文化部「建置國家語言資料庫先期規劃研究」勞務採購案期中報告

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 1 月 30 日

目次

圖目錄

表目錄

簡介

民國 108 年 1 月，國家語言發展法公布，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。」而第八條又提到：「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。」

若以上述這幾點為基礎，目前台灣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台灣的語言多樣性。因此，未來台灣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、活動資訊等等，各種語言相關資料。

以下，本勞務採購案期中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的設計理念、使用到的工具技術、收集方式、應用層面等資訊；接著，再彙整台灣各專家學者對於本土語言資料現況、還有如何設置國家語言資料庫，所提出來的寶貴意見；最後，本期中報告會依照前述各項說明，提出規劃國家語言資料庫的用途以及使用對象、規劃國家語言資料庫的內容項目、研擬各種授權書及授權機制草案，等建議。

1、世界各國家語言資料庫的現況分析

本章節將對世界各國家語料庫的現況作簡單介紹。首先將針對英國、美國、日本、中國等較具代表性，或是與我國建制國家語言資料庫相關性較高的語料庫作較詳細的說明，接著再補充歐洲和其他地區的國家語料庫介紹。本章節特別針對英國、美國、日本、中國國家語料庫作詳細說明的原因如下：

首先，選擇英國國家語料庫是因為，這是世界上的第一個國家語料庫。另外，該語料庫語料選擇較為平衡，規模也達到一億詞，包含書面語和口語，且有詳細的標記，因此對於我國建制國家語料庫具有指標性的參考意義。

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，COCA），考量到這些語料庫不一定是官方所設，加上其相關資料的說明也不像開放美國國家語料庫齊全且詳細，因此這些語料庫最後並沒有被納入本章節作介紹。另外，雖同為英語相關語料庫，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。

選擇日本國立國語研究所的語言資料庫主要原因有二：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，這點和台灣未來想成立國家語言研究中心的目標剛好重合，因此日本的模式值得台灣作參考。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；不過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，這點也和台灣未來想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，是目前各國家語料庫中最值得我們參考的。

最後，選擇中國國家漢語語料庫是因為，中國國家漢語和台灣的華語、閩南語、客家語皆同屬漢語語系，因此該語料庫的內容設計也值得我國做參考。

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

英國是第一個設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，其設計原仍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 億單詞。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）的各種類形的書面（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。

該語料庫是由牛津大學出版社所領導的 BNC 聯盟（BNC Consortium），詞典出版商 Addison-Wesley Longman 和 Larousse Kingfisher Chambers，牛津大學計算服務（OUCS）的學術研究中心，蘭開斯特大學計算機語言學研究中心（UCREL）和大英圖書館的研究與創新中心等成員所創置。團隊找到了合適的文本並且確認使用許可後，即將資料轉換成機器可讀模式，並且添加各類標記。

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」之標準方面，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%）。每筆語料經過挑選過程後也以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 W. &R。Chambers），兩所大學（牛津大學和蘭開斯特大學）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，亦可自學；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，因為語料庫的建置是經過規劃的，是精華的部分。

1.2. 美國國家語料庫—The 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，因此促成本語料庫（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 2200 萬個詞，與其他英語的語料庫相比規模不算大，例如，現代美式英語語料庫(COCA)的總詞數已多達 5 億 6 千萬詞還再繼續擴增。

本語料庫一樣採用書面語料（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、以及學術用語。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 本語料庫之語料不大，但是其語料具有較多標記，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。

美國國家語料庫是個協作開發計畫（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。該語料庫網址為：http://www.anc.org/。

美國國家語料庫目前收錄了 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 萬個英語單詞的當代美語語料，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 個單詞，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 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（英語：National Institute for Japanese Language and Linguistics，簡稱 NINJAL）

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：

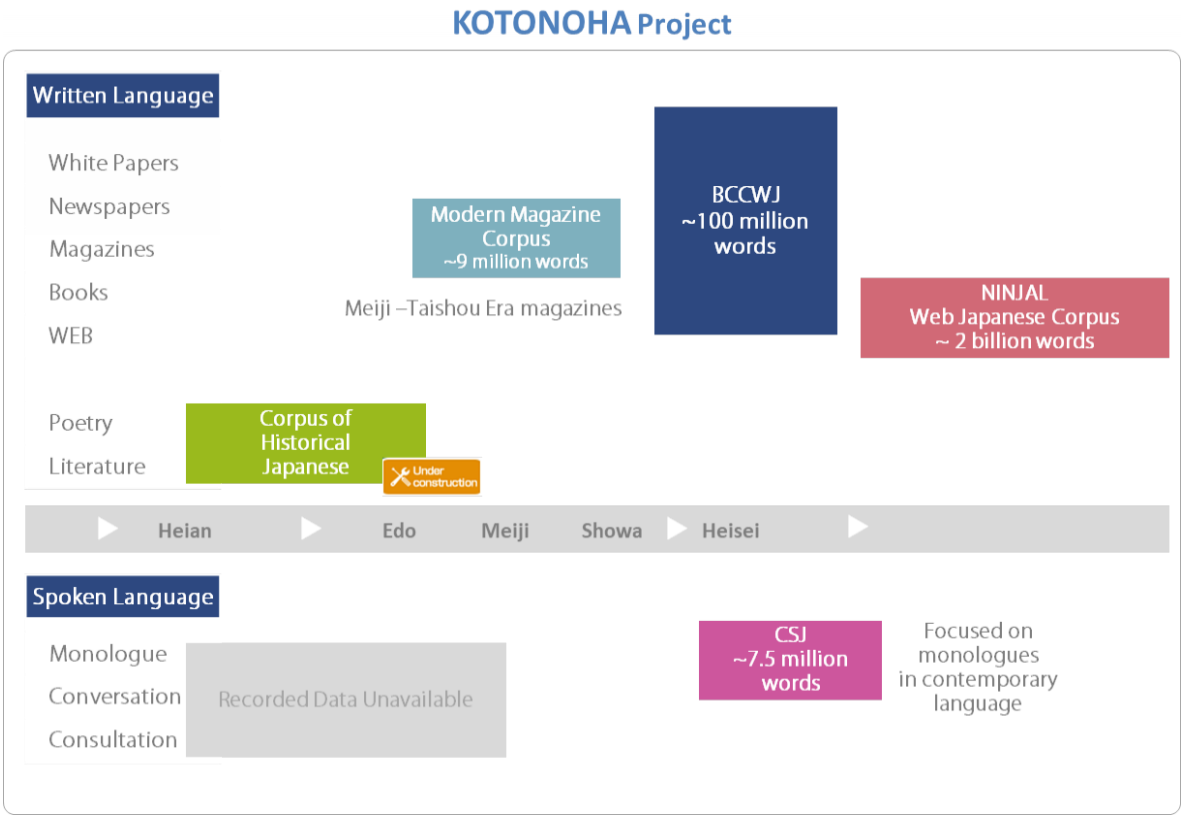


圖 1. KOTONOHA 計畫英語版和日語版簡圖之一

圖 2. KOTONOHA 計畫英語版和日語版簡圖之二

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。1.3.1 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）--再作進一步介紹。

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/。

1.3.1. 現代書面日語平衡語料庫（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)。各類別的字數語所佔比例如下表：

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…

在技術方面，日本國家語料庫採用 XML 格式來編譯。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（圖片位址：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。

1.4. 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設置的，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，這是因為網站上並沒有提供之後加入新語料的更新數據。另外，這五大類別又可以再細分成 56 個小類別，如下圖表是中國現代漢語語料庫標計類別。

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%

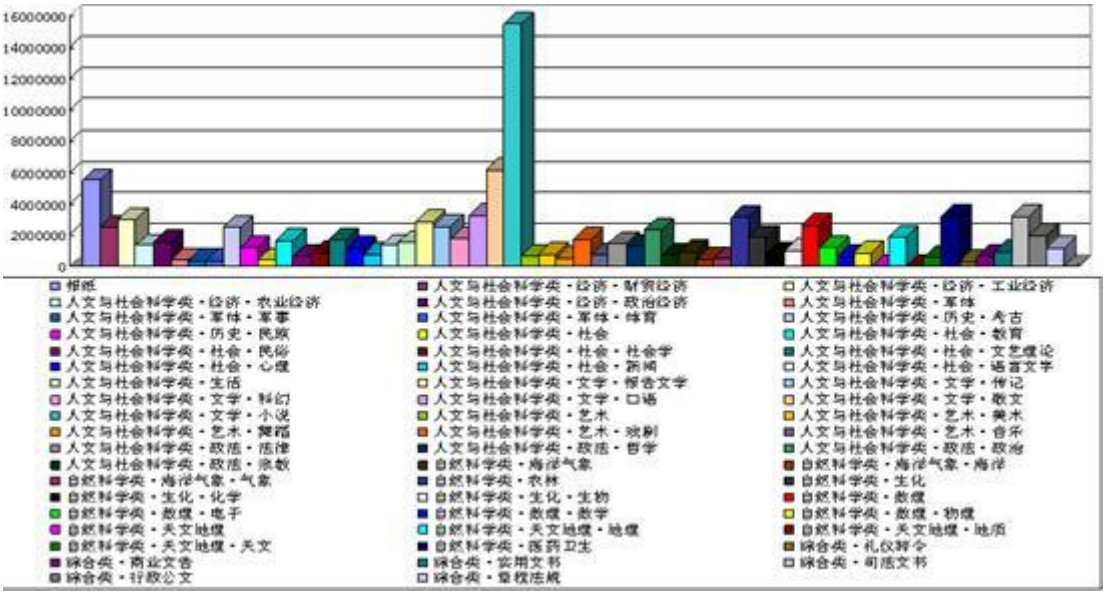


圖 4. 中國現代漢語語料庫 56 個小類別（圖片位址：http://corpus.zhonghuayuwen.org/Resources/cccorpusintro.files/image002.jpg）

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。完整的標記如下表，語料標記圖例如下圖。

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. 歐洲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 萬個單詞，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 億個單詞。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) 該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，如下圖。

