語料庫數位化工作流程指南

一、前言

語言是人類表達與溝通的重要媒介之一，試想若是缺少語言這樣的工具，世界會變得如何？如果各位使用過 Google 的語言選項就可以知道，將近 117 種的語言洋洋灑灑的呈現在網頁上任君選用，何等便利。目前世界上現存語言已知的有三千多種，在新的語言如世界語誕生的同時，也有許多的語言凋零當中。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。談到語料庫，一般人或許感到納悶與陌生，簡單來說，語料庫在語言學上指大量的文本，經整理與格式標記，由數位的方式處理與保存，再加以應用。

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，甚至結合影像以影像辭典的類型呈現，不僅在研究分析上給予很大的助益，在語言學習上也有極大的貢獻。舉例來說，手語是聽障者日常使用的語言，而世界各地的手語就如同口語一般都有其一套語法系統與結構。歷時四年的「台灣手語參考語法」，以 2,300 個影像說明 4,500 個手語辭彙，由中正大學戴浩一教授領軍並於 2006 年上線，這是全球首部有系統而完整的手語參考語法，民眾不但可以自學標準手語，想了解台灣手語的外籍人士也可以透過英文版學習；「台灣手語參考語法」也可以提供啓聰學校與從事特殊教育相關人員的教學、參考資料。

這些研究成果除學術價值外，還有人文、實用及應用價值。在人文方面，手語研究可以反映語言與文化的多元性，保存珍貴的語言資料，可促使普羅大眾瞭解聽障者社會與文化，進而尊重並欣賞聽障者文化與價值體系。

本指南參照國科會「數位典藏國家型科技計畫」內語言主題小組建置之語料庫爲基礎，包含說明語料庫建構技術，並與國外相關語料庫比較其差異；另以計劃實際執行經驗值，探討建置語料庫時面臨的挑戰以及發展，供目前或未來想進行語言典藏的人員參考。除此之外，也期望藉此能夠建立大眾對語言典藏的瞭解，進而一同加入典藏工作，擴大並豐富目前語料庫的典藏量。

二、語料庫構成技術

（一）概述

語言學是以研究人類語言爲對象的學科，牽涉範圍相當廣泛，從生理、心理、物理、數學、地理、哲學、美學、社會學、歷史學、民族學、人類學、工程學各方面等都與語言相關。傳統上，語言學是文化人類學的分支，但現在語言學逐漸自成一格。語言學研究句法和詞語等語言的描述，也研究語言的發展史。而爲研究用途的資料，以往只能以單一物品的方式保存，如書籍、報章雜誌…等文本，甚至採集口語資料用的錄音帶等，這些資料在日積月累之後必定會囤積相當的數

量，在使用與保存方面都耗費心力，於是結合計算機語言的語料庫就成爲一有效的解決方案，並成爲語言學理論研究、應用研究和語言工程上不可或缺的基礎資源。

語料庫通常指爲語言研究收集，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與記錄語言的實際使用狀況。透過語料庫觀察和掌握語言事實，可以分析以及研究語言系統的規律性。

計算語言學於數理語言學興起後約 20 年也應運而起，此時電子計算機已發展到第四代，成爲語言學家的得力助手。計算語言學的目的，是闡明如何利用電子計算機來進行語言研究，其項目有資料統計、情報檢索、研究詞法及句法、文字識別、語音合成、編制機器輔助教學、機器翻譯等等。由於電子計算機儲存量龐大，計算能力精確，作業效能較高,又能用於撰稿、修改、儲存文稿和各種資料，對於語言研究有很大的助益。

語料庫具多種類型，確定類型的主要依據研究目的和用途，這一點往往能夠表現在語料採集的原則與方式。語料庫原則上可以分爲四種類型:
(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；(4)專用的(Specialized) : 只收針對某一特殊用途的語料。 除此之外，按照語料的種類劃分，語料庫也可以分爲單語(Monolingual)，雙語(Bilingual)和多語的(Multilingual)。

語料庫與語言訊息處理有著某種與生俱來的聯繫，以往不了解語料庫方法的時候，在自然語言理解和發展，機器翻譯等研究中，分析語言主要的方法是基於規則，但對於用規則無法表達或不能涵蓋的語言事實，計算機就很難處理。語料庫出現之後，人們利用它對於自然語言進行調查與統計，建立統計語言模型，研究和應用基於統計的語言處理技術，在訊息檢索、文本輸入和整理語料的自動分詞和標注，到語料的統計和檢索，自然語言訊息處理的研究都爲語料的加工提供了關鍵性的技術。

