語料庫建置入門工作流程指南

數位典藏與數位學習國家型科技計畫　拓展台灣數位典藏計畫

出 版 序

「數位典藏國家型科技計畫」於西元2002年開始執行，衆多機構計畫與公開徵選計畫的工作夥伴紛紛加入我們的團隊，進行種類繁多而又數量鉅大的數位化工作，第一期五年計畫於民國2006年圓滿結束。次年，即與「數位學習國家型科技計畫」整合爲「數位典藏與數位學習國家型科技計畫」（TELDAP, http://teldap.tw/ ），以「呈現台灣的文化與自然多樣性」爲總體目標，持續拓展各方面重要數位資源，並更有系統地往教育、研究與產業等面向推廣數位成果；同時，還準備更積極結合民間力量，推動相關產業的成長，既藉以保存我國重要文化資產，也加速創造數位時代新文化。

作爲「數位典藏與數位學習國家型科技計畫」的分項計畫，我們也由第一期「內容發展分項計畫」改名「拓展台灣數位典藏計畫」(http://content.teldap.tw)，更積極地拓展數位內容來源，向民間公私立單位甚至是個人收藏，廣泛徵集有關檔案、考古、語言、地理、族群、藝術、民間生活與動物、植物等數位化計畫，並希望能更好地整合這些自然與人文不同性質的數位內容，製作成兼具趣味性與啓發性的數位素材，既供民衆免費下載進行教育與研究之用，也便利廠商與公私典藏者發現彼此在商業加值方面的合作機會。「拓展台灣數位典藏計畫」與「數位典藏與數位學習國家型科技計畫」其他分項計畫的相互協力，將加速我國數位內容由典藏保存跨入教育、研究與商業加值的過程，以利呈現台灣的文化與自然多樣性，並讓更多國內外民衆體會並珍視我國歷史文化之富盛與自然生態之茂美。

在典藏與加值數位內容的同時，無論是於「內容發展分項計畫」或是於「拓展台灣數位典藏計畫」時期，本計畫同仁都持續調查與記錄公私立機關與公開徵選計畫等工作夥伴從事各類物件數位化的工作流程及相關技術，並結合各項符合國際標準的數位化技術與工作流程資訊，編撰一系列「數位化工作流

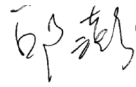
程叢書」。自西元2005年以來，我們即先精選諸如瓷器、書畫、古籍等單一種類的數位化物件，綜合不同典藏計畫從事此項單一物件數位化的工作經驗，並輔以國內外相關理論與實務成果，陸續撰寫了21冊不同主題的數位化工作流程指南（這21冊內容都可自「拓展台灣數位典藏」網站的「數位化書籍」欄位下載全文電子檔）。

自2008年以來，我們即持續修訂擴充這套「數位化工作流程叢書」，希望增加流通管道，以供更多博物館、圖書館、機構與個人參考。我們的準備工作，主要分爲修訂既有「精選物件」指南以及新撰「共通原則」指南兩方面；前者指的是修訂既有的21冊工作流程指南，特別是針對數位化新技術與規範的引進、更實用的軟硬體設備以及數位內容保護機制等層面做修訂，預訂每年修訂出版七本專書，並於三年內全部出版完畢。至於新編的「共通原則」指南，則重點放在導入數位資訊「生命週期」與品質管理等關鍵概念，以「跨物件」而非單一精選物件爲探究對象，採用共通原則做爲架構該指南的數位化工作流程內容；這裏所謂的共通原則，指的是諸如專案規劃、整合性工作流程、影像資料、影音資料、文字資料、色彩管理、委外製作和數位內容保護與授權等，這八個共通原則都成爲我們調查、研究與撰寫指南的主題內容，預計三年出版八本指南。

精選物件指南與共通原則指南之間，其實具有一種相輔相成的關係。共通原則指南著重在分析數位化工作的各項重要主題，引導讀者對數位化的利弊得失做通盤而深入的思考。精選物件指南則描述特定物件的數位化實務與技術，便利讀者針對單一物件，選擇最合適、最有效益的數位化工作流程。透過這套「數位化工作流程叢書」的出版，相信可爲更多有志投入數位化工作的單位與個人，提供一套富有整體性思惟並且又能循序漸進的實用指南。要特別強調的是：這套叢書的主要立論基礎，仍在於多年來陸續加入我們的機構與公開徵選計畫工作團隊多年累積的各種寶貴經驗，這些經驗讓更多的數位內容可以用更精緻的品質以及更合宜的成本來製作、展示與維護，從而豐富我國數位典

藏與數位學習事業。在陸續出版這套「數位化工作流程」叢書的同時，我們要感謝接受訪問的工作夥伴以及參與寫作的同仁，也衷心感謝協助我們審查與諮詢數位化工作流程指南的所有學者專家。最後，也盼望讀者隨時給我們指正與建議，讓我們的工作可以做的更好。

數位典藏與數位學習國家型科技計畫
拓展台灣數位典藏計畫・數位內容建置與整合子計畫

計畫主持人  敬誌

中華民國 99年2月10日

編 者 序

本書是「數位化工作流程：語言主題小組」(2006)的增訂本，原書由時任語言主題小組召集人的鄭錦全院士統籌規劃、助理賴佳旻負責主要的編輯、撰寫工作，共58頁，前言簡介語言主題小組之各計畫，主文爲參與語言主題小組的計畫團隊之經驗分享，包括現代漢語平衡語料庫數位化工作流程簡介、語言分布GIS系統建置數位化工作流程簡介、台灣手語影像辭典數位化工作流程簡介，以及閩南語兒童語料庫數位化工作流程簡介等。本次改版新增後設資料與語料庫相關國際標準、語料庫建置流程、語料庫與數位學習應用、延伸議題等章節與附錄。此外，原書主文各實例僅配合新版體例更新少許文字，另增語言典藏二期子計畫「台灣國語口音之社會分布典藏」的建置流程簡介，實例也不限參與語言主題小組的計畫，特邀中央研究院語言學研究所黃居仁研究員帶領的研究團隊中文詞彙網路小組提供建構「中文詞彙網路」等資料庫的經驗。

本書是參與初版與增訂本編寫的所有工作人員以及各研究團隊的共同成果。本次增訂，特別感謝提供資料與文稿、參與討論的各計畫研究團隊，尤其是國立中正大學語言學研究所戴浩一教授與蔡素娟教授更新、審閱「台灣手語影像辭典」以及「閩南語兒童語料庫」的簡介、中央研究院語言學研究所中文詞彙網路小組提供文稿分享經驗、中央研究院語言學研究所曾淑娟副研究員審閱「台灣國語口音之社會分布典藏」的相關文稿，中央研究院語言學研究所齊莉莎研究員「台灣南島語典藏」研究助理瓦歷斯．浦亞先生提供後設資料、跨資料庫檢索與部份設備之資料，並參與繪製數位化流程圖之討論。此外，「拓展台灣數位典藏計畫」辦公室同仁於撰寫期間也提供許多建議，在此一併感謝。

本次增訂由編者負責全書架構與各章節大綱擬定、邀稿、部份文稿撰寫以及全書審閱、文字編輯，「語言、影音與新聞主題」小組助理詹景勛先生則擔任大部份新增文稿的撰寫以及全書的圖文編排等工作。

最後，本書從編者接下編務到定稿不到三個月，限於時間與經驗，必然仍有許多疏漏不足之處，敬請方家不吝指教。

中央研究院語言學研究所助研究員
蕭素英 敬誌
中華民國99年1月26日

目 錄 | CONTENTS

前言

Introduction

一、什麼是語料庫？

語料庫是龐大且具有組織架構的語言資料庫，語料庫可以只收錄單一語言，也可以囊括多種語言，語料內容則涵蓋文字、手語、聲音等領域。語料庫爲語言學研究的重要成果，也是研究工具，通常做爲語言統計分析、語言學術研究等用途，對一般使用者而言，則是學習語言的工具之一。

語言是人類表達與溝通的重要媒介之一，語言學是以人類語言爲研究對象的學科。目前世界上現存語言已知的有三千多種，在新的語言誕生的同時，也有許多語言在凋零。如何保存這些凋零或是發展中的語言，語料庫就是一個很好的選擇，也是現今語言學研究結合資訊科技的結晶。語料庫通常指爲了語言研究而收集並採用數位形式保存的語言材料，由自然語言或口語的樣本構成，用來表達特定語言或是語言轉變。經過科學標注並具有適當規模的語料庫能夠反映、記錄語言的實際使用狀況。透過語料庫觀察、掌握語言事實，可以研究分析語言系統的規律性，是語言學理論研究、應用研究以及語言工程重要的基礎資源。

按照語料的種類劃分，語料庫可以分爲單語(Monolingual)，雙語(Bilingual)和多語的(Multilingual)。近年語料庫的類型逐漸多元，從以往的單語語料庫，到現今的多語語料庫，甚至結合影音呈現，不僅對於語言的研究分析有很大的助益，對於語言學習也有極大的貢獻。

語料庫與語言訊息處理有密切的關係。未使用語料庫方法之前，在自然語言處理和機器翻譯等研究中，分析語言的主要方法是基於規則，但對於規則不能表達或無法涵蓋的語言事實，電腦就很難處理。語料庫出現之後，人們可利用語料庫來調查、統計自然語言，建立統計模型，研究自然語言處理技術。另一方面，自然語言訊息處理的研究也爲語料的加工提供了訊息檢索、文本輸入、自動分詞和標注、語料的統計和檢索等各方面的關鍵技術。

語料庫的功能主要涉及三個層面，一是語料庫的規模，二是語料的分布，三是語料加工的程度。規模大小關係到統計數據是否可靠，語料的分布涉

及統計結果的適用範圍，語料加工的深度則決定這個語料庫能爲使用者提供什麼樣的語言學訊息。

根據語料採集的原則與方式，語料庫可以分爲以下四種類型：

1. 異質的(Heterogeneous)： 無特定的語料收集原則，廣泛收集並原樣儲存各種語料。
2. 同質的(Homogeneous)：只收集同一類的語料。
3. 系統的(Systematic)：根據預先確定的原則和比例收集語料，使語料具有平衡性和系統性，能夠代表某一特定範圍的語言事實。布朗語料庫(Brown Corpus)於六〇年代於布朗大學建立，是世界上第一個根據系統性原則採集樣本的標準語料庫，具一百萬詞規模。「中央研究院現代漢語平衡語料庫」簡稱「研究院平衡語料庫」(Sinica Corpus)則是世界上第一個有完整詞類標記的漢語平衡語料庫。
4. 專用的(Specialized)：只收針對某一特殊用途的語料。

語料加工主要指文本格式處理和文本描述兩項工作，文本格式處理是對於已採集的語料文本進行整理，轉成格式一致的電子文本，例如資料庫格式、XML格式等。

文本描述是說明每一篇語料樣本的屬性或特徵，包括篇頭描述和篇體描述。篇頭描述說明整篇語料樣本的後設資料屬性，例如語體、內容所屬的領域、作者、出版時間與發行出版社……等，篇體描述是在文本裡添加各種屬性標記，如詞語切分標記、詞類標記、語法特徵標記、語意訊息標記、言談標記……等。漢語文本語料庫的加工一般是從詞語切分（斷詞）、詞類標記，到語法、語意屬性標記循序漸進，所標注的訊息增加，語料加工的深度也就相對增加。

沒有篇體描述訊息的語料叫做素語料，漢語的文本素語料只能以字爲單位進行檢索與統計，而經過詞語切分處理的語料，就能夠以詞爲單位進行檢索、統計和定量分析，如果還加注了詞類標記，那麼可以獲得的訊息就更多

了。語料的標注如果由人來執行，當然能夠保證其準確性，但速度很慢，對於大規模的語料來說，人工標注顯然緩不濟急，不符需求，因此大規模的語料加工往往需要藉助自動化技術來進行詞語切分、詞類標注等語料加工。

二、本書章節說明

本書包括引言、後設資料與相關國際標準、語料庫建置流程、語料庫建置實例、語料庫與數位學習、延伸議題、結語、附錄等八個單元。「後設資料與相關國際標準」 介紹了泛用的後設資料標準「都柏林核心集」(Dublin Core, DC)、「開放語言典藏社群」(Open Language Archives Community, OLAC)以DC爲基礎所制定的適用於語料庫的後設資料OLACMS、OLAC採用的跨資料庫檢索網路協定 OAI-PMH以及語言代碼國際標準。「語料庫建置流程」將語料庫建置工作分爲語料數位化、語料庫系統建置、後設資料建立三個部份，簡要說明建置語料庫的流程。「語料庫建置實例」收錄文本、口語、影像、語言地理資訊等類型的語料庫建置經驗，供目前或未來其他單位或計畫進行語料庫建置時參考。「語料庫與數位學習」介紹了兩個將語料庫加值應用於教學網站之實例。「延伸議題」則探討語料庫建置涉及的數位內容保護、人力與設備成本等議題。

撰文：蕭素英、李佩瑛

壹、後設資料與相關國際標準

Metadata and related international standards

後設資料(Metadata)在數位典藏領域中最常見的解釋是「資料中的資料」(Data about Data)。以數位相機所拍攝的照片爲例，拍攝完的每一張照片都是一筆數位檔案，除了影像資料外，這張照片檔案內還會有EXIF後設資料，上面記載拍攝日期、時間、地點、光圈、快門、焦距、鏡頭以及白平衡設定等多項資料。

根據數位典藏與數位學習國家型科技計畫後設資料工作組（以下簡稱後設資料工作組）的解釋，後設資料的定義爲：[1]

> 後設資料(Metadata)是一組結構與標準化的背景資料，包括描述性、結構性與管理性三大類型，以及語義性、語法性與詞彙性三大屬性，用來描述每個數位典藏品的內涵與特徵，以便數位典藏品能夠在數位化環境或系統中，達到最佳化資源探索的效能，並能有效率而精準地被探索、呈現、管理、控制與執行相關功能，且順利地與其他數位典藏品進行資源互通與共享，最後還能達成數位典藏品的永久保存目的。

由此可知，每一件要永久保存的數位典藏品，背後都應該擁有一組後設資料，這些後設資料是確保數位化成果能被有意義地永久保存，有效率地被搜尋利用。以功能導向而言，後設資料有三種類型：[2]

（一）描述性後設資料：用以描述一項文件或資源的內涵與關聯性，以便於發現與辨識資源，例如：書目記錄與本章之後將介紹的Dublin Core。

（二）結構性後設資料：給予數位典藏品實質的結果，以便於瀏覽、檢索和呈現上述的資源，例如書的章節結構、具翻頁功能的電子全文，全文與相關影像的連結。

（三）管理性後設資料：爲了長久的管理、使用與觀看數位化資源的相關資源，如檔案格式、數位化解析度、智財權管理資訊等。

---

1 數位典藏與數位學習國家型科技計畫後設資料工作組網頁，計畫簡介：http://metadata.teldap.tw/introduction/introduction-frame.html。

2 沈漢聰，《數位典藏技術彙編》電子書，數位典藏國家型科技計畫，2004年，ch.9-1。

物件的後設資料可以隨計畫需求而調整，因此後設資料內容有大有小，各類型物件的編目規範也有所差異。語料庫的後設資料必須顧及許多層面，爲了追求資料的健全完整，在後設資料欄位的制訂上會採用多種國際標準，比如Dublin Core都柏林核心集、OLACMS、ISO語言代碼等標準；如果語料庫要做到跨資料檢索，或是與國外進行資料交換，則要使用OLAC開放語言典藏社群所推薦的網路協定標準OAI-PMH。

語言學研究的目的是調查瞭解語言行爲模式和分析各種語言，執行研究工作時，語言的發音人、語料搜集地點等都是探討項目之一，所以每一個語言樣本的背後都要擁有詳細的描述資料，以做爲語言研究的基礎資訊，對於語料庫而言，後設資料的內容屬於較龐大的類型。

語料庫工作團隊在進行語料收錄之前，最好先按照計畫需求先完成後設欄位的制訂，仔細考量收錄語料時所必須記錄的資訊，避免完成語料收集工作並且離開調查地點之後，才發現資料掛一漏萬，屆時要再次進行語料收錄、田野調查，不僅費時費力且浪費計畫經費。

後設資料工作組強調，後設資料的規劃與實施是數位典藏工作的基礎建設，未來語料的檢索功能和語料的完整性與實用性，都端看後設欄位的詳細與否。有鑑於此，語料庫計畫團隊制訂後設資料欄位時，務必多花功夫，以求完整、全面。以下介紹泛用型的後設資料元素集「都柏林核心集」(Dublin Core)以及幾個與語料庫後設資料相關的國際標準。

一、都柏林核心集

1995年，OCLC(Online Computer Library Center)與NCSA(National Center for Supercomputing Application)聯合召開第一屆會議，會議上集合了圖書館界、資訊科學界的各領域專家 ，制定了一套專爲網路資源而設計的後設資料元素集。這套元素內容依據會議地點美國俄亥俄州Dublin而命名，稱爲Dublin Core（簡稱DC），目前Dublin Core已經成爲國際標準，後續發展及規格內容由Dub-

lin Core Metadata Initiative（簡稱DCMI）組織管理。

Dublin Core的規範力求簡單而有效，目前廣泛使用於數位典藏物件的後設資料上。Dublin Core的每個欄位都可以選擇性或重覆性使用，大部分的欄位也有一套限制性的細項可選用，可以進一步的表達完整的意義。每個元素欄位可以採任意排序呈現，著錄的規則也可按照計畫需求來訂定。非強制性的特色讓Dublin Core易於掌握及使用，但並不一定適用於所有的物件，對於意義與概念複雜的典藏物件更是如此。

目前Dublin Core欄位有兩種層級，較簡單的Dublin Core欄位中，採用15個元素欄位來描述數位典藏物件；至於完整Dublin Core欄位，則是在15個元素欄位中，再細分修飾語欄位，欄位內容包含使用對象、出處以及版權所有者等，更利於資料被搜尋使用。

15個Dublin Core欄位其下包含的修飾語又分Element Refinement元素精緻化、Element Encoding Scheme元素編碼表兩種，使用的原則有三大項：

1. 一對一原則：Dublin Core一次只描述一個數位典藏品，內容相同但屬於複製本或不同版本的物件，在Dublin Core元素中的創作者(Creator)、貢獻者(Contributor)等欄位的內容會不同。
2. 簡化原則：元素欄位可以不使用任何修飾語，僅保留資料值。
3. 適當的資料值：隨著物件不同，填入元素欄位或是修飾語欄位內的內容也會不同，應仔細斟酌填寫才能發揮後設資料的效用。

Dublin Core 有簡單容易制訂的特性，未經專業訓練的使用者也能制訂後設資料，甚至可以自行發展編輯器；此外，Dublin Core 的彈性大，內容可依需求延伸、選擇，同時又具有可重複性及可變性，符合多樣類型的數位典藏需求；最後，Dublin Core是以英文爲發展基礎，易於國際上通用，是其強力優勢，因而成爲國際間普遍應用的後設資料標準。

表1-1、Dublin Core一覽表[3]

Element	Definition	Qualifiers	
		Element Refinements	Element Encoding Schemes
Title	A name given to the resource	Alternative	
Creator	An entity primarily responsible for making the content of the resource		
Subject and Keywords	The topic of the content of the resource		LCSH MESH DDC LCC UDC
Description	An account of the content of the resource	Table of Contents Abstract	
Publisher	An entity responsible for making the resource available		
Contributor	An entity responsible for making contributions to the content of the resource		
Date	A date associated with an event in the life cycle of the resource	Created Valid Available Issued Modified	DCMI Period W3C-DTF
Resource Type	The nature or genre of the content of the resource		DCMI type vocabulary
Format	The physical or digital manifestation of the resource	Extent Medium	IMT
Resource Identifier	An unambiguous reference to the resource within a given context		URI
Source	Reference to a resource from which the present resource is derived		URI

3 Dublin Core元素清單，數位典藏與數位學習國家型科技計畫後設資料工作組網頁 http://metadata.teldap.tw/standard/standard-frame.html。

Element	Definition	Qualifiers	
		Element Refinements	Element Encoding Schemes
Language	A language of the intellectual content of the resource		ISO 639-2 RFC 1766
Relation	A reference to a related resource	Is version of Has version Is replaced by Requires Is part of Has part Is referenced by References Is format of Has format	URI
Coverage	The extent or scope of the content of the resource	Spatial Temporal	DCMI point ISO 3166 DCMI box TGN DCMI Period W3C-DTF
Rights Management	Information about rights held in and over the resource	AccessRights License RightsHolder	

二、語料庫後設資料相關國際標準

（一）OLACMS 後設資料元素集

「開放語言典藏社群」(Open Language Archives Community, OLAC)是一個由個人或組織所組成的國際性合作協會，成立於2000年12月，目前的主要協調人爲Steven Bird與Gary Simons ，中央研究院鄭錦全院士是諮詢委員，中央研究院語言學研究所黃居仁研究員是顧問。

鑑於全世界許多組織都需要使用到語言資源，例如語言學家、工程師、檔案管理相關人士、軟體發展商和出版商等，大部分的使用者都希望透過單一介面就能取得所需的資源，包含描述語言的相關資訊、用來查詢語言的工

具等，但是不同的語言資源散佈於網路各處，使用者難以一次就找到所需的資源，因此OLAC設立兩個目標，[4]一是針對語言典藏發展一套一致性的實踐指引；二是發展具有互通性的語言資源儲存器與服務中心。

爲了完成這兩項目標，OLAC以Dublin Core Metadata Initiative與Open Archives Initiative（簡稱OAI）所制訂的兩個標準作爲基礎，以達到與國外資料庫進行資料交換、跨資料檢索的目的。

後設資料上，OLAC以Dublin Core的15個元素欄位進行修改，制訂出一套更詳細的後設資料欄位，即爲OLACMS，欄位如表1-2：

表1-2、OLACMS元素欄位

欄位元素	中文	欄位元素	中文
Contributor	貢獻者	Language	語言
Coverage	涵蓋範圍	Publisher	出版者
Creator	創造者	Relation	關聯性
Date	日期	Rights	權利管理
Description	資料描述	Source	來源
Format	資源格式	Subject	主題
Format.cpu	資源cpu格式	Subject.language	主題使用語言
Fomat.encoding	資源編碼格式	Title	資源標題
Format.markup	標誌語言	Type	資源型態
Format.os	作業系統需求	Type.functionality	軟體資源的功能
Format.sourcecode	程式語言	Type.linguistic	語言學上的資源型態
Identifier	資源識別碼		

OLACMS採用四個屬性做更詳細的特性定義，另外還包含一個langs附屬屬性。

1. refine：用來識別較仔細的意義與特性。
2. scheme：規範各元素內容文字是已經標準化的名稱。
3. code：用來標記後設資料中，OLAC特有的標誌系統。
4. lang：每個OLACMS中必有的屬性，註明元素欄位使用的語言。
5. langs：屬於元素的屬性，規範後設資料被閱讀時所採用的語言。

4 張如瑩，〈語言開放典藏社群簡介及語言座標計畫參與狀況〉，語言典藏子計畫，數位典藏國家型科技計畫網頁http://www2.ndap.org.tw/newsletter06/news/read_news.php?nid=888。

（二）跨資料庫檢索網路協定

OLAC也爲有意進行跨資料檢索的語料庫計畫提供了一套解決方案。爲了促進資料庫之間的相互搜尋，OLAC採用Open Archives Initiative（簡稱OAI）[5]所制訂的網路協定--OAI-PMH，透過此協定內容，使用者可以不分系統、應用程式、領域、語言的限制，在網路上搜尋資料，包含後設資料中所登錄的內容也可供搜尋。

OLAC透過OAI-PMH的協定，到各個資料提供者(Data Provider)，也就是語料庫中抓取資料，然後在OAI 服務提供者(Service Provider)中建立一個索引。一旦有使用者在網路上搜尋資料時，就可以快速的看到完整而豐富的索引結果。

如果語料庫計畫團隊想要與OLAC進行跨資料庫檢索，有兩種方法，一種是由語料庫計畫團隊自行架設OAI Data Provider的伺服器，供OAI Service Provider定期抓取資料。第二種是語料庫計畫團隊依照OLAC建議的XML延伸性標誌語法，將語料庫資料製作成相關的文件，提供給OAI Service Provider。