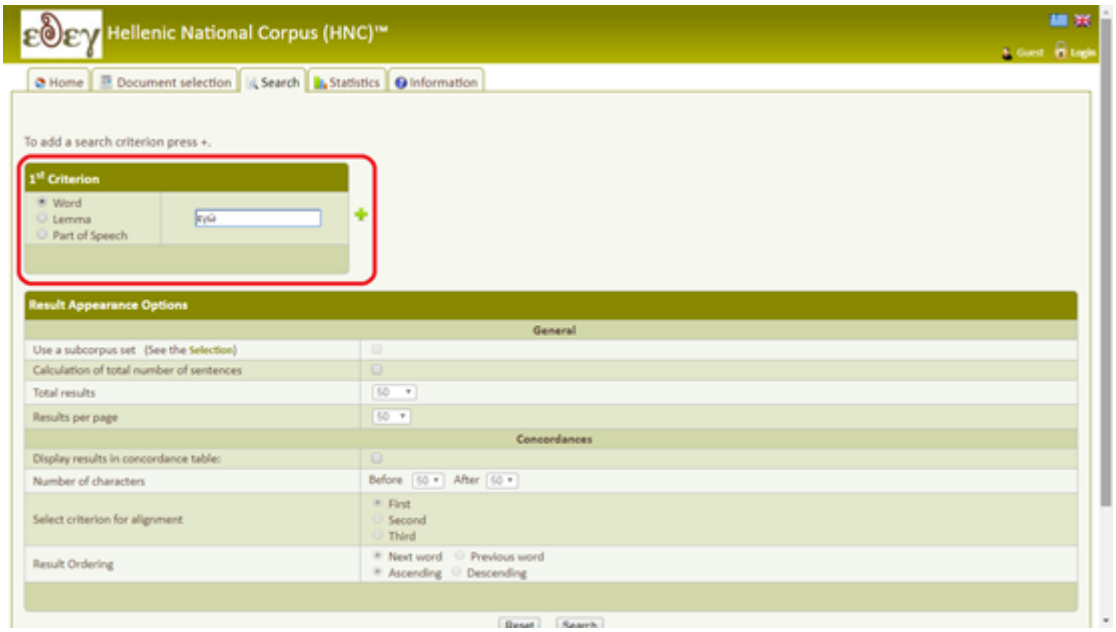


圖 6. 希臘國家語料庫查詢頁面之一

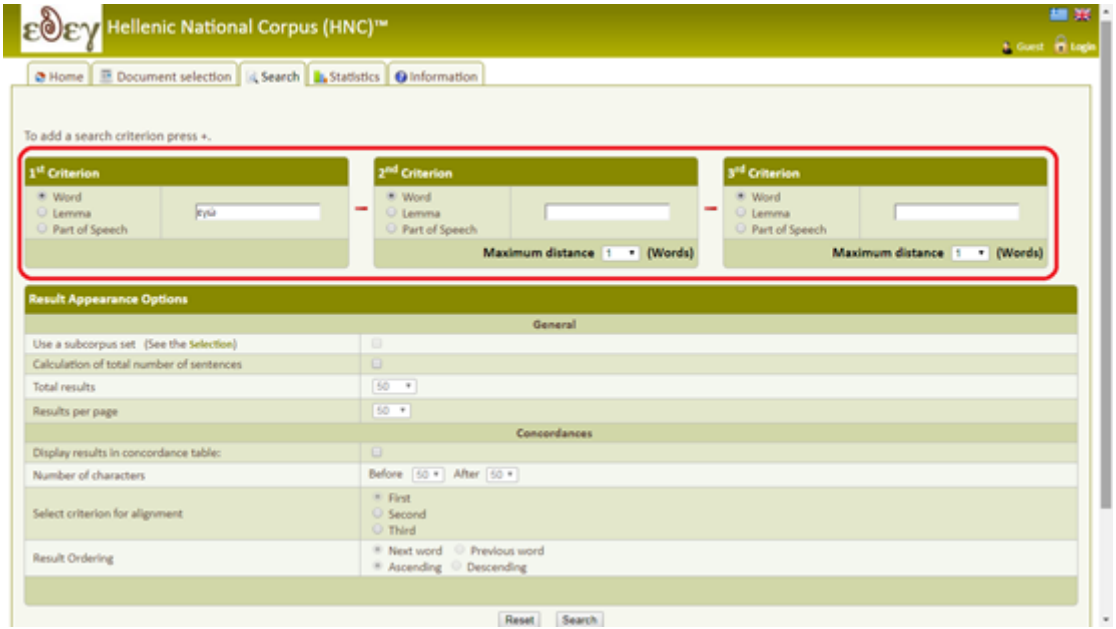


圖 7. 希臘國家語料庫查詢頁面之二

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。前包含 1.876 億個單詞。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（CorCenCC – National Corpus of Contemporary Welsh）

該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，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/。

1.6. 其他設有國家語料庫的國家

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)、除了詞彙和文法，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**	**詞佔比**

		Number of tokens	**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形態學研究之外，還可以作為形態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字，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，然後再做增減。語料庫的每個字形（wordform）都會被標上四種形態學的相關資訊，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 形態學詞典作為基礎。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，將多方來源的語料搜集，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，這樣子的建置架構與內容值得擁有多種國家語言的台灣來學習。

從 2012 年起，澳洲國家語料庫揀選了 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 年代搜集的語音資料，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，其中逐字稿的文字已與時間相互對照，也標有說話者，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，不應再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「資料交換的搜集與服務格式（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. 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. 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搜集，但值得一提的是，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 萬個詞單字量，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. 其他

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)
http://web-corpora.net/AlbanianCorpus/search/
(3) 克羅埃西亞國家語料庫(Croatian National Corpus)
https://web.archive.org/web/20060424031437/http://hnk.ffzg.hr/
(4) 喬治亞國家語料庫(Georgian National Corpus)
http://gnc.gov.ge/gnc/page
(5) 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

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

(7) 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

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

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

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

(11) 土耳其國家語料庫(Turkish National Corpus (TNC))
https://www.tnc.org.tr/

2、國外手語資料庫的分析

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手語為國家語言之一，而手語資料庫因為該語言的特性，在建置上須考量圖片、影像檔及文字說明等特殊處理。關於手語資料庫的規劃，現階段可諮詢中正大學語言所的手語實驗室，其已有多篇碩博士論文產出，亦釋出手語資源，網址為：http://tsl.ccu.edu.tw/web/。

在國外手語資料庫部分，美國設有國家單位「美國國立手語與手勢資源中心（National Center for Sign Language and Gesture Resources, NCSLGR）」，而康乃狄克大學的手語資料庫則搜集各年齡層的語料，且因手語的轉寫（transcription）方式還未有一致的原則，該實驗室製作了 ID 查詢表（ID glosses），讓不同轉寫方式間可以互相轉換，並提出 ID 查詢表應成為實務上的最佳典範（best practice），由於手語資料庫的語料搜集更是龐大的工程，其建置理念值得參考。奠基於 ID 查詢表的概念，全球手語資料庫（Global SignBank）於 2018 年發表了操作手冊，可參考其步驟及說明。另外，ASL-LEX 雖然不是語料庫，而是詞彙庫，但該網頁以資料視覺化的方式呈現手語詞彙，反映該語言象似性（iconicity）很高的特色，並加上受試者對各詞彙的評分，可延伸至心理語言學的範圍。

2.1. 美國國立手語與手勢資源中心（National Center for Sign Language and Gesture Resources (NCSLGR)）

美國手語研究計畫（American Sign Language Linguistic Research Project，縮寫為 ASLLRP）由布朗大學執行，該計劃的目標主要是：(1) 研究美國手語的結構，包括語意、句法及韻律 (2) 與資工背景的學者合作，開發手語的辨識與產生器 (3) 開發多媒體工具，以供手語研究使用，能夠存取與分析一手的手語語料。

在該計畫的第三項目標之下，該中心建造了手語與手勢資源語料庫（National Center for Sign Language and Gesture Resources corpus，縮寫為 NCSLGR corpus），網址為：https://www.bu.edu/asllrp/ncslgr-for-download/download-info.html。該語料庫收集了手語資料，資料形式為影片檔，且顧及不同的拍攝角度、正面近拍，並附有標記的 XML 檔案。迄今，該語料庫已有 1,866 個詞形（word type）以及 11,854 個字（word token）；以句子的單位來看，該語料庫含有 1,002 個即席敘事中的句子，以及 885 個獨立的句子，搜羅不同句構的句子。雖然已有 XML 檔案，但尚未將其轉換為可以搜尋的模式，不過也因為採用了 XML 的檔案形式，資料已經可以讓訪問者下載、遵循一致的共識進行資料處理，甚至有 Python 的程式碼壓縮檔以供使用，連結為：https://www.bu.edu/asllrp/ncslgr-for-download/signstream-xmlparser.zip。

奠基於上述的語料庫資源，該中心又進而建立了美國手語影像詞典（American Sign Language Lexicon Video Dataset，縮寫為 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負責攝錄，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。

值得一提的是，此兩個語料庫的標記處理使用的是 SignStream® 軟體，並匯集到高立德大學（Gallaudet University）所建置與維護的資料庫存取介面（Database Access Interface）系統內。

2.2. 康乃狄克大學手語與語言習得實驗室（Sign Linguistics & Language Acquisition Lab at University of Connecticut）

康乃狄克大學手語與語言習得實驗室主要計畫為「手語的學習、標記、典藏與資料分享（Sign Language Acquisition, Annotation, Archiving and Sharing）」，合稱為 SLAAASh 計畫，美國手語資料庫（ASL SignBank）為此計畫的成果之一。因為該計畫期望手語的資源可以共享，因此除了搜集各個年齡層的手語語料之外，該計畫亦需要處理當初錄製手語人員的同意，例如兒童的手語資料當初是在該實驗室的其他計畫下搜集而成的，由於此份語料搜集時的目的已有不同，因此必須向當時的受試者取得同意，進一步讓該資料能夠有延伸的用途，進行焦點團體訪談（focus group），以取得社群的支持，並擴及至未來可能納入的手語者、家屬與聾人社群的更多成員。

