 年後屬於 public domain，如果是將（閩南語）電影的音檔部分存成獨自的檔案，不會影響原本的著作權狀態。

洪惟仁教授

在討論已數位化的資料時，洪教授提到《台日大字典》網站之建置者應有相關著作權，在取用資料方面可與建置者洽談，是否能夠另存一份於文化部，以防網站斷線。

楊允言教授

楊教授於2003年開始其研究計畫，但語料都是沒有授權過的。楊教授也提到程俊源教授執行的教育部計畫，是詞頻統計與分析的研究，資料量已達到幾百萬的規模，來源為朗讀稿、台語連續劇、早期的歌曲，書面及口語各一半，書面是教育部出版品，臺語文學獎的資料也已授權給教育部。

賴惠玲教授

國家語料庫是國家級的語料，因此所有的語料上線後都是對外開放的，所有的語料都需要經過授權。賴教授在執行客委會語料庫建置計畫前也累積非常多語料，但在國家級的層次上仍然碰到授權的問題，書面文本不論是公私部門、出版社、個人，所有的授權程序都不能少，比語料產出本身要複雜很多。

因為胡萬川教授蒐集了許多閩南語、客語的故事集，閩南語部分有台中葫蘆墩的資料等。此外，賴教授亦親自與胡萬川老師重新取得授權、簽署授權書。胡老師覺得這不是他一人的成果，無法代表台中市政府或文化局的工作同仁，因為他是東勢客語故事集的 editor，我們查了條文才知道他可以代表授權，而且語料庫也非私人用途，而是予以公部門使用，才由胡萬川老師代表授權。

蔡素娟教授

蔡教授在建置閩南語兒童語料庫時，已將家長同意書留存下來，向家長說明使用項目，亦有電台的授權給計劃主持人的文件，至於轉寫屬於蔡教授及團隊的智慧財產權，可與中正大學一起同意，貢獻語

料。理想情形是如科技部條文所述，若計畫成果欲予政府部門無償使用，相關單位可來文到中正大學，由學校簽文。

張俊盛教授

張教授表示授權有範圍，也有可不可轉讓的問題，意即能不能給所有的研究者，甚至出版社，都是有範圍的。

4. 如何保存逐漸流失的國家語言

各專家學者們也針對母語的保存與推廣進行了討論。以下將針對這兩點作進一步說明。考量到現今臺灣各族群母語流失情況嚴重，在討論如何保存逐漸流失的母語時，不少專家學者都一致提出了將現有語料進行數位化典藏的重要性，另外也有專家提出一些關於語言調查與語料收集的見解，如下：

湯愛玉教授

首先，湯教授提到，將現有語料進行數位化應用是必要的，而且這項工作可以配合 AI 技術一起發展。另外，湯教授也提到，為了傳承的需要，國家語料庫應該要收集不同年齡層之影音資料，並將之作數位典藏，如此才能看到語言磨損的程度（attrition）。

郭志忠博士

郭博士也認為，將錄音資料數位化，可以延長其保存時間。然後針對錄音資料的收集與處理，郭博士也另外提出了一些見解。首先，郭博士認為，在自然情境下發生的聲音資料（spontaneous），才是最值得被收集的語料。然後，郭博士也提出了 SNR（訊雜比）的概念，訊雜比（SNR，signal to noise ratio）就是定義什麼東西是訊號，什麼不是，例如只要符合訊雜比（例如：15dB 以上）就可納為語料，其餘不符標準的聲音訊號，就可以忽略。定義訊雜比的優點是，讓語料收集

者不一定要進錄音室也可收集到一定品質的語料。郭博士認為，這個做法不但比較能夠收錄到更自然的（spontaneous）聲音資料，也比較尊重受訪者的意願，因為受訪者們不一定能夠抽空特地到錄音室錄語料。

齊莉莎教授

齊教授也認同上述郭博士的看法。齊教授表示，錄音資料固然有其效用，但很難收集，因為很多人會拒絕接受錄音。例如像是原住民的耆老，他們多居住在都市以外的地區，很難請他們特地離開家園、到別的地方錄音/影。而即使都市裡有不少原住民的年輕人，他們也可能已經不太會說族語了，所以無法邀請他們來錄音/影。

蔡素娟教授

蔡教授進一步區分數位典藏和數位應用，蔡教授認為這兩項工作應該要分開處理進行，才比較有清楚有效率。此外，除了現有語料的數位化之外，未來國家語言研究中心還可以透過「任務編組」的方式，來進一步進行語料收集、語料處理、各項調查作業等事項。

