庫博中文語料庫分析工具的數位人文價值

闕河嘉*

數位時代下電子媒介的力量無所不在，舉凡電子文本的普及、資料庫的開發、電子傳播資訊容易獲取等特色，如何快速有效地分析大量的電子形式文本是發展「庫博中文獨立語料庫分析工具」（此後簡稱「庫博」，CORPRO）的動機。猶記 2016 年蔡英文總統就職演說結束後，網路上紛紛立即出現記者們對蔡總統就職演說的分析報導。譬如，就職演說中詞彙頻次的統計以及用詞文字雲；詞彙的頻次反映出年輕人、經濟、年金改革等受到新政府重視；關注演說中涉及兩岸關係的詞彙，並引述相關句子評論蔡總統的立場。這個案例不僅說明分析文本詞彙使用特色的重要，對於其他領域的延伸，能夠獨立操作分析各類型文本的工具將有助於個人理解社會文化現象。由於篇幅限制不在此詳述庫博的功能，本文是分享筆者以庫博從事數位人文研究和教學的心得。

庫博是一款特別針對中文使用的特殊語境，以語料庫語言學（corpus linguistics）為基礎的電腦輔助文本分析軟體工具。庫博能針對大量文字文本資料，進行詞頻統計分析、關鍵詞脈絡索引，詞彙之間的共現關係、和其他參照的文本比較出顯著常用與低度使用的詞彙等功能[1]。撰寫這些功能對於任何自然語言實驗室的學生都是基本能力，但是對於人文社會領域的學者或學生卻是無法獨自處理的。特別是當面對大量文本時，這個資工實驗室的雕蟲小技卻極可能帶給社會人文研究重要的突破。《紅樓夢》研究中的經典大哉問「後四十回究竟是曹雪芹的原稿，還是高鶚的續書」就是一個著名的例子。新紅學大師胡適以考證學認為後四十回是高鶚偽托，蔡元培認為《紅樓夢》和曹雪芹的文化內涵、精神關係以及時代背景、思想內容、文化特徵密不可分，因此反對以考證學斷定高鶚的偽托。白先勇則從寫作者的觀點及經驗，認為《紅樓夢》劇情前後貫徹和人物語調一致的程度，補書有其高難度，是以後四十回是曹雪芹的原

* 國立臺灣大學生物產業傳播暨發展學系副教授

[1] 有興趣者可參閱筆者的專文，闕河嘉、陳光華（2016）。

稿，高鶚應該只是參與修補的工作。相較於傳統的研究方法，當代學者以語料庫語言學研究方法則從文字風格角度出發，比較前八十回和後四十回兩個語料庫，發掘了前後文風的許多差異特徵，補充說明後四十回不是曹雪芹原稿的證據。

早期的語料庫語言學研究方法之應用在於語言學研究，包括了英語為第二語言的教學研究。學生英文學習時往往受到原來母語的影響，因此研究者可以藉由分析相當數量的學生英文作文，找出某特定文化背景學生的英文表達特色，進而提出有效的教學法。同時，也有已經建置好的英文語料庫，內含了大量的好的英文材料，譬如經典文學、新聞及雜誌，學生可以輸入關鍵詞，從內建文本中搜尋出包含該詞彙的所有句子、段落、甚至是文章，藉由觀察這些句子的使用脈絡瞭解到詞彙的正確用法。有趣的是，當今華文當道之時，語料庫研究反而回過頭來應用在分析中文學習的文本，想必在未來也將蓬勃發展。

近幾年，語料庫語言學研究方法應用擴展到英語系的人文社會研究領域，由英國的蘭開斯特大學語言學和英語系的麥克內瑞（Tony McEnery）和貝克（Paul Baker）主導[2]。語料庫語言學開始應用到人文社會研究，主要是因為有許多跨國且涵蓋時間廣的英文語料庫的產生，譬如英國國家語料庫（The British National Corpus, BNC）、美國開放國家語料庫（The Open American National Corpus, OANC）等。這些語料庫中網羅了各式新聞報導，成為研究語料庫的蒐集來源，研究者得以進行某概念在這些新聞中的展現分析，包括比較報紙之間的立場、歷時性分析、傳達的意識型態……等。其他研究領域有觀光研究、公衛、社工等，而研究議題有分析大眾媒體如何傳達觀光目的地的意象給消費者、美國主流新聞報導北韓的意識型態、癌症病患在臉書支持社群表現的性別差異、英國新聞媒體在難民報導的態度、媒體中氣候變遷、基改食品、核能開發爭議等議題。