在語料標記上，雖然該資料庫尚未有如同美國國立手語與手勢中心龐大的軟體系統，但該資料庫考量到手語目前尚未有一致的書寫紀錄系統，而提出應該將各個書寫最小的單位編號，給予 ID，目前已有 ID 匯編表（ID gloss system）。除了康乃狄克大學，荷蘭奈梅亨拉德堡德大學（Radboud University）的 NGT（Nederlandse Gebarentaal

Corpus）、倫敦大學學院（University College London）BSL（British Sign Language）以及芬蘭于韋斯屈萊大學（University of Jyväskylä）的 FinSL（Finnish SignBank）皆使用了 ID 彙編表系統。

2.3. 全球手語資料庫（Global SignBank）

全球手語資料庫的前身為荷蘭奈梅亨拉德堡德大學（Radboud University）的 NGT 資料庫（Nederlandse Gebarentaal Corpus），網址為：https://signbank.science.ru.nl，該資料庫所搜集的語料庫皆有各自的授權方式，並將開發過程存於 GitHub 上，連結為：https://github.com/Signbank。

2018 年，該研究團隊發表了全球手語資料庫的操作手冊 (Crasborn et al., 2018)，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，該資料庫的系統即可即時更新最小對立體（minimal pair）的資料。

2.4. ASL-LEX

ASL-LEX 是美國聖地牙哥州立大學（San Diego State University）、塔夫茨（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。

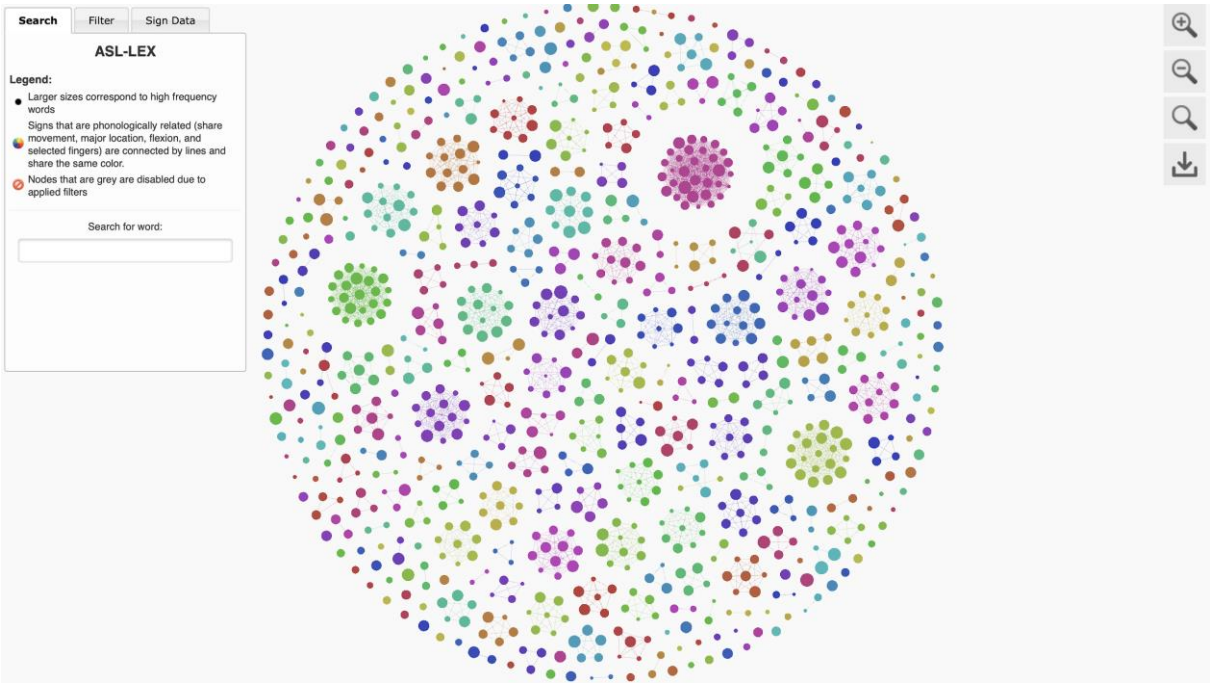


圖 8. ASL-LEX 將手語詞彙視覺化，每個原點代表一個詞彙。

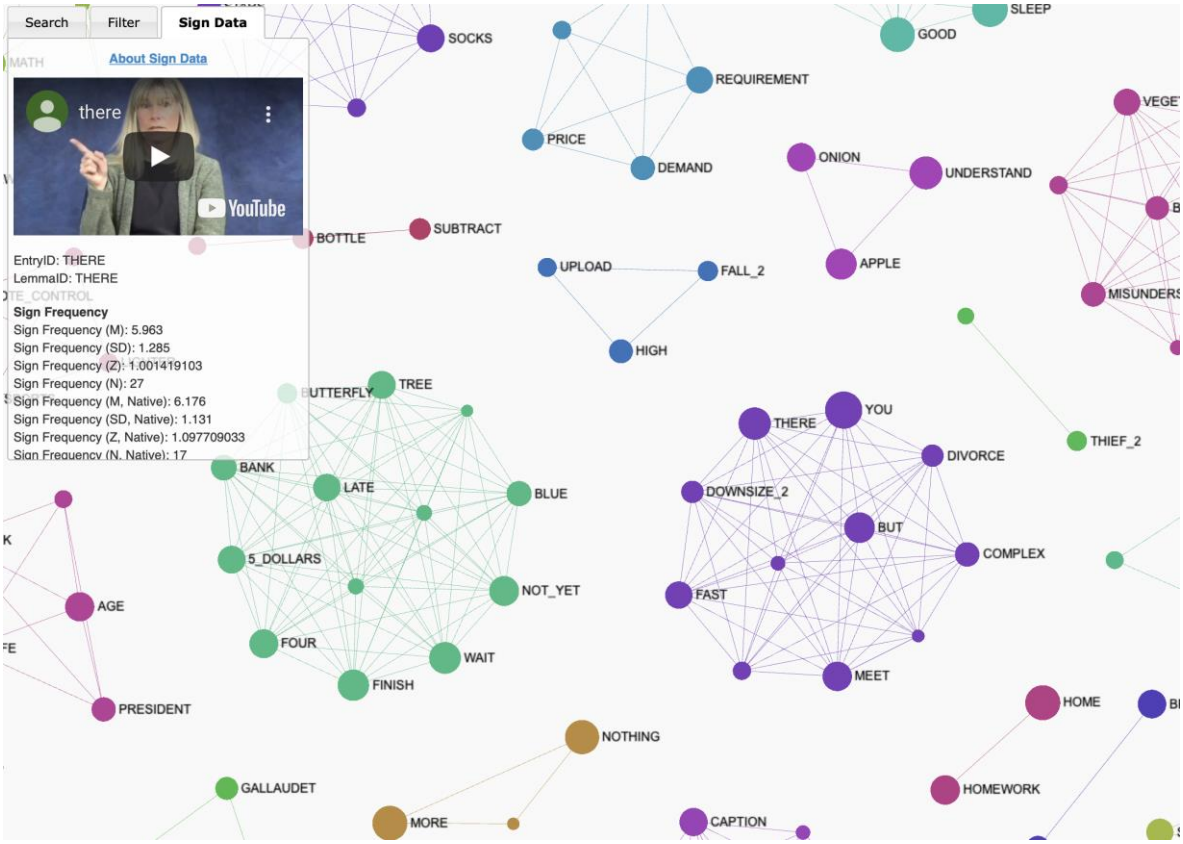


圖 9. 在 ASL-LEX 中檢索「there」一字的結果畫面。

該研究計畫搜集了 1,000 個手語手勢，並附上 (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) 影像檔時間長度 (4) 是否為複合詞等標記資訊，這些標記資訊顯示出該詞彙庫承載著心理語言學的背景，從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。整個資料庫可以 csv 檔下載，包括標記資料、英文翻譯、鄰近詞、評分實驗結果等。

其中，象似性（iconicity）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。該資料庫的像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），這些受試者都不會手語，以避免手語使用者有時直接將英文字母打出來，卻認為該詞彙具有很高的象似性。

在資料視覺化方面，每個節點代表一個詞彙，原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，包括手指姿勢（selected fingers）、手指是否彎曲（flexion）、動作（movement）及手勢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，在檢索時除了點擊特定原點查看該詞彙的詳細資料及鄰近詞外，亦可設定檢索條件顯示特定手指姿勢、動作等，了解哪些詞彙符合該檢索條件。

3、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 Howey)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。以下僅列舉芬蘭、美國國內語料庫和 Mozilla 公司《同聲計劃》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. FIN-CLARIN

CLARIN（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. 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。

3.2. 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. ONAC

貢獻語料的流程（Contribute Texts）

(1) 確認語料內容格式是否符合 ANC 所訂定的條件（詳見 8.2.1.2 小節），並閱讀許可協議。許可協議內文如下：

Grant of license

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.