張永利教授

針對蔡教授提到的「任務編組」，張教授進一步提出可以將人員分成好幾組來處理各自的工作，如著作權組、資訊組、調查研究組（又可細分成普查組和收集資料組）、認證組……等等。

5. 語言地圖

洪惟仁教授

洪教授在語言調查已有三十餘年的經驗，認為語料蒐集後須整理音檔、粗資料後才能公開或收錄，例如：有些已轉寫成國際音標，有些則無。洪教授提到可參考卜溫仁教授（Warren Brewer）的著作

《Mapping Taiwanese》、張屏生教授的詞彙調查研究。洪教授於 2017 年亦執行了文化部的委託計畫，進行臺灣語言使用的語言地理分析。

針對客語的方言調查，有涂春景教授、張屏生教授的研究。客語方言調查為多點調查，沒有網狀調查，客家的語言地理學（研究）很少人做，張教授在屏東做的調查很詳細。在原住民族語部分，小川尚義的地圖應該比較完整，也對原住民族語有鑽研；鄭仲樺也調查了 90 幾個點，李壬癸教授也有相關的研究。

洪教授提到，語言地圖電子化是浩大的工程，先以文獻整理的方式建置分享系統與介紹。若再將語言地圖與語言調查的概念結合，除了語言調查的報告本身，也可放上所蒐集的語料音檔。針對時空等後設資料記錄，洪教授提到有些較早的原始資料沒有相關資訊，可能會有無從得知的情形發生。

張永利教授

原住民族語應該還沒有像洪惟仁教授以鄉為單位、那麼大範圍的調查，因為方言實在太多了，可能是 40 幾種，也可能更多，有散落的調查。

另外，張教授提到記音的正確性（accurary）問題與審閱校訂的必要，如第一次諮詢會議中齊莉莎教授所說，族語 e 樂園已有很多原住民資料。因族語不像閩客語，語言就有十幾種，方言非常多，須仰賴過去學者累績的研究，把舊有的資料整理，是國家語言資料庫可有所發揮之處，並非單純將舊有資料上傳，必須為正確性高的資料。

宋麗梅教授

原住民方言調查，以官方分類是 16 族 42 個方言，但不是很全面的分類。原住民族語言研究發展基金會今年初成立，其中一項業務是資料庫的規劃，執行原民會交辦的業務，文化部是主要的統籌者，協

助跨部會的討論，把族人、學者手中的舊資料、音檔整理，將歷史資料數位化。

在執行面來說，宋教授說明基金會與原語會（原住民族語發展學會）、原文會（原住民族文化事業基金會）將會密切合作，也可與原民台討論如何能夠有系統地取得影音資料以進行數位化。基金會是在年中制定下一年的計畫，從公部門的角度來規劃執行，若是能在第二年（前三年）制定郭博士提到的 SOP、格式、哪些詞彙需要收進來，是以字為本還是句子等，就能找相關的單位執行。

高照明教授

未來國家語言調查應補充大範圍、全台的調查，希望藉此擴增國家語言資料庫的內容，語言地圖與GIS規劃的發想其實是謝舒凱教授在去年中研院的研討會提到的想法。

劉昭麟教授

在處理數位人文資料時，近幾年常融合時空分析，同地方的語言也會因為時間而改變，加上空間的維度，或即使不做時空的分析，我們也需要有時間的考量，未來若有餘力就能做更精細的分析，而不會讓地理上的記音變得很混淆。

謝舒凱教授

因為聽到法令通過，謝教授提到長遠規劃語言與文化保存的願景，且考量這些資料較難取得，大多可能無法以個人團隊進行，希望由國家語言資料庫的力量往這個方向前進。在規劃安排上需要考量是幾年的規劃、哪個時間點可以做到什麼事情，理解需要整理舊有的資料，且語言一直在變動，而一個語言的生命需要長期持續的調查。

郭志忠博士

郭博士提到需要與國家語言調查執行單位協調、搭配，討論制定優先順序（語種方言）、檔案格式（時空資訊）、工作 SOP，並善用科技與群眾外包的力量，例如：在田野調查實際到訪時，如果設備有時空資料記錄、將 GPS 打開可自動化，郭博士同意洪惟仁老師所說，只是紀錄大地名是不夠清楚的，地理座標連結到地理資訊系統上也較容易。