計算機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。

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屬性，例如語體、內容所屬的領域、作者、出版時間、發行出版社…等，篇體描述是在文本裡添加各種語意學屬性標記，對於漢語語料庫來說，常見的有詞語切分標記、詞性標記、專有名詞標記，還有針對語法特徵標記，如子句標記或語意訊息標記…等。對於漢語語料庫的加工，一般是從詞性切分、詞性標記，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的增加。

通常沒有篇體描述訊息的叫做生語料，對漢語的生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。如果還加注了詞性標記，那麼可以獲得的訊息就更多了。語料的標注如果由人來作，當然能夠保證其準確性，但是對於處理大規模的語料來說，顯然人工標注較爲不切實際。因此每個大規模的語料庫加工往往需要藉助自動化的方式，其中詞語自動切分，詞性自動標注就成爲眾所矚目的語料加工技術。

（二）操作實例— 台灣兒童語料庫數位化工作流程

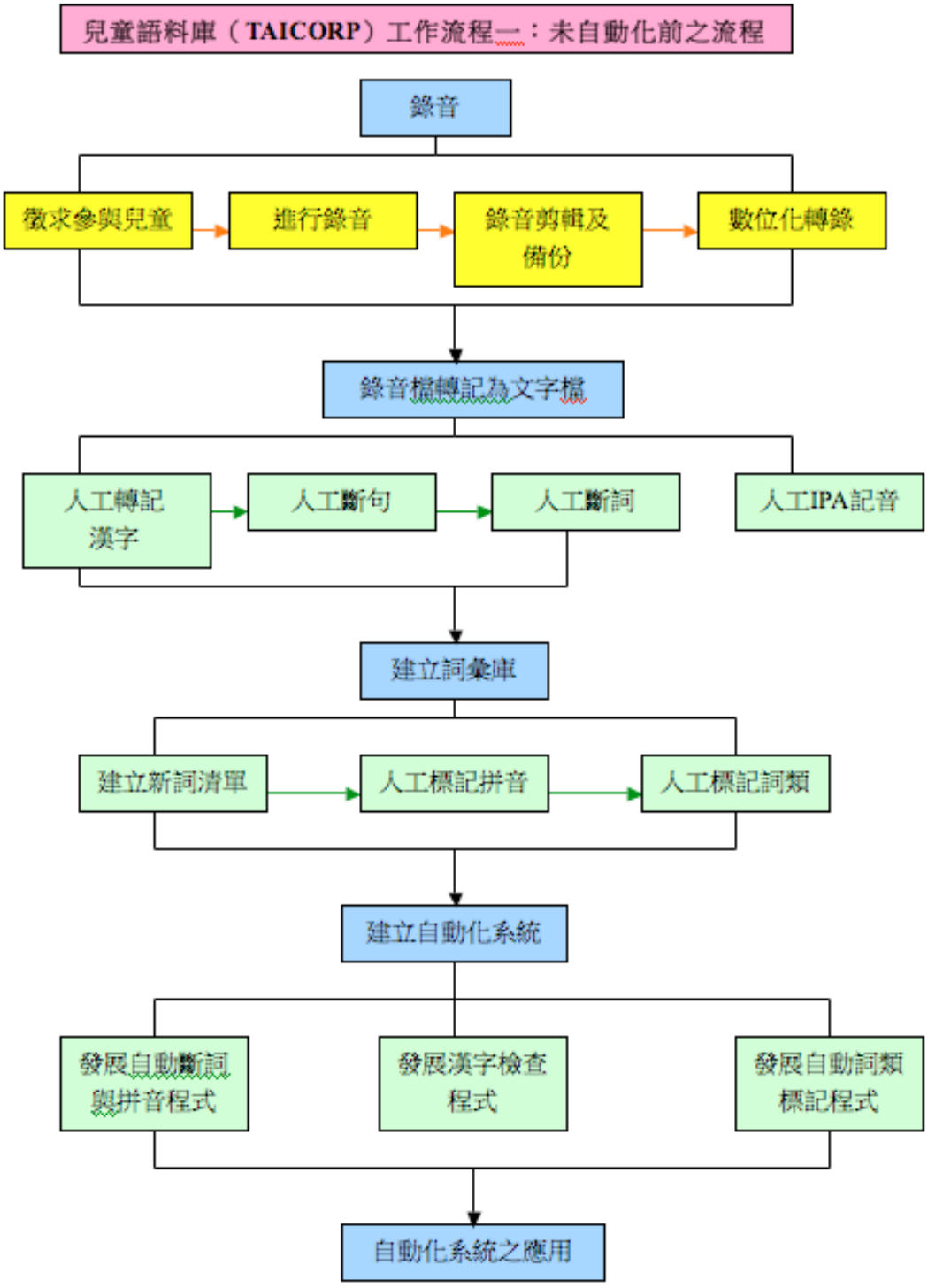
「台灣兒童語料庫」（Taiwan Child Language Corpus，簡稱TAICORP）是將所收集之台灣兒童口語錄音語料，依照世界標準的兒童語料交換系統 Child Language Data Exchange System（簡稱 CHILDES， MacWhinney and Snow 1985, MacWhinney 1995）格式，建構成語料庫。其主要目的在（1）提供國內外學者語料共享的便利性與語料分析工具；（2）藉由標準規格的設定，使台灣兒童語料的收集能更有系統、更有效率，並且快速地涵蓋台灣地區所有語言，並設立相關網站，開放國內外學者使用。