（三） 語言代碼國際標準

ISO 639 系列是國際標準組織所訂定的語言代碼，分爲六個部份。[6] ISO

5 〈檔案管理局97年工作成果--工作分項領域知識資料彙整〉。取自檔案管理局網頁 http://wiki.archives.gov.tw/index.php?option=com_content&view=article&id=556&Itemid=107。

6 ISO 639 的六個部份包括：ISO 639-1:2002 Codes for the representation of names of languages -- Part 1: Alpha-2 code; ISO 639-2: 1998 Codes for the representation of names of languages -- Part 2: Alpha-3 code; ISO 639-3: 2007 Codes for the representation of names of languages -- Part 3: Alpha-3 code for comprehensive coverage of languages; ISO 639-4 Codes for the representation of names of languages -- Part 4: General principles of coding of the representation of names of languages and related entities, and application guidelines （尚未出版）; ISO 639-5: 2008 Codes for the representation of names of languages -- Part 5: Alpha-3 code for language families and groups; ISO 639-6: 2009 Codes for the representation of names of languages -- Part 6: Alpha-4 code for comprehensive coverage of language variants。

639-1 是第一部份，於2002年出版，使用兩個字母編碼，用來標示世界上主要的語言，註冊機構爲 Infoterm (International Information Center for Terminology)[7]。ISO 639-2 是第二部份，於1998出版，使用三個字母來表示語言、大語言(macrolanguage)、語系以及語言集合，其中大語言是數種密切相關語言的泛稱；此外，有 mis, mul, und, zxx 等四個特殊代碼以及使用者自行定義的保留碼區(qaa~qtz)，"mis" 表示「未被編碼的語言」(Uncoded languages)，"mul" 表示內容包括多種語言，且不一一標示，"und"表示「未確定的語言」(Undetermined Language)，"zxx" 表示「沒有語言內容」(No Linguistic Content)，使用於系統要求一定要標示語言，但內容不含語言訊息的情況。 ISO 639-2 的註冊機構是美國國會圖書館。ISO 639-3 是語言開放典藏社群(OLAC)目前推薦使用的語言代碼國際標準，於2007年出版，延伸ISO 639-2，但不包括語系、語言集合，目標是以三碼涵蓋所有語言，包括現存、絕跡、歷史、古老與人工的語言，美國國際語言暑期學院(SIL International)[8] 於2002年起參與ISO 639-3標準的制定，已將SIL 語言代碼整合進入新的標準，並自Ethnologue 第十五版起使用該標準。SIL International 也是 ISO 639-3的註冊機構。 ISO 639-5 於2008年出版，延伸ISO 639-2 中的語言集合，以三碼描述語系、語族、語群或是具有共同性質的語言之集合（如手語、混合語、人工語），註冊機構也是美國國會圖書館。 ISO 639-6 於2009年11月出版，試圖以四碼描述全世界所有之語言文字變體，由於才出版不久，除了主要參與制定的機構GeoLang Ltd 外，[9]採用的單位很少。

---

7 Infoterm: http://www.infoterm.info/。

8 SIL的主要工作項目包含語言發展、學術研究、語言能力培訓、語言傳播媒材開發、翻譯、技術語言發展等項。SIL 出版的 Ethnologue: Languages of the World，在第14版之前使用SIL自訂的語言代碼。

9 Geolang: http://www.geolang.com。

表1-3、 部份語言與語言集合之國際標準代碼[10]（製表：蕭素英）

英語名稱	中文名稱	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	備註
Altaic languages	阿爾泰語系	tut		tut		語言集合
Amis	阿美語		ami			
Amis, Nataoran	豈蘭阿美語		ais			
Artificial languages	人工語言	art		art		語言集合
Atayal	泰雅語		tay			
Austro-Asiatic languages	南亞語系	aav				語言集合
Austronesian languages	南島語系	map		map		語言集合
Babuza	巴布薩語		bzg			
Basay	巴賽語		byq			
Bunun	布農語		bnn			
Buriat	布里雅特語		bua	bua		大語言
Buriat, China	巴爾虎布里雅特蒙古語		bxu			
Buriat, Mongolia	蒙古國布里雅特語		bxm			
Buriat, Russia	俄羅斯布里雅特語		bxr			
Chinese	中文、漢語		zho	zho/chi	zh	大語言
Chinese, Gan	贛語		gan			
Chinese, Hakka	客語		hak			
Chinese, Huizhou	徽語		czh			
Chinese, Jinyu	晉語		cjy			
Chinese, Late Middle	近代漢語		ltc			
Chinese, Literary	文言文		lzh			
Chinese, Mandarin	官話		cmn			
Chinese, Min Bei	閩北語		mnp			
Chinese, Min Dong	閩東語		cdo			
Chinese, Min Nan	閩南語		nan			
Chinese, Min Zhong	閩中語		czo			

10 資料來源 Languages of Taiwan，《Ethnologue : Languages of the World》，Ethnologue:Web，網頁http://www.ethnologue.org/show_country.asp?name=TW，2010年1月21日查詢 ；ISO 639 http://en.wikipedia.org/wiki/ISO_639，2010年1月21日查詢; ISO 639 Code Tables http://www.sil.org/iso639-3/codes.asp，2010年1月21日查詢，List of ISO 639-5 codes http://en.wikipedia.org/wiki/List_of_ISO_639-5_codes，2010年1月21日查詢。

英語名稱	中文名稱	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	備註
Chinese, Old	古漢語		och			
Chinese, Pu-Xian	莆仙語		cpx			
Chinese, Wu	吳語		wuu			
Chinese, Xiang	湘語		hsn			
Chinese, Yue	粵語		yue			
Creoles and pidgins		crp		crp		語言集合
Daur	達斡爾語		dta			
Dongxiang	東鄉語		sce			
English	英語		eng	eng	en	
English, Middle (1100-1500)	中古英語		enm	enm		
English, Old (ca. 450-1100)	古英語		ang	ang		
Esperanto	世界語		epo	epo	eo	
Formosan languages	台灣南島語族	fox				語言集合；階層關係 map:fox
German	德語		deu	deu/ ger	de	
Germanic languages	日耳曼語族	gem		gem		語言集合；階層關係 ine:gem
Indo-European languages	印歐語系	ine		ine		語言集合
Japanese	日語		jpn	jpn	ja	
Jurchen	女真語		juc			
Kalmyk~Oirat	卡爾梅克語、衛拉特語		xal	xal		
Kanakanabu	卡那卡那富語		xnb			
Kavalan	噶瑪蘭語		ckv			
Ketangalan	凱達格蘭語		kae			
Kitan	契丹語		zkt			
Korean	韓語		kor	kor	ko	
Kulon-Pazen	巴宰語		uun			
Manchu	滿語		mnc	mnc		
Mon-Khmer languages	孟高棉語族	mkh		mkh		語言集合；階層關係 aav:mkh
Mongolian	蒙古語		mon	mon	mn	大語言
Mongolian, Classical	古典蒙古語		cmg			
Mongolian, Halh	喀爾喀蒙古語		khk			
Mongolian, Middle	中古蒙古語		xng			

英語名稱	中文名稱	ISO 639-5	ISO 639-3	ISO 639-2	ISO 639-1	備註
Mongolian, Peripheral	內蒙古蒙古語		mvf			
Mongolian languages	蒙古語族	xgn				語言集合；階層關係 tut:xgn
Oirat, Written	書面衛拉特語		xwo			
Paiwan	排灣語		pwn			
Papora-Hoanya	巴布拉洪雅語		ppu			
Puyuma	卑南語		pyu			
Qiang, Northern	北部羌語		cng			
Qiang, Southern	南部羌語		qxs			
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			
Tangut	西夏語		txg			
Taroko	太魯閣語（賽德克語）		trv			
Thao	邵語		ssf			
Tibetan	藏語		bod	bod/tib	bo	
Tibetan, Amdo	安多藏語		adx			
Tibetan, Classical	古典藏語		xct			
Tibetan, Khams	康巴藏語		khg			
Tibetan, Old	古藏語		otb			
Tibeto-Burman languages	藏緬語族	tbq				語言集合；階層關係 sit:tbq
Tsou	鄒語		tsu			
Tungus languages	通古斯語族	tuw				語言集合；階層關係 tut:tuw
Turkic languages	突厥語族	trk				語言集合；階層關係 tut:trk
Uighur	維吾爾語		uig	uig	ug	
Uighur, Old	古維吾爾語		oui			
Yami	達悟語（雅美語）		tao			
Yugur, East	東部裕固語		yuy			
Yugur, West	西部裕固語		ybe			

撰文：蕭素英、詹景勛，致謝：瓦歷斯．浦亞

貳、語料庫建置流程

Procedures for Building

本章概述語料庫的建置流程，下一章將以實例介紹口語、文本、手語等不同類別的語料庫。語料庫建置可分爲語料數位化、系統建置、後設資料建立等三大部分，如圖2-1。[11]前章已介紹後設資料，系統建置則與語料性質與建置目的密切相關，因此本章主要討論語料數位化的流程。

規劃語料庫首先要根據建置目的決定收錄的內容，並依此訂定語料的數位化規格、使用的設備、語料加工的標記集等。爲語音辨識與合成研究而建置的語音資料庫收錄的可能是在錄音室錄製的高品質聲音檔，加工的標記可能是語音的聲學參數，而爲了歷史語言研究所建置的文獻語料庫收錄的語料是代表各時代語言的文獻文字檔，加工標記可能包含文獻出處、分詞標記與詞類語意訊息等。

語料庫要收錄的語料可能是書面的資料，也可能須再調查採集；有些書面資料可能有電子文字檔，有些可能只有手寫或印刷的紙本。決定收錄的內容之後，需要考慮資料授權的問題。根據中華民國著作權法，著作財產權存續於著作人生存期間及其死亡後五十年；法人爲著作人之著作，其著作財產權存續至著作公開發表後五十年；就原著作改作之創作爲衍生著作，亦享有獨立於原著作的權利；製版權自製版完成存續十年。

收錄文本資料的語料庫需要獲得各文本著作財產權所有人的授權。古籍雖然任何人都可以自由利用，收錄於語料庫時仍須注意所採用版本的版式可能仍擁有製版權，校勘、註釋等衍生著作的著作財產權也可能仍存續。若收錄的語料是由發音合作人提供，考量著作財產權以及個人肖像權、個人資料與隱私保護等學術倫理議題，需要取得發音合作人的同意授權書。

---

11 本圖由詹景勛製圖；後設資料之部份參考「數位典藏與數位學習國家型科技計畫後設資料工作組」以及「拓展台灣數位典藏－內容建置與整合子計畫」辦公室提供之資料，影音收錄之部份流程由閩客語典藏研究助理余瓊怡小姐參考李道明〈影音檔案數位化之規劃與流程〉（http://content.ndap.org.tw/index/?dl_id=76 ，2009年1月31日下載）與實際經驗製作初稿，「台灣南島語數位典藏」研究助理瓦歷斯．浦亞曾參與討論，在系統建置、資料備份方面提供許多建議，謹此致謝。

此外，依據著作權法，受雇人於職務上完成之著作，其著作財產權歸雇用人享有；[12]出資聘請他人完成之著作，其著作財產權依契約約定歸受聘人或出資人享有，[13]未約定者其著作財產權歸受聘人享有 ，語料庫建置可能牽涉錄影、錄音、攝影、文字轉記、翻譯、系統設計建置、工具程式開發等工作，均爲著作權法保護之「著作」，從事這些工作的若爲委外或勞務承攬的工作人員，依法可能取得著作財產權，簽訂契約時應特別注意。

影音語料錄製完成後，除了存檔保存，須剪輯成合適的段落，刪除不適用或含個人隱私等敏感內容的片段，[14]或以模糊處理、靜音的方式處理保護隱私。[15]剪輯處理之後再輸出永久典藏格式檔案保存，[16] 並轉出較低規格的公開

12 中華民國著作權法第11條：「受雇人於職務上完成之著作，以該受雇人為著作人。但契約約定以雇用人為著作人者，從其約定。依前項規定，以受雇人為著作人者，其著作財產權歸雇用人享有。但契約約定其著作財產權歸受雇人享有者，從其約定。前二項所稱受雇人，包括公務員。」

13 中華民國著作權法第12條「出資聘請他人完成之著作，除前條情形外，以該受聘人為著作人。但契約約定以出資人為著作人者，從其約定。依前項規定，以受聘人為著作人者，其著作財產權依契約約定歸受聘人或出資人享有。未約定著作財產權之歸屬者，其著作財產權歸受聘人享有。依前項規定著作財產權歸受聘人享有者，出資人得利用該著作。」

14 如手語辭彙庫這類根據腳本拍攝的影片，同一辭彙可能拍攝多次，再剪輯出適用的片段。

15 如荷蘭的 IFA 對話影音語料庫(IFA Dialog Video corpus)這類自由對話語料庫，雖已請發音人避免提及姓名等敏感內容，仍不免有一些不適宜的片段須刪除後才能釋出。

16 永久典藏的格式是考量當前技術、共通性、處理速度、儲存空間與成本等種種因素後，能容許的最高規格。數位典藏與數位學習國家型科技計畫目前推薦的影訊永久典藏格式是MPEG-2，資料傳輸率8M/sec.，聲音格式為不經壓縮的WAVE格式，取樣頻率 44.1KHz, 16-24bit。

格式檔案[17]，進入文字轉記階段。紙本資料可以先掃描，使用文字辨識軟體轉爲文字或直接人工輸入文字。重要的原典建議將掃描的書影一併典藏，讓使用者查閱時可以連結至原文獻圖檔。[18]

文字檔需要二次校對，第一次校對的重點是檢查內容是否與底本相符，第二次校對同時檢查篇名、頁碼等出處標記(Markup)是否正確。實務上，一校可以由兩人同時輸入或轉記同一份文件，再以電腦工具程式自動比對，因爲兩人在同一處同時犯錯的機率不高，可以很快找出錯誤的地方加以校正。

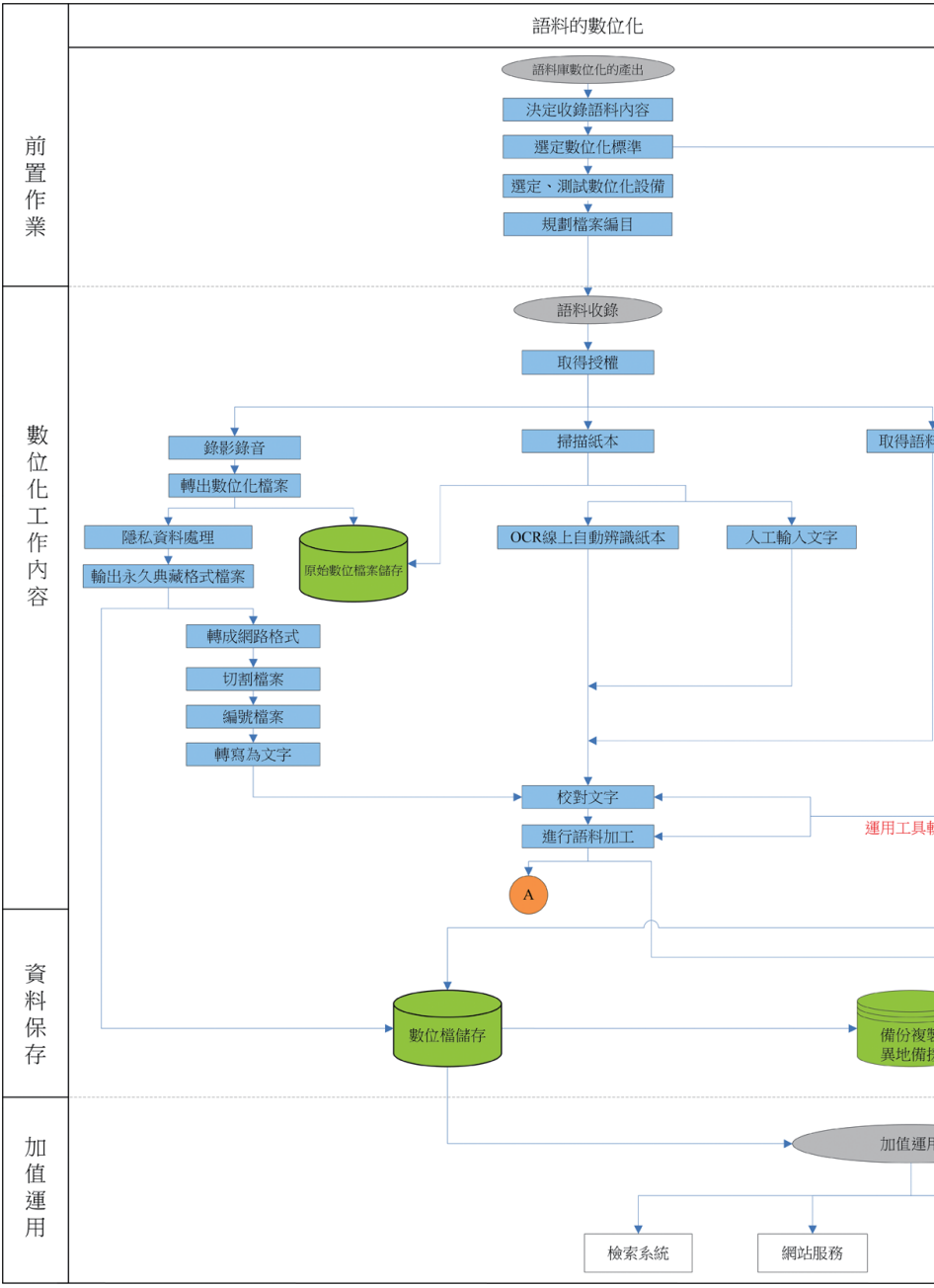
校對完成的檔案可以開始標記等加工工作。不同用途的語料庫使用的標記集也不同，如語音語料庫可能會使用韻律標記集，對話語料可能使用言談標記集，一般語料庫可能使用詞類、構詞語法標記……等。單以人工標記語料不但耗時費力，也難以維持一致的品質，通常需要設計、開發自動處理程式與人工校正的工具界面。

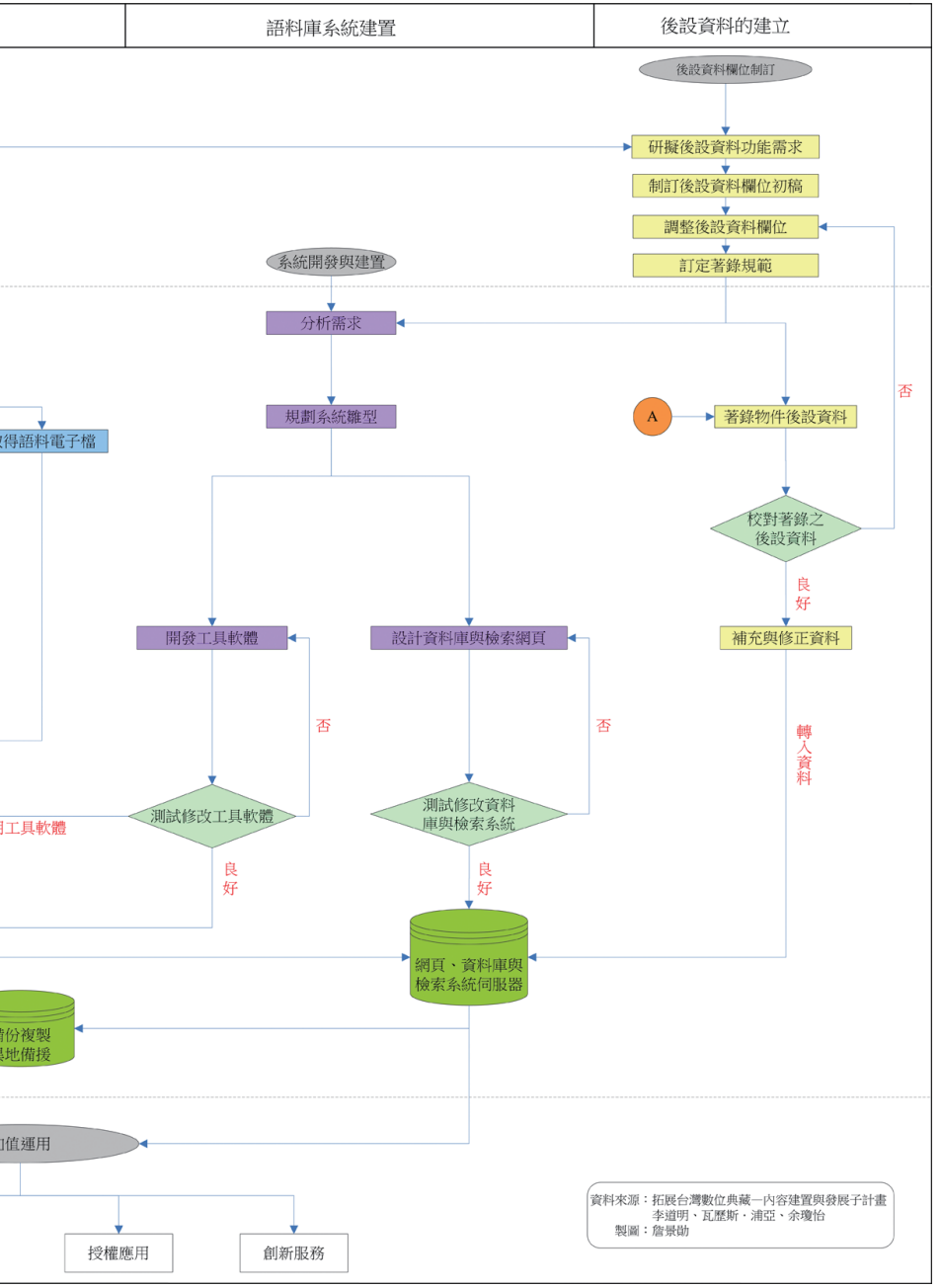
語料庫系統建置包括架構資料庫、設計檢索系統、開發工具軟體與工作界面、建置網站等工作，這些工作在與料庫規劃初期就要同時展開，配合語料處理的進度按部就班完成。最後，語料庫建置完成上線後的管理維護相當重要，因此，檔案備份、異地備援、系統安全等問題也應一併規劃、切實執行。

---

17 公開格式通常經過壓縮，可能有多種版本供不同網路頻寬、平台的環境使用。

18 如「閩南語典藏－歷史語言與分布變遷資料庫」(http://southernmin.sinica.edu.tw)提供了《荔鏡記書影》、「閩客語典藏」(http://minhakka.ling.sinica.edu.tw/)的《廈英大辭典》、《增補廈英大辭典》、《台日大辭典》等的查詢結果均提供至原典相應頁面圖檔的連結。


語料的數位化
前置作業
數位化工作內容
資料保存
加值運用
語料庫數位化的產出
決定收錄語料內容
選定數位化標準
選定、測試數位化設備
規劃檔案編目
語料收錄
取得授權
錄影錄音
轉出數位化檔案
隱私資料處理
輸出永久典藏格式檔案
原始數位檔案儲存
轉成網路格式
切割檔案
編號檔案
轉寫為文字
掃描紙本
OCR線上自動辨識紙本
人工輸入文字
校對文字
進行語料加工
A
數位檔儲存
檢索系統
網站服務


語料庫系統建置
後設資料的建立
後設資料欄位制訂
研擬後設資料功能需求
制訂後設資料欄位初稿
調整後設資料欄位
訂定著錄規範
系統開發與建置
分析需求
規劃系統雛型
A
著錄物件後設資料
否
校對著錄之後設資料
良好
補充與修正資料
轉入資料
開發工具軟體
設計資料庫與檢索網頁
否
否
測試修改工具軟體
測試修改資料庫與檢索系統
良好
良好
網頁、資料庫與檢索系統伺服器
值運用
授權應用
創新服務
資料來源：拓展台灣數位典藏—內容建置與發展子計畫
李道明、瓦歷斯．浦亞、余瓊怡
製圖：詹景勛

參、語料庫建置實例

Samples for Building

語言是語料庫收藏的主要內容，但是語言的呈現擁有多種形式，例如文字、手語與一般口語等，針對這些不同的呈現方式，在進行數位典藏計畫時就必須以不同的數位化媒介與工作流程來對應。

