客家語語料庫簡介與進化

謝杰雄

「語料庫」是非常重要的資源，對語言保存與教學、翻譯、數位應用等層面也是非常必需的資源寶藏，世界各種主要語言幾乎都有語料庫，世界上最大的語言資料庫，可以說是 GOOGLE，打開 google 翻譯系統，裡面就有 141 種語言，包括毛利文、苗文都有。客家語長期以來一直沒有一個綜合型、專業型、平衡型甚至是學習者語料庫都沒有，對客語的保存、分析與教學、現代化等方面帶來許多不便，也由於語料不易蒐集，加上正確度偏頗，對所分析的結果，常常出現許多例外或偏差，甚至不少研究者，為符合其所預期的理論或結果，常常有研究者自省製造句子的情形。客家語早年有私人建置的《山哈客語語料庫》和《政大客語語料庫》，但是規模都非常小，加上語料取得不易，客語用字又多次變動，也是維護和建置不易的主因，目前該兩種小型語料庫，均已經關閉。

鑒於科技的成熟，各種語料庫技術和演算法的進步，客家委員會自 2017 年 12 月委託國立政治大學團隊建置《臺灣客語語料庫》，也是目前台灣地區最大、唯一的客語語料庫，屬國家型語料庫，於 2022 年 10 月中旬正式上線。網址：https://corpus.hakka.gov.tw/

本語料庫是以書面語料、口語語料為主的語料庫，書面語料達 600 萬字、口語語料達 40 萬字，涵括四縣、海陸、大埔、饒平、詔安、南四縣六種腔調，語料蒐集涵蓋 1990 年代迄今之臺灣客語書口文本，並由受過教育訓練之工作人員進行資料清

<table>
<tr><td></td><td>理，包含用字轉寫校訂、轉寫標記加註、斷詞標記標示。</td></tr>
<tr><td>(00:49:45-00:49:46) 男： 麼个…
(00:49:47-00:50:46) 男： 毋係喔…該大家都有…佢毋單淨講下定定…毋係講吾阿嫂个嫁妝…就
(00:50:16-00:50:46) 男： 愛喊佢合<CS-ja>でんちく</CS-ja
(00:50:23-00:50:24) 男： 麼个…
(00:50:30-00:51:14) 男： 係啊係啊…毋係…佢這兜就…佢嫁除了客家莊…該央時大家就常勤儉肚囥等在該唱山歌…在該聽山歌…非等著這個安到討親…該無話講…
(00:51:16-00:57:27) 男： 對哦…係哦…無毋著…所以這隻…</td><td>口語語料另提供音訊播放功能，供使用者依語輪點選時間戳記聽取音訊內容。</td></tr>
<tr><td>後設資訊
語料名稱 106年全國語文競賽客家語朗讀文章
單元名稱 幸福个味緒
出版年份 2017
腔調 四縣腔
文類 散文
主題 生活
載體 網路文字資料
客家菜/N ，/PU 對/P 好嘴斗/VS 个/GE 佢/PN 來/VA 講/VA 毋單只/阿公/N 帶/N 佢兜/PN 老嫩/N 大細/VS ，/PU 去/VA 客家/N 餐廳/N</td><td>書面語料檢索關鍵詞之後可以顯示該筆語料的後設資訊，包括語料來源、出版年份、所屬腔調、文題、主題等等，也可顯示分詞之後的詞性標記。</td></tr>
<tr><td>飛 天頂 -2 5.731%
天頂 飛 1 5.731%
天頂 口頭 2 3.91%
像 天頂 -1 2.969%</td><td>檢索書面語料可以顯示與關鍵詞的共現詞和共現值，並依據共現值高低排序，對語言教學和編製客語教材有一定的參考價值。</td></tr>
</table>

《臺灣客語語料庫》的建置，代表客語保存、研究、數位化應用等進入另一個里程碑，但是不可避免的是，就一個現代化語料庫而言，《臺灣客語語料庫》還有許多要努力的空間。

首先是語料來源的平衡性問題，基本上《臺灣客語語料庫》現有語料以散文、小說、故事語料、俗諺語等為主，缺乏如科學、經濟、運動、政治等等專門性語料，這是由於客語書寫者一般均不會接觸這樣的議題所致；其次是語料量的問題，這個語料庫雖標榜有600 萬字語量，然而包含六腔，平均而言每個腔調是 100 萬左右，對現代語料庫而言，嚴格言之這是一個小型，甚至是一個微型語料庫，還有許多待努力的空間，主要還是因為客家語沒有純客語書寫的雜誌、報紙，甚至與其他語言，如英語、華語、日語對應的翻譯語料，可資比對和研究。三是學習者語料庫缺乏，所謂學習者語料庫是把學習者的學習狀況所呈現的語料蒐集起來，建置的語料庫，可以分析學習者的準確度和偏誤類型，做為編製教材和教師改善教學的參考，過去要做這種學習者學習呈現分析，多半只能透過現場觀察和分析極少量的學習者文本，以致分析結果並不準確，建置客語的學習者語料庫是一個有待努力的方向。四是專業性語料庫也值得努力，比如法律、經濟、政治、海洋等專題性語料庫，就現階段而言也可以建置山歌主題、俗諺語主題、客家產業之類等專業性語料庫，對教學和研究會很有幫助。

無論如何，今天有《臺灣客語語料庫》的建置，是一個劃時代的大事，這裡僅就〈客家語語料庫與教學、應用〉做一些簡介和說明，與期待下一階段的持續和努力。