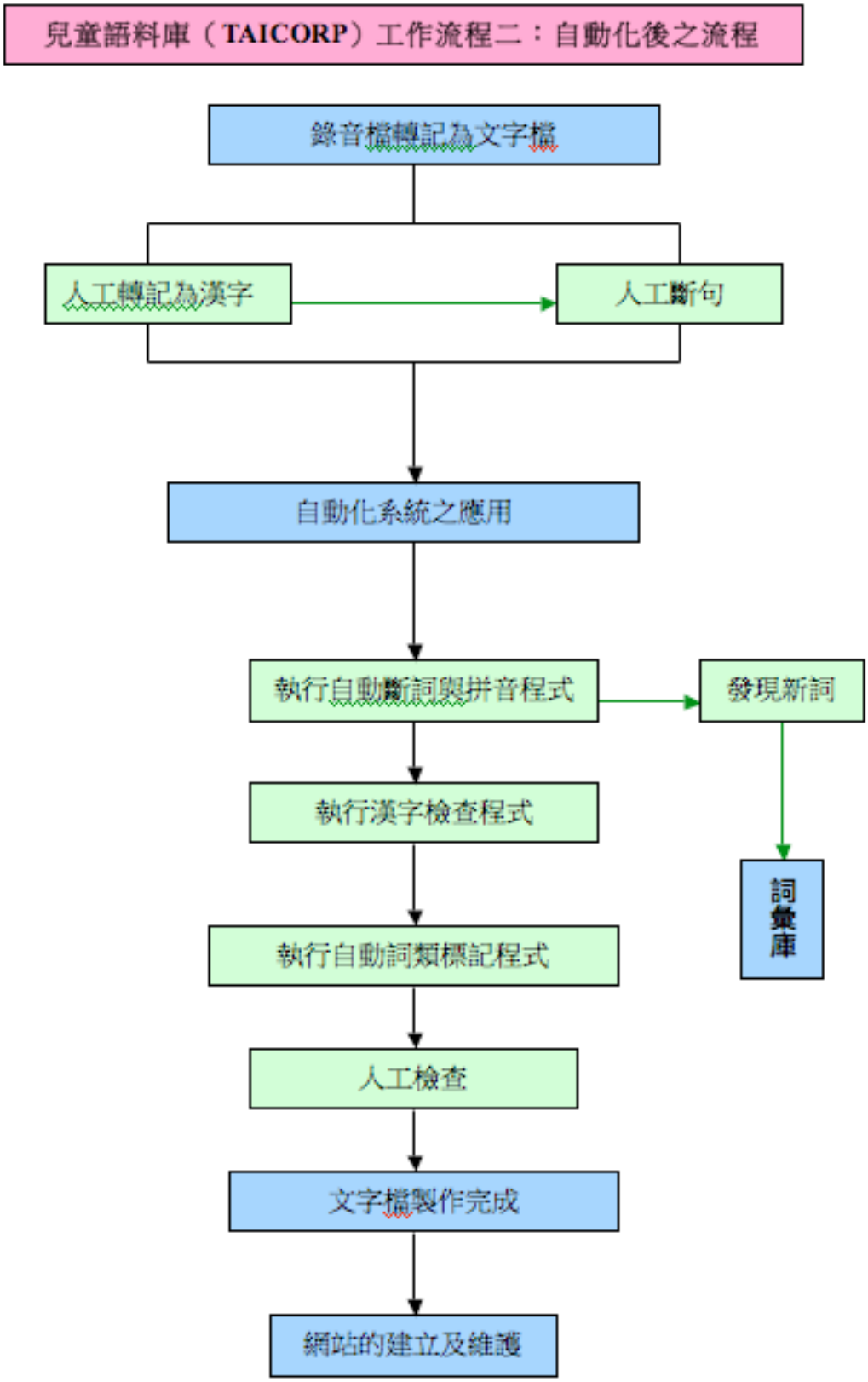
在新生代普遍使用國語的時代背景之下，台灣閩南語兒童語言所學得的語料相較之下顯得彌足珍貴。此語料庫可提供語音學、音韻學、構詞學、句法學、語意學、語用學等不同層面的語言學與兒童語言習得研究，也可提供語音工程方面的研發與應用。本計畫由國立中正大學語言學研究所蔡素娟教授主持，從1997年10月開始錄音，經轉記、標記、格式化等過程，歷時將近九年。共收錄431人次錄音檔案，錄音總長共約330 小時。文字檔共約五十萬句，一百六十多萬詞。