文字化的語言通常出現在文獻或是器物之上，數位化典藏時可能要進行文獻的翻拍、掃描、文字輸入等手續，有些文本可能已經有電子版本，口語則是採用錄音的方式來記錄，接著進行語音轉寫、標記等過程，至於手語則必須依賴影像來傳達，語料加工的方式也與前兩者大相逕庭。

本章分別收錄文本、口語、影像與語言分布等幾個語料庫類型的建置實例，希望藉由這些計畫的經驗分享，使讀者更瞭解各式語料庫的製作流程，參考運用。

文本方面收錄「中央研究院現代漢語平衡語料庫」以及「中文詞彙網路」等中文詞彙知識檢索系統的建置實例。前者介紹世界上第一個有完整詞類標記的漢語平衡語料庫，後者包含詞網、漢字知識本體、詞頻分布等幾個以詞語爲主要內容的語料庫檢索系統之簡介。

口語方面收錄三個實例。「語言分布GIS地理資訊系統」除了一般辭彙的田野調查之外，還結合地理資訊系統來研究語言分布的情形。「台灣兒童語料庫--閩南語兒童語料庫」定期錄音以追蹤兒童語言的發展，而「台灣國語口音之社會分布典藏」則以街頭問卷訪談的方式收錄口語。

在影像方面，「台灣手語影像辭典」利用錄影的方式建立出手語的教學、查詢與釋義的檢索網站，在內容的建置流程上就與文本、口語有所差異。國內外也有一些口語語料庫除了聲音以外，還收錄影像，這類影音口語語料庫除了標記聲音外，還可能標記手勢、表情等非口語訊息。中央研究院語言學研究所曾淑娟副研究員的「新世紀語料庫－多媒體的語言呈現與典藏」，即屬於此類，但限於時間匆促，本書未能收錄該計畫建置的流程，有興趣的讀者可以自行參考該計畫網站。[19]此外，一般影音口語資料庫由於授權與隱私問題，通

19 新世紀語料庫－多媒體的呈現與典藏 http://mmc.sinica.edu.tw/。

常不釋出影像檔，而只釋出標記檔，但荷蘭的 IFA 對話影音語料庫(IFA Dialog Video Corpus)[20] 致力於建置公開的影音語料資源，釋出所有影像、聲音、標記檔與相關文件，讀者也可參考。

希望這些實例能讓讀者更清楚不同類型語料庫的建置流程，供有意參與語料庫建置的機構與計畫參考。

一、中央研究院現代漢語平衡語料庫數位化工作流程簡介

製作日期：2005/10/13
更新日期：2010/01/25

語料庫爲本(Corpus-Based)的研究是近年語言學及計算語言研究的一個重要發展，其影響更遠及文學及社會學的計算研究。以理論語言學或自然語言處理研究來說，語料庫所擔負的功能是在無窮衍生的語言事實中抽出一個具有代表性的樣本。這個樣本不能太大，否則便失去了抽樣的意義與優點；又不能太小，否則即無法提供足夠的訊息，也無法提供大量素材進行統計研究或作爲測試語料。因此，語料庫構建的第一個大問題是：如何以有限的語料代表複雜的當代語言全貌？[21]

「中央研究院現代漢語平衡語料庫」簡稱「研究院平衡語料庫」(Sinica Corpus)，是世界上第一個有完整詞類標記的漢語平衡語料庫。這個語料庫由中央研究院資訊科學研究所陳克健研究員與語言學研究所黃居仁研究員共同帶領的「中央研究院詞知識庫小組」完成。該小組自1990年前後便開始致力於

20 van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2009). "Promoting free Dialog Video Corpora: The IFADV Corpus Example," in M. Kipp et al. (Eds.): Multimodal Corpora, LNAI 5509, pp. 18–37, 2009. 該語料庫網址為：http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/。

21 詞庫小組.1995.《研究院語料庫的內容及說明》，中文詞知識庫小組技術報告 #95-02，南港，中央研究院。

中文語料庫的收集，[22]至1994年止已收集有近二千萬字之現代漢語語料及超過五百萬字之古代漢語語料。[23]有了這些處理中文語料庫及大量處理電子詞庫詞條的經驗爲基礎，在1994年分別得到了中央研究院「中文資訊」跨所研究群之專案計畫及國科會計畫補助，乃開始著手進行現代漢語平衡語料庫的建構。爲兼顧理想與實用性，初步目標定爲兩百萬詞，爲傳統小規模平衡語料庫之兩倍，最終目標定爲五百萬詞。1996年開放供各界使用，1997年開放的研究院語料庫3.0版已達到五百萬詞的預計規模。2001年國家型數位典藏科技計畫展開，詞庫小組認爲應持續收集近年之語料，使語料樣本能完整呈現二十世紀臺灣使用漢語的全貌，因此以新五百萬詞爲目標進行知識典藏工作，目前介面已升級至4.0版，提供更完整的語料條件檢索功能。

數位化工作流程說明：

該計畫的數位化作業，大致依照下列六項步驟進行，依序分別爲：一、詞類分析、定義及確定；二、選擇語料文本來源；三、程式抓取電子語料；四、程式自動分合詞及詞類標記；五、人工詞類檢查；六、匯入語料庫。茲分別介紹如次。

（一）詞類分析、定義及確定：

分詞規範的研擬分爲兩種方式進行，一方面是邀請台灣知名的學者專家召開討論會，就其專業領域的角度，對分詞規範的大方針進行討論；另一方面

22 Huang, Chu-Ren and Keh-jiann Chen. 1992. A Chinese Corpus for Linguistics Research. In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France.

23 Huang, Chu-Ren 1994. Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In Matthew Chen and Ovid Tzeng Eds. In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change. Pp. 165-186. Taipei: Pyramid.

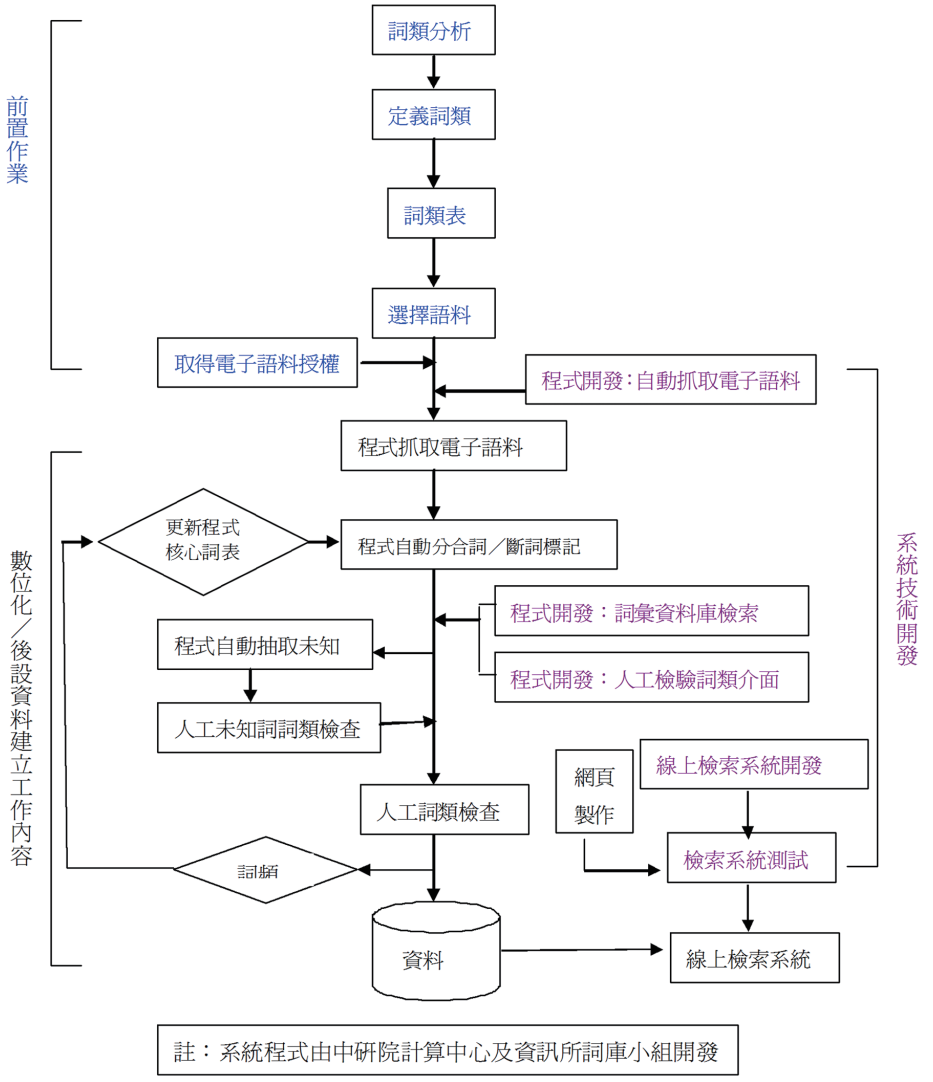


圖3-1-1、現代漢語平衡語料庫工作流程圖
圖片提供者：中央研究院語言學研究所　盧秋蓉小姐

則是中央研究院詞庫小組根據分詞規範，實際從事語料分析，從上百萬的語料中，整理出分詞標準的細節規定。然後，於1998年舉行分詞規範公聽會，1999年中文分詞原則正式通過爲國家標準，編號CNS14366。[24]中央研究院詞庫小組再依此規範進行詞類分析、定義及確定之工作。

《資訊處理用中文分詞規範》有下列兩個突破：(1)提出分級的觀念及確立信、達、雅三級的標準。最容易達到的「信級」訂爲基本資料交換的標準；以技術上較難，但自動分詞程式仍可達到的「達級」作自動翻譯、資訊檢索等自然語言處理的標準；至於最需要人工分詞才能達到的「雅級」則視爲電腦處理、理解中文之最高目標。(2)把分詞規範分成不變核心（分詞單位定義及基本原則），以及可變準則（輔助原則）。在確定分詞規範架構後，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。[25]

（二）選擇語料文本來源

平衡語料之抽取以自中央研究院詞庫小組現有之語料（近二千萬字之現代漢語語料）中取得爲優先，但也同時透過不同管道取得不同文體、內容之語料。以下依來源之不同種類大致列舉。

1. 交換取得之語料：此項包括經由合作計畫交換取得的，如中國時報，洪建全基金會，師大國語中心。或是由計算語言學會內部之語料作共同體(Consortium)間交換語料而得，如由致遠科技及台大取得。
2. 直接向版權所有單位取得：慷慨提供該計畫版權語料做學術研究用的有：天下雜誌社，國語日報社，資訊傳眞雜誌社，「女人女人」製作單位，「伴我成長」製作單位，「我們一家都是人」製作單位以及許多中研院內的單位等。另有舊金山州立大學畢永峨，清大郭賽華，交大劉美

24 「資訊處理用中文分詞原則」國家標準 ，檔案編號CNS14366。

25 「資訊處理用中文分詞規範」設計理念及規範內容。

君，輔大楊承淑等多位教授提供他們轉寫(Transcribe)的口語資料。

3. 由公共區域取得的公共資料：大部份由聯合新聞網、中時電子報及電子佈告欄(BBS)或蕃薯藤等萬維網中取得。

（三）程式抓取電子語料

使用程式CKIP Corpus&Spider1.4.6a抓取線上電子語料。助理使用電子語料抓取程式，需先選擇語料來源，再選取欲匯入語料庫之文章。由於語料來源媒體之分類並不一致，而現代漢語平衡語料庫分類爲六類：文學、哲學、藝術、科學、社會、生活，故需將文章重新分類，以便匯入（圖3-1-2至圖3-1-6）。

目前，以主題爲準，訂出平衡語料庫的內容比例爲：文學20%、哲學10%、藝術5%、科學10%、社會35%、生活20%，根據此參考值爲基準選取語料。

圖3-1-2、抓取電子語料工作畫面。（示範者：邱智銘）

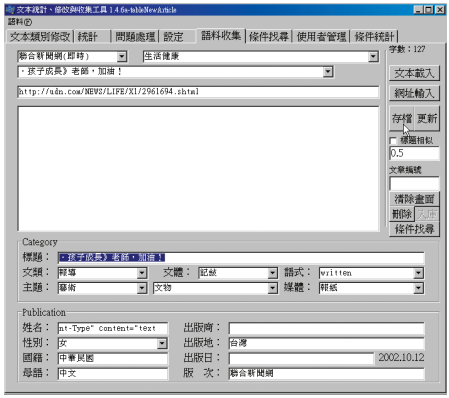


圖3-1-3、語料收集畫面

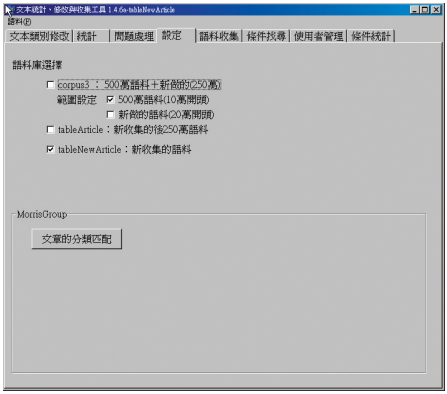


圖3-1-4、確認需匯入之語料庫位置

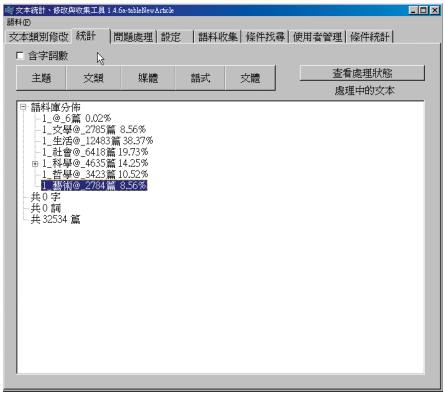


圖3-1-5、目前語料庫收集文章情形

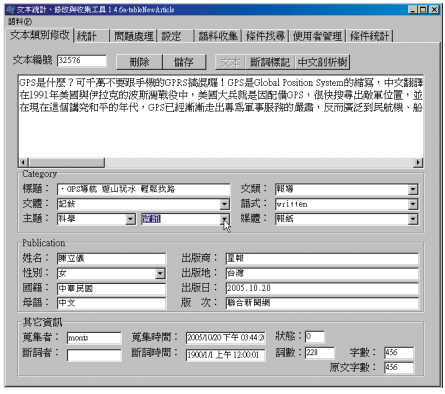


圖3-1-6、語料修改重新分類的畫面

（四）程式自動分合詞及詞類標記

語料選取完畢，接下來的工作是標記詞類，但是在這之前，還要先爲語料做斷詞工作，唯有每個詞區隔非常明確之後，才能標記詞類。目前機器自動斷詞的正確性約達95%。

基本上，自動斷詞的步驟是以中研院辭典中的八萬目詞爲基礎，切分爲一個一個獨立的詞。未列在辭典中的成分，則以字爲單位，一一切分開。然後佐以構詞律對衍生性強的詞綴及數字組合成分進行結合詞彙的工作。而目前分詞的原則是採用中央標準局委託中華民國計算語言學學會研擬的《中文資訊處

理分詞規範》國家標準草案的原則切分。

機器自動斷詞是使用CKIP Tag Tool V1.8a系統，該程式即是一個協助詞類標記檢查的輔助工具，輸入欲執行自動斷詞之語料的文本編號，執行自動斷詞後，程式會將斷詞後之語料顯現於語料本文下方欄位（圖3-1-7、圖3-1-8）。

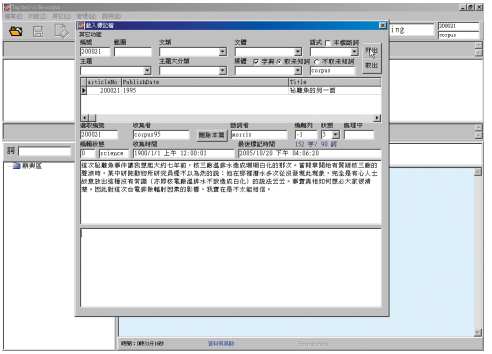


圖3-1-7、選取欲執行自動斷詞之語料

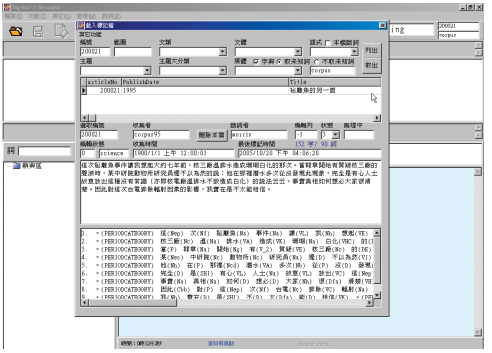


圖3-1-8、自動斷詞執行完畢

（五）人工詞類檢查

使用程式自動斷詞及詞類標記後，由於斷詞會因文章內容而導致詞彙不同的切斷方式，故爲避免斷詞與文義不符，再由助理以人工方式作詞類檢查。

在進行人工確認時，會利用中文斷詞編輯介面。系統進行人工確認，每次以一句爲單位，並列出上下句供參考。確認無誤後，再以上下鍵移動繼續進行人工詞類檢查之工作（圖3-1-9）。

若發現斷詞不恰當，於需修改的詞彙上點選，即可進行重新修改（圖3-1-10至圖3-1-12）。

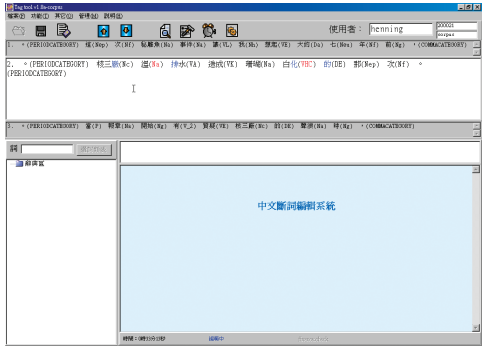


圖3-1-9、中文斷詞編輯系統

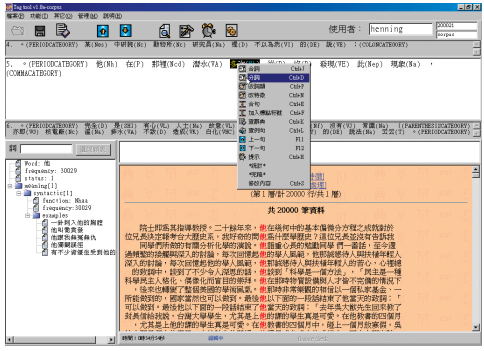


圖3-1-10、修正斷詞畫面

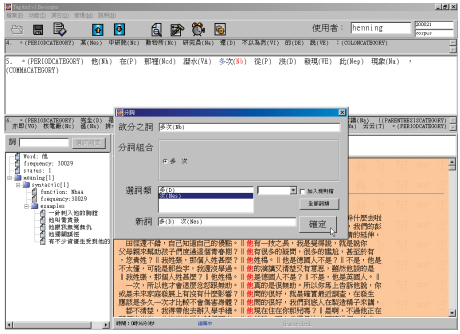


圖3-1-11、輸入欲修正之詞彙及斷詞方式

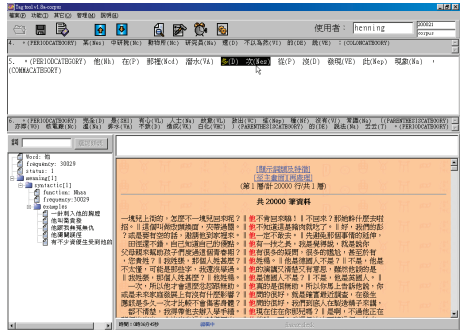


圖3-1-12、斷詞修正完畢

（六）匯入語料庫

人工詞類檢查進行完畢後，再將完成詞類斷詞和標記之語料，以網路傳送至中央研究院計算中心，再由該單位匯入現代漢語標記語料庫。

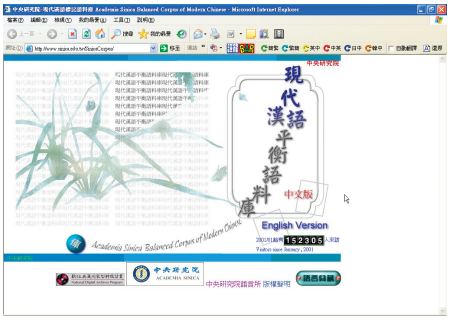


圖3-1-13、現代漢語平衡語料庫網頁

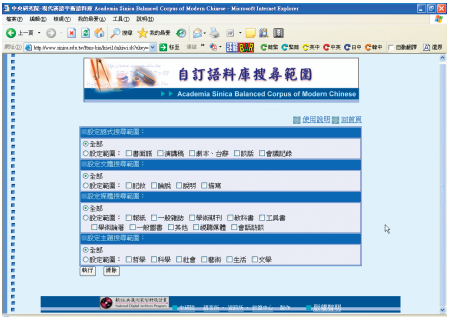


圖3-1-14、現代漢語平衡語料庫搜尋網頁

製作單位：數位典藏國家型科技計畫 內容發展分項計畫
中央研究院語言學研究所 語言典藏計畫
文字撰寫：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻
中央研究院語言學研究所語言典藏計畫助理 盧秋蓉、邱智銘
圖片拍攝：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠
圖文編輯：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、陳秀華
致謝：感謝中央研究院語言學研究所「語言典藏」之計畫主持人 鄭錦全院士與共同主持人黃居仁老師及 陳克健老師和助理盧秋蓉小姐、邱智銘先生撥冗指教及協助拍攝與提供資料，特別致謝。

二、中文詞彙知識檢索系統之建置流程

製作日期：2010/01/25

中央研究院語言學研究所的中文詞彙網路小組(Chinese Wordnet Group)，結合分析詳盡的中文詞彙詞義資料與網路科技的技術，初步開發了中文詞彙網路(Chinese Wordnet)，以利於提供中文詞彙詞義的相關訊息，便於從事中文詞彙詞義的研究所需。

我們嘗試以中央研究院中文詞彙網路小組所分析完成之詞義資料爲語料，運用系統結構化分析與設計方法建構符合需求的中文詞彙網路，也由於在實際應用上，詞彙知識庫屬於語言處理研究基本之參考資料，因此本資料庫可預期成爲中文語言處理與知識工程不可或缺的基底架構。

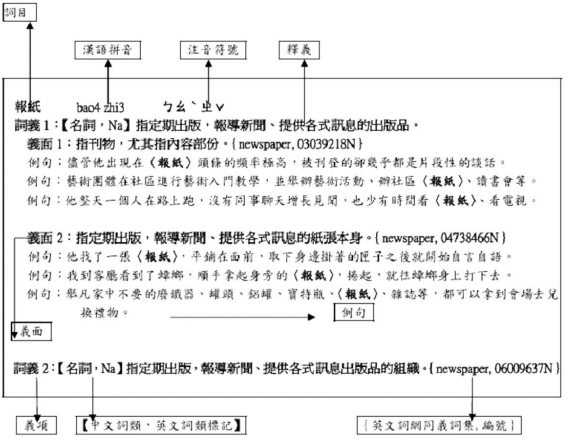


圖3-2-1、中文詞彙條目內容範例

一般全文檢索系統，只能以所檢索的標的文件所含有的文字資訊進行檢索，無法就其字詞義或周邊相關資訊進行檢索，這種檢索功能顯然不能滿足語言研究的需求。由過去相關研究中可以整理出語言學研究所需要考量的各項詞彙資訊，因此我們以中文詞彙爲研究對象，經過嚴謹的分析研究後，對每一個中文詞彙呈現出詞目、詞義、領域、釋義、語義關係、英文對譯、例句、附註等內容。經過嚴謹分析的詞彙資訊，除可有系統性地保存詞彙知識外，更可滿足多元的語言學相關研究使用。

中文詞彙網路小組的研究成果，從2003年初起，至2009年6月止，累積的成果共有9,362個詞形，25,173個詞義。中文詞彙知識檢索系統之開發則將

上述累積之工作成果依照結構化系統分析與設計方法在網際網路上建構一工作平台提供相關研究人員查詢使用，除了研究成果共享之目的外，更希望藉此作爲中文詞彙知識網路研究之基礎架構。