此外，郭博士贊同 crowdsourcing（群眾外包）的方式，能夠同時、平行留存各地的方言，甚至也可主動設計，以互動的方式蒐集用語、音的差異。如果發音人在 app 或 browser 上提供語料，也可紀錄所在地點與當時的時間，將地理資訊紀錄自動化。

賴惠玲教授

賴教授提到，建議不用將時間訂得那麼固定，也可能是兩年、兩年、兩年的規劃，不一定是三年、三年，因為委員們已從經驗提出許多執行面的規劃，這個計畫是屬於大方向的規劃，我們已有共識之下，年限可以有彈性。

6. 國家語料庫的建置細節與閩南語語料庫規劃

高照明教授

《國家語言發展法》通過後，明定須建置國家語言資料庫，國家語言包括華語、閩南語、客語、原住民族語、臺灣手語、閩東語，這個計劃要做的是國家語言資料庫，華語的部分已建置很多語料庫，除了張教授提到的國教院語料庫，更早以前有中研院一千萬詞的平衡語料庫，接下來是由賴惠玲老師正在執行的客委會計劃，大概是一千六百萬字左右，亦有口語的語料，加起來約是一千萬詞。族語 e 樂園、宋

麗梅老師的台大南島語語料庫、齊莉莎老師也有很多資料，亦有小規模的語料庫。

宋麗梅教授

原住民族語發中心也有滿多資料可以參考，所有的業務資料都會轉到原住民族語言（研究）發展基金會[10]。

張俊盛教授

張教授建議可參考英國等成功案例，除了一億詞的規模，也須平衡與取樣，最好能夠代表語言的剖切面，每個方面都具有代表性。英國國家語料庫，沒有包括 Welsh，而是另外建置 CorCenCC – National Corpus of Contemporary Welsh。台語或閩南語，也是另外做一個類似 CorCenCC 的國家級的語料庫，以 Welshi 為例是 1 千萬詞。

在語言名稱方面，張教授認為閩南語、客語、原住民族語都是高度政治性的，並不是為了哪個族群伸張，或是拯救瀕危的語言。我的母語是臺語，但我的立場不是要為台語做些什麼，而是從英國國家語料庫的建置過程及臺灣的現狀來看，台語滲透在日常生活中，但純粹的台語可能很多人看不懂，摻雜不同的語言。（語料）來源可能是出版品，灰色地帶有網路社群的文字，要不要納入可能還需要討論，這是我們現在的語言，但很多人覺得不登大雅之堂，需不需要斷代也是另一個問題。

在執行方面，張教授表示英國國家語料庫的主導單位是英國國家圖書館，我們也有國家圖書館，也是一個可能性，參與的人大部分是出版界，也有英國的大學出版社，可以協助提供文字（語料）。不贊

[10] 針對原住民族語言研究發展中心的計畫，宋麗梅教授提到六年計劃已於 2020 年 3 月 31 日結束，可至語發中心的官網查詢研究報告案，往後相關語言研究之延續會由原住民族語言研究發展基金會接手繼續，並建置新網站，目前原住民族語言研究發展中心的網址為 http://ilrdc.tw/。

成 crowdsourcing，因為沒有權威性，也不贊成現在開始錄製資料，應以現有資料為主，在一兩年內可以完成，我們已經慢了英國國家語料庫至少二十年，但我們有經驗可參考，有科技可運用，出版界也將作品數位化了，我們做的時間應該不是六年、七年的時間。

張學謙教授之書面建議[11]

(1) 臺灣語料庫計畫應該愛是臺灣在地、社區導向的語料庫計畫，建議增加會當予各族群的民眾，用社區參與的語料庫收集方式。按怎邀請民眾參與，會當參考威爾斯的 Corpws Cenedlaethol Cymraeg Cyfoes 的例。

(2) 語料庫除了典藏以外，需要推廣伊的應用，特別是語言教學，透過真實語料，學習語言、使用語言，同時嘛會當透過語料庫進行臺灣社會、文化的探索。

(3) 各族群 kah 伊的語言名稱愛斟酌。既然是語料庫計畫，著愛用語料去探討臺灣人實際使用閣族群語言名稱的狀況，袂當講一套（語料庫反應社會的語言使用），做一套（無照社會名稱使用習慣，用政治考量掩崁社會語言事實）。咱用 google 搜尋的結果去揣，結果顯示閩南語語料庫：22,100 項結果；台語語料庫：114,000 項結果。建議採用民間多數的用法，通少愛用並列的方式呈現。

