拾肆、LDC 所發行的中文語料庫以及 Sketch Engine 語料庫檢索

收集大量的語料曠日廢時，且牽涉版權問題，如果加上標記語料所花的人力和物力更為可觀，美國賓州大學的 Linguistic Data Constorium (LDC)經常發行各式語料，包括 Chinese Gigaword、中文新聞和句法樹庫以及語音的語料庫。這些語料庫有些已經標注詞性、語法結構、和語義角色等訊息的，購買 LDC 的語料庫，可以大幅減少語料庫收集和開發的時間。

另一個減少語料庫開發時間的方法是使用 Sketch Engine 的服務，透過繳交個人年費約 1 千 2 百元，可以檢索 Sketch Engine 裡面 10 幾億繁體和簡體中文語料庫的內容，也可以上傳不超過 50 萬詞的語料，由 Sketch Engine 來自動分詞，建立索引，並產生關鍵詞和搭配語檢索程式。

拾伍、 結論與建議

語言的研究傳統上是純人文的研究，1950 年代後期電腦發明之後，計算語言學及自然語言處理技術這兩門新興學科誕生，語言的研究開始與科技有密切的關係。1980 年代起由於電腦軟硬體技術的突飛猛進與價格快速下降及個人電腦的日漸普及，電腦輔助語言教學系統逐漸普及。1990 年代中期以後網路興起，網頁及機讀資料的普遍使得語料庫語言學這門新興學科快速興起。語料庫，資訊檢索，

計算語言學，數位學習這幾門新興學科與英語教學形成一個橫跨語言，教育，及科技的新興研究專題。

語料庫及語言科技的重要性可以從下列事實可以看出來。英國牛津大學於1990 年代初整合英國數個研究機構發展出一億詞英文的英國國家語料庫(BNC)及檢索的工具，其它包括朗文出版社，牛津大學出版社，劍橋大學出版社，Collin出版社與伯明罕大學合作的Collins Cobuild Project，及倫敦大學為了辭典編纂學及英文文法的研究，也紛紛建置大型語料庫。我國語料庫的建設首推由中研院詞庫小組陳克健教授及黃居仁教授在1990 年代初期開始的漢語平衡語料庫及之後的句法樹庫，這些資源奠定台灣在語料庫語言學厚實的基礎，並培養了相當多年輕的語料庫語言學家。美國雖然在1960 年代即有Brown Corpus，但在英國發展了英國國家語料庫十多年之後，也開始美國國家語料庫的建設。除了英國，台灣，美國，歐洲國家，日本，大陸，幾乎世界各國都努力建置大型語料庫。

語言科技與語料庫息息相關，語言機率模型的建立需要大量語料。語料庫及語言科技相關研究早已經成為世界級學術重鎮與資訊大廠鎖定發展的重點科技。史丹福大學，柏克萊大學，麻省理工學院，卡耐基美崙大學，約翰霍普金斯大學，哥倫比亞大學，馬理蘭大學，劍橋大學，愛丁堡大學，東京大學，京都大學，北京大學無一不設有自然語言處理及語言科技的相關學程及大型研究計畫。除了學界之外，IBM 及AT&T 對於自然語言處理及語言科技的研究已經累積數十年的經驗。微軟在華盛頓州西雅圖總部及北京微軟研究院都有多組研究人員從事與語言科技產品的開發。Google 則於數年前開始網羅自然語言處理的學界菁英。語言科技對人文教育的影響最直接的一個例子就是主辦並設計托福考試的ETS 前幾年已經利用語言科技開發出英文作文自動評分系統，由於實驗顯示這套系統與專家的評分高度一致，ETS 於幾年前已將兩個專家評分減為一人評分，另一個由電腦系統取代，如果兩者差距大於某一級距才由第二位專家評分。語言科技對於國

防科技也息息相關，美國國防部先進研究計畫總署 DARPA 每年均有大筆經費支持語言科技的研究作為戰略及反恐情報收集與分析。此外，美國國家標準局 NIST 幾年前開始舉辦語言科技相關技術的競賽和評比，凡此均足以證明語言科技已經無遠弗屆，甚至無所不在。語料庫是語言科技的基礎，語言科技是語料庫的應用，兩者密切相關，缺一不可。

1988 年中研院中研院資訊所陳克健研究員及語言所黃居仁研究員成立詞知識庫小組，並規劃我國大型語料庫的建立。24 年來在他們的努力下奠定了我國大型語料庫的發展的基礎，目前除了已經完成具有詞性標記的 1 千萬平衡語料庫及數萬句的中文句法樹庫之外，還有 8 萬目詞具有語法訊息的詞庫，並且擴充了 HowNet 的語義，此外在中文分詞、詞性標記、句法剖析、語義分析的技術也打下了堅實的基礎。為了進一步促進語言科技相關產業，發揮最大的綜效，我們建議整合國科會、經濟部、教育部的經費與資源，結合產官學的力量，以語料庫語言科技為核心，研究並開發以下相關技術 1．具有自然處理技術的文本及網頁的資訊及知識擷取系統，能辯識文章裡面所包含的人、事、時、地、物資訊，並能理解文章中某些特定的語意。2. 中英及中日雙向的大型機器翻譯系統 3.能夠對於中英文作文自動偵錯及評分的系統 4.中文及多語辭典的半自動編纂系統,如半自動編纂搭配語。5.能夠自動回答問題的問答系統。這些系統所整合的技術有助於帶領相關產業開發出數百億產值的市場，開拓知識經濟新的科技服務業。