(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。

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. 使用格式正確的 XML。

f. 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，有關著作權議題可以參照以下網站頁面： Frequently Asked Questions、A brief intro to copyright、10 Big Myths about copyright explained。

3.2.2. MASC（**Contribute Data and Annotations**）

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語料和標記工具（Data Download、Tool Downloads），再按網站建議的格式進行加註。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 進行連繫。

3.3. 群眾外包—「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，故每一種語言約需搜集來自不同人、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 10. Common Voice 語言計畫規格：以台灣腔華語為例

圖 11. 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。

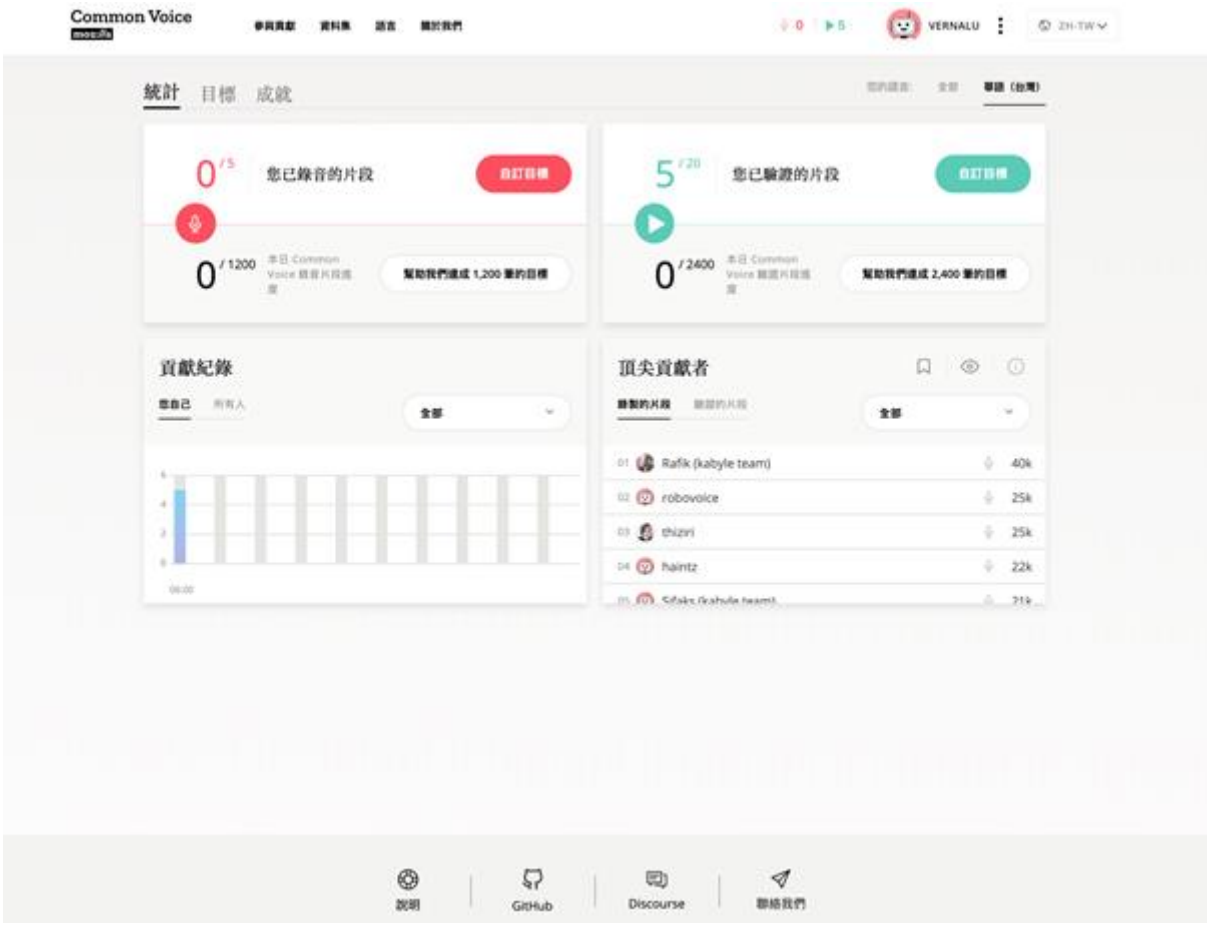


圖 12. 貢獻者在網站上創建帳號之後，就可以擁有自己錄音和驗證的所有記錄

圖 13. Common Voice 音檔資料收集流程

圖 14. 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 15. 志願者朗讀隨機跳出之例句，錄音之後等待他人驗證

3.4. 群眾外包—亞馬遜 MTurk 平台

Munro et al. (2010) 利用亞馬遜 MTurk 平台，發現可以使用群眾外包形式進行下列的主題，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元的回答、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，尤其是在已搜集的資料呈現偏態（skew）之時。此外，Ortega-Santos (2019) 亦透過 MTurk 平台搜集西班牙語系國家的語料，該計畫利用兩週的時間搜集了 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。

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。

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搜集所需的時間，且語料來源可能更加廣泛。

4、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，是有經驗可作為參考的。

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）：

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：姓名標示-相同方式分享 4.0 國際（ShareAlike 4.0 International License）規範使用者可自由以任何媒介或格式修改、重製及散佈某資源，並得用以商業用途。惟使用者須遵守以下條件：(1) 須依任何原授權人要求之方式標示出其姓名 (2) 須有著作權聲明 (3) 須表明是否有修改過該資源，並標示出做出修改的地方 (4) 若要分享此修改過之資源，須同樣使用此一授權方式（或此授權方式之後續版本）授權給其他的使用者，且不得用任何方法限制此授權方式保障之大眾對此資源的使用權。）

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。以下為 Nabu 系統的使用流程圖：

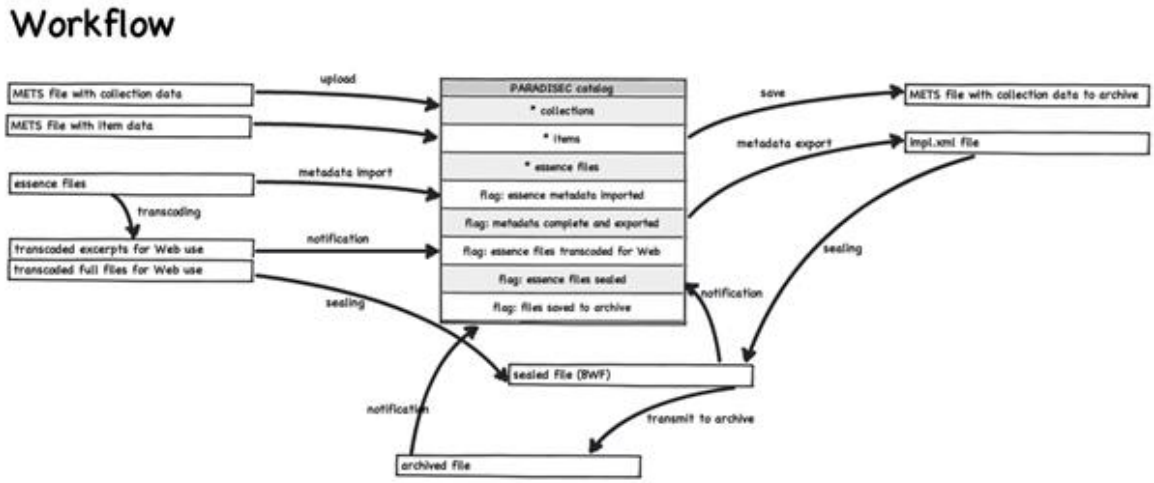


圖 16. Nabu 系統的使用流程圖

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 3.5 章節）、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，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 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「資料交換的搜集與服務格式（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 17. PARADISEC 使用 OAI-PMH 及 OLAC 架構下的 API 串接內容

上圖中的 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）」、「語料貢獻者搜集語料時的角色（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。

4.2. 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（下圖），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）。

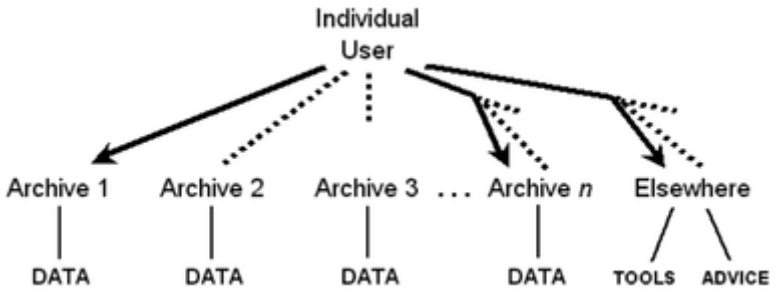


圖 18. 語言資源的分散讓使用者難以查詢（Bird & Simons，2003）

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）：

貢獻者（contributor）：貢獻此資源的人

範圍（coverage）：地理與時間上的範圍

創造者（creator）：實際創造此資源的人，如原住民語的發音人

日期（date）：資源創造循環中的重要日期

描述（description）：關於資源內容的描述

格式（format）：數位格式

指稱（identifier）：如網路連結或 ISBN 等明確的資源指稱

語言（language）：資源內容的語言

出版者（publisher）：使該內容公開的出版者

相關資源（relation）：相關資源的連結

相關權利（rights）：此資源的權限人

來源（source）：此資源的來源

主題（subject）：此資源的主題，以關鍵字描述

名稱（title）：此資源的名稱

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，亦能夠對於台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. 語言代碼國際標準

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（荳蘭阿美語），以下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

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	荳蘭阿美語		ais			
Atayal	泰雅語		tay			
Austronesian languages	南島語系	map		map		語言集合
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	台灣南島語族	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	台灣自然手語		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			

5、專家諮詢會議重要結論

本章節是 2019 年 11 月 11 日召開的專家諮詢會議上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母語、語料庫跨語言檢索與數位加值應用。最後，在 5.5 小節將針對各專家的意見們作統合與補充建議。

5.1. 關於資料的收集

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台灣本土語言的資料等。

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本土語言資料納入國家語言資料庫，這樣的做法會比較恰當。

章教授整裡 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，關於第 5 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搜集 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。

郭志忠教授

除了一般的書面語口語資料外，郭教授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，郭教授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搜集多少。

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，除了想辦法搜集新的資料以外，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。

5.2. 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。

5.3. 如何保存逐漸流失的母語

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。考量到現今台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）。

郭志忠教授

郭教授也認為，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，郭教授也另外提出了一些見解。首先，郭教授認為，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，郭教授也提出了 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。郭教授認為，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教授的看法。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。

5.4. 語料庫跨語言檢索與數位加值應用

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。

郭志忠教授

首先，為了要促進國人認識並且學習各個國家語言，郭教授提出了跨語言查詢的概念。郭教授舉例，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、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！

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。

5.5. 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本土語言資料，各專家學者們認為能收集到的資訊畢竟有限，因為可能民間還散落了不少語言資料，但大家並不完全清楚。因此，後來專家學者們在此項議題的討論上，就轉而著重在如何收集資料，還有收集資料時可能會碰上的著作權議題這兩點上。

綜合前述芬蘭、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，對於台灣國家語言資料庫如何收集資料的建議，可以整理出以下結論。

(1) 首先，可以採用張永利教授和章忠信教授的建議，先開設一個公開的工作坊（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。
(2) 建立一個存放和整理 metadata 的資料庫，並把從(1)的工作坊收集到的資訊通通納入。
(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。
(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。