1. 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、自動化系統之應用等五個方面，共細分二十三項步驟依序進行，分別介紹如下。

閩南語兒童語料數位化工作流程圖如下：





（1）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」。

A. 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。簡述如下：具備語言學碩士知識背景，並以閩南語爲母語。透過每星期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。

B. 徵求閩南語家庭之兒童：目標選定中正大學附設托兒所、幼稚園及鄰近鄉鎮，徵求來自以閩南語爲母語之家庭，並年齡在一歲至三歲之間的幼兒。陸續選出共14名兒童。
(a) 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。
(b) 排定錄音時間：聯絡家長，並排定錄音時間表。

C. 進行錄音：
(a) 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。下圖左起爲迷你光碟片、專業用耳機、專業用麥克風、迷你光碟隨身錄音機。
(b) 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。
(c) 錄音期間：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。

D. 錄音剪輯及備份
(a) 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。
(b) 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。

E. 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格式）。所使用之轉錄軟體爲GoldWave Digital Audio Editor（GoldWave Inc. 研發）。

（2）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」。

A. 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，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》如下圖由左至右。轉記平台爲CHILDES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：330*10=3,300小時。

B. 人工斷句：由於台灣兒童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，將自然語言切分成個別獨立意義的句子。

C. 人工斷詞：由於目前無閩南語斷詞標準，故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。

D. 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4.5*330=1485小時。

（3）建立詞彙庫

錄音以人工轉記爲文字需耗費相當大的人力，因此最終還是要建立自動化系統以降低人力，而系統的建立則需要詞彙庫作基礎。「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」。

A. 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。

B. 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。

C. 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。

（4）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」。

A. 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。

B. 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。

C. 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，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（例：???）呈現。

（5）自動化系統之應用

A. 執行自動斷詞與拼音程式：將語句切割成詞，並標注拼音。

B. 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。

C. 執行自動詞類標記程式：標記詞類。

D. 人工檢查：檢查程式輸出檔，如欲標記的詞不只一個詞類，則檢查其自動標記是否正確。

（5）網站的建立及維護

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。

A. 網站之建立及維護：爲語料庫建立專門網站，以供全球各地學者研究之用。完成最後檢測之後，網站則開放外界瀏覽。

三、語料庫國內外比較與介紹

（一）國內—漢語平衡語料庫

「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。由於加詞類標記的漢語語料庫是史無前例的嚐試，這個語料庫是由中央研究院資訊所、語言所共同指導的詞庫小組完成的。該小組由陳克健（資訊所）、黃居仁（語言所） 兩位研究員主持，自1990年前後，便開始致力於中文語料庫的收集（Huang & Chen 1992），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（Huang 1994）。由於有了處理中文語料庫，及大量處理電子詞庫中詞條的經驗（陳克健等 1991, Chen 1994），中央研究院詞知識庫小組認爲，應有足夠的實質與人力條件來進行耗時費力的漢語平衡語料庫建構。

因此，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，1996年經計算中心設計規劃完成 WWW版，開放供各界使用，1997年開放的研究院語料庫3.0版已達到五百萬目詞的預計規模。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。

（二）國外—British National Corpus

British National Corpus（以下簡稱 BNC）爲一英語平衡語料庫，廣泛收錄 20 世紀後半的文本與口語資料，其中文本約佔九成，包含全國與地方性的報紙、各種類別的期刊、學術論文、已出版或未出版之書信與手稿…等；口語部份約佔一成，包含大量非正式的日常對談、較正式的商業與政府會議、甚至於廣播節目與聽眾來電，日常對談的部份則徵求義工錄製而成，對談內容跨各年齡層、地區與階層。