(4) 社會的日常口語、文字需要收入，通好是多媒體的方式。原住民語、客語、台語攏有文字，需要透過語料庫呈現國家語言的讀寫世界。建議加上：簡訊、email 收集、blog 的文件收集。

(5)《國家語言發展法》是臺灣語言正義的起步。臺灣的語料庫計畫需要回應語言正義的呼籲。轉型正義的相關文件、口語記錄需要收入

[11] 張學謙教授因無法與會，以書面提出建議。

語料庫。其他關係語言歧視、語言迫害的記錄，相關的語言自傳、讀寫的自傳嘛應該納入考量。

由於閩南語語料庫目前沒有專責機構負責，且已有許多閩南語語料散落各處，專家們討論如何將各來源之語料經過處理，以便往後提供檢索之功能。

首先，第二次專家諮詢會議中討論了閩南語語料的蒐集，相關列表請參考「2 關於資料的收集」一節。

高照明教授

以中研院平衡語料庫為樣本，並與客委會客語語料庫比較來看，希望是一千萬詞的規模，除了閩南語語料的蒐集之外，希望每一句能夠有華語的對應。閩南語語料的用字規範以教育部為主，未來執行單位亦須提出分詞標準及詞性集，皆奠基於中研院的標準之上，如果有不同的部分附加原因。

此外，高教授亦提出詞性數量的問題，希望有折衷的方法，因為太多詞性不利於機器學習，大概 10 類詞類，若是 30 幾種詞性可能需要幾百萬詞的語料才能產生自動分詞的程式。此分詞標準、詞性集，甚至是程式本身，都可以開放予使用者下載，預留空間給相關上下游的語音辨識與合成、機器翻譯等資訊發展，以及 e-learning 等的整合，釋出 open data，甚至團隊需要有整合相關的人員，如此一來便可建立一個語料處理工具的生態圈。

洪惟仁教授

《閩南語常用詞辭典》是否可以作為詞性標記的參考呢？詞類數目夠不夠？

江俊龍教授

江教授回顧《閩南語常用詞辭典》編撰的時空背景，提到當時著重的是釋義與例句的編寫，因此詞類數量可能不足夠，建議參考中研院的詞性標記集、目前正在建置的客委會語料庫等，據此提出新的詞性標記標準。

曾淑娟教授

回應詞類數目多寡的問題，中研院 CKIP 的詞性集分為三個階層，中間階層的版本約莫有 35 個詞類。除了中研院 CKIP 的詞性集，曾教授研究的 modeling 也採用了 universal POS 的共用詞性集。對於使用目的而言，CKIP 的詞性集除了應用於機器學習，也適合語言分析。建議執行單位提出有階層區分的詞性集，可先組成小組，研擬出 35 個詞類的草稿，再從這個較細的版本，濃縮出較精簡的詞性集，往後也毋需作出大更動。

由於 CKIP 詞性集是基於華語提出的，曾教授認同諮詢委員們的看法，認為華語的詞類標記標準無法直接、完全套用至閩南語（、客語或是原住民族語），必須根據各個語言的語法修正。在執行（implement）時，團隊編制須有閩南語的專家，與操作模型的成員搭配。

另一大主題是語音辨識與合成，曾教授提議可以將語言使用調查與口語語料蒐集結合，如此的考量是不必分成兩項工作，錄音的處理與應用也可同時進行，但最重要的是臺灣語言分布調查能夠讓我們了解語言分佈的概況、基礎母體組成等。

在言談標記方面，曾教授認為不需要在最初就標記言談功能，且依照研究需求，依據的理論都不同。主要是轉寫時，要把言談相關的詞語一併記錄，但是不需要類似 Du Bois 的詳細標記。

曾教授認為，轉寫工具統一是很重要的，可以 PRAAT 格式進行，日後可以很容易加工，製作其他標記。

至於書面語及口語語料的比例，曾教授認為口語比例需要再下修，建議可以朗讀口語多一些，自發性對話少一些。

郭志忠博士

郭博士贊成階層性的詞性集這個作法，也接著提到語音技術的發展。閩南語的語音辨識資源不若華語那麼多，以往分詞與詞性標記對語音技術來說很重要，但近五年到十年來，因為有了深度學習的發展，越來越少使用文字分析得出語言參數，再將其轉換並合成語音。過往的技術常被稱作 text-to-speech（TTS，文字轉語音），現今的技術則為 end-to-end TTS，但 end-to-end 技術可能會是更大的黑盒子，因為很難校正合成錯誤，因此語言分析處理仍有需要。