以上的(1)到(4)是第一階段，即先盡量收集並整合各項現有資料，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，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。

另外，雖然目前已經辦過專家諮詢會議，蒐集了許多寶貴建議，但難免還有疏忽之處；而且即使未來想要再次舉辦會議來收集資料，也不一定能確保各專家學者皆能撥空參與。因此，建議之後可以考慮採用視訊會議，或者線上問卷的方式，來收集專家學者們的意見。在線上問卷的部分，可以依照領域分別編列一系列不同主題的線上問卷，然後再請各專家學者們填寫。相信如此可以蒐集到更完整的意見，提供未來建置語料庫作參考。

6、盤點各本土語言資料庫

以下整理的資料可以列入國家語言資料庫。本章節先依華語、閩南語、客語、原住民語、手語區分各語別資料，接著再分別以線上資料、紙本資料小節來做呈現。線上資料包括語料庫、線上辭典等；紙本資料則包括詞典、傳說歌謠集、學習教材、童話集、得獎作品等。

6.1. 華語

6.1.1. 線上資料

(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詞庫小組所建置。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，口語語料庫一千萬六百萬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發展的華語語音語料庫，已建置 2 千多個小時的語料庫且仍在持續建構中。

(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 個小時聲檔與文字轉記檔及標記。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，相關說明可參考作者本人的介紹網站：http://www.ling.sinica.edu.tw/v3-3-1.asp-auserid=20.htm。

(6) 本土語言資源網：為教育部所整理製作的網站，裡面整合了華、閩、客、原等台灣各本土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。

(7) 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現當代。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。類似的彙編計畫如美國的「美洲哲學組織圖書庫（The Library of the American Philosophical Society, APS）」，該網站提供搜尋功能，可

依照關鍵字檢索，或是點選美國地圖中的區塊，查看關於該地區的作品條目。

6.1.2. 紙本資料

(8) 各種華語學習資源：如（a）教育部重編國語辭典修訂本（b）教育部國語辭典簡編本（c）教育部國語小字典（d）教育部異體字字典（e）教育部成語典...等等。

6.2. 閩南語

6.2.1. 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。該計畫語料的主要來源為網路上的新聞報導，

原文為華語，並翻譯成閩南語拼音。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) 臺灣白話字文獻館： 本計畫為台灣師範大學台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/。

(5) 台語文語詞檢索：由台中教育大學台灣語文學系楊允言所設置。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。

(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、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。

(7) 台語文記憶：由台中教育大學台灣語文學系楊允言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。

(8) 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。

(9) 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。

(10)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。

(11)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。

(12)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。

(13)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/。

(14)本土語言資源網：為教育部所整理製作的網站，裡面整合了華、閩、客、原等台灣各本土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。

(15)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/。

6.2.2. 紙本資料

(16)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。

(17)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，2-台中縣民間文學集 13 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。

(18)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、102 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。

(19)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。

(20)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。

6.3. 客語

6.3.1. 線上資料

(1) 國立政治大學客語口語語料庫：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，亦追求發音人的性別平衡。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：除了東勢（大埔）腔調，四縣、海陸、饒平及詔安等腔調亦已各累積了最少 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) 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。

(5) 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/。

(6) 客英大辭典：傳教士 D. MacIver 及 M. C. MacKenzie 所編撰的版本，初版在 1905 年問世，增訂版於 1926 年發行，其特色是紀錄了嘉應州（今梅州縣）與潮州一帶的客語使用。目前中央研究院的閩客語典藏網站有提供線上版辭典查詢，網址為：http://minhakka.ling.sinica.edu.tw/bkg/hakyin/。

(7) 客家委員會客語認證詞彙資料庫、政府開放資料庫：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/。

(8) 本土語言資源網：為教育部所整理製作的網站，裡面整合了華、閩、客、原等台灣各本土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，目前由台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，字幕卻是翻譯過的華語字，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/。

6.3.2. 紙本資料

(10)《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。目前國家圖書館有存放該套書籍。

(11)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、102 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。

(12)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。

(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)具有代表性的灣客語文學作品。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。

6.4. 原住民語

6.4.1. 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。後又承蒙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。網址為：http://203.66.168.190/。

(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) 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/。

(4) 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。

(5) 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/。

(6) 台灣南島語數位典藏：由中研院語言所齊莉莎所建立，包括語料庫、語言地理系統及書目資料庫等。已建立魯凱語、賽夏語、泰雅語、鄒語、阿美語、布農語、排灣語、卑南語、巴宰語、卡那卡那富語及西拉雅語語料庫。不過目前該典藏網站屬於關閉狀態。

(7) 本土語言資源網：為教育部所整理製作的網站，裡面整合了華、閩、客、原等台灣各本土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。詳細資訊可參考原委會網站的原視新聞與原視節目專區，網址為：http://www.ipcf.org.tw/。

6.4.2. 紙本資料

(9) 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）。

(10) 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。

(11) 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。

6.5. 臺灣手語

臺灣手語因為性質的關係，需要以多媒體影片來呈現。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以及臺灣手語電子資料庫。

6.5.1. 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。於 2001 年在國科會支持下開始收集詞項並錄影。第三版線上辭典已收錄 3500 個詞項，並且增加「位置」及「手形」的查詢功能，也可以按照中文筆畫查詢。網址為：http://tsl.ccu.edu.tw/web/browser.htm。【臺灣手語線上辭典】也提供英文版 http://lngproc.ccu.edu.tw/TSL/indexEN.html。

(2) 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，目前已經完成的資料庫有兩個：「臺灣手語地名電子資料庫」及「臺灣手語姓氏電子資料庫」，可作為日後追溯語源的參考。網址為：http://signlanguage.ccu.edu.tw/index.php。本資料庫有 APP 可以下載。

(3) 臺灣常用手語辭典：國立臺灣師範大學特教中心開發之手語辭典，其應用程式網址為：http://signlanguage.moe.edu.tw/conversation，因為臺灣手語的發展歷史，可將其分為中文文法手語（以中文文法為基礎）及自然手語（由聾人社群長年自然發展而成），該辭典建置時特別計算出兩者所佔比例，共 9615 條詞彙，修訂版刪除使用頻率較低的詞彙，並附上中文同義詞及英文翻譯。

(4) 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，在檢索設計上，可依照「主手」、「副手」及「位置」分類搜尋。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。

(5) 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，但為呈現臺灣手語的語法等面向內容，語料以句子而非單詞的方式呈現，網址為 https://www.twsl.cc。

6.5.2. 紙本資料

(6) 手語畫冊：教育部於 1987 年出版之教材，手語歌比賽以此為準，較偏向聽人的中文手語詞彙。

(7) 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，亦有複合詞、片語及會話與文法練習，以自然手語為主，希望補足聽人領導的中文手語不足之處。

(8) 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。

7、規劃國家語言資料庫的用途以及使用對象

以下根據文化部對於本案的相關文件草擬國家語言資料庫的目標，用途，及使用對象。

7.1. 目標

(1) 激發國人對本土語言文化的興趣與熱愛，並關注瀕危的國家語言。

(2) 永續保存本國語言文化資產。

(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。

(4) 促進未來研究發展及加值應用。

7.2. 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。

7.3. 使用對象

學術界，教育界，社會大眾，以及對臺灣本土語言有興趣的國際人士。不過關於「對臺灣本土語言有興趣之國際人士」，葉茂林委員也建議這部分需要再補充說明對應做法，如建立與國外研究者之交互授權機制等 。

7.4. 應用與推廣

期望使用者可下載國家語料庫的語料，朝向公開資料（open data）的方向發展，另也開發語料庫工具及 API 等，讓使用者對語料能有更多元深入的處理與應用。借鏡澳洲國家語料庫，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）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。除此之外，葉茂林委員表示，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。其他推廣或可嘗試開發手機版程式 app，增加其使用時機，以及舉辦競賽，激發大眾對國家語料庫的想像與創意等，例如何信翰委員認為閩南語等本土語言與長照服務的結合是一實用的發想主題。

8、規劃國家語言資料庫的內容項目

本章節為一些規畫國家語料庫內容項目的建議。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；閩南語和手語語料庫因為沒有專責機構負責，因此必須由文化部建置並維護。另外，關於國家語言資料庫的建置原則，本報告在此提出六項建議：

(1) 分階段依據重要性及時效性，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的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。

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。」

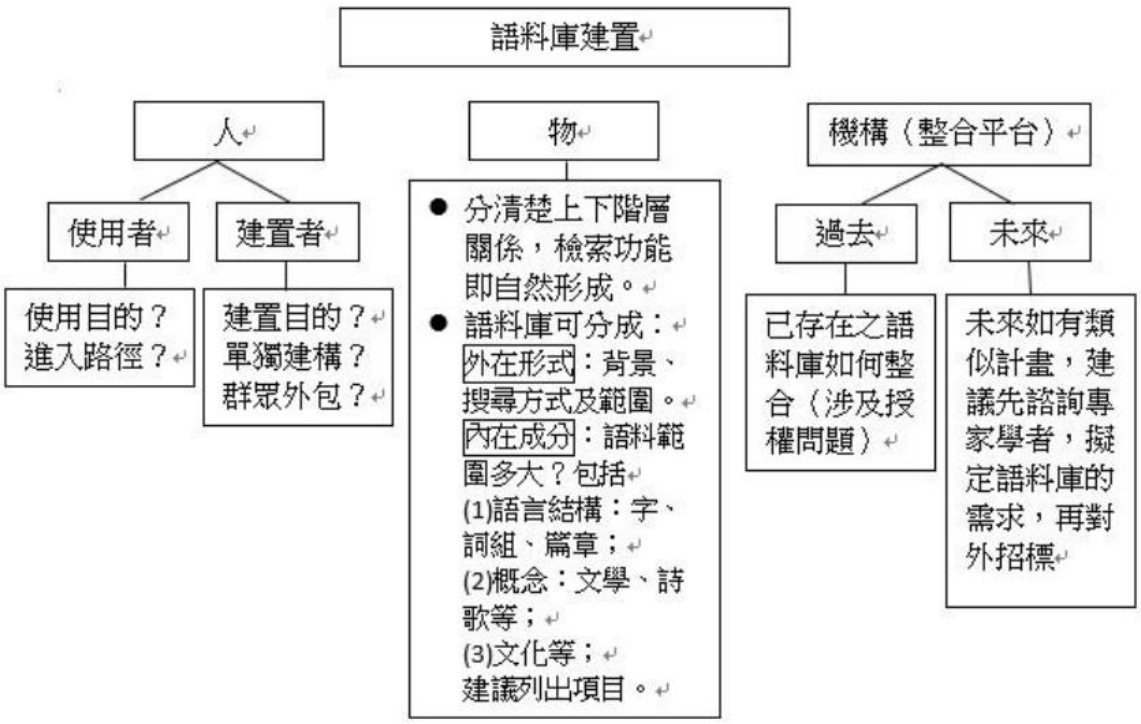


圖 19. 張榮興委員建議之語料庫架構規劃

8.1. 逐漸消失的母語

為了要讓國人瞭解並關注瀕危的國家語言，國家語言資料庫可以先收集彙整公共電視台，客家電視台，及原住民族電視台的母語流失相關主題的紀錄片，並放到網站上給民眾參考。如果前述的資料不夠，建議可再製作其他短的紀錄片。除了記錄片，另外也可以放入類