語料庫語言學成為社會科學的研究方法並非沒有社會學科理論依據的。符號學研究把社會文化視為一連串符號（sign）的組成，並認為任何的符號都是由符徵（signifier）和符指（signified）所組成。符指是符徵的意義內涵，而符徵是符指的載具，兩者是符號的一體兩面。符指存在於社會文化中的文本中，而符號學研究就是藉由分析文本瞭解其中賦予某符號的意義，針對特定的文本進行系統性分析得以揭示此特定文本蘊含某概念的意義、反映的價值觀、甚至進而批判其中可能傳達的意識型態。例如筆者過去兩年的研究曾蒐集了主流新聞報紙、獨立新聞媒體在過去 12 年間論及「基改」和「有機」這兩個新興農業概念的

[2] 麥克內瑞和貝克所主持的 ESRC Centre for Corpus Approaches to Social Science（CASS）為蘭開斯特大學極為重要的研究項目之一，http://cass.lancs.ac.uk/。

報導，分析這些報導傳達對基改的立場，以及「有機」的新聞報導所隱藏的基本框架，甚至也可以探究文本反應出的社會迷思。

現階段社會科學的應用語料庫語言學研究往往結合批判論述分析（critical discourse analysis），在提出研究見解時佐以語料庫研究得到的相對「客觀」數據。批判論述分析需要研究者反覆仔細檢閱研究資料，以漸漸形塑出研究者的見解。但由於需要精讀資料，研究僅能限縮欲仔細檢閱的文本量，因此提出的研究論點容易被批判過於「主觀」或所謂的「採櫻桃謬誤」（cherry picking），專挑對自己有利的內容進行研究。雖然現在檢索系統可以「全文檢索」網羅了周延的研究文本，避免了文本代表性的問題，但所得到的文本量往往超越研究者所能負擔。因此有必要藉由語料庫研究方法，探究大量文本整體所呈現的特徵，譬如詞彙和詞彙之間的緊密關係程度，來彌補批判論述分析研究方法「客觀」數據資料不足的根本限制。而且語料庫分析工具的關鍵詞脈絡索引功能也可以條列出研究文本中某詞彙的所有上下文，明確指出關鍵詞所在位置，研究者能以這些位置為中心擴展閱讀關鍵詞的上下文脈絡，更有效地檢視詞彙在脈絡中的意義。

語料庫語言學分析方法可能提供新興研究領域的出現。以語料庫分析為研究方法的研究模式大致區分為語料庫為本（corpus-based）和語料庫導向（corpus-driven）兩類研究思考。前者著重在針對研究主題文獻既有的論點、假說、問題，以語料庫此研究方法所獲得的資料和發現進行對話；後者則不做任何預設或假說，純粹以發掘語料庫本身所展現的特徵為主，說明研究對象或研究主題可能有的意義發現或浮現的新興議題。一般而言，單純的語料庫導向（corpus-driven）的研究案例可見於資工學者跨界人文學科研究的論文。在沒有學科背景的條件下，對於有價值的史料文本，盡情地提出史料中的有趣模式或特徵，或者偏好文本分析技術開發的議題，如官吏的遷移史分析。人文社會科學界雖然也有偏向單純語料庫導向的研究，如文學家鄭文惠[3]教授針對文學中的顏色（紅、白、藍色）探勘可能隱含的意義，但這樣的研究旨趣仍和原學科領域發展密切相關，因此研究發現對於原學科領域社群提供了有意義的價值。前述紅學的經典疑問的數位人文研究是以語料庫為本的研究（corpus-based）案例。然人文社會學科的研究不太可能僅單一採用語料庫導向（corpus-driven）的研究方法，除了針對研究文本進行試探性探勘，大部分的研究皆是混和著使用。

目前臺灣數位人文研究的範疇似乎偏重於技術導向的分析和發現。數位人文研究根本上可以說是電腦工程技術和人文社會學科研究的結合。目前稍具平

[3] 鄭文惠（2016）。

衡關係的發展是以合作團隊的方式，資工學者和社會人文學家作為夥伴關係，一方探究研究議題的發現，另一方執行文本探勘的實際操作。但要跨越兩個差異甚大的學門條件下，共同合作傾全力投入發表學術影響力高的國際期刊論文非常困難，畢竟文本探勘的基本功能對資工學界來說是微不足道的能力，自然無法產生具備資工價值的論文。在此論文發表掛帥的年代，能有兩學界雙贏的研究誠屬不易。資工學者進行人文學研究，大抵是從技術導向的學科領域學者決定人文學科研究題目，詮釋研究發現的意義和價值。就如此界定研究意義的事實而言，數位人文學科領域儼然成為了資工學界擴張的領域。現在，數位人文研究的趨勢強調基於既有的數位典藏基礎的加值應用研究，或以「觀測數位經濟創新之走向」在人文學科教學研究中運用大數據及數位科技工具[4]。我們似乎可以嗅出當今臺灣「數位人文」研究和教學是以經濟價值為導向的發展[5]。