中文詞彙知識檢索系統之系統分析與設計可分爲功能模組分析及資料模組分析兩個主題進行探討。

（一）功能模組分析：

依照研究目的，中文詞彙知識檢索系統透過網際網路提供方便的環境讓使用者進行詞彙知識的檢索，系統包含了儲存詞彙資料的詞彙資料庫以及網路使用者介面，在網路使用者介面上的主要功能模組分爲詞彙查詢及詞彙索引兩部份，詞彙查詢功能提供使用者以輸入關鍵字的方式進行詞彙知識查詢，考量詞彙查詢上的彈性，本系統加入模糊查詢模組，當使用者輸入的關鍵字查詢條件無法對應到精確的詞彙資料時，系統便自動呼叫調整機制，轉換查詢條件使其對應至相關的詞彙資料，如此可大大增加查詢上的彈性，提供使用者更方便的使用環境。其次，在索引詞類方面，本系統提供了44種精確的詞性分類供使用者點選索引。系統中的資料傳遞可由資料流程圖（圖3-2-2）表示：

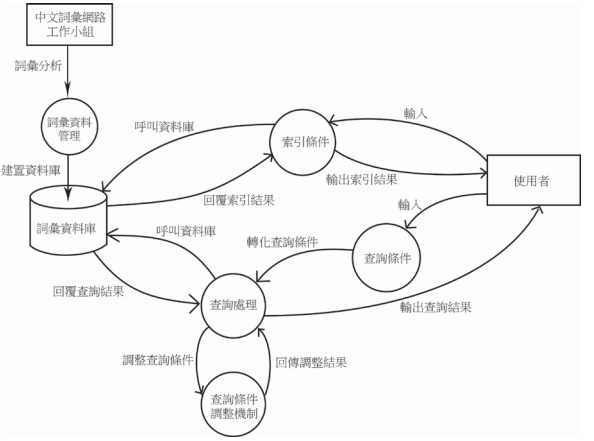


圖3-2-2、中文詞彙知識檢索系統資料流程圖

（二）資料模組分析：

在詞彙資料庫中包含了9,362筆精確分析的中文詞彙資料，這些資料可以透過不同的查詢條件進行展示，圖3-2-3爲查詢功能之實體關係模型資料模組分析，圖中包含了查詢條件、詞彙資料與調整條件等三個實體類型，以及查詢處理一個關係。透過實體關係圖可以了解查詢條件與詞彙資料之間多對多的對應關係，代表查詢條件可以同時查詢多筆詞彙資料，而每筆詞彙資料也可被多個查詢條件所呼叫，因此查詢條件與詞彙資料兩個實體之間存在多對多的關係。然而在查詢條件無法搜尋到精確符合的詞彙資料時系統將自動調整查詢條件，因此圖中呈現出查詢條件進行模糊比對時調整條件實體所依據的屬性項目。

中文詞彙知識檢索系統之設計簡述如下：

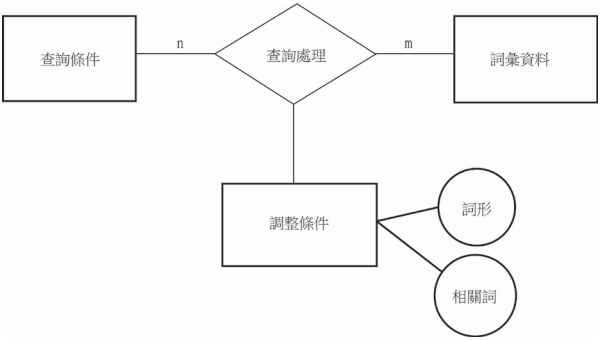


圖3-2-3、中文詞彙知識檢索系統實體關係圖

中文詞彙知識檢索系統因應網路使用環境特性選以Microsoft Windows Server搭配Access資料庫以及Active Server Page程式爲開發系統之作業環境，具有使用方便且相容性高的優點，因此不但可以使整個系統開發過程更爲順暢，同時亦兼顧了後續維護的可行性。

中文詞彙知識檢索系統在設計階段參酌使用者角度與系統功能發展角度共同建立起系統架構與操作流程，詳細描述系統範圍內相關之資料結構以及操作步驟，特別是設計一套整合式即時查詢的方式，提供系統使用者一個整合查詢介面快速查詢以及瀏覽有興趣的各個詞義資訊如圖3-2-4所示。系統提供的

查詢範圍，有：中文詞彙、釋義內文、英文對譯、中文詞彙模糊查詢、注音、漢語拼音等，使用者可依不同訊息或不同需求來選擇查詢的方式。主要的出發點是能對詞彙與語義相關連的內容，做廣泛而有效的檢索，也是藉著檢索的比對，來確保釋義語言及語義區分的一致性及強健性。在查詢結果之呈現上，以詞彙編號爲主鍵由資料庫中提取出詞目、詞義、領域、釋義、語義關係、英文對譯、例句及附註等項目依序排列，透過瀏覽器可清楚呈現給使用者。

圖3-2-4、中文詞彙網路查詢介面

圖3-2-5、詞彙查詢結果畫面

上述依照結構化系統分析與設計方式所開發之中文詞彙知識檢索系統可有效作爲相關研究之詞彙知識來源。以「比得上」爲例做說明，若選擇「中文詞彙」爲查詢範圍，則詞彙查詢結果如圖3-2-5所示，會將查詢詞彙的相關訊息表示在介面上；相同地，若以「比得上」當關鍵詞，選擇「釋義內文」爲查詢範圍，則資料庫內所有詞彙釋義內文有「比得上」的詞彙將會顯示在介面上，同理可運用在其他查詢功能。

在圖3-2-5中，查詢的詞彙「比得上」，提供了漢語拼音、國語注音、釋義、語義關係、英文對譯、例句等訊息，其中，「語義關係」的部份，除了標

示兩詞彙的關係，如：同義詞，尚可做進一步的連結，以讀取「比」的訊息；「英文對譯」的部份，則可以連結至SinicaBOW的網頁(http://bow.sinica.edu.tw/)，以讀取「比得上(Compare)」的相關訊息。此外，介面上提供的例句，則是取自中研院平衡語料庫裡，最具代表的實際語料。

在查詢過程中，若精確的關鍵字條件找不到相符資料時，系統將對使用者輸入之關鍵字查詢條件自動轉換成模糊條件查詢並且在結果畫面上顯示「你查詢的詞彙，目前尚未分析其詞義，以下以模糊查詢給予參考」的字樣，提醒使用者可點選查詢相關詞彙資料。如圖3-2-6以查詢「一點鐘」爲例。

圖3-2-6、模糊查詢結果畫面

在經過上述階段完成系統開發後，爲永續經營中文詞彙網路系統並持續提供詳盡的中文詞彙知識，未來著眼於資料內容的更新管理，在系統與資料來源間制訂一套定期的資料存取、使用者互動交流、更新與同步化之機制。進行自然語言處理研究經常需要對詞彙語義進行深入探討，分析詳盡的詞彙資料除了本身所包含的知識價值之外，更可提供相關應用研究最精緻之素材。本研究以中央研究院語言學研究所中文詞彙網路研究小組近年來豐厚之詞彙分析研究

成果爲基礎，將9,362筆深入分析之Synset資料建構於中文詞彙網路上，以人性化整合查詢介面透過網際網路呈現，除了提供相關研究人員以及有興趣的使用者查詢檢索外，更希望藉此系統作爲進階中文詞彙知識研究之系統化實作參考基礎，進而達到研究成果共享與學術交流之目的。

除了上述的中文詞彙網路之外，詞彙網路小組多年來致力於各種詞彙網路的研究，與架構多種以詞彙資料爲基礎、使不同時空的典藏知識內容可以轉換成互通訊息的知識本體資料庫。近年來陸續建構的資料庫如下：

1. 中文詞彙網路(Chinese Wordnet)：

   中文詞彙網路爲一詞彙知識庫系統，提供完整的中文詞義(Sense)區分資料。收錄的詞條以現代漢語通用語詞爲範圍，提供各詞目完整正確的訊息。在詞義理論與認知研究方面，本詞彙網路的詞彙知識庫系統可成爲基本參考資料；在實際的應用上，這個資料庫則可望成爲中文語言處理與知識工程不可或缺的基底架構。

   （網址：http://cwn.ling.sinica.edu.tw/ 免費開放使用）

   （對內版：http://140.109.150.20/）

2. 中英雙語知識本體詞網(Sinica BOW)：

   本資料庫爲全世界第一個知識本體詞網；以Princeton WordNet架構爲基礎，並以以台灣地區的語言使用爲經驗基礎。提供的訊息包含中英雙語跨語言資訊轉換、語言資訊與概念架構（知識本體）的連結、詞義的區分與詞義關係的連結以及使用領域，在使用語言與詞彙資料的基礎上，提供了知識運籌的基本架構(Infrastructure)。讓不同來源的典藏知識內容，可以轉換成互通的 (Inter-Operable) 訊息。

   （網址：http://bow.sinica.edu.tw/ 免費開放使用）

3. 中文詞彙特性速描系統(Chinese Word Sketch)：

   本系統是一個結合了鉅量語料庫的語法知識產生系統。除了一般的關鍵詞及語境查詢外，更提供了詞彙特性速描(Word Sketches) 、語法關

係以及同近義詞分析等自動產生的語法知識。「中文詞彙特性速描系統」與十四億字的LDC Chinese Gigaword語料庫結合後，提供了絕大部分中文詞彙實際使用的規則性描述，可應用於辭典編撰、華語文教學、語言學研究與自然語言處理。

（網址：http://bow.sinica.edu.tw/ 免費開放使用，需先申請使用帳號）

4. 漢字知識本體(Hantology)：

中文的漢字書寫系統跨越三千年，其所隱含的知識表達系統可說是最穩定、表達知識也最豐富的知識本體。漢語知識本體中表達的知識包括字形結構、意符、聲符、古音、中古音、現代音、字義、異體字關係以及詞彙衍生，並包括不同時期形、音、義的變化和關係，透過漢字知識本體的建構可系統性的表達漢字知識。

（網址：http://hantology.sinica.edu.tw/ ，目前尚未開放使用）

5. 遷台後歷屆總統元旦及國慶文告資料庫(Taiwan Presidential Corpus)：

本語料庫蒐集了1955至2007年間四位總統在國慶日及元旦的演說文告。語料庫中每個句子都做了斷詞的處理與詞類的標記，以方便使用者分析。本語料庫是特別爲中文政治語言分析所設計而成的，可作爲臺灣政治語言的一個代表性樣本。

（網址：http://140.109.19.114/president/ 目前僅供院內IP使用）

6. 中文詞彙詞頻分布系統：

本系統以中文十億詞語料庫(Chinese GigaWord)爲基礎，可提供查詢各詞詞頻差異，用於探討詞彙的使用情形，亦可做爲兩岸及其他華語地區的詞彙對比研究工具。

（網址：http://140.109.150.156/sinica/cwordfreq/ ，目前僅供院內IP使用）

三、語言分布GIS地理資訊系統建置數位化工作流程簡介

製作日期：2005/12/05
更新日期：2010/01/25

本節介紹中央研究院語言學研究所語言典藏第一期子計畫「 閩南語典藏－歷史語言與分布變遷資料庫」在語言分布地理資訊系統的建置流程。

閩南語和客家話是漢語的主要方言，是重要的語言資產，主要分布於福建南部、廣東、台灣與東南亞，但受到學校教育、媒體大量使用國語（普通話）的影響，這二種語言能使用的人口有越來越少的傾向，成爲相對的弱勢語言，亟待研究與保存。

臺灣人口流通量大而頻繁，語言接觸日益密切，語言生態丕變，方言中的「地區變體」與「社會變體」之消長分合，變化快速。近年方有學者開始積極調查繪製臺灣地區語言地圖，然電子語言地圖的繪製還在起步階段，展現語言分布變遷情況的語言地圖更付之闕如。

該計畫爲中央研究院「語言典藏」分項計畫「漢語典藏與典藏架構」的五個子計畫之一，以大眾文學之劇本、歌仔冊二種文體爲範圍，建立閩南語、客家語語料庫。並以閩客雜居的新竹縣新豐鄉爲對象，調查居民用語，研究閩客用語交互之影響。從歷史語言與語言分布兩點切入，結合文獻語言與生活語言，進行語言標誌，建置閩客語語料庫、詞彙庫與語言分布地理資訊系統，爲學界提供有力的研究工具。

由於新竹縣新豐鄉是閩客雜居的鄉鎮，所以該計畫以新竹新豐鄉爲範圍，進行語言分布的調查研究，發展語言分布地理資訊系統。

數位化工作流程說明：

該計畫的數位化作業，大致依照下列六項步驟進行，依序分別爲：（一）訪談及錄音；（二）街路定位及數位化地圖；（三）問卷輸入；（四）錄音備份；（五）圖層製作；（六）匯入資料庫。茲分別介紹如次。

（一）訪談及錄音

以新竹縣新豐鄉爲範圍，進行語言分布的調查研究，調查以訪談的方式進行，藉由問卷記錄訪談者平日對談所使用的語言，並且製作圖卡及字卡，請受訪者依照圖卡及字卡發音並錄音。但因在訪談過程中，受訪者不易同意訪談，故先拜訪鄰長，再由鄰長陪同前往，取得受訪者信任，進行訪談與錄音（圖3-3-1至圖3-3-4）。

圖3-3-1、訪談及錄音之情形（示範者：呂奇蓉、鄭月霞、郭彧岑）

圖3-3-2、訪談所使用之圖卡範例（攝影：蕭素英）

寫字	關門
查某	迫迌
血	桃園

圖3-3-3、訪談使用之閩南語字卡範例

朋友	講話
胃	肥
時	四

圖3-3-4、訪談使用之客家語字卡範例

在問卷設計上細分爲受訪者對不同對象所使用之語言，例如與長輩及晚輩談話的語言或有不同，問卷調查以「一戶家庭」爲語言分布調查的單位，以家庭最常使用的語言作爲該家庭的語言（圖3-3-5及圖3-3-6）。

新豐鄉閩客方言地理資訊系統建置基本資料表

訪問時間：__94__年____月____日____時____分　GIS地理資訊：
地址：　　村　鄰　　路／街　段　巷　弄　　號之　　樓之
姓名：　　出生：　___年___月___日　　性別：□1.男；□2.女
出生地：□1.新豐鄉；□2._____
職業：□1.農林漁牧；□2.工；□3.商；□4.教；□5.公；□6.其他；□8.無（退休）
教育程度：□1.無；□2.國小；□3.國初中；□4.高中職；□5.專科；□6.大學以上
族群：□1.客家人；□2.閩南人；□3.大陸各省市；□4.原住民；□5.其他____
1. 請問您住的這個地方，地名叫什麼？
2. 您在本地住多久了？
3. 您住過其他地方嗎？
4. 您會哪些語言？　□1.客語　□2.閩南語　□3.國語　□4._____ 流利程度：□流利　□流利　□流利　□流利 □還可以　□還可以　□還可以　□還可以 □會聽不會說　□會聽不會說　□會聽不會說　□會聽不會說
5. 您工作時常用什麼語言？（若複選，依常用順序標示1, 2, 3, 4等） □1.客語　□2.閩南語　□3.國語　□4._____
6. 您祭祖的時候使用什麼語言？（若複選，依常用順序標示1, 2, 3, 4等） □1.客語　□2.閩南語　□3.國語　□4._____
7. 您跟長輩對話使用什麼語言？（若複選，依常用順序標示1, 2, 3, 4等） 跟爸爸：□1.客語　□2.閩南語　□3.國語　□4._____ 跟媽媽：□1.客語　□2.閩南語　□3.國語　□4._____ 跟祖父：□1.客語　□2.閩南語　□3.國語　□4._____ 跟祖母：□1.客語　□2.閩南語　□3.國語　□4._____ 跟外祖父：□1.客語　□2.閩南語　□3.國語　□4._____ 跟外祖母：□1.客語　□2.閩南語　□3.國語　□4._____

圖3-3-5、使用之問卷範本－第一頁（問卷設計：蕭素英）

8. 您跟配偶對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____
配偶是：□1. 客家人； □2. 閩南人；□3. 大陸各省市；□4. 原住民；□5. 外籍____

9. 您跟兄弟姐妹對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
跟哥哥：□1. 客語 □2. 閩南語 □3. 國語 □4. _____
跟姊姊：□1. 客語 □2. 閩南語 □3. 國語 □4. _____
跟弟弟：□1. 客語 □2. 閩南語 □3. 國語 □4. _____
跟妹妹：□1. 客語 □2. 閩南語 □3. 國語 □4. _____

10. 您跟子女對話使用什麼語言？ （若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____

11-1. 您跟媳婦對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____
媳婦是：□1. 客家人； □2. 閩南人；□3. 大陸各省市；□4. 原住民；□5. 外籍____

11-2. 您跟女婿對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____
女婿是：□1. 客家人； □2. 閩南人；□3. 大陸各省市；□4. 原住民；□5. 外籍____

12-1. 您跟孫子女對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____

12-2. 您跟外孫子女對話使用什麼語言？（若複選，依常用順序標示 1, 2, 3, 4 等）
□1. 客語 □2. 閩南語 □3. 國語 □4. _____

謝謝！現在有一些簡單的字詞，請您說一說。我們會錄音以便分析。

1. 請從 1 數到 10
2. 請看圖卡或字卡說一說

圖3-3-6、使用之問卷範本－第二頁（問卷設計：蕭素英）

（二）街路定位及數位化地圖

因爲在地理資訊系統上製作語言分布圖，需要每個訪談地點的地理座標，因此將資料整理爲地理資訊系統所需之檔案，有以下二種方法。

第一種方法爲採用衛星自動定位，衛星定位儀器曾經採用GARMIN eTrex Vista，由於誤差值較大，現在採用更先進設備Trimble ProXH，精準度在一公尺以內，該儀器是在現場取得X及Y座標，以無線藍芽傳輸該地點的地理座標至筆記型電腦，電腦藉由程式直接寫到該家戶的通訊地址。但此方法在建築物密集的地方，難取得衛星訊息（圖3-3-7及圖3-3-8）。

圖3-3-7、使用之衛星定位儀器（左為現在使用之儀器）

圖3-3-8、現場實地進行定位（示範者：鄭錦全、張智傑；拍攝者：黃菊芳）

第二種方法爲利用已經向量化之航照地圖來尋找坐標，其工作是由二位助理擔任，一位助理使用ArcView軟體開啓已向量化之航照圖，於航照圖上找出受訪者住家位置，點選取得該位置X及Y座標。然後將該地點之X及Y座標唸給另一位助理聽，再由該助理負責將其資訊輸入另一電腦（圖3-3-9及圖3-3-10）。

第二台電腦所顯示的是從戶政機關取得的受訪者通訊錄所製作成的Microsoft Excel檔案，藉由航照圖定位，將Ｘ、Ｙ座標輸入。之後Microsoft Excel檔案再轉成DBF檔案，以ArcView軟體開啓，以便製作語言分布之空間資訊系統（圖3-3-11）。

圖3-3-9、確認受訪者門牌地址及地理位置
（示範者：林千慧、郭彧岑；拍攝者：黃菊芳）

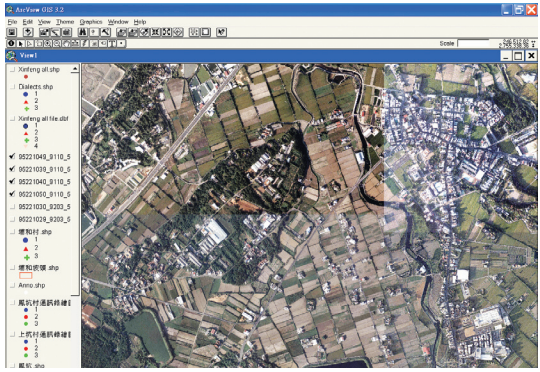
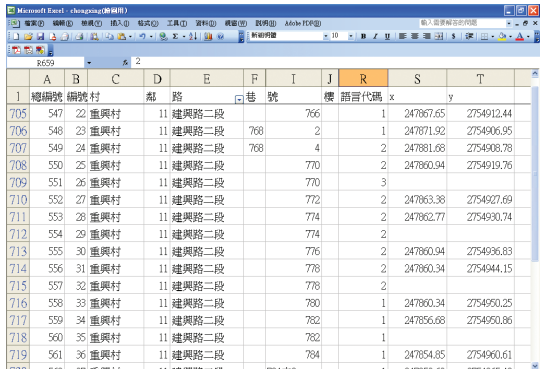


圖3-3-10、使用ArcView軟體開啓之航照圖



	A	B	C	D	E	F	I	J	R	S	T
1	總編號	編號	村	鄰	路	巷	號	樓	語言代碼	x	y
705	547	22	重興村	11	建興路二段		766		1	247867.65	2754912.44
706	548	23	重興村	11	建興路二段	768	2		1	247871.92	2754906.95
707	549	24	重興村	11	建興路二段	768	4		2	247881.68	2754908.78
708	550	25	重興村	11	建興路二段		770		2	247860.94	2754919.76
709	551	26	重興村	11	建興路二段		770		3		
710	552	27	重興村	11	建興路二段		772		2	247863.38	2754927.69
711	553	28	重興村	11	建興路二段		774		2	247862.77	2754930.74
712	554	29	重興村	11	建興路二段		774		2		
713	555	30	重興村	11	建興路二段		776		2	247860.94	2754936.83
714	556	31	重興村	11	建興路二段		778		2	247860.34	2754944.15
715	557	32	重興村	11	建興路二段		778		2		
716	558	33	重興村	11	建興路二段		780		1	247860.34	2754950.25
717	559	34	重興村	11	建興路二段		782		1	247856.68	2754950.86
718	560	35	重興村	11	建興路二段		782		1		
719	561	36	重興村	11	建興路二段		784		1	247854.85	2754960.61

圖3-3-11、輸入之Microsoft Excel檔案部分信息

（三）問卷輸入

將問卷輸入電腦，並建置檔案，便於使用者在使用語言分布空間資訊系統時，藉由點選資料，能更了解新竹新豐鄉語言分布之情況（圖3-3-12及圖3-3-13）。

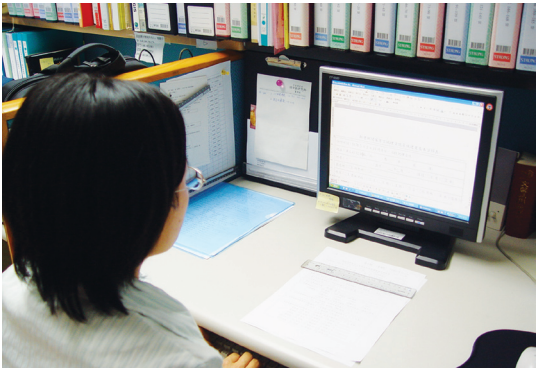


圖3-3-12、問卷輸入之工作狀況（示範者：黃菊芳）

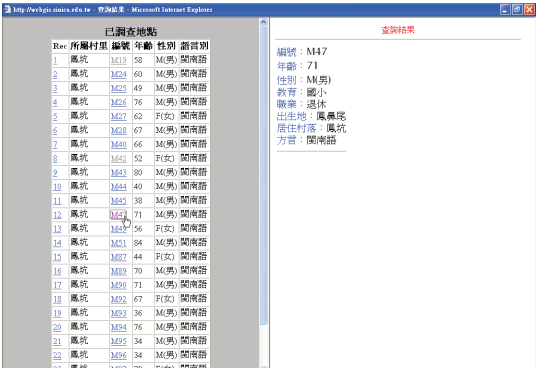


圖3-3-13、語言分布空間資訊系統之問卷資料點選（WebGIS 格式）

（四）錄音備份

將第一階段請受訪者依照圖卡及字卡朗讀所產生之錄音檔案儲存備份，方便於日後可隨時轉為使用所需之格式。未來會呈現於語言分布空間資訊系統，讓使用者能更了解新竹縣新豐鄉客家及閩南語的腔調及口音之獨特性。