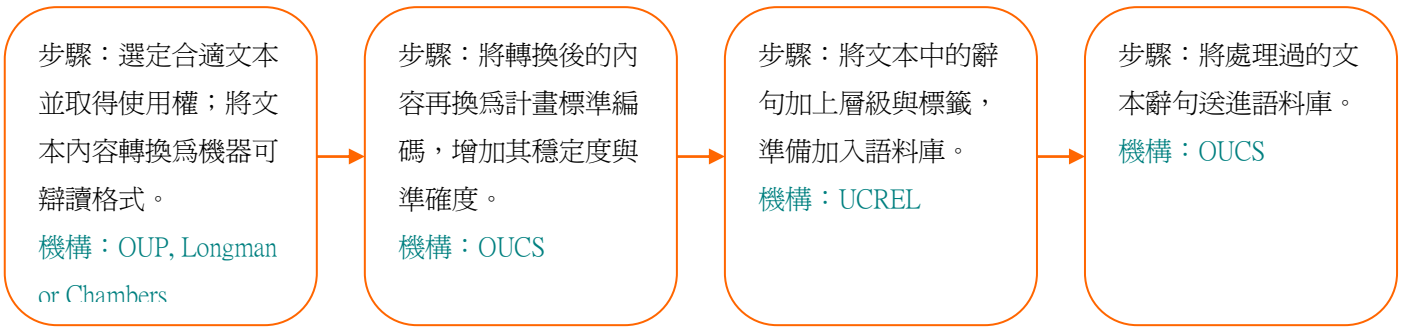
以語料庫的類型來說，BNC 爲單語語料庫，收錄以現代英式英語爲主之語料，而非歷史性之英語，內容方面則不設限，平衡並多元收錄各式不同語料。

Table 1. Composition of the BNC World Edition

Text type	Texts	Kbytes	W-units	S-units	percent
Spoken demographic	153	4206058	4.30	610563	10.08
Spoken context-governed	757	6135671	6.28	428558	7.07
All Spoken	910	10341729	10.58	1039121	17.78
Written books and periodicals	2688	78580018	80.49	4403803	72.75
Written-to-be-spoken	35	1324480	1.35	120153	1.98
Written miscellaneous	421	7373707	7.55	490016	8.09
All Written	3144	87278205	89.39	5013972	82.82

BNC 是由產學界共同組成集團運作，產業界包括牛津大學出版、朗文出版(Addison-Wesley Longman) 、樂思出版(Larousse Kingfisher Chambers)；學界則包括牛津大學計算中心(Oxford University Computing Services, 簡稱 OUCS)、蘭卡斯大學計算語言中心(University Centre for Computer Corpus Research on Language, 簡稱 UCREL)與大英圖書館研究與創新中心(British Library's Research and Innovation Centre)。

語料庫建立須經過幾個關鍵步驟，由不同單位進行，並記錄每個階段，存於 OUCS 的資料庫，建立程序如下 ：



BNC 於西元 1991 年開始建立，並於 1994 年建置完成，最初的版本於 1995 年二月發行，並提供歐洲學者研究使用。

目前 BNC 有各種不同的版本提供使用，但網站上依然開放大眾查詢，對於學術研究或是一般語言學習都相當的有幫助。

基本而言，不論是漢語平衡語料庫或是 BNC，都是以平衡抽樣試圖表現當代語言的全貌，因此包含各種類別的文體，因此對於語言研究來說，是相當重要的參考樣本。以下也順帶介紹一些典型的平衡語料庫：

（1）布朗語料庫(Brown Corpus)
於六 O 年代由 Francis 與 Kucera 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。文本選自於 1961 年美國出版的美式英文普通語體，共 15 種題材，500 個樣本，並每個樣本不少於 2000 詞，並於 1964 年完成。布朗大學並於 1961 年出版了當代英語詞頻辭典。約到七 O 年代，由 Greene 與 Rubin 設計了 TAGGIT 詞性標注系統，包含詞類標記 81 種，規則 3300 條，自動標注之準確率約 77%。

（2）LLC 口語語料庫(London-Lund Corpus of Spoken English, LLC)
由倫敦大學著名語言學家 Randolph Quirk 於 1960 年代建置之口語語料庫，包含 2000 小時之對話與廣播等口語素材，結合 1975 年 Jan Svartvik 在 Lund 大學進行的 Survey of Spoken English (SSE)，SSE 於 1981 年完成，於是建置成 London-Lund Corpus of Spoken English。LLC 其中有 87 個文本，每個文本約 5000 詞，最終規模爲五十萬詞。所採集的口語語料主要分爲五大類：面對面交談、電話交談、討論、採訪以及辯論；還有未經准許的公開評論、論證、演講；經准許的公開演講。