似聯合國瀕危語言的相關國際網站連結，讓民眾了解台灣各語言的瀕危程度。UNESCO Atlas of the World's Languages in Danger 網址為：http://www.unesco.org/languages-atlas/。

8.2. 臺灣的國家語言以及地理分布

建議以洪惟仁教授 2019 年出版的兩冊專書為基礎（需請求作者授權），即*臺灣社會語言地理學研究：臺灣語言的分類與分區I和臺灣語言地圖集II*，來介紹臺灣語言的分類與分區，目前洪惟仁教授已經同意將這兩冊專書中的語言地圖授權給文化部。在語言地圖（linguistic map、linguistic atlas）的資料展示，可參考日本國立國語研究所的官方網站，除了方言語料庫外，亦專闢一區塊「語言地圖」，分頁網址為 https://www.ninjal.ac.jp/english/database/type/maps/，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。

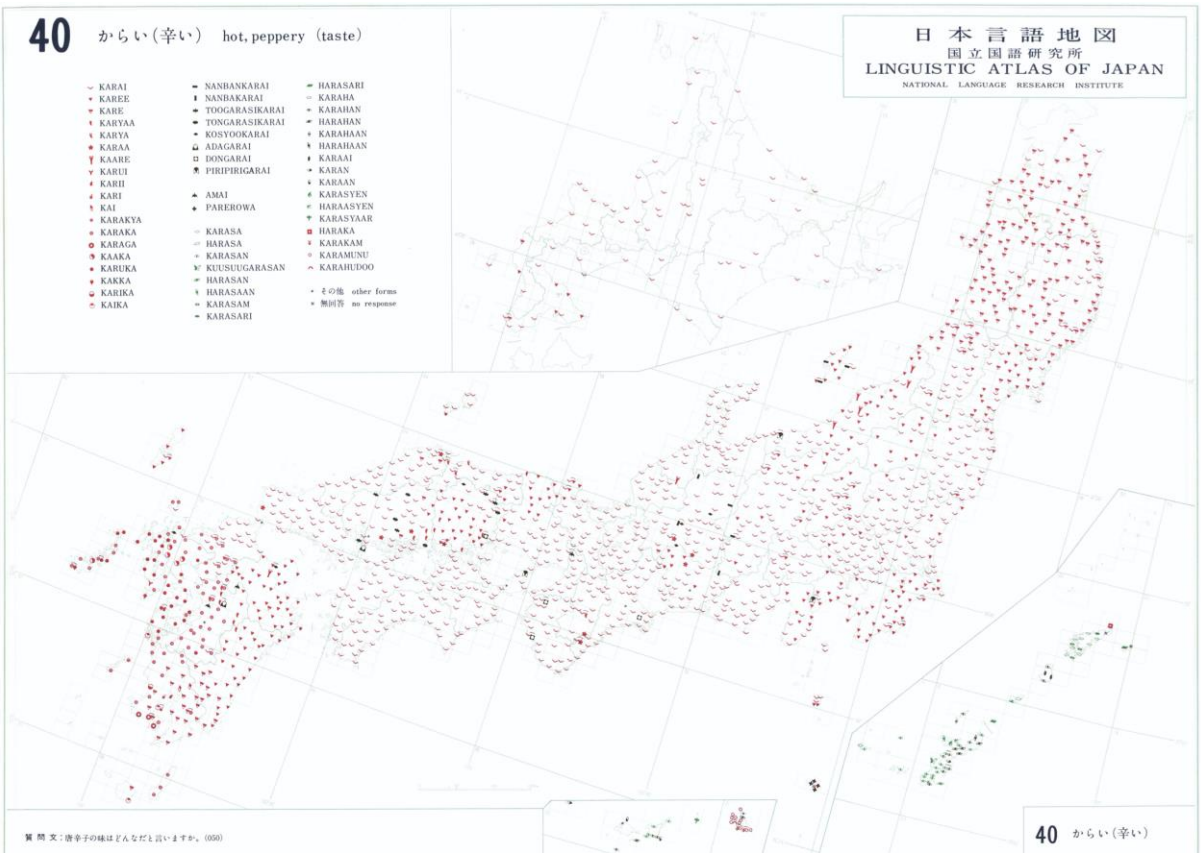


圖 20. 日本國立國語研究所收藏的方言語言地圖，以「辛い」一詞為例。

8.3. 國家語言調查報告

國家語言資料庫可以設專區，專門收錄教育部本土語言調查報告（網址：https://mhi.moe.edu.tw/newsList.jsp?ID=5）、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（網址：https://www.hakka.gov.tw/Content/Content?NodeID=626&PageID=37585）、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、及未來國家語言調查的相關報告等。

8.4. 國家語言語料庫檢索系統

收集第六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言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。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。

就第六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。因此綜合上述各點，建議之後建立國家語言資料庫時，可以先以目前仍無專責機構主導，且已有不少相關語料散落在民間或各機構的閩南語為優先。

江敏華委員在期中報告審查意見表中提到：「就第六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。

最後，江敏華委員提到，考量到未來國家語言資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搜集方面，語料來源者的問題，若是從新聞搜集手語語料，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。最後，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 。

8.5. 國家語言多媒體檢索系統

臺灣手語因為性質的關係，需要以多媒體影片來呈現。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以及臺灣手語電子資料庫。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，授權方式需再與學校研發處談細節。此外，所有包含影音資料都可以透過多媒體檢索系統來檢索內容。除了臺灣手語資料外，包括各國家語言音檔，及具有代表各族群文化、歷史、風俗、藝術的民謠、歌曲以及含有國家語言的重要儀式、典禮、或有關語言的紀錄片也應該收錄在國家語言多媒體檢索系統。

8.6. 國家語言學習資源及跨語言核心詞彙

收集各單位的國家語言學習資源，例如，辭典、有聲書等，並建立國家語言核心詞彙對照表，包括各國家語言的詞彙表、詞彙、例句、華語解說、及音檔。另外，應該以建立跨語言核心詞彙和相關例句並提供跨語言檢索。此外，也可以提供各語言相關的輸入法工具。

考量到活力較低的國家語言（如閩、客、原住民語等）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。

8.7. 臺灣本土語言研究參考文獻

本區可以列出臺灣各本土語言的相關研究，如洪惟仁(2004)《續修台北縣志.卷三住民志.第二篇漢人語言》、張屏生、蕭藤村、呂茗芬（2010）嘉義縣方言志(上下冊) (電子書) 等。接著，可以先將所有相關的學術著作依據語言別分類，然後再參考日本國立國語研究所官方網站的架構來呈現這些資料（先至日本官網》點選「資料庫」標題》點選「研究主題」底下的「方言語語言的多樣性」，即可看到各種日語方言相關資料，網址為：https://www.ninjal.ac.jp/english/database/subject/diversity/）。此外，根據江敏華委員的建議，須區分語言資料庫（languae database）與語料庫（corpus）的概念，因此除了語言別，亦可加入「語言資料庫或語料庫」的分類方式。