（五）圖層製作

將第二階段搜集整理後的Microsoft Excel檔案匯入ArcView軟體，再利用ArcView軟體製作成各種圖層，以供研究使用。

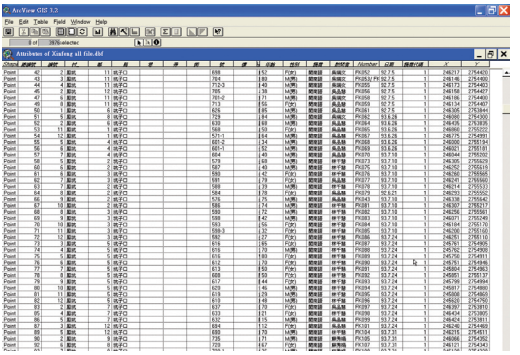


圖3-3-14、將Microsoft Excel檔案匯入ArcView軟體（部分私人信息省略）

將語言分布狀況製作成點狀圖，可分爲客家語－四縣腔、客家語－海陸腔、閩南語及其他語言，分別以不同符號作爲標示；再與其他相關資料分別製作成圖層，其他相關資料包含：已調查地點、建物、道路、河流、土地利用、鄉鎮界及航照影像等資料。分別製作圖層目的是爲了讓不同的使用者，能依據自身的使用需求，選擇點選所需的圖層介面。使用者可以藉由下載免費的ArcRead，加以使用已製作好的語言分布的情況（圖3-3-15、圖3-3-16）。

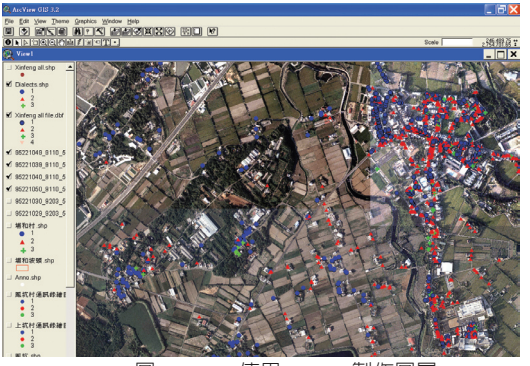


圖3-3-15、使用ArcView製作圖層
（藍色表閩南語、紅色表客家語－海陸腔、綠色表其他）

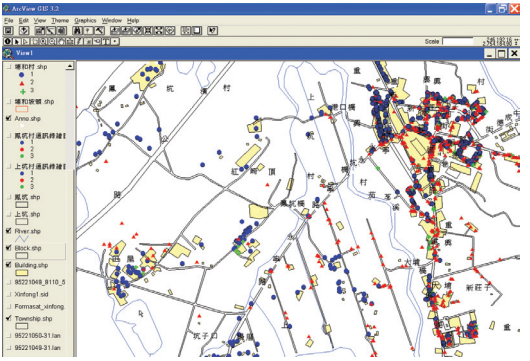


圖3-3-16、以道路交通圖呈現語言分布

此外，將第二階段街路訂位及數位化地圖和第三階段問卷輸入所搜集整理後的數位化地圖及問卷資料電子檔之Microsoft Excel資料，交由中央研究院計算中心製作圖層並轉檔。轉檔爲WebGIS，讓使用者能連結至WebGIS上的語言調查之空間資訊系統（圖3-3-17）。

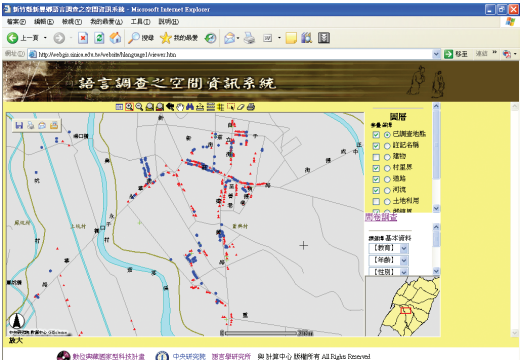


圖3-3-17、語言分布空間資訊系統之網路使用介面

（六）匯入資料庫

最後，將中央研究院語言學研究所的「閩南語典藏－歷史語言與分布資料庫」與中央研究院計算中心的WebGIS－「語言分布空間資訊系統」加以連結，讓使用者瀏覽「閩南語典藏－歷史語言與分布資料庫」時，能直接連結至語言分布空間資訊系統使用。

製作單位：數位典藏國家型科技計畫 內容發展分項計畫
中央研究院語言學研究所 語言典藏計畫
文字撰寫：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻
中央研究院語言學研究所
語言典藏計畫助理 盧秋蓉、黃菊芳、郭彧岑
圖片拍攝：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠
圖文編輯：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、陳秀華

致謝：中央研究院語言學研究所「語言典藏」之計畫主持人 鄭錦全院士與助理盧秋蓉小姐、黃菊芳小姐、郭彧岑小姐撥冗指教及協助拍攝與提供資料，特別致謝。

四、台灣手語線上影像辭典數位化工作流程簡介

製作日期：2005/11/16
更新日期：2009/12/03

國立中正大學語言學研究所「台灣手語之研究—音韻、構詞、句法與影像辭典」計畫的目標是對台灣手語做一個最完整的描述及分析，包括編纂一部有學術與實用價值的參考語法書，以及製作一部架設在網際網路上的數位影像辭典。此國科會支持之專題計畫執行期間爲民國九十年八月一日至九十四年十二月三十一日。

該計畫是由戴浩一教授負責統籌、規劃、執行和督導。參考語法中詞彙與句法的部分由戴浩一教授和張榮興教授負責，音韻、構詞的部分由蔡素娟教授和麥傑教授負責。辭典的編纂由蔡素娟教授負責，台灣手語線上辭典第一版於民國九十七年七月正式上網出刊，由電機系陳自強教授則督導研究生呂嘉雄進行數位影像辭典的網站架設；第二版於民國九十八年九月出刊，由語言所研究生余瓊怡進行網站維護。

本影像辭典目前約收錄3,000個詞項（包括中文及英文兩個搜尋介面及解說），辭典內容會陸續擴增。以下數位化工作流程主要介紹數位影像辭典之編纂與網站架設相關流程。

數位化工作流程說明

台灣手語數位影像辭典的數位化作業，大致依照下列八項步驟進行，依序分別爲：（一）收集手語詞彙；（二）準備錄影材料；（三）錄影；（四）影像轉檔；（五）影像剪輯；（六）詞彙影像之文字描述；（七）資料庫建置；（八）網站架設。茲分別介紹如次。

（一）收集手語詞彙

收集《台北市勞工局手語翻譯培訓教材第一冊》、《手能生橋》一、二冊、及《台灣手語參考語法》（國立中正大學語言所編纂）中的手語詞彙，包括單詞和複合詞，作爲拍攝影像辭典之詞彙清單（圖3-4-1）。

圖3-4-1、手語詞彙參考教材及語料採集圖片

（二）準備錄影材料及設備

將所收集的手語詞彙輸入成Microsoft Word檔，加入各詞之英文翻譯，建立中英文對照之詞彙清單，並加註手語動作之描述（圖3-4-2）。

隨後將詞彙清單轉檔製作成Microsoft PowerPoint簡報檔。每個詞項做成一張簡報，以便擔任示範的手語顧問可以依照所列詞項，依序錄影。爲使手語顧問在錄影時易於辨認，故檔案須爲黑底白字，一目瞭然（圖3-4-3）。

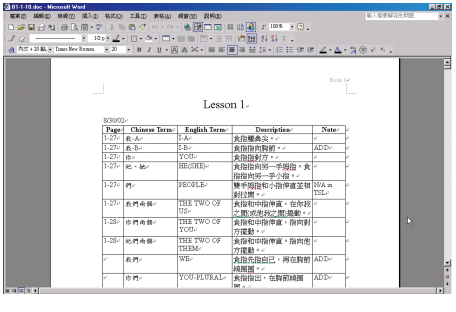


圖3-4-2、手語詞彙表

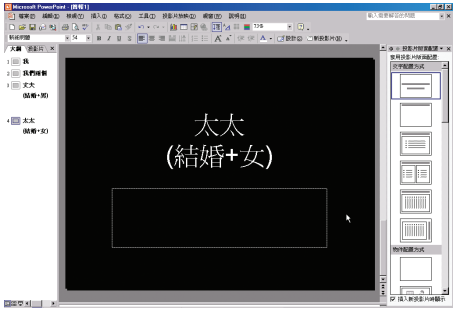


圖3-4-3、詞彙之簡報檔

（三）錄影

首先，將存有詞彙簡報檔的電腦準備就緒，然後架設錄影器材（註：前期錄影所使用的器材主要爲DV攝影機，隨著數位科技發展，現已改由數位攝影機攝錄影像）。錄影方式以固定之黑色布幔爲背景，手語顧問之服裝爲紅色上衣，以使影像色彩對比明顯。錄影前助理與手語顧問討論拍攝內容，以確認欲拍攝之詞彙（圖3-4-4、圖3-4-5）。

圖3-4-4、錄影背景及手語顧問之服裝。（示範者：顧玉山）

圖3-4-5、工作人員與手語顧問討論拍攝內容。（示範者：蘇秀芬、顧玉山）

由於每個詞彙，可能會有多種手語形式，研究助理會與手語顧問討論各種形式之差異（例如，該形式屬於台灣北部方言、南部方言、或僅是近義詞等）。原則上，一個詞彙如有多種形式都會全部錄製。

拍攝過程通常需要兩位工作人員、一位手語顧問及一位手語翻譯員。一位工作人員執行簡報詞彙清單的播放，決定拍攝的進度；另一位工作人員比對手語詞彙之正確性，以及控制影像錄製的品質。錄製過程中由手語翻譯員協助溝通。由於對於影像品質及清晰度有相當高的要求，而且不時需要與手語顧問確認，因此同一個詞往往一再重錄，以求詞彙的正確性及最佳影像品質（圖3-4-6、圖3-4-7）。

圖3-4-6、實際拍攝現場
（示範者：左起蘇秀芬、顧玉山、吳佩蘭）

圖3-4-7、手語翻譯員協助工作人員與手語顧問討論
（示範者：左起顧蕭月霞、顧玉山、蘇秀芬）

（四）影像轉檔

使用CyberLink PowerVCR影音捕手軟體將DV帶上的影像轉檔成MPEG檔，再用Quick Time Player將MPEG檔另存成檔案格式較小的MOV檔儲存於電腦伺服器中，以減少上線後使用者瀏覽影片等待的時間（圖3-4-8）。

圖3-4-8、轉檔之電腦畫面

（五）影像剪輯

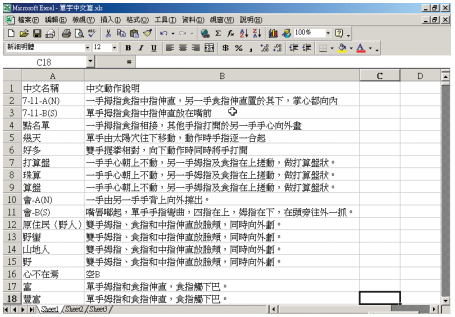
利用CyberLink PowerVCR影音捕手軟體內之「剪輯功能」把影像檔中的單詞或複合詞逐一切割出來。由於錄製過程中無法達到百分百的時間控制，在剪輯過程中常會發現兩個詞的影像相連，無法完整的切割出來，此時必須重新錄製。另外，爲求最佳影像品質，因此切割影像時需要反覆檢視，而一個詞項的影像長度也需多次修改，以求最適宜長度。因爲影像長度太短，則不能完整呈

現；長度太長，檔案容量較大，不但佔用較多硬碟空間，影片顯現的等待時間也會加長（圖3-4-9）。

圖3-4-9、影像剪輯（示範者：蘇秀芬）

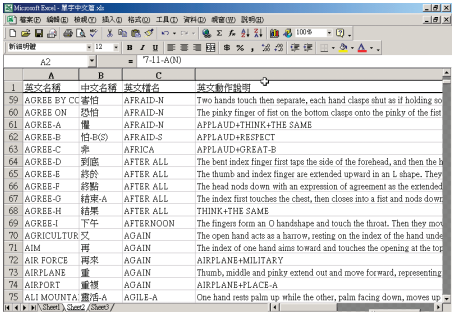
（六）詞彙影像之文字說明

每一個手語詞彙都有文字說明，以描述其動作。手語詞彙的中文說明也都翻譯爲英文，並將中、英文說明以Microsoft Excel列表（圖3-4-10、圖3-4-11）。



	A	B	C	D
1	中文名稱	中文動作說明		
2	7-11-A(N)	一手拇指食指中指伸直，另一手食指伸直置於其下，掌心都向內		
3	7-11-B(S)	單手拇指食指中指伸直放在嘴前		
4	點名單	一手拇指食指相接，其他手指打開於另一手手心向外畫		
5	燒天	單手由太陽穴往下移動，動作時手指逐一合起		
6	好多	雙手握拳相對，向下動作時同時將手打開		
7	打算盤	一手手心朝上不動，另一手拇指及食指在上撥動，做打算盤狀。		
8	珠算	一手手心朝上不動，另一手拇指及食指在上撥動，做打算盤狀。		
9	算盤	一手手心朝上不動，另一手拇指及食指在上撥動，做打算盤狀。		
10	會-A(N)	一手由另一手手背上向外擦出。		
11	會-B(S)	嘴唇閉起，單手手指彎曲，四指在上，拇指在下，在頭旁往外一抓。		
12	原住民（野人）	雙手拇指、食指和中指伸直放臉頰，同時向外劃。		
13	野蠻	雙手拇指、食指和中指伸直放臉頰，同時向外劃。		
14	山地人	雙手拇指、食指和中指伸直放臉頰，同時向外劃。		
15	野	雙手拇指、食指和中指伸直放臉頰，同時向外劃。		
16	心不在焉	空8		
17	富	單手拇指和食指伸直，食指觸下巴。		
18	豐富	單手拇指和食指伸直，食指觸下巴。		

圖3-4-10、詞彙之中文說明



	A	B	C
1	英文名稱	中文名稱	英文動作說明
59	AGREE BY CC	害怕	AFRAID-N
60	AGREE ON	恐怕	AFRAID-N
61	AGREE-A	懼	AFRAID-N
62	AGREE-B	怕-B(S)	AFRAID-S
63	AGREE-C	非	AFRICA
64	AGREE-D	到底	AFTER ALL
65	AGREE-E	終於	AFTER ALL
66	AGREE-F	終點	AFTER ALL
67	AGREE-G	結果-A	AFTER ALL
68	AGREE-H	結果	AFTER ALL
69	AGREE-I	下午	AFTERNOON
70	AGRICULTUR	又	AGAIN
71	AIM	再	AGAIN
72	AIR FORCE	再來	AGAIN
73	AIRPLANE	重	AGAIN
74	AIRPORT	重複	AGAIN
75	ALI MOUNTA	靠近-A	AGILE-A

圖3-4-11、詞彙之英文說明

撰寫手語詞彙的文字描述時，需要不同的工作人員再進行比對，必要時與計畫主持人、共同主持人及研究助理進行討論後，才確認最適合的文字說明。英文版詞項文字描述則由計畫共同主持人或英文專家再度確認。

每個詞彙除了手形的描述外，還有方位或接觸身體的位置、方向性、擺動方式、臉部表情等描述。此外，該手語形式屬於台灣北部用法或南部用法，或其形成動機也會在必要時加以描述。

（七）資料庫之建置

透過PhpMyAdmin介面管理線上辭典資料庫，將步驟六所製作完成之中文版詞彙表中的詞項加注漢語拼音、中文筆劃數等資料。每個詞項都包含影像檔、中文解說和英文解說，並完成所有詞彙的影像檔、文字說明，及每個詞項的影像與文字的連結，資料庫版面如下圖（圖3-4-12、圖3-4-13）。

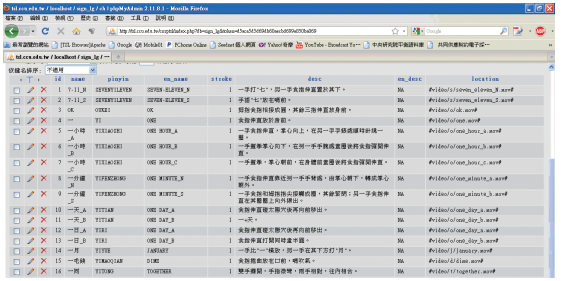


圖3-4-12、中文版資料庫

圖3-4-13、英文版資料庫

（八）網站架設

所有詞彙的影像檔、文字說明，及每個詞項的影像與文字的連結，都在上述步驟七完成，同時以此網站中文筆劃數順序、英文A-Z字母順序逐一確認資料的正確性。

本辭典檔共有「台灣手語－中文」和「台灣手語－英文」兩個介面（圖3-4-14）。中文版之搜尋功能可以用中文字首之筆劃數（一至十九劃）、漢語拼音或關鍵字輸入。英文版則以A-Z字母順序或關鍵字搜尋（圖3-4-15）。

圖3-4-14、台灣手語線上影像辭典首頁

圖3-4-15、中文版之搜尋功能網頁

以詞項「交通」爲例，「交通」之「交」字爲六劃，在六劃之類別下，點選「交通」此詞項，得出網頁左方詞項「交通」之台灣手語形式，及右方之中文描述（圖3-4-16）。

圖3-4-16、台灣手語線上影像辭典中「交通」之搜尋結果

而進入英文版介面，可以輸入詞項「TRAFFIC」（註：英文輸入不分大小寫），得出網頁左方詞項「TRAFFIC」之台灣手語形式，及右方之英文描述（圖3-4-17）。

圖3-4-17、台灣手語線上影像辭典中「TRAFFIC」之搜尋結果

台灣手語影像辭典數位化的流程基本上是一環扣一環，每一個步驟都影響著成果的好壞，因此在制定流程時，必須經過縝密的思考，並且將實際執行的結果透過無數次的討論進行修正與改進。自台灣手語線上辭典第一版於民國九十七年七月正式上網出刊以來，感謝來自各方的建議與回饋，給予本團隊相當多的鼓勵，也感謝各研究成員長期的努力與付出，更要感謝國科會長期的大力支持與協助，才能讓「台灣手語線上辭典」有今日豐碩的成果。未來我們會在台灣手語線上辭典的拓展上持續深耕、與時俱進，希冀台灣手語研究在語言學這個領域裡受到國際重視，也期盼能對台灣的聽障教學提供實質幫助。

製作單位：數位典藏國家型科技計畫 內容發展分項計畫
國立中正大學語言學研究所 台灣手語之研究：音韻、構詞、句法與影像辭典
文字撰寫：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻

國立中正大學語言學研究所手語研究計畫主持人 戴浩一
國立中正大學語言學研究所手語研究計畫共同主持人 蔡素娟
國立中正大學語言學研究所手語研究計畫助理 蘇秀芬、陳欣徽
圖片拍攝：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠、陳秀華
圖文編輯：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、陳美智、陳秀華

致謝：感謝國立中正大學語言學研究所「台灣手語之研究」之計畫主持人 戴浩一教授、共同主持人 蔡素娟教授及助理蘇秀芬小姐撥冗指教和協助拍攝與提供資料，手語顧問顧玉山及手語翻譯員顧蕭月霞配合拍攝影像辭典工作情形，特別致謝。

五、閩南語兒童語料數位化工作流程簡介

製作日期：2005/12/14
更新日期：2010/01/25

國立中正大學語言學研究所「台灣兒童語料庫」Taiwan Child Language Corpus（簡稱TAICORP）是將所收集之台灣兒童口語錄音語料，依照世界標準的兒童語料交換系統 Child Language Data Exchange System（簡稱 CHILDES; MacWhinney and Snow 1985, MacWhinney 1995）格式，建構成語料庫。其主要目的在（1）提供國內外學者語料共享的便利性與語料分析工具；（2）藉由標準規格的設定，使台灣兒童語料的收集能更有系統、更有效率，並且快速地涵蓋台灣地區所有語言。語料庫最終將設立網站，開放國內外學者使用。

在新生一代普遍使用國語的時代背景之下，台灣閩南語兒童語言習得的語料彌足珍貴。本語料庫可提供語音學、音韻學、構詞學、句法學、語意學、語用學等不同層面的語言學與兒童語言習得研究，也可提供語音工程方面的研發與應用。本計畫由國立中正大學語言學研究所蔡素娟教授主持，從1997年10月開始錄音，經轉記、標記、格式化等過程，歷時將近九年。共收錄431人次錄音檔案，錄音總長共約330 小時。文字檔共約五十萬句，一百六十多萬詞。

數位化工作流程說明：

該計畫的數位化作業，大致依照下列五項步驟進行，依序分別爲：（一）錄音；（二）錄音檔案轉記爲文字檔；（三）建立詞彙庫；（四）建立自動化系統；（五）自動化系統之應用；（六）網站的建立與維護等六個方面，共細分二十三項步驟進行，茲分別介紹如次。

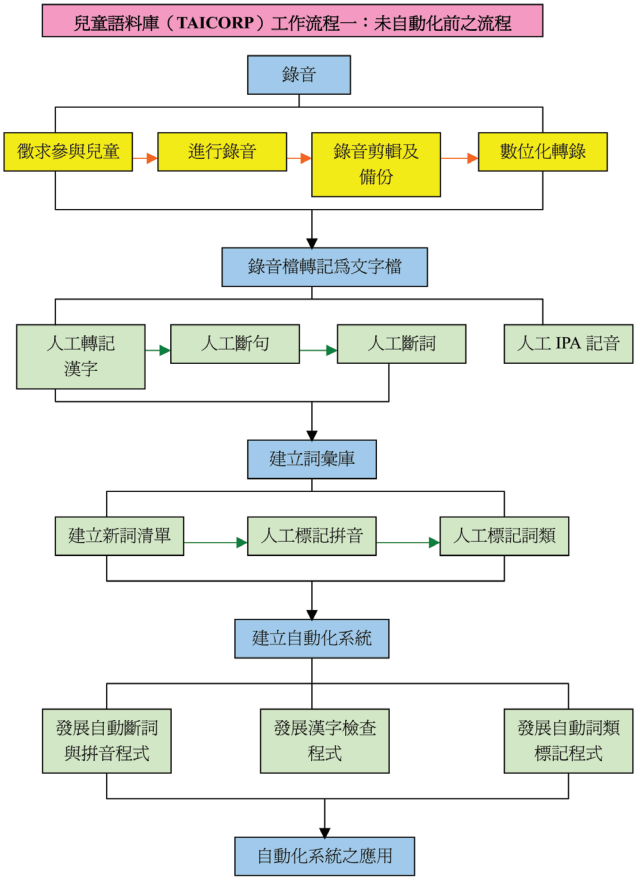


圖3-5-1、兒童語料庫未自動化前工作流程圖

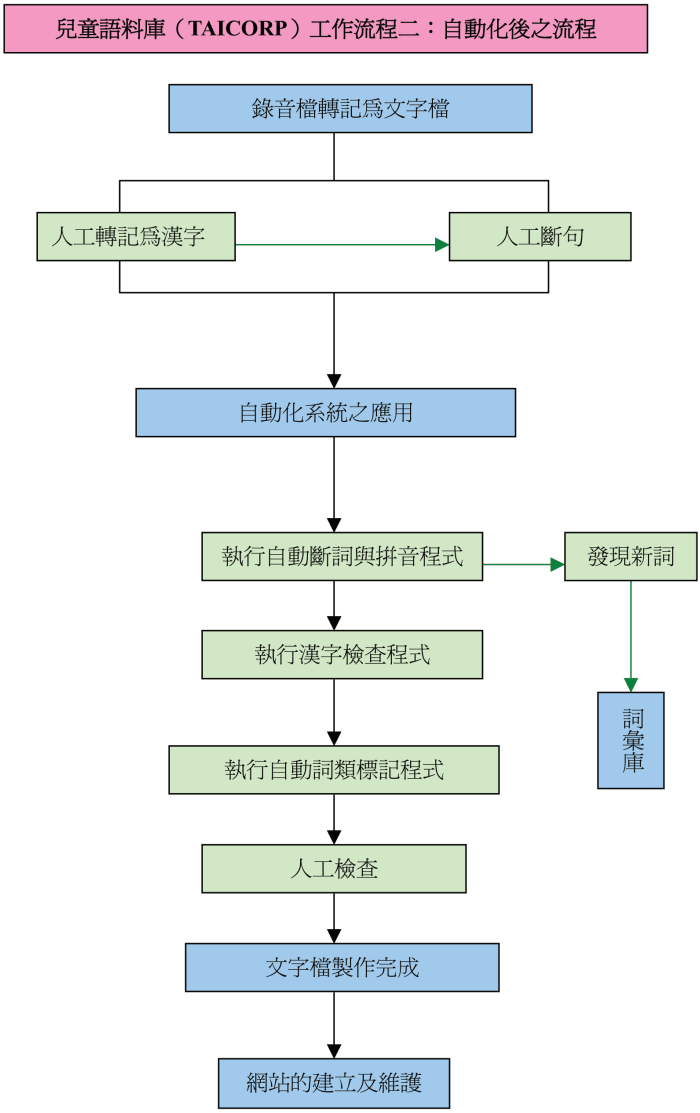


圖3-5-2、兒童語料庫自動化後工作流程圖
製作日期：2005/11/17
流程圖提供者：國立中正大學 語言學研究所 蔡素娟教授

（一）錄音

「錄音」部分分爲五個步驟進行，分別爲「訓練研究助理」、「徵求參與兒童」、「進行錄音」、「錄音剪輯及備份」、「數位化轉錄」。

1. 訓練研究助理：由計畫主持人訓練研究助理，最核心的研究助理有三名。需具語言學碩士級背景知識，並以閩南語爲母語。每星期透過三到六小時的討論會，訓練助理，瞭解閩南語音韻及書寫系統、閩南語詞彙、句法、語意及詞類標記系統、CHILDES系統及兒童語言習得相關文獻；並熟悉IPA國際音標記音。
2. 徵求說閩南語家庭之兒童：目標選定中正大學附設托兒所、幼稚園及鄰近鄉鎮，徵求來自說閩南語家庭，年齡在一歲至三歲之間的幼兒。陸續共選出14名兒童。
   (1) 以海報及網路發布廣告；利用幼稚園家長日到場對家長說明，徵求說閩南語家庭的兒童。
   (2) 排定錄音時間：聯絡家長；並排定錄音時間表。
3. 進行錄音
   (1) 準備錄音器材：錄音器材選擇方便攜帶、機動性強、容量較大、易長期保存語料之錄音器材（圖3-5-3）。