筆者極力呼籲「人文精神」（humanism）在數位人文研究應有的地位。狹義上大學裡人文學科的領域涵蓋了人類學、哲學、宗教、文學、藝術、歷史等，廣義上則包含了社會學科、傳播學、政治學、教育學、法律學等。然而，我們不能忘記人文學（humanities）的根本精神在於以觀察、分析及理性批判來探討人類情感、道德和理智，理解、反思社會文化現象並提出回饋社會的觀點。於是，以創新加值或知識經濟為導向的數位人文研究或教學，並不是人文主義或人道主義精神。以數位工具處理數位化材料看似人文學科的研究，以數位工具導入人文社會科學教學創新，培育具備邏輯思考、問題解決與實作能力之跨領域創新人才看似人文學科的一環，但是這些研究不足以稱為具備有人文思想的普遍價值。

李歐梵在接受訪談時以中國五四運動為例說明「人文精神」。當時的知識分子扮演啟蒙者的角色，救助大眾、教育大眾、或是用古典人文主義來訓練學生。這樣的人文精神包含了人文主義和人道主義，對知識分子而言是一種絕對的普世價值，也是知識分子應有的啟蒙責任。中國的知識分子自覺他的啟蒙任務就是為了一個廣義的國家、民族、社會，站在一個中心地位，影響社會。正因為知識分子的言論和國家、民族、社會密切相關，其在中國社會文化裡始終有著舉足輕重的地位。而知識分子的研究根本上就是一種啟蒙式人文主義的體現。因此，以知識分子人文主義精神來看，數位人文研究不應限縮為一種以文學或歷史學科為材料的人文學科研究而已，更應發展為具有知識分子角色的人

[4] 見國立政治大學數位人文團隊，「教育部數位人文創新人才培育計畫」計畫簡介，http://www.dhcreate.nccu.edu.tw/about.html。

[5] 法鼓山投入佛學經典研究非關經濟收益，對於社會科學研究的價值貢獻更為廣泛深刻。

文學科研究。這要看領導數位人文發展的主政者是否認同中國社會知識分子這般角色在數位人文學科發展的應有地位與價值。

固然電腦工程技術和數位文本的便利帶給人文社會學研究語料庫語言學時的研究方法，但是不同學科之間隔行如隔山，人文社會學科的知識累積仍需要有國學知識、社會科學理論作為基礎。再提白先勇討論《紅樓夢》作者爭議作為例子，他不苟同學者認為後四十回的文字功夫、藝術價值不及前八十回，甚至後四十回有幾處感人的文采比前八十回還猶有過之。白先勇在反對胡適認為高鶚偽書之際，和胡適讚賞後四十回的悲劇下場安排的觀點卻是一致的。這說明了要詮釋「客觀」的詞彙使用分析仍需有深度的學科背景知識。數位人文研究發展仍以穩固原社會科學知識領域為主體，而以數位分析工具為輔。

對於任何分析中文文本的語料庫分析工具，如何以正確的詞彙判讀文本（即斷詞技術）仍是根本關鍵的問題。這也是中文的語料庫語言學為方法的數位人文研究遲至近幾年才出現的原因。研究者曾研究新聞媒體字現「基因改造」食品和作物的報導，其中瘦肉精、萊克多巴胺、校園午餐搞非基運動……等專有名詞都非任何既有詞庫內含的詞彙。如果軟體工具沒有提供研究者修正詞彙的機會，即使電腦計算能力再怎麼高強，研究分析呈現的數據對於社會科學研究學者仍無參考價值。

近幾年原為英文語料庫的套裝軟體 Antconc，WordSmith 也發展出可以識別中文碼的能力，基本上仍是套用原來英文分析的軟體，外加一些輔助軟體協助提升中文斷詞正確性。既有英文語料庫分析工具發展來分析中文文本的方式，大抵在軟體中增加了「補充詞典」的功能。庫博語料庫的分析工具提供的文本分析功能和 WordSmith 或 Antconc 大致一樣，不同之處在於庫博比較著重於修正詞彙功能便利性的設計上。中文詞彙的複雜度遠非英文軟體開發者所能想像到的，除了研究領域的專有名詞之外，還有人名、同義卻不同文（如台北市、臺北市）的詞彙。為求提升斷詞的正確性，庫博的詞彙功能修正採取辭典校正和研究分析結果互動觀察的設計。這個互動特色的詞典建構功能只能由使用者親自操作體會了。