8.8. 語料庫後設資料(metadata)和國家語言資訊處理工具

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。

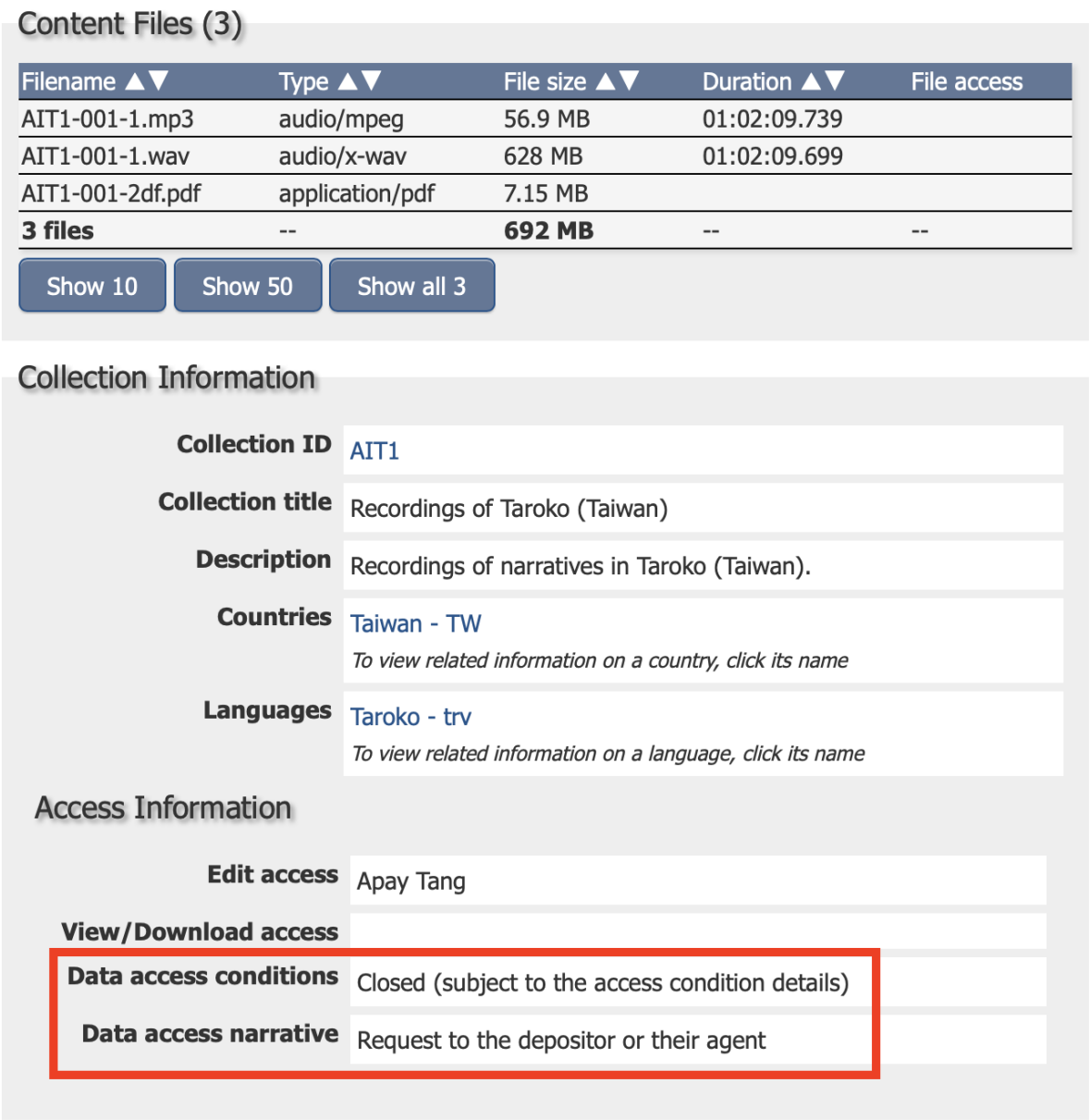
最後，為了推廣國家語言資料庫之利用，該資料庫必須易於使用（如，使用較通用的檔案格式和檔案編碼），也易於機器讀取，以增進學術或商業界使用的意願。資料說明也應完整且正確，並提供英語等外語版本說明。授權方式應儘量簡單而友善。然後，也可透過論文發表的方式讓學術界同儕認知該語言資料庫的存在，進而引起利用該資料庫進行研究的興趣及可能性。

考量到上述資料儲存、資料存取與資料推廣之原則，還有參考第4章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 有點像是資料交換的方法，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。



Content Files (3)

Filename ▲▼	Type ▲▼	File size ▲▼	Duration ▲▼	File access
AIT1-001-1.mp3	audio/mpeg	56.9 MB	01:02:09.739	
AIT1-001-1.wav	audio/x-wav	628 MB	01:02:09.699	
AIT1-001-2df.pdf	application/pdf	7.15 MB		
3 files	--	**692 MB**	--	--

Show 10 Show 50 Show all 3

Collection Information

Collection ID	AIT1
Collection title	Recordings of Taroko (Taiwan)
Description	Recordings of narratives in Taroko (Taiwan).
Countries	Taiwan - TW *To view related information on a country, click its name*
Languages	Taroko - trv *To view related information on a language, click its name*

Access Information

Edit access	Apay Tang
View/Download access	
Data access conditions	Closed (subject to the access condition details)
Data access narrative	Request to the depositor or their agent

圖 21. 澳洲 PARADISEC 數位典藏計畫中的資料存取（access information）的資訊頁面

圖 22. 澳洲 PARADISEC 數位典藏計畫中資料庫的介紹頁面

此外，葉茂林律師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，都是非常寶貴的意見。

綜合以上想法，國家語言資料庫的內容項目規劃可以下方樹狀圖表示：

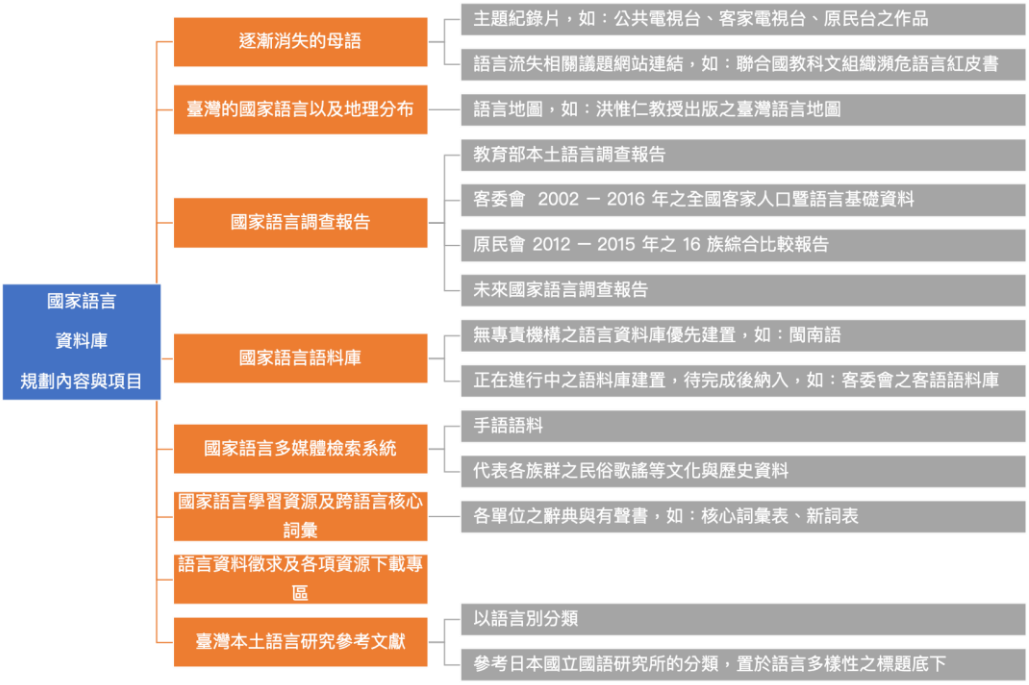


圖 23. 國家語言資料庫規劃內容與項目樹狀圖

9、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。

9.1. 臺灣國家語言資料庫之使用者條款草案

一、

（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

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。 臺灣國家本土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

o ID：需要對用戶進行身份驗證或標識。

o 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）。

o 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。

o FF：存取、使用該資料需要付費。

o 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

o BY：必須註明出處，即作者身份。

o NC：內容僅可用於非商業目的。

o 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。

o 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。

o LRT：內容僅能適用或應用於語言研究及/或技術開發。

o 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

o NORED：不允許您重新散佈資料。

o 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本土語言資料庫散佈修改後的版本。

o SA：允許在類似條件下重新散佈資料。

o 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

o *：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，惟臺灣國家本土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本土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，或對臺灣國家本土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。

9.2. 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），並一併交付本資料相關之使用、利用權、及/或散布權（Rights of Distribution）等之權利和義務。

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。除因雙方之嚴重違約而解除或終止本契約之情形外，乙方於本契約終止後仍保有一切本契約授權之使用、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、其他基於智慧財產權法之權利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，則乙方得依其對前述主張或通知之認定合理與否，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而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，除本契約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、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範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。除根據本協議第十條提前終止之情況以外，本協議在雙方約定履行各自義務之時間範圍內皆為有效。另，本協議書之以下條款，於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、參考文獻

Australian National Corpus Incorporated. (2012). Australian National Corpus. Retrieved from http://www.ausnc.org.au/

AWS Public Dataset Program. Retrieved from https://aws.amazon.com/opendata/public-datasets/

BNC Consortium. (2007). The British National Corpus, version 3 (BNC XML Edition). Retrieved from http://www.natcorp.ox.ac.uk/

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.

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., Haugh, M., Peters, P., & Fallu, M. (2012, January). The Australian National Corpus: National Infrastructure for Language Resources. In LREC (pp. 3295-3299).

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).

Crasborn, O., Zwitserlood, I., van der Kooij, E., & Schüller, A. (2018). Global SignBank manual.

FIN-CLARIN | Kielipankki. Retrieved from https://www.kielipankki.fi/organisaatio/fin-clarin/

Gippert, J., Meurer, P., & Tandashvili, M. (2012). Georgian National Corpus. http://gnc.gov.ge/gnc/page

Global SignBank. Retrieved from https://signbank.science.ru.nl

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.

Hungarian National Corpus. Retrieved from http://corpus.nytud.hu/mnsz/index_eng.html

Ide, N., & Suderman, K. (2002). Open American National Corpus | Open Data for Language Research and Education. Retrieved from 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. Retrieved

from 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. Retrieved from https://ucnk.ff.cuni.cz/cs/

Jeff Howe - Crowdsourcing. Retrieved from https://www.youtube.com/watch?v=F0-UtNg3ots

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

Ministry of Education and Culture (Finland). (2013). FINNA. Retrieved from https://www.kiwi.fi/display/Finna/In+English

Morozova, M., Rusakov, A., & Arkhangelskiy, T. (2012). Albanian National Corpus. Retrieved from albanian.web-corpora.net

Ministry of Education and Culture (Finland). (2013). FINNA. Retrieved from https://www.kiwi.fi/display/Finna/In+English

Morozova, M., Rusakov, A., & Arkhangelskiy, T. (2012). *Albanian National Corpus*. Retrieved from albanian.web-corpora.net

Mozilla Foundation. (2017). Common Voice by Mozilla. Retrieved from https://voice.mozilla.org/

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. Retrieved from http://nkjp.pl/index.php?page=0&lang=1

National Institute of the Korean Language Republic of Korea. Retrieved from https://www.korean.go.kr/front_eng/main.do

Ossetic National Corpus. Retrieved from http://corpus.ossetic-studies.org/search/index.php?interface_language=en

Pacific and Regional Archive for Digital Sources in Endangered Cultures. Retrieved from http://www.paradisec.org.au/home.html

Peters, P. (2009). The architecture of a multipurpose Australian National Corpus. *Haugh et al.(eds).*

The CorCenCC project team. (2016). CorCenCC – National Corpus of Contemporary Welsh. Retrieved from http://www.corcencc.org/

The Institute for Bulgarian Language. (2001). *Bulgarian National Corpus*. Retrieved from https://dcl.bas.bg/bulnc/en/

The Institute of Russian language, Russian Academy of Sciences. (2004). Russian National Corpus. Retrieved from http://ruscorpora.ru/en/

Semantic Versioning. Retrieved from https://semver.org/

Semantic Web Research Center (SWRC). Retrieved from http://semanticweb.kaist.ac.kr/home/index.php/Home

Sign Linguistics & Language Acquisition Lab at University of Connecticut. Retrieved from https://slla.lab.uconn.edu/

Slovak National Corpus. Retrieved from https://korpus.sk/index_en.html

Tatar National Corpus. Retrieved from http://tugantel.tatar/?lang=en

Thai National Corpus. Retrieved from http://www.arts.chula.ac.th/ling/tnc/

The Abkhaz National Corpus. Retrieved from http://clarino.uib.no/abnc/page

The Balanced Corpus of Contemporary Written Japanese (BCCWJ). Retrieved from https://pj.ninjal.ac.jp/corpus_center/bccwj/en/

The Institute for Language and Speech Processing (ILSP / "Athena" R.C.). (2019). Hellenic National Corpus. Retrieved from http://hnc.ilsp.gr/

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).