圖3-5-3、錄音器材，左起為迷你光碟片、專業用耳機、專業用麥克風、迷你光碟隨身錄音機。

(2) 進行錄音訪談：至兒童家中進行訪談錄音。錄音爲週期性，寒暑假亦不間斷。二歲以下者，每週訪談一次；二至三歲者，每兩週訪談一次；三至四歲者，每二至三週訪談一次。每次訪談約1至2小時不等，實際錄音時間40至60分鐘。談訪錄音期間從1997年10月至2000年5月。共錄音431人次，約330小時。訪談方式爲：錄下兒童在家長或保姆陪同下，在自己家中的日常對話。錄音的內容除了自然言說，還藉助圖畫簿、故事書、玩具、布偶、剪紙、摺紙或其他遊戲，引發兒童主動說話。

4. 錄音剪輯及備份

(1) 錄音剪輯：由助理將錄音光碟中不相關的錄音或太長的空白錄音刪除，將錄音切割爲較小段落，在光碟中標記段落編號；於光碟中輸入錄音日期、檔名。每1小時的錄音約需耗時1.5小時剪輯。總工作時間：1.5×330小時=495小時（圖3-5-4）。

(2) 錄音備份：使用迷你光碟錄音座及迷你光碟隨身錄音機進行迷你光碟備份製作（圖3-5-5）。

圖3-5-4、進行錄音剪輯
（示範者：謝沛諭）

圖3-5-5、進行錄音備份
（示範者：謝沛諭）

5. 數位化轉錄：將迷你光碟錄音檔轉爲較不佔空間之MP3格式，以方便儲存。於日後可隨時轉爲語音分析所需之格式（如*.wav）。所使用

之轉錄軟體爲GoldWave Digital Audio Editor（GoldWave Inc.研發，見圖3-5-6）。

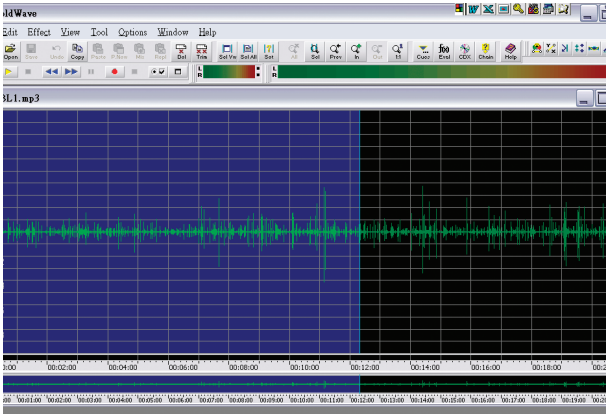


圖3-5-6、進行數位化轉錄

（二）錄音檔轉記爲文字檔

「轉記」分爲四個步驟，依序爲：「人工轉記漢字」、「人工斷句」、「人工斷詞」、「人工IPA記音」。

1. 人工轉記漢字：由於閩南語的漢字書寫系統目前並沒有定案，再加上有許多本字無法確定，或者有音無字的情形，因此有必要訂定文字轉記的原則。故在進行文字轉記前，首先需確立閩南語書寫系統，本計畫所參考的辭典主要有四本，依優先順序排列爲：《臺灣閩南語辭典》、《台灣話大辭典》、《廈門方言詞典》、《閩南語詞彙》，如圖3-5-7由左至右。轉記平台爲CHILDES

圖3-5-7、閩南語辭典

兒童語料交換系統。每1小時錄音需要花約10小時不等的時間轉記成文字檔。總工作時間：錄音330小時×10=3,300小時。

2. 人工斷句：由於本語料庫之語料爲口語語料，助理需參考言談分析之斷句原則，將自然言談切分成獨立意義句子。
3. 人工斷詞：由於目前無閩南語斷詞標準，故本計畫根據中華民國計算語言學學會所訂定之「資訊處理用中文分詞規範調查研究及草案研擬」，將語句切分爲獨立意義，且扮演特定語法功能的字串。
4. 人工IPA記音：採語音轉記 (Phonetic Transcription) 的方式詳細轉記兒童實際發音。在音段方面，以Unicode IPA符號記音，參考書目爲《Handbook of the International Phonetic Association》(1999)；聲調採用五度標音法。每小時的錄音約需花4.5小時記音。共錄音330小時×4.5=1,485小時（圖3-5-8）。

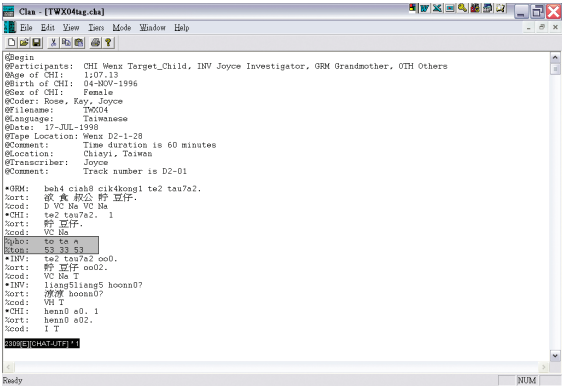
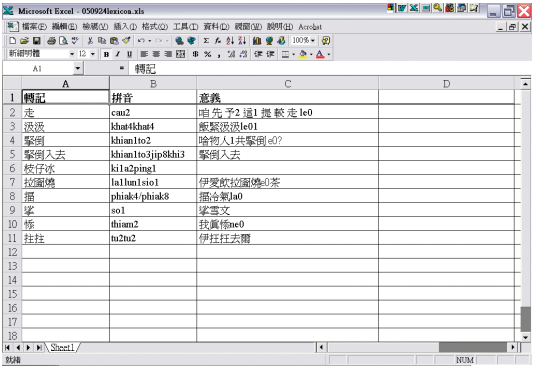


圖3-5-8、完成Unicode IPA記音之文字檔

（三）建立詞彙庫

錄音以人工轉記爲文字很費人力，因此最終目標還是要建立自動化系統。而自動化系統的建立需要詞彙庫作基礎。「建立詞彙庫」依序分爲三個步驟進行：「建立新詞清單」、「人工標記拼音」、「人工標記詞類」。

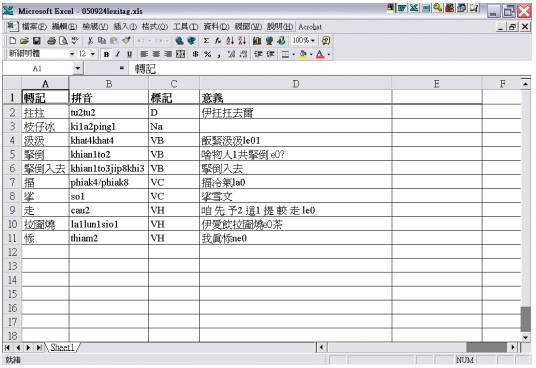
1. 建立新詞清單：以轉記好之文字檔中之所有詞彙建立清單，經由人工確認詞彙清單中的漢字與詞典是否一致。
2. 人工標記拼音：根據教育部於民國八十七年所公佈之「閩南語羅馬拼音第二式」人工標記詞彙清單中的漢字之拼音（圖3-5-9）。



	A	B	C	D
1	轉記	拼音	意義	
2	走	cau2	咱先予2這1提較走1e0	
3	汲汲	khat4khat4	較緊汲汲1e01	
3	緊倒	khian1to2	啥物人1共緊倒e0?	
5	緊倒入去	khian1to3jip8khi3	緊倒入去	
6	校仔冰	ki1a2ping1		
7	拉圈燒	la1lun1sio1	伊愛飲拉圈燒0茶	
8	摒	phiak4/phiak8	摒冷氣la0	
9	寧	so1	寧雪文	
10	慘	thiam2	我眞慘ne0	
11	拄拄	tu2tu2	伊拄拄去簡	

圖3-5-9、人工標記拼音

3. 人工標記詞類：參考中央研究院詞庫小組《詞類標記原則》以及《CANCORP: The Hong Kong Cantonese Child Language Corpus》(Lee and Wong，1998)、《台灣閩南語動詞分類研究 》（曹逢甫，1996） 等相關文獻。採用中研院詞庫小組的詞類標記，但是僅限於46個簡化標記，以避免詞類劃分過細時產生主觀強制性的歸類（圖3-5-10）。



	A	B	C	D	E	F
1	轉記	拼音	標記	意義		
2	拄拄	tu2tu2	D	伊拄拄去簡		
3	校仔冰	ki1a2ping1	Na			
4	汲汲	khat4khat4	VB	較緊汲汲1e01		
5	緊倒	khian1to2	VB	啥物人1共緊倒e0?		
6	緊倒入去	khian1to3jip8khi3	VB	緊倒入去		
7	摒	phiak4/phiak8	VC	摒冷氣la0		
8	寧	so1	VC	寧雪文		
9	走	cau2	VH	咱先予2這1提較走1e0		
10	拉圈燒	la1lun1sio1	VH	伊愛飲拉圈燒0茶		
11	慘	thiam2	VH	我眞慘ne0		

圖3-5-10、人工標記詞類

（四） 建立自動化系統

「建立自動化系統」以上述詞彙庫爲基礎。分爲三個部分：「發展自動斷詞與拼音程式」、「發展漢字檢查程式」、「發展自動詞類標記程式」。

1. 發展自動斷詞與拼音程式：將輸入之句子或整個文字檔案，根據本計畫修訂「資訊處理用中文分詞規範調查研究及草案研擬」所撰寫之「閩南語斷詞原則」及詞彙庫之詞項，根據長詞優先之準則，與詞彙庫比較。若所輸入之漢字與詞彙庫一致，則以黑色呈現，並在其後標注拼音；若所輸入之漢字尚未建立於在詞彙庫，則以藍色呈現。此程式除了斷詞及標注拼音之外，還可以將新詞納入詞彙庫（圖3-5-11）。
2. 發展漢字檢查程式：目的爲求漢字與詞彙庫所列之標準之一致。搜尋之方式有三：一爲輸入閩南語羅馬拼音、二爲輸入可能之漢字、三爲輸入國語之相對詞；透過此三種任一，皆能擷取出詞彙庫中含有該詞之詞條。但若該詞未建立於詞彙庫中，查詢後則不顯示（圖3-5-12）。
3. 發展自動詞類標記程式：以人工標記詞類之文字檔作爲基礎，

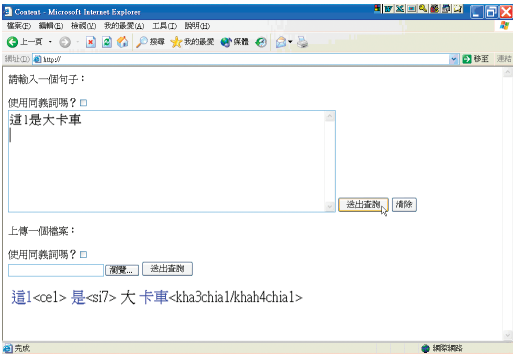


圖3-5-11、自動斷詞與拼音程式

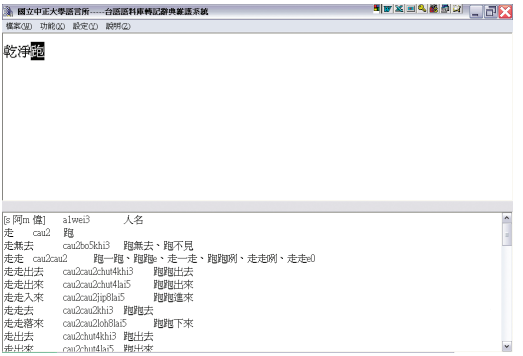


圖3-5-12、漢字檢查程式

發展自動詞類標記程式。將輸入之句子（已完成斷詞工作），自詞彙庫中擷取出其詞類標記；當該詞有多個詞類標記時，程式則以頻率最高之標記爲優先考量並標記之。若該詞在詞彙庫中未標記詞類，則以三個問號（???）呈現（圖3-5-13）。

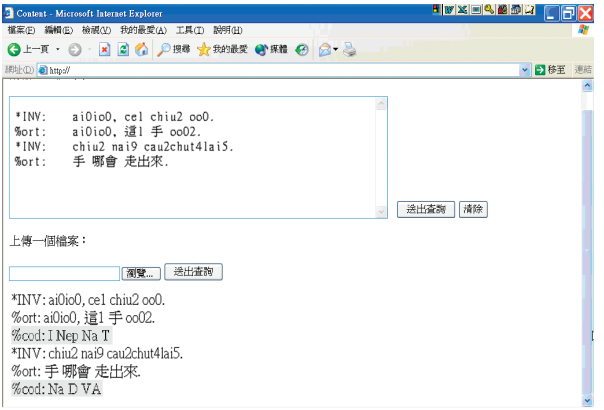


圖3-5-13、自動詞類標記程式

（五）自動化系統之應用

1. 執行自動斷詞與拼音程式：將語句切割成詞，並標注拼音。
2. 執行漢字檢查程式：檢查漢字與詞彙庫所列之標準是否一致。
3. 執行自動詞類標記程式：標記詞類。
4. 人工檢查：檢查程式輸出檔，如詞有不止一個詞類，則檢查其自動標記是否正確。

（六）網站的建立及維護

1. 網站架構及內容之編纂：計畫主持人與研究助理討論網站內容及所呈現之介面。網站內容包含語料庫簡介、資料庫、使用手冊、相關程式以及相關網站之連結。
2. 網站之建立及維護：爲語料庫建立專門網站，以供世界各地學者研究之用。完成最後檢測之後，網站將開放給外界瀏覽（圖3-5-14）。

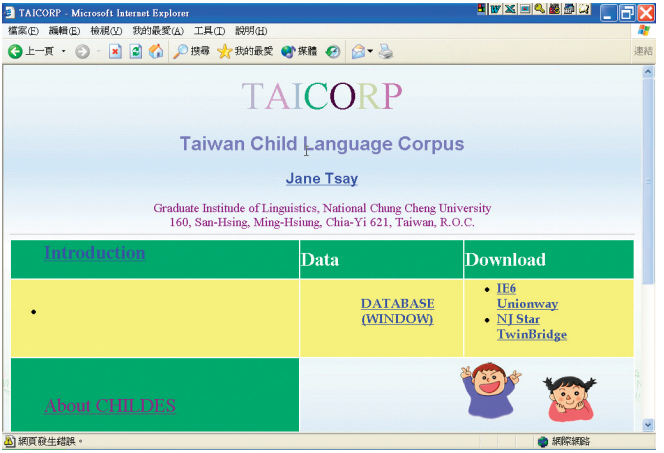


圖3-5-14、網站首頁

製作單位：數位典藏國家型科技計畫 內容發展分項計畫
國立中正大學語言學研究所 台灣兒童語料庫計畫

文字撰寫：國立中正大學語言學研究所台灣兒童語料庫計畫
助理 謝沛諭

數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻

圖片拍攝：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、林淑惠、陳秀華

圖文編輯：數位典藏國家型科技計畫 內容發展分項計畫
語言主題小組助理 賴佳旻、陳美智、陳秀華

致謝：感謝國立中正大學語言學研究所「台灣兒童語料庫」之計畫主持人 蔡素娟教授、前任助理黃婷鈺小姐、劉慧娟小姐及現任助理謝沛諭小姐撥冗指教及協助拍攝與提供資料，特別致謝。

六、社會語音語料庫

製作日期：2010/01/25

中央研究院語言學研究所語言典藏第二期子計畫「臺灣國語口音社會分布典藏」主持人爲曾淑娟副研究員，該計畫主要目的是以社會語言學爲主體，輔以聲學語音學工具進行社會語音學的研究。資料庫的建立可以提供結構化的資料，以促進系統化的研究。社會語言學的研究方法整理社會，經濟，教育與語言背景資料。數位錄音的技術有系統的自動處理語音內容與標記。計算語言學的資料運算方法讓語音與社會語言學的資料能有效的整合與分析。我們希望爲在臺灣所使用的國語口音建置數位內容資料庫，藉以記錄語音的社會性、區域性與語言特質。臺灣的國語，因爲受到多語環境的影響，在詞彙與口音上與其他使用現代漢語的區域相比自有其獨特之處。利用社會、政經、網路使用等指標，配合數位錄音的語音聲學分析，本子計畫在臺灣各主要縣市收集並典藏自然語音，一方面記錄語言的使用情形與社會變遷的連動性，另外一方面也達到原音重現的典藏目的。

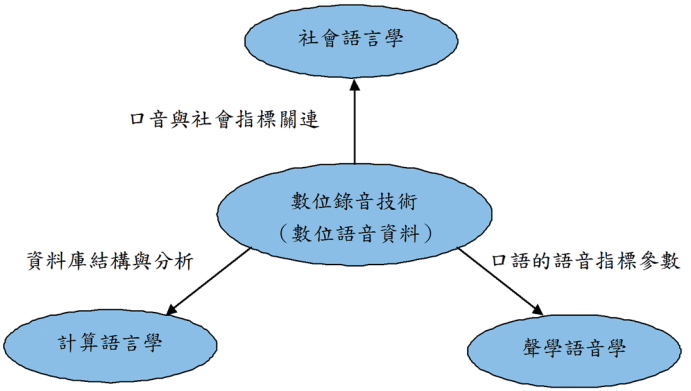


圖3-6-1、社會語音語料庫研究關係圖

數位化工作流程說明：

「台灣國語口音社會分布」的數位化作業，依照下列七項步驟進行。依序分別爲：一、擬定採樣地點與對象；二、錄音設備；三、發文縣市政府；四、問卷訪問；五、數位錄音內容文字轉記；六、建立問卷內容資料庫；七、網頁製作與維護。茲分別介紹如次。

（一）擬定採樣地點與對象

1. 採樣地點：全台各縣市，選擇當地民眾較多、非學生聚集及周圍噪音較少地區，如：郵局、大型公園、文化中心、圖書館
2. 單次採樣時間：三天
3. 採樣人員：2人爲1組，共四組
4. 採樣人數：每組目標採樣人數爲30人，四組共120人，有效人數至少100人/地點
5. 作業方式：一人負責主導提問與錄音，另一人則負責記錄錄音相關資料：姓氏、性別、錄音地點、職業註記。
6. 採樣對象年齡：20~40歲
7. 採樣對象性別：不拘

（二）錄音設備

圖3-6-2、錄音記錄器、麥克風

1. 數位錄音採樣系統：

   記錄器：SONY Hi-MD MZ-RH1

   麥克風：SONY ECM MS907

2. 錄音格式：

   聲音輸入爲Hi-SP雙聲道，聲音輸出格式爲wave音訊，單聲道取樣速率441KHz，16位元。

（三）發文縣市政府

錄音出訪前，需要請中央研究院語言學研究所行政人員發文至相關的縣市政府，以確保計畫人員能得到協助與安全。

（四）問卷訪問

1. 語言背景：受訪者本身語言使用情形，以及與家庭成員間語言使用情形。
2. 社會經濟背景：由小學起到最高學歷爲止的求學經歷，及超過半年以上的工作經歷，目前月薪等。地區以鄉鎮市爲單位。
3. 網路使用及國際觀：是否經常使用網路，使用網路的目的和行爲，瀏覽的網站的語言類別，點選網路新聞的類型，有沒有出國經驗。

圖3-6-3、戶外問卷訪問

4. 語言自我評價：詢問受訪者根據本身語言使用經驗及習慣，是否覺得自己說國語時受到當地口音的影響，及有哪些音不完整或容易分辨不清。

（五）數位錄音內容文字轉記

所有採集的數位語音資料會以PRAAT軟體做爲轉寫工具，將採訪的內容依問卷問題回答作爲切分的原則，將錄音內容與語音信號對齊。

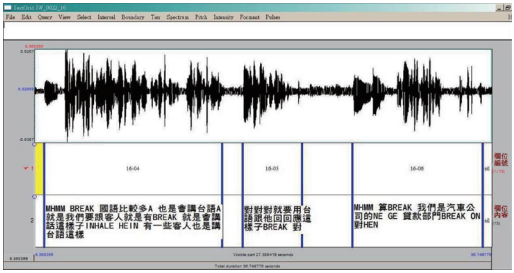


圖3-6-4、錄音內容文字轉記

（六）建立問卷內容資料庫

每位受採訪者，會建立一個問卷內容資料，依問卷問題順序將問卷調查的內容鍵入。

姓氏	求學所在地_國小	求學所在地_國中	求學所在地_高中職	求學所在地_大學（二技）	求學所在地_研究所	身份別	工作_總年資	工作_地區一	工作_地區一_年資	工作_地區一_行業	工作_地區一_語言	工作_目前月薪
許	台中市	台中市	雲林縣	高雄市		工作	3年	台中縣	3年	科技業製造生產	閩	3萬-5萬
蕭	台中市	台中市	台中縣/雲林縣斗六市			工作/進修	4年	台中市	4年	汽車公司貸款部	漢.閩	3萬以下
李	台中縣烏日鄉	台中市	台中縣大里市	台南縣永康市		工作	10個月	台中縣烏日鄉	10個月	製造業	漢.閩	3萬以下
鄭	台中市	台中市	台中縣烏日鄉			工作	10年以上	台中縣市	10年以上	幼稚園.安親班	漢	3萬
張	台中市	台中市	台中市	台北市		待業						
林	台北縣	台北縣	高雄縣			工作	7年	台北市	7年	軍人	漢	3萬-5萬
張	台北縣	台北縣	台北縣	台北市		工作	2年	台北市	2年	服務業	漢	3萬以下

圖3-6-5、問卷資料庫

（七）網頁製作與維護

該計畫網頁「新世紀語料庫」包含曾淑娟副研究員主持建置的所有語料庫，不僅止於數位典藏計畫。其中台灣國語口音之社會分布計畫的內容，因爲語料收集還未告一段落，因此還未上線。不過，其他語料庫的內容也已經包含大部分進行語音語料庫收集應該注意的事項。

肆、語料庫與數位學習

Digital Learning

最早的學習者語料庫是八〇年代末期所建立的朗曼學習者語料庫(Longman Learners' Corpus)。九〇年代中期，比利時魯汶大學 Centre for English Corpus Linguistics 的Sylvaine Granger建立了國際學習者英語語料庫(International Corpus of Learner English, ICLE)，該語料庫是一廣泛國際合作的計畫，目前收錄超過二百萬詞、十四種不同母語背景的英文學習者語料。

將語料庫應用於學習有越來越多的趨勢，以下以中央研究院語言學研究所的「全球華語文數位教與學資源中心」以及國立成功大學外國語文學系的「成鷹計畫」爲例，說明語料庫學術資源如何加值應用於教學之上。