（3）朗文語料庫(Longman Corpus)
朗文語料委員會(Longman Corpus Commitee)建置，由 1981 年 1 月開始至 1990 年 11 月歷經九年多完成。語料選自 19 世紀末至 20 世紀初的英語語料，知識性文本(informative)佔 60%，想像性文本(imaginative)佔 40%，並廣泛橫跨十種領域，如：自然、純科學、應用科學、社會科學、國際事務…等，規模約 2,800 萬詞。

四、語料庫建置問題

語料庫建置多半是爲研究用途開發，並視相關需求建置，因此不僅在語料收集方面或是技術開發上都需嚴謹考量，才能建構出符合質與量的語料庫。

建構一個中文的平衡帶詞標記的語料庫，包括語料的收集，語料的整理（包含語料的清潔，爲語料分類，加詞類標記等等），人工的校訂。從早期的建構經驗中，由於缺乏合適的工具，因此遭遇以下困難: (1) 早期以檔案形式作爲語料的最小單位，一份檔案通常包含數十篇不同的文本，文本的格式屬性以符號配合文字在文本之前表示，這樣以檔案爲單位的架構對整體語料的管理及統計相當不便，同時對人工校對的工作分配而言，也相對失去彈性。(2) 大量的語料蒐集，維護，分類，校訂交由個人以檔案的方式處理，並無統一的處理介面，形成管理上的紊亂。(3) 過去使用自行開發的系統(Chen, Liu 92) 將語料加以斷詞標記，卻發現由於文本當中未知詞的存在，使得系統的斷詞表現大幅下降，而必須事後靠大量的人力來加以合分詞。(4) 人工校正時，由於斷詞及詞類標記時常有歧異現象發生，校

正者沒有工具立即檢驗相關的用法或範例，造成判斷上的困難，使得有時候斷詞標記的校對品質因人而異，這些問題除了造成管理上的困難之外，同時在人工校正的過程中花費大量的人力及時間，在斷詞標記的一致性上也不易維持。

五、語料庫加值應用與展望

語料庫或許對一般大眾來說是一陌生且深奧的，但事實上，諸如語言翻譯（機器翻譯）、雙語教學系統…等這些早就為普羅大眾所熟悉的名詞， 它們的共同根源都是來自於語料庫。

（一）機器翻譯

以線上即時翻譯來說，它的前身即為機器翻譯，Machine Translation (簡稱 MT)為一種電腦應用系統，可以將文章由一種自然語言翻譯成另一種自然語言。MT 並非新興技術，其構想起於 40 年代末期，由於科學家、工程學家、經濟學者、企業家…等人有閱讀大量文件或使用非母語溝通的需要，如遇此種情形，具有翻譯能力的人往往供不應求，而機器翻譯正好可以紓解這樣的供需。再者，學者專家一向有去除語言障礙能促進國際之間的合作與和平的理想，機器翻譯於焉誕生。

在這樣的構想還頗為模糊的時期，Warren Weaver(1894－1978)可謂機器翻譯的先驅，他於1947年寄給電腦控制學家Norbert Wiener的信件，以及與英國放射結晶學家Andrew Booth的對話中首先提出機器翻譯的構想，並在兩年之後撰寫了闡述相關理念的備忘錄「Translation」，並成為日後的The Weaver memorandum（1949），堪稱當時較為具體兼具代表性的文章。

機器翻譯雖然是由簡單的概念而來，但其背後的運作方式卻是相對的複雜，需透過文法、語義學、語法、片語…等分析，經拆解成符號後再重新組合。這種類型的機器翻譯需要龐大的辭彙，包含形態學、語法規則與語義資訊，但單一的機器翻譯形式並不能完全滿足需要，於是逐漸產生因應各式需求的機器翻譯形式。

1954年由美國喬治城大學與IBM合作的實驗，成功的將超過60句俄文翻譯成英文，雖然只簡單使用六種文法規則與250種字彙，無疑這也展現了機器翻譯的可行性，同時啓發了全球對於機器翻譯的興趣，尤其是當時的蘇聯。