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. Melbourne: University of Melbourne, pp. 144–59. http://hdl.handle.net/11343/40961

Turkish National Corpus (TNC). Retrieved from https://www.tnc.org.tr/

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

九年一貫台語教學資源網。取自 http://www.taiwanwe.com.tw/

中央研究院 iCorpus 臺華平行新聞語料庫。取自

http://icorpus.iis.sinica.edu.tw/

中央研究院中文詞彙特性速描系統。取自

http://wordsketch.ling.sinica.edu.tw/

中央研究院漢語平衡語料庫。取自 http://asbc.iis.sinica.edu.tw/

台大臺灣南島語多媒體語料庫。取自 http://203.66.168.190/

台語文記憶。取自 http://ip194097.ntcu.edu.tw/memory/TGB/mowt.asp

台語文語詞檢索。取自

http://ip194097.ntcu.edu.tw/TG/concordance/form.asp

台灣現當代作家研究資料庫。取自 http://cw.nmtl.gov.tw/

本土語言調查報告- 本土語言資源網。取自

https://mhi.moe.edu.tw/newsList.jsp?ID=5

全國客家人口暨語言基礎資料| 客家委員會全球資訊網。取自

https://www.hakka.gov.tw/Content/Content?NodeID=626&PageID=37585

林宜靜. (2018, January 12). 百冊看文學！「台灣現當代作家研究

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

客英大辭典查詢。取自 http://minhakka.ling.sinica.edu.tw/bkg/hakyin/

客家委員會客語認證詞彙資料庫。取自 https://wiki.hakka.gov.tw/

国立国語研究所。取自 https://www.ninjal.ac.jp/

语料库在线。取自 http://corpus.zhonghuayuwen.org/

原住民族語文學創作作品集 - 教育部語文成果網。取自

https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=480

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。取自 https://m-dictionary.apc.gov.tw/

族語 e 樂園。取自 http://klokah.tw/

教育部悅讀越懂閩客語電子報。取自

https://epaper.edu.tw/learning.aspx?classify_sn=6

教育部臺灣客家語常用辭典。取自 https://hakkadict.moe.edu.tw/

教育部臺灣閩南語常用詞辭典。取自

https://twblg.dict.edu.tw/holodict_new/default.jsp

國立中正大學臺灣閩南語口語語料庫。取自

http://lngproc.ccu.edu.tw/Corpus/

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。取自 https://coct.naer.edu.tw/cqpweb/

曾淑娟 Shu-Chuan TSENG - 中央研究院語言學研究所。取自

http://www.ling.sinica.edu.tw/v3-3-1.asp-auserid=20.htm

閩客語典藏。取自

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。取自 http://tsl.ccu.edu.tw/web/browser.htm

臺灣手語電子資料庫。取自 http://signlanguage.ccu.edu.tw/index.php

臺灣白話字文獻館。取自 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。取自

https://language.moe.gov.tw/result.aspx?classify_sn=46&subclassify_sn=460&thirdclassify_sn=481

臺灣閩南語我嘛會每日一詞。取自

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。取自

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。取自

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。取自

https://language.moe.gov.tw/result.aspx?classify_sn=42&subclassify_sn=446

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_

蘭嶼達悟語口語資料典藏網。取自 http://yamiproject.cs.pu.edu.tw/

附錄一、第一次專家諮詢會議記錄

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

齊莉莎教授

張永利教授

湯愛玉教授

章忠信教授

郭志忠博士

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

<關於資料的收集>

（張永利教授）

- 先開一個 **workshop**（小型會議）作為一個平台，區分不同的 session，並邀請各界的人們討論相關問題，分享、交流各式相關資訊。
- 專家會議能搜集到的想法和資料畢竟有限，應做適度擴充，所以可以廣邀各界（專業）人士提供看法和見解。
- workshop **最主要的目的是收集非學術界的、未出版的、流落在民間的資料**，如小型語料庫、字典，以及前人已經蒐集到的關於各種台灣本土語言的資料等。

（章忠信教授）

- 收集資料本身並不困難，困難之處在於如何**數位化**收集到的資料。因數位化之前必須徵求資料擁有者的同意，導致資料數位化這一步驟會有一定的困難度。
- 策略：廣收資料。買得到的就用買的，買不到的就問看看對方是否能夠提供。
- 開一個平台，分享「誰做過什麼」之類的不涉及著作權的部分，作為初步收集資料；收集著作權的部分，能做多少是多少。
- 必須釐清：要收集的到底是東西本身，還是數位化後的資料？
- 目前可以做一個 **metadata**，上面標註各資料來源（i.e.「誰做過什麼」），如此一來可以達成收集資料的目的，也不會有侵權的問題。
- 分階段進行，現階段先盡可能地搜集資料（metadata），之後再思考是否需要數位化蒐集到的資料，並參照經濟利益、法律問題等去做相關的處理。

（湯愛玉教授）

- 原住民語言部分的田野調查可設置專門的推廣人員進行。

（可以設在語言調查組底下）

（郭志忠教授）

- 原住民語的語料有多少就儘量搜集多少。
- 歌謠、民間歌曲等語料，可以利用押韻等，判別原有讀音。

（齊莉莎教授）

- （原民語料庫）除了想辦法搜集新的資料以外，也要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。
- 在校對詞表時，常會出現連 native speaker 之間也出現歧異看法的時候，此困難必須想辦法解決。

以客委會語料庫的 1 萬 5000 篇語料作為標竿，那麼原住民語國家語料庫的規模應該為何？如何平衡？

（齊莉莎教授）

- 收集原住民語的困難在於，一個族當中，不同地區的人可能會各自發展出次方言，並且隨著時間演進，次方言間的差異也隨之增大。所以，在收集的過程中，要不斷的標記地區、確認該方音的來源。
- 結論：完整了解所有原住民族語之（次）方言是幾乎不可能的。

（張永利教授）

●（原住民語的）語料庫平衡可參考宋麗梅老師的原住民語料。

<關於著作權問題>

（張永利教授）

- 可以成立類似中研院的「**智財組**」，專門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了一個專門處理著作權的組。）

- 只要不侵害作者的著作人格權，應該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。

（翁聖賢律師）

- 關於著作權應考慮以下幾點：

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 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。

- 找握有著作權的 assignee 洽談。

（章忠信教授）

- 語音轉文字的話，若是有 annotation（智慧的投入），就算是利用別人的作品而成的著作。
- **著作**：有創作的成分就算，所以口語也是著作的一種；沒有創作的成分的話只能算是「重製物」，所以有給提示稿的錄音也是重製。在此之上，錄音應該不是「著作」，只是一個「錄製成果」，所以原則上講話的人才握有內容的權利。

<關於未來國家語言研究中心之業務>：語言調查、蒐集、典藏；

整合現有資料、跨語系查詢

（郭志忠博士）

- 萌典已經做到某種程度上的**跨語系查詢**（但不包含原住民語）。
- 要成立語言中心，必須釐清需要聘用多少人力、乃至於多少人事預算……等等；如果是要做出文化部標案的規格的話，則需要做更細的發想。

（蔡素娟教授）

- 中心應該先有架構，還要先分組別，以利進行往後搜集資料以及調查之作業。

（然中心目前還停在發想草案的階段）

- 在開啟計劃之前，可以先思考如何分期進行，較大面向地去做探討和釐清。
- 「**任務編組**」是很好的想法，但是在分組之前，必須先釐清一些工作細項，如：應該收集什麼樣的語料？資料要做何種處理？
- **數位典藏**和**數位應用**應該分開處理。

（張永利教授）

- 將人員分成好幾組（e.g. 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），各自處理各自的工作。

（湯愛玉教授）

- **語料數位應用**是必要的（**可以配合 AI** 發展）；並且，為了傳承的需要，應該收集不同年齡層之影音資料，典藏（or 數位化）後保存起來做交流，如此才能看到語言磨損的程度（attrition）。
- 建議中心可以排出閩、客、原三大種語言各自的**詞（頻）表**，如此對語言認證也有幫助。

（齊莉莎教授）

- 錄音資料固然有其效用，但很難收集，因很多人會拒絕接受錄音。尤其原住民的耆老，很難請他們離開家園、到別的地方錄音/影，而年輕人也可能已經不太會說族語了。

（郭志忠博士）

- 根據目的不同而收集的錄音資料，各自其實都有不同的作用；但對錄音資料而言，共同的重點就是 **SNR（訊雜比）**。
- 訊雜比（SNR，signal to noise ratio）：要定義什麼東西是訊號，什麼不是，只要符合訊雜比（例如：15dB 以上）就可納為語料，不一定要進錄音室（除語音合成必須在 studio 裡進行），如此也較尊重受訪者的意願。
- 將錄音資料**數位化**，可以延長其保存時間。
- 自然狀況下的聲音資料（spontaneous）才是最值得被收集的語料。