一、全球華語文數位教與學資源中心

「全球華語文數位教與學資源中心」是數位學習國家型科技計畫「兼具教學與研究功能的全球華語文數位教與學資源中心」之計畫成果，由中央研究院語言學研究所鄭錦全院士主持。

中央研究院執行數位典藏計畫已有多年，在語料庫方面累積了豐碩的成果，但這些資源多著眼於學術研究需求，一般華語文教師與學生使用較爲困難。「全球華語文數位教與學資源中心」建置之目的即爲整合這些語料庫以及延伸資源，提供易於使用的學習工具與教學資源，建構一個兼具教學與研究功能的數位教與學資源中心。

圖4-1、全球華語文數位教與學資源中心首頁

這項計畫有兩個主要目標，一是以「一詞泛讀」的理論爲基礎，幫助學生加快學習詞語的用法；二是提供華語文教師編寫教材所需要的語言信息理據。

「一詞泛讀」的學習模式是「全球華語文數位教與學資源中心」的核心理念，藉由龐大資料庫，使用者搜尋一個詞語時，就能獲得這個詞語出現的相關句子，瞭解該詞出現的語言環境以及與不同詞語的搭配組合，因而更能掌握該詞的用法，進而加快學習語言的速度，這種「針對一個詞語廣泛閱讀」的方法對於成人外語學習者尤其有效。

「全球華語文數位教與學資源中心」使用的語料庫包括中央研究院語言學研究所的「上古漢語語料庫」、「近代漢語語料庫」、「現代漢語平衡語料庫」等三個語料庫，再加上「國立編譯館國小國語課本語料庫」，以及與元智大學合作建置的「唐詩三百首語料庫」等；此外，中央研究院有英國國家語料庫(British Nation Corpus) 的使用授權，因而「一詞泛讀」的學習模式也能提供給英文學習者檢閱英文詞語的用法。

這些語料庫的內容如下：

（一）「上古漢語語料庫」：《論語》、《孟子》、《大學》、《莊子》、《老子》等古籍。

（二）「近代漢語語料庫」：《紅樓夢》、《西遊記》、《水滸傳》、《儒林外史》等章回小說。

（三）「現代漢語平衡語料庫」：各類題材的現代漢語，500萬詞（20多萬句，約14萬筆詞條）。

（四）「國立編譯館國小國語課本語料庫」：5萬多詞。

（五）「唐詩三百首語料庫」：約7千筆詞條。

（六）「英國國家語料庫」：英文一億詞標記語料庫。

該網站的所有頁面都提供中英文對照連結，便於外國語言學習者使用，並區分「語言學習區」與「語言教學資源區」兩大區塊。「語言學習區」提供學習者線上中、英文「一詞泛讀」的學習；「語言教學資源區」則方便教師搜尋所需的素材。

爲了提供利於學習的模式，該計畫在「語言學習區」中依照句子長短、詞頻高低和詞語語意類別等因素計算出句子的難易度，並在查詢結果中提供「由簡入繁的閱讀模式」，將查詢結果依難易度排列，方便學習者自由選擇閱讀的難易度。另外亦提供「隨機提取」模式，由系統隨機提取查詢結果，難易不一；「近義詞」模式的內容則是取自《同義詞詞林》，可將意義相近的詞語依照近似層級高低排列。

圖4-2、一詞泛讀查詢頁面

「語言教學資源區」提供現代漢語、近代漢語及上古漢語語料庫、唐詩三百首、宋詞三百首等語料庫的詞頻統計以及文本標記閱讀。使用者可查閱個別語料庫的詞頻排序、個別詞的頻率、個別頻率的詞、累積詞頻等信息，教師可依據詞頻統計提供的訊息得知詞語的數量與頻率，從而決定詞語學習的先後安排，而文本標記閱讀則提供詞語的詞類標記。

「全球華語文數位教與學資源中心」網站整合豐富的語料庫資源，基於學習理論提供使用者界面幫助學習者有效掌握詞語的用法，並可依照難易程度

循序閱讀，也將客觀的統計數據提供給從事華語文教學的教育者參考，對於學習者或是教育者雙方都是利多。

圖4-3、語言教學資源查詢頁面

二、國立成功大學成鷹計畫與CANDLE前瞻性英文學習中心

「國立成功大學提升全校英語能力計畫（簡稱成鷹計畫）」由成功大學教務處委託外文系規劃執行，從2006年起爲提升成大學生的英語能力，購買英語教學網路平台、建立網路英語能力檢測系統、並建立網路多媒體互動英語學習課程。此計畫希望能鼓勵英語教師提昇本身應用資訊科技的能力，並以該能力運用在線上英語教學教材，讓學生在上課時能同時增進外語能力及電腦科技應用知能。

計畫內容包括：

（一）線上英語能力檢測系統

1. 建立網路測驗系統軟硬體設備。

2. 完成編寫英語能力檢驗題庫及分級。
3. 測試線上英語能力檢測系統並評估及改良。
4. 開放檢測供全校學生（免費）及社會人士（可收費）修習使用。

（二）多功能英語資源教室

1. 規劃教室之功能及購置軟硬體視聽設備。
2. 完成教室之設置並啓用以服務學生。
3. 規劃資源教室與課程之整合。
4. 全系教師視聽媒體教學專業成長。

（三）線上英語課程

1. 規劃線上語文課程內容及實施方式，完成軟硬體設備建置。
2. 提供學生可選擇之線上課程，讓學生不受時空的限制，進行線上學習。
3. 線上課程實施評量及修訂，學生可依評量的結果，選擇適當的課程學習。
4. 增加課程供全校學生（免費）及社會人士（可收費）修習使用。
5. 課程內容融合聽、說、讀、寫四種語言技能的訓練，題材取自與日常生活相關的食、衣、住、行、育、樂六大主題。更可加入當下流行的元素及話題，提供豐富多元的課程內容讓學生能夠藉由學習語言連結比較中西文化。
6. 因應政府擬定95學年欲實施之政策，線上學習之課程承認其學分數，更可獲得學位，經由網路學習來獲得學分和學位已成爲時代的趨勢。

其中屬數位英語教材之CANDLE(Corpus and NLP for Digital Learning of English)系統乃由國立清華大學劉顯親教授「前瞻性數位英文學習中心」研發團隊從2003年至2006年國科會數位學習國家型計劃推動下所製作之數位英語學習教材。

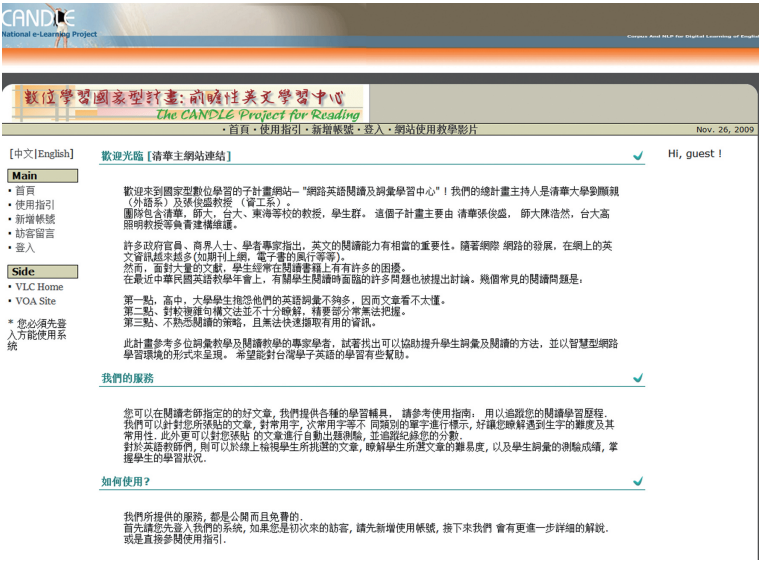


圖4-4、CANDLE首頁頁面

該計劃利用先進之語料庫及自然語言處理工具來建立網路電腦系統內之學習支援，並建立一學習中心CANDLE以協助英語學習。根據學生英文程度，提供合宜之聽、說、讀、寫、文化、翻譯之教材，以及合適的練習題目以精練其英語技能。除一般英文語料庫，CANDLE尚包括大量運用中英雙語之「光華雜誌」語料庫，其內容主要報導現代台灣之各方面資訊；雙語語料庫在計算機學界是極具前瞻性之研究議題，系統中採用雙語語料庫，讓成大學生在學習系統中善用學習者之母語長處及原有之本國背景知識學英文，這是學生心理及系統上之「電腦化」學習支援。現階段CANDLE系統提供了學生聽、說、讀、寫的練習以及全文翻譯與檢索的功能。

「成鷹計畫」雖然不是以全語料庫應用的教學網站，但是將語料庫與教學計畫做結合，整合資源後進行教學利用，也是語料庫學術資源的加值利用方式之一，依舊值得借鏡參考。

伍、延伸議題

Extended Issues

一、數位內容保護

「數位典藏與數位學習國家型科技計畫」已經耕耘多年，參與各計畫有許多數位化產出，而且成果陸續增加之中，對於擁有成果的計畫而言，除了開發新的加值應用，創造新價值之外，保護既有的資源也是重要的環節，若是多年來投入的成果被隨意剽竊，難免打擊士氣，而且對於數位化典藏的大環境也有長遠的影響。近年來，數位內容的保護機制已有不少成果，除了對數位化典藏產出進行保護外，整個數位化典藏的過程也都可納入保護機制之中，從數位化工作、資料傳遞、使用狀態追蹤等，都有相關的整合技術。

目前完整的數位內容保護概念是數位版權管理（Digital Rights Management，簡稱DRM），其內容保護的方式結合了硬體以及軟體兩者，在軟體上限制數位內容的存取權限、次數，在硬體上限制儲存媒介，兩者相互配合下，讓使用者擁有一段可使用數位內容的生命週期，在週期之內可追蹤與限制數位內容的存取、複製、使用狀況，生命週期結束之後數位內容即無法使用。

至於數位版權管理爲什麼會興起發展？其原因列舉如下：

1. 保護智慧財產權

   數位版權管理的許多技術，其前身都是來自反盜版概念，所以技術發展的目的都是爲了保護數位產出的智慧財產權。在數位化的時代中，許多數位化的內容已經是無形的財產，保護這些智慧財產不被濫用，有利於典藏單位將數位內容產出利用於其他加值項目上。

2. 保護隱私權與機密內容

   資料每進行一次傳輸，就多一分被竊盜的風險，爲了防止資訊被從中攔截，因此發展出資料加密，特定存取軟硬體等保護技術。許多敏感單位便大量使用這些技術來保護機密內容。

3. 創造新商機

   數位版權管理的機制建立之後，也建立了一套數位內容的使用模式，此模式有利於套用在商業應用之上，受到使用限制的數位內容

可以改以商品的型態，提供相關的數位資訊與服務給客戶使用。

4. 統一標準

從商業角度來看，當數位內容越來越多，數位版權管理也越發重要，許多業者看到未來的發展潛力，相繼投入相關技術的開發，搶攻市場。而統一的標準利於開發者發展相關技術，也利於使用者使用，因此能吸引廠商與消費者的加入，整合性的數位版權管理也隨著興起發展。

透視這些數位版權管理的興起原因後，不難理解數位版權管理的目的主要是保護智慧財產權，防止數位內容在沒有授權的情形下無限制散布，即使受到授權使用，也必須能追蹤使用狀態，以確保無盜用情形。[26]有效的保護財產，就有利於數位內容產出，繼續發展未來願景，同時開發數位內容的加值應用。

數位版權管理目前常用的技術包含數位浮水印、公開金鑰與數位版權描述語言，接下來就概略介紹這三種技術。

1. 數位浮水印

「數位浮水印」技術是指將代表作者的識別標誌、圖騰等植入圖片或是影像等數位影像檔案中的技術。數位浮水印可以作爲著作版權認定的依據，若發生著作權糾紛時，數位浮水印可作爲著作權擁有者的證明。正因如此，將代表自己或是單位的數位浮水印加在數位內容上，可擁有一定的嚇阻作用，意謂版權所有，請勿侵權使用。

數位浮水印適合照片圖檔、音訊檔、影像檔等數位檔案使用。數位浮水印依照可見程度，分爲顯性與隱性兩種。顯性浮水印是可見的，因此具有第一線的嚇阻效用；而隱性的浮水印則無法用肉眼

26 陳心渝、李政宏、邱一航、林韋伶，〈數位版權管理機制實作—以數位典藏管理系統為例〉，《第四屆典藏技術研討會論文集》，2005年9月，頁93~100。

察覺，具有版權的保護作用，一般在數位典藏計畫中所指的數位浮水印指的是後者。

數位浮水印的設計須考慮以下重要因素：[27]

(1) 透明度(Transparency)：浮水印不能影響到閱聽的品質。

(2) 強健性(Robustness)：浮水印即使遭到攻擊，仍能存在於數位內容之中。

(3) 安全性(Security)：植入的浮水印必須具有不可偵測的特性，即使知道了浮水印的架構，也必須要擁有相對應的金鑰才能移除。

(4) 容量(Capacity)：能加入浮水印的多寡，這條件通常和透明度的要求背道而馳。

(5) 複雜度(Complexity)：嵌入與移除浮水印所需的時間與難度，以及抽取浮水印時是否需要原始來源資料或相關資訊比對(blindness)。

(6) 可逆性(Invertibility)：原始資料是否可藉移除浮水印回復。

(7) 明確性(Unambiguous)：必須明確標示版權所有人。

各個數位典藏計畫無論是自行研發浮水印技術或是購買商業化浮水印技術，都可以考量以上的條件作爲防盜技術的標準。但隨著科技技術的發展，數位浮水印技術的防侵盜版權功用也日益減弱，只有版權宣示作用，絕對無法保障內容不會被盜取，已屬於較爲消極的防範措施。

2. 金鑰

27 綜合參考以下兩文之資料。蕭人豪、林欣慧、林金龍、林麗虹，〈數位浮水印技術發展現況：以典藏計畫為例〉，《第三屆數位典藏技術研討會》2004年8月，頁163~169；Steinebach, M., J. Dittmann & E. Neuhold "Digital Watermarking - Common watermarking techniques, Important Parameters, Applied mechanisms, Applications, Invertible watermarking, Content-fragile watermarking," http://encyclopedia.jrank.org/articles/pages/6725/Digital-Watermarking.html ，2010年1月27日下載。

金鑰是一種加密技術，利用密碼學的技術嵌入數位內容之中，藉以限制檔案的存取、複製行爲。依照設計不同主要有兩種方法：

(1) 對稱式加密法

這是傳統的加密方法，在加密與解密的兩端各擁有一把私密金鑰(Private Key)，若是其中一端要與其它人進行檔案傳遞等作業，必須雙方各自產生一把私鑰，才能解除限制。對於團體間而言，此方法較缺乏效率，但是能有效保護檔案安全。

(2) 非對稱式加密法

這是公開金鑰(Public Key)的加密方法，在加密與解密的兩端必須擁有兩把金鑰，一把是公諸於世的公開金鑰，一把爲個人私密的金鑰。加解密的動作必須仰賴成對的金鑰才能完成，利用公開金鑰加密後，可由私密金鑰解碼，其用意是利用金鑰間的不可逆性來防止有人心人士推算密碼演算法以竊取檔案。

3. 數位版權描述語言

數位版權描述語言指作者與使用者之間，對於數位內容使用的權利、義務範圍的描述語言。目前以XrML(eXtensible Rights Markup Language)最爲常見，這是國際標準組織作爲數位版權描述語言的標準，可供數位化內容的數位版權管理、後設資料管理、內容管理、內容傳遞等服務，此外也可作各式媒體的內容版權管理標準語言，如電子書、數位出版、廣播、音樂等，已有許多廠商採用。數位版權的管理可以將數位內容資料加上版權簽章資訊，以控制數位內容的流通與拷貝，除了XrML外，其他數位版權描述語言與相關組織還有ODRL(Open Digital Rights Language)、EBX(Electronic Book Exchange)與MPEG(Moving Picture Experts Group)…等。

隨著語料庫資源的多樣化，語料庫所面臨的數位版權問題也各有不同。古籍文獻語料庫的資料並無版權限制，人人都可以使用，但查詢介面、資料庫

系統卻是計畫單位辛辛苦苦建立，大部分的語料庫開放查詢時，只能呼籲使用者註明文獻查詢出處，別完全抹殺計畫單位的貢獻。也有一些計畫採用不提供全文的策略，讓查詢結果只出現局部段落，一方面達到語料庫的資料查詢用意，另一方面又保護資料的著作權所有人或自己的辛苦成果。[28]

隨著多媒體型式的語料庫產生後，數位內容保護的相關技術顯得更爲重要，未來，當語料庫要走向加值應用時，更必須納入數位版權管理的概念，以確保數位內容的珍貴性。

二、人力與設備成本分析

語料庫數位化工作一般來說不像文物或藝術品等物件的數位化典藏工作般，需要使用到相當昂貴的機器設備，因此設備成本的支出只佔整體計畫經費的小部分，而因語料庫的數位化工作耗時費力，人力成本的支出將會佔計畫經費的較大部分；此外，田野調查收錄語料所需要的旅費、食宿、膳雜與人事開銷等，也是一筆龐大的支出，因此當計畫主持人規劃計畫細節與撰寫計畫書時，也要費心進行經費規劃。

由於語料庫的類型多元，並不是每一種語料庫都需要進行田野調查，因此本書在此不詳述田野調查如何規劃經費，僅就語料庫數位化過程之中的人力分析與設備使用進行介紹。

（一）人力成本分析

國立中正大學語言學研究所蔡素娟教授所主持的台灣兒童語料庫—閩南語兒童語料庫計畫，在執行期間以工作流程調查表（表5-1）記錄下計畫執行期間所需要的人力與工作時數，本節將以此爲範例，瞭解一項計畫所需的人力成本。

28 中研院現代漢語平衡語料庫的查詢結果即不提供出處或完整段落，部份原因即是未獲得語料著作權所有人的足夠授權。

「台灣兒童語料庫」建置計畫爲期三年（89.8.1至92.7.31），語料來源爲蔡素娟教授國科會專題研究計畫「台灣話聲調習得的發展之研究」（87.8.1至89.7.31，爲期三年）實地採集嘉義地區14名（9男5女）一歲多至三歲多的兒童玩遊戲或看書時的自由對話，總長度329小時又14分鐘，轉寫文字約230萬字，整個語料庫由資料收集到建置完成歷時長達六年。

以閩南語兒童語料庫所記錄的工作調查來看，該計畫的田野調查總共錄下431人次的錄音，錄音時間長達330小時；進行錄音剪輯時，每小時的錄音檔案必須投入1.5倍的時間，因此光是錄音剪輯就必須花費495小時的時間，換算每天8小時的工作日，約計需要62天，分攤至每位助理後，每個人也需要花費20天來進行錄音剪輯。錄音剪輯需要花費大量時間，台灣兒童語料庫所計算的1.5倍剪輯時間屬於極佳理想的狀態，一般進行錄音剪輯時可能需要花費更久時間。

當計畫進行到人工轉記漢字與人工IPA記音時，分別需要錄音時間的10倍與4.5倍時間來執行，計算之後，其總時間也長達4,785小時。以每天上班8小時，一週上班五天，且有三位助理分攤進行的狀況計算，也必須花費10個月的時間來執行。這僅只是工作流程調查表中所記錄的時間，其他包含人工斷句、人工斷詞、人工標記拼音、人工標記詞類等，都尚未列入計算內。

又如中央研究院資訊科學研究所王新民副研究員曾以三年時間建置國語新聞口語語料庫，收錄公共電視2001年11月至2003年6月計250小時的國語新聞節目，在使用程式自動比對、擷取公視網站上的相應文字資料作爲轉寫的文字底稿的情形下，兩位專任助理最後也僅能完成198個小時的文字轉寫與言談標記工作。[29]由此可知，若是未發展自動軟體，在語料庫建置的過程之中，必

29 Wang, Hsin-Min, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng. 2005. "MATBN: A Mandarin Chinese Broadcast News Corpus," Computational Linguistics and Chinese Language Processing 10.2, pp.219-236.

須花費大量的人力來進行詞語分析、標記等工作，若是計畫主持人有心建立更詳細的語料內容，則所花費的人工時間會更久。

語料庫數位化工作至目前爲止，仍然必須仰賴大量的人才與人力參與，才能順利完成工作，除去一般事務性人力成本後，執行工作的人力成本會是經費支出之中的一大部分，計畫主持人在規劃經費時，必須瞭解計畫需求，仔細計算人力成本需求。

（二）設備成本分析

語料庫的建置工作中，最常使用到的器材就是錄音設備；而每一個語料庫的最終目的就是建立一個資料庫伺服器，以提供使用者查詢，因此伺服器也是每個計畫必須購買的設備，以下將這些設備做簡單的介紹。

錄音設備可分爲類比式與數位式，類比式如卡匣錄音機(Cassette Recorders)，數位式如MD錄音機(Minidisc Recorder)、DAT錄音機(Digital Audio Tape Recorders) 以及固態錄音機(Solid-State Recorder)、硬碟錄音機(Hard-disk Recorder)以及電腦。目前市面上的錄音筆非常多，甚至是一般的MP3播放機就附有錄音功能，但是爲了收錄的語料品質能滿足語音聲學分析或典藏的需求，建議捨棄一般的卡匣錄音機或錄音筆，選用較高階的數位錄音硬體與外接麥克風。

1. 線性PCM錄音機

有些PCM錄音機本身就配備一對高感度的電容式麥克風，除了收音敏感度佳之外，還能夠收錄立體聲音場，收音

圖5-1、PCM錄音機[30]

30 圖片提供：台灣樂蘭企業股份有限公司。

品質明顯優秀；而錄音檔案上，PCM錄音機的錄音取樣率可以高達96kHz以上，相對於一般MP3檔案的取樣率僅爲44.1kHz，可知PCM錄音機的錄音效能更佳。PCM錄音機能直接將聲音儲存爲未經壓縮的高品質檔案（如WAV格式），也利於事後的監聽、判讀，以及後製處理。也因爲PCM錄音機的性能較佳，因此價格較爲高昂，約略在一萬至兩萬元之間。

2. 數位錄音座

圖5-2、數位錄音座[31]

數位錄音座的價格更昂貴，約莫等同於一台配備高級的筆記型電腦。 此處所指的數位錄音座並非是音樂錄音工作所使用的器材，而是比較小型且適合語言收錄的器材。此類數位錄音座通常必須與筆記型電腦搭配使用，在科技進步之下，目前數位錄音座的體積已經大爲縮小，重量約莫在1公斤上下，對於田野調查工作者而言更便於攜帶。

數位錄音座可以用USB介面與筆記型電腦連結，收錄資料時可以直接存入筆記型電腦之中，而某些機型也提供記憶卡插槽，可以不須與電腦連結，直接將錄音資料收錄至記憶卡內即可。數位錄音座最大的優勢是提供多組錄音輸入孔，接上多組麥克風後，可同時進行多人發音的語言收錄。此外，數位錄音座的錄音品質也可以達到線性PCM規格，儲存檔案可以選擇16-bit或是24-bit，而取樣率可高達

31 同前註。

192kHz以上，整體錄音品質更勝PCM錄音機。

3. 麥克風

麥克風依照構造可以分爲兩大類，分別是Dynamic Microphone動圈式麥克風與Condenser Microphone電容式麥克風。[32]動圈式麥克風採用線圈、振膜、永久磁鐵組合，一般到KTV消費時，拿在手上高唱的麥克風就屬此類，特色是造價成本低、聲音溫潤，缺點則是體積較大，靈敏度低，高低頻表現較不理想。

另一種電容式麥克風是以電容隔板造成電壓變化的方式來記錄音訊，優點是體積小且靈敏度高，適用於高感度錄音；不過電容式麥克風需要以穩定電壓驅動，有些產品需要額外電池供電。電容式麥克風的高靈敏度適合語料收錄，輕便的特性適合外出使用，對於語料庫計畫而言，建議使用電容式麥克風。