在1953年艾森豪(Dwight D. Eisenhower, 1953-1961)上任之後，由私人翻譯Leon Dostert主導關於翻譯方面的事務，也曾經於中情局(Central Intelligence Agency)服務，他曾被邀請至喬治城大學設置語言與語言學機構，替政府訓練語言學以及翻譯相關人才。他在參加1952年於麻省理工舉辦的第一次MT會議之後，由原本對機器翻譯的存疑轉為熱衷，積極想實現與展現機器翻譯的可能性，他找來舊

識、同時也是IBM的創辦人Thomas J. Watson一同展開跨機構合作。基於政治因素，實驗展示以俄翻英爲主，但只運用了六種規則、250個字彙與有限的句型，並以IBM原提供美國國防部使用的IBM701系列電腦進行運算。

IBM701的辨讀方式是靠讀卡機，所謂的卡片上有80個欄位，可用欄位共72個，需先鍵入並儲存於中間磁鼓記憶體(intermediate drum storage)才能辨讀。實驗展示由一位對俄文並不熟悉的女性操作員以英文字體鍵入”Mi pyeryedayem mislyi posryedstvom ryechi.”，電腦經60000次的運算處理過後以打字方式輸出“We transmit thoughts by means of speech.”， 接著她又鍵入一連串的字彙“Vyelyichyina ugla opryedyelyayatsya otnoshyenyiyem dlyini dugi k radyiusu.”輸出的結果則是“Magnitude of angle is determined by the relation of length of arc to radius.”

屏除實驗中途曾有兩次當機之外，此次的實驗展示可以說相當成功的表現了機器翻譯的可能性，也引起當時媒體爭相報導，試想只要靠著這樣的機器，便可以將自己完全不熟識的語言轉換成自己的語言，對於一般大眾這也無疑是劃時代的創舉。

對於俄國人來說，這樣的展示也頗具威脅性，於是也從史達林(Joseph Stalin, 1879-1953)死後開始進行機器翻譯的實驗，並於1956年初展示相關成果，系統依循IBM-Georgetown的模式。喬治城大學則在1956年初獲得一筆國家科學基金，展開大規模的俄翻英研發，並組織了超過20位研究人員，1957年研究人員由原先的二大組轉換成自由競爭的方式而細分爲四個不同項目，藉以延伸各種不同的研究方式可能性。但這些研究在1964年ALPAC出現之後則進入機器翻譯的黑暗時期。

後續的十年有許多不同的政府機構與學術團體致力於MT的研究與開發，如IBM替美國空軍完成的俄翻英系統。其他的學術團體如麻省理工、哈佛大學、柏克萊大學…等則致力於理論研究，也開發出早期的人工國際語言與轉換系統(e.g. MIT與Cambridge Language Research Unit, CLRU)。

但1964年由美國政府贊助的機構ALPAC(Automatic Language Processing Advisory Committee)卻於1966年撰寫的一份報告中扼殺了MT的發展，報告中指出MT並未能正確有效的翻譯，與人工翻譯相較之下成本爲其二倍，並沒有迫切發展的需要。自此之後MT於美國發展趨緩，而加拿大、歐洲等地則因當地的語系較繁雜逐漸產生需要，與當初美國針對俄文與技術層面的發展不盡相同。

後期的MT則逐漸全球化，在80年代左右則有跨國合作的商用MT系統出現，如Systran這樣的電腦翻譯軟體也廣泛的被國際組織與企業採用。在80年這樣的市場熱潮領導了人們對於MT的一些省思與注意，無論是結合人工智慧與新的語言學

理論，MT的前景令人期待，但最終都是希望能提供人際溝通之間便利的工具，與文化之間的融合了解。

（二）機器翻譯—以 SYSTRAN 爲例

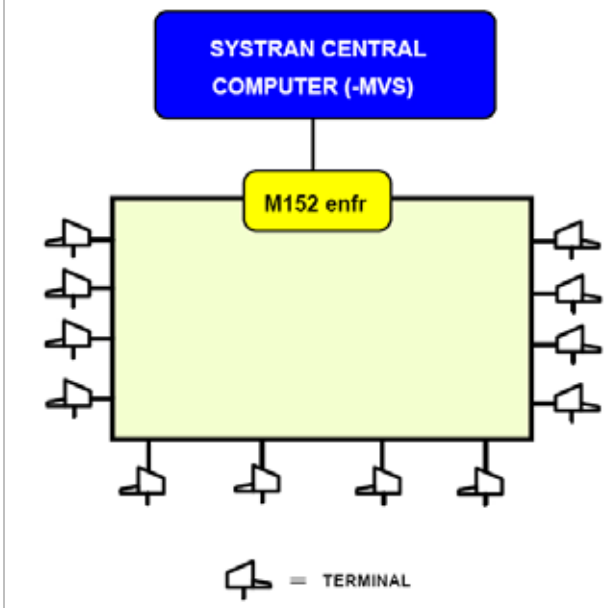
科學昌明不僅僅帶來工業上的進步，也逐漸的帶全球化，甚至演變成全球本土化的趨勢，各式文本的流通與人際傳播對於無暇聘請專業翻譯人才的機構、企業來說成爲課題，於是使用由機器翻譯進化的翻譯軟體就成了一種迅速的解決方案。