此外，分詞與詞性標記對於不同語種的轉換來說很重要，例如：華語和閩南語之間能夠翻譯，搭配分詞與詞性標記，以提升翻譯和語言轉換的品質，借助華語內容量（content）較大的優勢。閩南語內容量不足一直是臺語 TTS 的困難之處。在機器翻譯亦然，受限於 data resource 夠不夠，如果語料較少，可採用 knowledge-based/rule-based translation；語料較豐富時，可修正機器學習的結果，並進一步協助技術發展。

劉昭麟教授

分詞與詞性標記在語言教學上不可或缺，因為教育和 end-to-end 技術不一樣，若能夠做到底層的詞性標記，可以交給下一代更多。

賴惠玲教授

賴教授以客委會語料庫建置經驗分享口語語料處理的困難，包括蒐集、轉寫與詞性及言談標記等，所要投入的人力資源比書面多很多，

認為口語語料的比例可作些許調整，客語語料庫在執行過程發現蒐集來的語料偏向某些語類，因為早期的資料是熱愛閩南語、客語的文字創作者所投入完成的，不似華語從一開始就有書寫的文字，每天也有大量的報紙文字，因此要蒐集那麼多字詞的語料，口語與書面的比例無法以 50%、50%看待，因為轉寫需要時間及受過訓練的轉寫人員，在用字規範上，也需要經過處理及專家的討論。

蔡素娟教授

蔡教授於 1997 年開始建置兒童語料庫，處理語料的轉記、閩南語的詞碼建立，因為很多時候閩南語沒有對應的文字、本字，差不多有兩萬詞，蔡教授已有一二十年的經驗了。

蔡素娟教授提到，兒童語料庫有詞性標記，言談部分有一些 correction、repetition 等基本的言談標記，其他標記可由個人研究再處理，我們可以做的是 discourse 研究者覺得 crucial 的標記，或是分階段性完成。

張俊盛教授

國家語料庫的規模大，所有的註記都是自動的，包括 tokenization 和詞性，但轉寫的成本不低。

7. 語料庫跨語言檢索、核心詞彙與數位加值應用

除了保存逐漸流失的母語，專家學者們也針對如何推廣母語學習進行了討論。

郭志忠博士

首先，為了要促進國人認識並且學習各個國家語言，郭博士提出了跨語言查詢的概念。郭博士舉例，目前萌典（網址：https://www.moedict.tw/）已經做到某種程度上的跨語言查詢，即在網

站內進行字詞搜索時，搜索結果會同時顯現多種語言的結果。不過，目前萌典並不包含原住民語的跨語言查詢，倘若未來國家語言資料庫能納入華語、閩語、客語、原住民語、閩東語、臺灣手語等各語言的跨語系查詢，對於各國家語言的推廣與學習應該有很大的幫助！

郭博士提到，在做語音辨識或合成技術時，都會碰到詞彙的問題。以漢語來說，可能就是一字詞到最多四字詞，大部分是二字詞或三字詞，不是以詞頻來看待，雖然常用詞也有很多應用，因為常用詞不一定是核心詞彙，例如：Named Entity（命名實體）等專有名詞、人名的使用率很高，但不能算是核心詞彙，而是不限領域的詞彙，其他的專有名詞依此組合，像是工業技術研究院一詞不是核心詞彙，但工業、技術、研究院可能是，這部分的詞彙也是滿需要的。

湯愛玉教授

除了提到的語料數位典藏，湯教授提到數位化過後的語料又可再進一步做數位應用，以利大眾研究與學習。例如，湯教授建議未來國家語言研究中心公佈閩、客、原等各自語言的詞（頻）表供民眾參考，如此對於想考取語言認證的民眾應該會有一定幫助。

江俊龍教授

閩南語和客語都有漢字書寫不一致的現象，像是客語「打嘴鼓」在閩南語也有相對應的詞彙，如果文化部有跨部會的規劃，可針對兩個語言漢字書寫不一致的部分預先規劃。

劉昭麟教授

劉教授也認為詞彙對應的概念很好，且為了保存各國家語言的樣貌，除了以漢字書寫閩南語，也可採用拼音標示，因為閩南語大部分的書寫其實是為了標音。如果能夠取得原始的音檔，也可使用拼音標示。