根據收錄聲音的靈敏度差異，麥克風的設計可略分全指向(Omni-directional)、單指向(Cardioid)與雙指向(Bi-directional)等類型，全指向型會收錄周遭的許多聲音，雙指向型則收錄前後兩方向的聲音，而單指向型只收錄一個方向的音源。語料庫計畫收錄語言時，一支麥克風只專注收錄發音人的語音，周邊的聲音干擾必須越少越好，以利後續的判聽作業，因此單指向的麥克風較適合收錄語料使用。

選擇麥克風還要注意接頭規格是否符合錄音筆與錄音座使用。依照響應頻率、靈敏度、抗阻等規格差異，麥克風的售價價差很大，以3.5mm接頭接上PCM錄音筆的迷你麥克風與領夾型麥克風，售價約在2,000~4,000元左右；規格更佳但體積稍大，[33]可桌立或是手持的麥克風，售價約在4,000~6,000元左右；而規格最佳的頂級麥克風，其

32 麥克風，維基百科，網頁：http://zh.wikipedia.org/zh-tw/。

33 此種麥克風的體積還是小於手握式的動圈式麥克風，重量也較輕。

售價則高達一萬元以上。

如果搭配PCM錄音筆使用，建議選擇領夾式麥克風，以避免手持麥克風與錄音筆的不便；如果與錄音座一起使用，那可桌立與手持的麥克風最爲合適。

圖5-3、可桌立與手持的電容式麥克風[34]

4. 伺服器(Server)[35]

以硬體方面來說，伺服器是指專門儲存數位資源的電腦硬體；以軟體而言，也泛指用來管理數位資源並提供使用者服務的電腦軟體，例如檔案伺服器、資料庫伺服器與應用程式伺服器三種。在此，本文要介紹的是用來儲存數位資源的設備。

伺服器和一般桌上型電腦有許多的差異，伺服器是給數位資源擁有者使用，數位資源透過網路提供給一般桌上型電腦使用者。伺服器的硬體耐用度是針對24小時不休息的使用狀態而設計，爲了應付許多使用者的需求，運算能力也比一般桌上型電腦更爲優良，一般來說，常見的伺服器，大略可分爲三類，其體積、型態與效能等也有一些不同的地方，主要有直立式伺服器、機架式伺服器 (Rack-Mount Server) 與刀鋒式伺服器(Blade Server)這三類。

34 圖片提供：台灣樂蘭企業股份有限公司。

35 圖片提供：IBM。

直立式伺服器是入門的機種，適用於一般小型公司，外觀近似一般桌上型電腦，兩者容易混淆。一台直立式伺服器所佔用的空間和一般電腦相似，但是其配備採用比一般電腦穩定許多、工作效率也較佳的CPU與記憶體，足以負荷長時間不間斷的工作；因體積的限制，直立式伺服器的硬碟擴充性不強，擴充數量與一般桌上型電腦相近。直立式伺服器是大部分小型的典藏計畫會選購的類型，入門機型的價位約莫在五萬至六萬元之間，功能比較齊全強大的機型，價格約莫在十萬元以內。

圖5-4、直立伺服器

選購直立式伺服器時，也一定要增購UPS不斷電系統，以保障資料儲存與傳輸上的安全。目前市面上的UPS不斷電系統主要有三種，分別是Off-Line離線式、On-Line在線式與Line Interactive在線互動式，三種UPS不斷電系統的價差很大，但是伺服器建議使用最安全的On-Line在線式不斷電系統，價位約在一萬元上下。

當伺服器的需求達到十台直立式伺服器以上時，佔用的空間將相當龐大，可使用機架式的伺服器以節省空間。機架式伺服器爲扁

平狀，最小機架單位以1U，[36]一台機架式伺服器的大小約1U到5U不等，爲了有效管理機架式伺服器，以及善用空間，此種伺服器必須裝在機櫃內使用，一個全高的機櫃約有42U的空間。

機架式伺服器的優勢是擁有極大彈性的擴充性，效能比直立式伺服器優秀；不過保養負擔也相對沉重，因爲機櫃內安放了多台伺服器，散熱性顯得相當重要，必須24小時開啓空調來克服散熱問題，同時機櫃的放置地點，保管方式、保管人員也要仔細安排。

圖5-5、1U大小的機架式伺服器

圖5-6、安裝於機櫃內的機架式伺服器

36 機架單位，維基百科網頁http://zh.wikipedia.org/zh-tw/机架單位。機架單位由美國電子工業聯明制訂，用來標定伺服器等設備的單位，高為44.45mm，寬482.6mm。

大型機房內的機架式伺服器後方會佈滿排線，即使有空調協助散熱，溫度仍然驚人，除了散熱外，佈滿的排線也增添了管理與維護上的困難，所以刀鋒式伺服器應運而生。

刀鋒伺服器需與刀鋒基座搭配，基座提供電源、風扇、網路等功能，基座上的插槽則可以插上刀鋒伺服器。訊號連接以插槽取代排線，而且擁有類似熱抽取的功能，更利於管理與維護；在伺服器數量相等的情況下，刀鋒伺服器的散熱性也更好。

圖5-7、刀鋒伺服器

機架式伺服器的效能好，但是價格較直立式伺服器昂貴許多，一組全高的機櫃價格大約六萬元左右（包括KVM螢幕），一台機架式伺服器（空機不含硬碟）的成本至少要八萬元左右，加上不斷電系統以及周邊配備，一組這樣子的設備其成本很容易就超過二十萬元。而先進的刀鋒式伺服器單價比一台機架式伺服器更爲昂貴，建置成本更不是一般計畫所能負擔，後續在人力上的維護成本，也是不容小看。

一般而言，中型企業或是具規模的組織才會建立機架式伺服器的機房，例如學校或是文教單位等，對於小規模的計畫單位來說，這種伺服器太過沉重；因此，除了自行購買直立式伺服器之外，向大單位租用代管伺服器也是可以選擇的方式。

撰文：詹景勛、李佩瑛

表5-1、台灣兒童語料庫工作流程調查表
單位：國立中正大學 語言學研究所　　數位化物件名稱：閩南語兒童語料　子計畫名稱：台灣兒童語料庫
主持人（負責人）（E-mail、Tel）：蔡素娟 教授 Lngtsay@ccu.edu.tw　05-2720411*31502
聯絡人（E-mail、Tel）：謝沛諭 astpph@ccu.edu.tw　05-2720411*21509

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
1	訓練研究助理（瞭解閩南語音韻及書寫系統；瞭解閩南語詞彙、句法、語意及詞類標記系統；瞭解CHILDES系統；熟悉IPA國際音標記音）	計畫主持人3名研究助理（具語言學碩士級背景知識；母語為閩南語）	錄音機（NT8,000/台） 錄音帶（NT150/片）		(1)《閩南語詞彙》一、二冊 楊秀芳, 教育部國語推行委員會, 1998. 《台灣閩南語語法稿》楊秀芳, 大安出版社, 1995. 《台灣閩南語方言記略》張振興, 文史哲出版社, 1993. Handbook of the International Phonetic Association, (1999) The CHILDES Project, Brian MacWhinney (1995)	每星期3-6小時的討論會與記音練習。		碩士級專任助理1名 NT34,000/月 碩士級兼任助理2名 NT6000*2/月

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
2	徵求說閩南語家庭的兒童（以海報及網路發布廣告；利用幼稚園家長日到場對家長說明，徵求說閩南語家庭的兒童）	3名研究助理（熟悉電腦網路應用；基本美工海報設計）	桌上型電腦3部(NT50,000/台)	(1) Microsoft OS: 98/2000/ XP Microsoft Office: Word/ Excel			目標選定中正大學附設托兒所、幼稚園及鄰近鄉鎮，徵求來自說閩南語家庭，年齡在一歲至三歲之間的幼兒。陸續共選出14名兒童。	
3	排定錄音時間（聯絡家長；並排定錄音時間表）	3名研究助理	桌上型電腦(NT50,000/台)					
4	準備錄音器材	3名研究助理	迷你光碟隨身錄音機(NT12,000/台) 專業用麥克風(NT8,000/支) 迷你光碟片(NT800/15片裝) 專業用耳機(NT3,000/副)			二週	選擇方便攜帶、機動性強、容量較大、容易長期保存語料之錄音器材。	

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
5	進行訪談錄音（至兒童家中進行訪談錄音。錄音為週期性，寒暑假亦不間斷。二歲以下者，每週訪談一次；二至三歲者，每兩週訪談一次；三至四歲者，每二至三週訪談一次）	3名研究助理（熟悉迷你光碟錄音機之操作；有耐心；喜歡與小孩互動）	迷你光碟隨身錄音機(NT12,000/台) 專業用麥克風(NT8,000/支) 迷你光碟片(NT800/15片裝）			每次訪談約1-2小時不等，實際錄音時間40-60分鐘。 錄音期間：1997年10月至2000年5月。 共錄音431人次，約330小時。	進行訪談中，錄下兒童在家長或保姆陪同下，在自己家中的日常對話。錄音的內容除了自然言說，還藉助圖畫簿、故事書、玩具、布偶、剪紙、摺紙或其他遊戲，引發兒童主動說話。	助理田調費NT200元/人次 訪談費NT200/人次
6	錄音剪輯	3名研究助理（熟悉迷你光碟錄音機之操作）	迷你光碟隨身錄音機(NT12,000/台) 專業用耳機(NT3,000/副) 迷你光碟片(NT800/15片裝）			每小時的錄音約需耗時1.5小時剪輯。 總時間：1.5*330小時。	光碟中輸入錄音日期、檔名。 將不相關的錄音或太長的空白錄音刪除。 將錄音切割為較小段落，並在光碟中標記段落編號。	

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
7	錄音備份（迷你光碟之備份製作）	3名研究助理（熟悉迷你光碟錄音機之操作）	迷你光碟隨身 錄音機(NT12,000/台) 迷你光碟錄音座(NT35,000/台) 迷你光碟片(NT800/15片裝) 耳機 (NT3,000/副) 光纖線 (NT2,000/條)			每片迷你光碟片約需2.5小時。 總工作時間：2.5*330(hr)=825小時		
8	數位化轉錄	多名研究助理（熟悉迷你光碟錄音機之操作）	桌上型電腦(NT50,000/台) 迷你光碟隨身錄音機1部 (NT12,000/台)	GoldWave Digital Audio Editor（GoldWave Inc. 研發）			將迷你光碟錄音檔轉為較不佔空間之MP3格式，以方便儲存。於日後可隨時轉為語音分析所需之格式（如WAV格式）	

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
9	閩南語書寫系統之確立 （由於閩南語的漢字書寫系統目前並沒有定案，再加上有許多本字無法確定，或者有音無字的情形，因此有必要訂定文字轉記的原則）	3名研究助理（熟悉電腦文書處理之操作；閩南語書寫系統之基本知識）	桌上型電腦 (NT50,000/台) 迷你光碟隨身錄音機 (NT8,000/台) 專業用耳機 (NT3,000/副)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	所參考的辭典主要有四本，依優先順序如下： 《臺灣閩南語辭典》董忠司, 五南圖書出版公司, 2001. 《閩南語大辭典》陳修, 遠流出版公司, 1998. 《廈門方言詞典》李榮, 江蘇教育出版社, 1998. 《閩南語詞彙》楊秀芳, 教育部國語推行委員會, 1998.			

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
10	人工轉記漢字（錄音檔轉記為文字檔）	多名研究助理（熟悉電腦文書處理之操作；閩南語書寫系統之基本知識）	桌上型電腦(NT50,000/台）	CHILDES之CHAT轉記平台	CHILDES（Child Language Data Exchange System; MacWhinney and Snow 1985, MacWhinney 1995）兒童語料交換系統	每1小時錄音需要花約10小時不等的時間轉記成文字檔。 總時間：330*10=3,300小時。		
11	人工斷句（將自然言談切分成獨立意義句子）	多名研究助理（具句法學及語意學等語言學相關背景知識）	桌上型電腦(NT50,000/台）	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	CHILDES兒童語料交換系統之斷句標準		本語料庫之語料為口語語料。需參考言談分析之斷句原則。	
12	人工斷詞（將語句切分為獨立意義、且扮演特定語法功能的字串）	多名研究助理（具語言學相關背景知識）	桌上型電腦(NT50,000/台）	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	中華民國計算語言學學會所訂定之「資訊處理用中文分詞規範調查研究及草案研擬」。			

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
13	人工IPA記音（採語音轉記(phonetic transcription)的方式。在音段方面，以Unicode IPA符號記音；聲調採用五度標音法）	多名研究助理（熟悉電腦文書處理之操作；熟悉國際音標；有語音學基礎）	桌上型電腦(NT50,000/台) 迷你光碟隨身錄音機(NT12,000/台) 專業用耳機(NT3,000/副)	(1) CHILDES之CHAT轉記平台 (2) Microsoft OS: 98/2000/XP (3) Microsoft Office: Word/Excel (4) Unicode IPA字型軟體	CHILDES兒童語料交換系統 Handbook of the International Phonetic Association (1999)	每小時的錄音約需花4.5小時記音。共4.5*330=1485小時。		
14	建立新詞清單（以轉記好之文字檔中之所有詞彙建立清單，經由人工確認詞彙清單中的漢字與詞典是否標準一致）	多名研究助理（具語言學相關背景知識）	桌上型電腦(NT50,000/台)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel				
15	人工標記拼音	多名研究助理（具語言學相關背景知識）	桌上型電腦(NT50,000/台)	(1) Microsoft OS: 98/2000/XP (2) Microsoft Office: Word/Excel	教育部於民國八十七年所公佈之「閩南語羅馬拼音第二式」。			
16	詞類標記系統之確立	多名研究助理（具句法學與語意學基本知識）	桌上型電腦(NT50,000/台)	Microsoft OS: 98/2000/XP Microsoft Office: Word/ Excel 「閩南語詞彙庫」	中央研究院詞庫小組「詞類標記原則」 CANCORP: The Hong Kong Cantonese Child Language Corpus, Lee and Wong (1998). 台灣閩南語動詞分類研究 曹逢甫 (1996).		採用中研院詞庫小組的詞類標記，但是僅限於46個簡化標記，以避免詞類劃分過細時產生主觀強制性的歸類。	

程序	工作內容	操作人員（數量、專業能力之要求）	硬體（名稱、版本、價格）	軟體（名稱、版本、價格等）	依循標準（技術規範、成品規格、品質要求…等）	耗時	總結（困難、缺失、特色…等）	成本估算
17	人工標記詞類	多名研究助理（具句法學與語意學基本知識）	桌上型電腦(NT50,000/台）	Microsoft OS: 98/2000/XP Microsoft Office: Word/ Excel (3)「閩南語詞彙庫」				
18	發展自動斷詞與拼音程式（斷詞及標注拼音）	1名程式設計師（熟悉電腦程式語言；具語言學基本知識）	桌上型電腦(NT50,000/台）	「閩南語詞彙庫」 Linux Operating System Visual C	本計畫自行研發		此程式除了斷詞及標注拼音之外，還可以將新詞納入詞彙庫。	程式設計費（按件計酬）
19	發展漢字檢查程式	1名程式設計師（熟悉電腦程式語言；具語言學基本知識）	桌上型電腦(NT50,000/台）	(1)「閩南語詞彙庫」 (2) Linux Operating System (3) Visual C	本計畫自行研發			程式設計費（按件計酬）
20	發展自動詞類標記程式	1名程式設計師（熟悉電腦程式語言；具語言學基本知識）	桌上型電腦(NT50,000/台）	(1) Linux Operating System (2) Visual C	本計畫自行研發		根據「閩南語詞彙庫」中該詞項之詞類標記。	程式設計費（按件計酬）

陸、結語

Conclusions

對於語料庫的應用，語言學家關心的是如何呈現該語言原來的面貌，而電腦科學家則希望能將語料加以組織及結構化，再導入資料庫技術，以應付使用者不同的檢索需求。因此，語言典藏數位化一方面將克服傳統紙筆技術的問題，另方面也可摒棄書面格式的語料輸出，而這些理想都必須有賴電腦關聯式資料庫的技術予以達成。

從書面格式的語料庫進展到關聯式資料庫，代表著複雜度的增加，但是也在資料的有效運用及操控性上獲得相對的回報。細究之下，複雜度的增加並不是眞實的，那些不同但相連結的資料表都可被認爲是與語言學家的專業知識更加密切關聯。資料庫理論無疑是如何設計欄位、記錄及資料表，正如同語言學要如何呈現單字、句子及文章一樣，在彼此之間建立一個緊密而有效的連結。

本書所介紹的語料庫即利用現代資料儲存與擷取技術，以電腦的資料結構將原始語料庫的檔案轉換成資料庫。其中，對於語料庫的結構化、與規格化，乃利用關聯式資料庫的精神，一方面將語料資料定義的更爲嚴謹，另方面對於資料與資料之間的連結也更爲明確。

參考文獻

References

專書

沈漢聰，《數位典藏技術彙編》電子書，數位典藏國家型科技計畫，2004年，初版。

van Son, R., Wesseling, W., Sanders, E., and van den Heuvel, H. (2009). "Promoting free Dialog Video Corpora: The IFADV Corpus Example," in M. Kipp et al. (Eds.): Multimodal Corpora, LNAI 5509, pp. 18–37, 2009.

期刊論文

陳心渝、李政宏、邱一航、林韋伶，〈數位版權管理機制實作—以數位典藏管理系統爲例〉，《第四屆典藏技術研討會論文集》，2005年9月，頁93~100。

蕭人豪、林欣慧、林金龍、林麗虹，〈數位浮水印技術發展現況：以典藏計畫爲例〉，《第三屆數位典藏技術研討會》2004年8月，頁163~169。

詞庫小組，《研究院語料庫的內容及說明》，中文詞知識庫小組技術報告#95-02，南港，中央研究院，1995年。

Huang, Chu-Ren and Keh-jiann Chen. A Chinese Corpus for Linguistics Research. In the Proceedings of the 1992 International Conference on Computational Linguistics (COLING-92). 1214-1217. Nantes, France. 1992.

Huang, Chu-Ren Corpus-based Studies of Mandarin Chinese: Foundational Issues and Preliminary Results. In Matthew Chen and Ovid Tzeng Eds. In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change. Pp. 165-186. Taipei: Pyramid. 1994.

Wang, Hsin-Min, Berlin Chen, Jen-Wei Kuo and Shih-Sian Cheng. "MATBN: A Mandarin Chinese Broadcast News Corpus," Computational Linguistics and Chinese Language Processing 10.2, pp.219-236. 2005.

網路資源

李道明，〈影音檔案數位化之規劃與流程〉，拓展台灣數位典藏計畫網站，2009年1月31日下載，http://content.ndap.org.tw/index/?p=843。

Steinebach, M., J. Dittmann & E. Neuhold "Digital Watermarking - Common watermarking techniques, Important Parameters, Applied mechanisms, Applications, Invertible watermarking, Content-fragile watermarking," http://encyclopedia.jrank.org/articles/pages/6725/Digital-Watermarking.html ，2010年1月27日下載。

附錄

Appendix

語料庫建置相關網路資源:

1. 中文詞彙網路：http://cwn.ling.sinica.edu.tw/。
2. 中文詞彙特性速描系統：http://bow.sinica.edu.tw/。
3. 中央研究院現代漢語平衡語料庫：http://dbo.sinica.edu.tw/SinicaCorpus/。
4. 中央研究院語言座標計畫：http://linganchor.sinica.edu.tw/。
5. 中英雙語知識本體詞網：http://bow.sinica.edu.tw/。
6. 台灣手語影像辭典：http://tsl.ccu.edu.tw/web/index.php。
7. 台灣南島語數位典藏計畫：http://formosan.sinica.edu.tw/。
8. 台灣樂蘭企業股份有限公司：http://www.rolandtaiwan.com.tw/roland/index.php。
9. 全球華語文數位教與學中心：http://elearning.ling.sinica.edu.tw/。
10. 前瞻性英文學習中心網頁：http://elearning.eng.ntnu.edu.tw/CANDLE/。
11. 國立成功大學成鷹計畫網頁：http://english.ncku.edu.tw/。
12. 新世紀語料庫：http://mmc.sinica.edu.tw/。
13. 閩南語典藏—歷史語言與分布變遷資料庫：http://southernmin.sinica.edu.tw/。
14. 閩客語典藏：http://minhakka.ling.sinica.edu.tw/bkg/index.php。
15. 數位典藏與數位學習國家型科技計畫後設資料工作組：http://metadata.teldap.tw/index.html。
16. 數位典藏技術彙編電子書：http://www2.ndap.org.tw/eBook/showContent.php。
17. AHDS Literature, Languages and Linguistics：http://www.ahds.ac.uk/litlangling/index.htm。
18. Brown Corups Manual：http://icame.uib.no/brown/bcm.html。
19. CGN--The Spoken Dutch Corpus project：http://www.tst.inl.nl/cgndocs/doc_English/topics/index.htm。

20. DoBES: Documentation of Endangered Languages： http://www.mpi.nl/DOBES/。
21. Dublin Core Metadata Initiative：http://dublincore.org/。
22. ELAN：http://www.lat-mpi.eu/tools/elan/。
23. GNU Operating System： http://www.gnu.org/。
24. GeoLang: The prime sponsor of the World Language Documentation Cengre：http://www.geolang.net/。
25. HCRC Map Task Corpus：http://www.hcrc.ed.ac.uk/maptask/。
26. IBM台灣網頁：http://www.ibm.com/tw/zh/。
27. IFA Dialog Corpus：http://www.fon.hum.uva.nl/IFA-SpokenLanguageCorpora/IFADVcorpus/。
28. IMDI: ISLE Meta Data Initiative：http://www.mpi.nl/IMDI/。
29. ISO-1 Registration Authority：http://www.infoterm.info/standardization/iso_639_1_2002.php。
30. ISO 639-2 Registration Authority： http://www.loc.gov/standards/iso639-2/。
31. ISO 639-3 Registration Authority： http://www.sil.org/iso639-3/。
32. ISO 639-5 Registration Authority： http://www.loc.gov:8081/standards/iso639-5/。
33. Open Language Archives Community：http://www.language-archives.org。
34. OLAC Metadata Set：http://www.language-archives.org/OLAC/olacms.html#Metadata elements。
35. SIL International：http://www.sil.org/。
36. SMIL: W3C Synchronized Multimedia Integration Language：http://www.w3.org/AudioVideo/。
37. TEI: Text Encoding Initiative： http://www.tei-c.org/index.xml。

38. The Odrl Initiative：http://odrl.net/。

39. Unicode Standard： http://www.unicode.org/standard/standard.html。

40. Wynne, M (editor). 2005. Developing Linguistic Corpora: a Guide to Good Practice. Oxford: Oxbow Books. Available online from http://ahds.ac.uk/linguistic-corpora/ [2010-01-08下載]。

41. XML: W3c Extensible Markup Language： http://www.w3.org/TR/2006/REC-xml11-20060816/。

42. XrML-The Digital Rights Language for Trusted Content and Services：http://www.xrml.org/about.asp。

國家圖書館出版品預行編目資料

語料庫建置入門數位化工作流程指南 / 李佩瑛等作.
--初版. --臺北市：數位典藏拓展臺灣數位典藏計畫，
民 99. 03 面；公分.
參考書目：面
ISBN 978-986-02-2784-0(平裝)

1. 文獻數位化 2. 文物典藏 3.語料庫 4.工作說明書
028. 026 99004583

語料庫建置入門 數位化工作流程指南

指導單位：行政院國家科學委員會
發 行 人：林富士
總 編 輯：邱澎生
執行編輯：林彥宏、林定立、林芳志、高朗軒
作　　者：李佩瑛、邱智銘、郭彧岑、曾淑娟、黃菊芳、詹景勛、蔡素娟、盧秋蓉、蕭素英、賴佳旻、戴浩一、謝沛諭、蘇秀芳、中文詞彙網路小組
審 稿 者：中央研究院語言研究所 蕭素英助研究員
發行單位：數位典藏與數位學習國家型科技計畫 拓展台灣數位典藏計畫
地址：115 台北市南港區研究院路二段128號
中央研究院歷史語言研究所
電話：886-2-2782-9555轉288
傳真：886-2-2786-8834
網址：http://content.teldap.tw
Email：content@gate.sinica.edu.tw
封面設計：禧恩股份有限公司 林秦華先生
排版印刷：禾古精緻印刷有限公司

中華民國99年3月初版
ISBN 978-986-02-2784-0