SYSTRAN可說是翻譯軟體的先驅，由匈牙利裔科學家Peter Toma發明，他於50年代末期移居加州，並於La Jolla成立了SYSTRAN，取自於System Translation的簡稱。公司成立之後曾在1969年爲美國空軍研發俄譯英的系統，並陸續爲美國國家空情局(US National Air Intelligence Center)研發出數套西歐語系的系統，並於南斯拉夫內亂時替美國政府研發出第一套塞爾維亞-克魯埃西亞文對英文(Serbo-Croation-to-English)的系統。

而SYSTRAN的專利技術也不止用於美國，繼1974到75年美國太空總署的阿波羅聯盟號計畫(Apollo-Soyuz Test Project)俄英系統之後，也爲未來歐盟所使用的系統奠定了良好的基礎，由最初的英譯法原型之後，陸續提供各種歐系語言的系統，至今已有17種翻譯系統於歐盟與其他歐洲相關機構使用。

以歐盟所使用的系統來說，當時仍需設置終端機與用戶端做連結，尚未進化爲線上軟體，用戶只須將需要翻譯的文件透過電子郵件的形式寄到一特殊信箱M152，經辨識需求的代碼之後，用戶會在數分鐘或半小時之內不等的時間內，同樣以電

子郵件的方式收到翻譯的文件。(圖一)



圖一：M152翻譯系統

在1992年SYSTRAN開始轉移技術至個人電腦，至1997年則發行了配合微軟視窗的專業版，並積極與企業合作，如：支援SEIKO的攜帶型翻譯機，或是與SONY合作研發線上遊戲軟體平臺，許多跨國企業也因爲整合的需要而使用SYSTRAN的翻譯系統，以免人工翻譯耗費時間，爭取更快速的運作。

列舉SYSTRAN的大型客戶如下：

- Bentley	- Mercedes-Benz
- Bombardier	- NEC(Japan)
- Chemical Abstract	- Phillip Morris
- Cisco	- Saint-Gobain
- Ford	- Sony
- France Telecom	- Toyota
- O.C.E.D.	- Dassault
- Daimler Chrysler	

在90年代網際網路逐漸起步之後，SYSTRAN讓一些逐漸擴張的網路社群意識到機器翻譯能夠在增強網路的功能與相容性，於是如當時較大型的入口網站Alta Vista就加入線上翻譯的服務，稱爲Babelfish，其他的入口網站如Lycos, Wanadoo, Free, Yahoo!, Google之後也陸續加入，而目前則以Google以獨特運算方式勝出而使用者眾。在資訊爆炸的時代，人們不須再花費龐長的時間學習異國語言，雖說人工翻譯的確有周全性，但翻譯軟體所能提供的迅速也確實是可以肯定的。

（二）學習者語料庫

最早的學習者語料庫是八 O 年代末期所建立的朗曼學習者語料庫(Longman Learners’ Corpus)。九 O 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，現存有超過二百萬詞，存有十四種不同母語背景的英文學習者語料，此外，香港科技大學也建置了類似的學習者語料庫 The HKUST(Hong Kong University of Science and Technology) Corpus of Learner English 。現代學習者語料庫常與學習者中間語(inter language)分

析連結並做比對，將學習者語言看成是一種規則系統並普遍存在於學習者之間。

以台灣爲例，國立成功大學的外國語文學系也建置了「成大英語學習網站及網路英檢系統計畫」，從 2006 年起爲提升全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。此類課程也將成爲成大之特色，對於校內學生以及修習校內課程之國際學生也是一項福音。

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「前瞻性數位語言學習中心」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數位英語學習教材。

根據該計劃的內容，利用最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準備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。

六、結論

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真實的，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、紀錄及資料表，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。

本文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、與正規化,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。雖然本文所述之議題可能已超過實際的技術問題,但透過南島語語料庫數位化計畫的嘗試，相信對於未來語料庫的研究將有著極深刻的影響。