高照明教授

以教育部詞典為基礎，加入平行語料，讓使用者輸入華語後能夠檢索到其他國家語言的內容。有關對照語料，客語有格林童話故事，希望能夠取得授權，將其納入翻譯語料。如此一來，不僅是詞彙對照，也可以句為單位對照。

此外，希望建立核心詞彙，同樣的詞彙、同樣的例句在不同的語言如何說。

張永利教授

原住民族語也需要有核心詞彙表，且因為文化不同，每個語言的詞彙不同，可能要參考謝舒凱老師做的 ontology，也就是概念式的詞彙，而非只是用詞彙來制定。

蔡素娟教授

蔡教授同意以概念定義詞彙，也表示臺灣手語已有很多資料。

張俊盛教授

張教授表示詞彙和例句是另一個概念，詞彙（辭典學）不是語料庫的問題，而是語料庫的應用。Kilgarriff 曾出版一本書專門討論 BNC，但有了語料庫，就有詞彙表，可以依此補充低頻詞。

8. 專家諮詢會議與文獻整理總結

關於國家語言資料庫要納入哪些國家語言資料、資料需要經過哪些處理，以及資料庫建置的優先順序與流程，各專家學者們已有諸多討論與重要共識。諮詢過程中已蒐集不少語言資料，對於資料的取得、授權也參考了客委會客語語料庫的建置經驗，進一步的語料處理與程序也有所討論，以下將會對專家提出的寶貴意見列點整理。

此外，綜合前述他國國家語料庫案例參考、芬蘭、美國和 Mozilla 的群眾外包作法，還有太平洋區域瀕危文化數位典藏計畫的建置方式，綜合出席專家諮詢會議的學者與書面建議，對於臺灣國家語言資料庫如何收集資料與後續處理的建議，可以整理出以下結論：

(1) 首先，張永利教授和章忠信教授建議先開設公開的工作坊（wordshop），廣收各式資料和意見，並且詢問各個資料提供者是否擁有資料著作權？願不願意主動提供資料？等。在蒐集專家委員們的寶貴建議後，目前語言資料庫及語料庫的可能資料已彙整至「伍、本國國家語言相關之語言資料庫」及「7.4 語言資料徵求及各項資源分享」兩章，並已展開相關授權意願初步詢問及了解須經過哪些處理方式，至於工作坊的開設，未來可再細分不同的 session 進行更細部的討論。

(2) 建立一個存放和整理後設資料（metadata）的資料庫，並把從(1)取得的語料之資訊通通納入。湯愛玉教授提到，資料應包含各個年齡層的語料，以觀察語言磨損（attrition）的程度。此外，各個國家語言的腔調、（次）方言的區分可能不是最完整的分類，例如：原住民族語 16 族 42 個方言是官方的分類，在後設資料的紀錄上應盡可能標示語料的腔調與方言。

(3) 成立一個專門處理著作權議題的單位，負責和(2)的資料庫理面提及的資料擁有者進行接洽，並商榷資料貢獻和著作權議題等等。賴惠玲教授從客語語料庫的建置過程提出授權相關的寶貴建議，認為授權的過程可能比語料產出本身複雜，需要預留足夠的時間取得授權，甚至須請智財權人重新授權。洪惟仁教授、翁聖賢律師提到，可考慮公開發表超過 50 年的閩南語電影，屬於公領域（public domain）的資料較無著作權爭議。

(4) 成立一個資訊小組，一旦(3)的著作權小組順利取得資料使用權後，資訊小組就可以開始對各種資料進行整理、格式統合、研發處理相關工具、並將資料放至國家語言資料庫上等各種技術性工作。資訊小組在整理資料時，建議可以參考太平洋區域瀕危文化數位典藏計畫的架構來整理資料。

(5) 因為現有的資料已具一定規模，優先順序應是先整合現有資料，再視已有資料擴增新資料，而整合現有資料的任務亦包括電子化、將相關檔案連結起來、檢審校對等前置作業。

(6) 不過，在了解現有資料的過程中也發現，對於較少資料的語言或語料，需要同時進行其他方式的資料收集，尤其是在回顧國外手語語料庫、資料庫的例子後，更應借助目前資訊技術的幫助、聾人社群對臺灣手語及文化的熟悉與了解，建置多媒體的臺灣手語語料庫。