庫博的設計出發點為打造出一款能讓人文社會學科背景的個人（研究者）方便且獨立操作的工具，可以盡情地「玩」（play）一群文本資料，重複試驗分析其中詞彙的可能關係。研究者能藉由「玩」研究相關的語料庫，發現對於研究問題有意義的資料。就前述《紅樓夢》的大哉問，筆者在臺大參與教授的「數位人文概論」課程中，即有學生進一步比較了《紅樓夢》後四十回和高鶚的《白馬嘯西風》文風展現的差異程度，發現兩者極為相似，因而推論《紅樓夢》後四十回可

能為高鶚代筆。同門課程中還有學生藉由庫博的詞彙分析功能分析 2015 到 2017 年之間臺灣四大報紙電子新聞文本，瞭解媒體如何再現《一例一休》新法案之社會大眾回應[6]，以及其他許多由學生期末報告改寫、投稿研討會並成功接受發表的論文[7]。為期一學期的課程教授即能有如此豐碩的研究成果產出，庫博之便利使用與容易上手可見一斑。

最後很重要的是，庫博文本分析工具或語料庫語言學終究是研究方法的一部分，研究者仍應該回歸研究的基本精神，確認語料庫研究是否為回答研究問題的最適方法。固然庫博這類語料庫分析工具能協助研究者瞭解研究文本語料庫的特徵，但是獲得文本特徵的方法沒有一定的步驟，因此需要研究者良好的研究設計。在研究語料庫中發現有意思的特徵需要研究者把玩庫博文本分析工具，至於如何玩得出神入化則端看研究者的巧思，充分運用限定的功能以玩出有意義的發現。

參考文獻

王章逸、闕河嘉（2016）。〈大埔之歌：臺灣主流報紙中的「土地徵收」〉，發表於「第七屆數位典藏與數位人文國際研討會」，臺北：臺灣大學，2016 年 12 月 1 日至 3 日。

吳孟家、葉克芸、陳品臻、簡翊淇（2017）。〈以語料庫分析取徑探究社會對《一例一休》新法案之態度：以 2015-2017 年臺灣四大報紙電子新聞為例〉，發表於「第八屆數位典藏與數位人文國際研討會」，臺北：政治大學，2017 年 11 月 29 日至 12 月 1 日。

高竹瑩、葉亦辰、黃志揚、林荷鎧（2016）。〈「移工」一詞在臺灣報刊雜誌之興起〉，發表於「第七屆數位典藏與數位人文國際研討會」，臺北：臺灣大學，2016 年 12 月 1 日至 3 日。

國立政治大學數位人文團隊，「教育部數位人文創新人才培育計畫」計畫簡介。取自 http://www.dhcreate.nccu.edu.tw/about.html

郭柏傑、闕河嘉（2016）。〈臺灣獨立媒體中的基改食品〉，發表於「第七屆數位典藏與數位人文國際研討會」，臺北：臺灣大學，2016 年 12 月 1 日至 3 日。

鄭文惠（2016）。〈情感現象學與色彩政治學：中唐詩歌白色抒情系譜的數位人文研究〉，收錄於《數位人文：在過去、現在和未來之間》（ISBN：978-986-350-198-5），臺北：臺大出版中心，2016 年 12 月。

闕河嘉、陳光華（2016）。〈庫博中文獨立語料庫分析工具之開發與應用〉，收錄於《數位人文：在過去、現在和未來之間》（ISBN：978-986-350-198-5），臺北：臺大出版中心，2016 年 12 月。

羅盤針、鄭碩、江安淇、曾博揚（2016）。〈以語料庫分析取徑探究臺灣新聞中的跨性別：以聯合知識庫為例〉，發表於「第七屆數位典藏與數位人文國際研討會」，臺北：臺灣大學，2016 年 12 月 1 日至 3 日。

[6] 吳孟家、葉克芸、陳品臻、簡翊淇（2017）。〈以語料庫分析取徑探究社會對《一例一休》新法案之態度:以 2015-2017 年臺灣四大報紙電子新聞為例〉，發表於「第八屆數位典藏與數位人文國際研討會」。

[7] 羅盤針、鄭碩、江安淇、曾博揚（2016）；王章逸、闕河嘉（2016）；郭柏傑、闕河嘉（2016）；高竹瑩、葉亦辰、黃志揚、林荷鎧（2016）。