(7) 在語言地圖與 GIS 的結合方面，洪惟仁教授表示願意授權相關資料，並邀請洪惟仁教授、張永利教授撰文介紹臺灣整體語言分布、原住民族語介紹（詳見「洪惟仁教授專文撰稿 – 臺灣的語種分布與分區」、「張永利教授專文撰稿 – 臺灣原住民族語言簡介」）。實行方面，在第一階段應先收集各個語言的現有資料並整理校對，以連結或檔案的方式典藏；第二階段可結合田野調查、國家語言調查的力量，蒐集較難取得或全台大範圍的資料，同時設計 GIS 系統，以互動、影音的方式呈現地理語音資料。

(8) 語料處理方面，曾淑娟教授認為中研院 CKIP 的詞性集可作為閩南語語料庫的基礎，並針對閩南語的語言特性調整。在詞類數量上，以 CKIP 中間階層的 35 個詞類為主，再濃縮成簡化的 10 類詞類，以提供機器學習與語言分析。委員們認為言談標記先從基本標記開始即可，轉寫、標記的工具力求一致，以便訓練標記人員。

(9) 在跨語言檢索方面，要處理書寫用字不一致的情形，建立對照詞表與華語對應。此外，可逐年釋出跨語言的核心詞彙，但並非僅以是否有共用詞彙而定，可參考語言認證、ontology 的作法。

以上的(1)到(6)是第一階段，即先盡量收集並整合各項現有資料，並針對較少資料的語言或語料先行收集，一旦第一階段的工作差不多到一個段落，而且國家語言資料庫也具有初步規模後，在擴增語料方面就可以開始參考並採用美國國家語料庫的協作開發計畫（Collaborative Development Project）、「同聲計畫」、還有太平洋區域瀕危文化數位典藏計畫的作法。

例如，之後可以在國家語言資料庫上上傳說明文件指引使用者將語料納入該典藏計畫。另外，國家語言資料庫也可以開發並釋出類似「同聲計畫」的網站或 APP，供民眾自行創建帳號，並貢獻語料，但必須有審核機制。

與此同時，語料校對、轉寫、標記將會是語料建置需要大量人力與時間的任務，包括確立分詞、詞性集及基礎言談標記的原則、選擇一致的轉寫與標記工具、開發輔助工具、資料與程式開源等，需要紀錄各個語料的處理完成項目。在法律方面，除了資料蒐集階段的授權處理，敏感資訊的去識別化是後續階段的另一課題。

9. 專家諮詢會議與會專家

第一次專家諮詢會議

時間：2019 年 11 月 11 日

地點：臺灣大學外語教學暨資源中心 207 研討室

出席專家學者：

蔡素娟教授（閩南語、臺灣手語）、齊莉莎教授（原住民族語）、張永利教授（原住民族語）、湯愛玉教授（原住民族語）、章忠信教授（著作權法）、郭志忠博士（華語、閩南語自然語言處理）

計畫主持人：高照明教授

協同計畫主持人：黃子桓博士

協同計畫主持人：翁聖賢律師

第二次專家諮詢會議

時間：2020 年 4 月 30 日

地點：受新冠肺炎影響，採 U meeting 線上會議

出席專家學者：

劉昭麟教授（華語自然語言處理）、曾淑娟教授（華語自然語言處理）、江俊龍教授（客語）、洪惟仁教授（閩南語）、郭志忠博士（華語、閩南語自然語言處理）

計畫主持人：高照明教授

協同計畫主持人：翁聖賢律師

第三次專家諮詢會議

時間：2020 年 5 月 18 日

地點：受新冠肺炎影響，採 Google Meet 線上會議

出席專家學者：

江敏華教授（客語）、吳靜蘭教授（原住民族語）、宋麗梅教授（原住民族語）、洪惟仁教授（閩南語）、張永利教授（原住民族語）、張俊盛教授（華語自然語言處理）、郭志忠博士（華語、閩南語自然語言處理）、曾淑娟教授（華語自然語言處理）、楊允言教授（閩南

語自然語言處理）、葉瑞娟教授（客語）、劉昭麟教授（華語自然語言處理）、蔡素娟教授（閩南語、臺灣手語）、賴惠玲教授（客語）、謝舒凱教授（華語自然語言處理）、張學謙教授（閩南語，書面意見參見註腳 11）

計畫主持人：高照明教授