

揭開中文詞彙的神祕面紗～ 語料庫科技在語文教學的應用（V）



圖片來源：臺灣華語文語料庫網頁

【語文教育及編譯研究中心副研究員 吳鑑城】

語料庫，顧名思義，是「語料」的「庫藏」。這個「寶庫」裡可能收藏著來自不同文本的語言素材，涵蓋範圍廣泛，包括（但不限於）古籍、現代小說、新聞文章、學術論文，甚至社群媒體上留言的書面語語料；也可以是保存著來自各種情境所產出的自然對話、演講、訪談，講課等語音（或其逐字稿）的口語語料。語料庫所涵蓋的龐大真實使用情境語言材料，蘊含著各種真實使用情境下的語言表達，反映了語言使用的多元面向。

近年來，大家耳熟能詳的大語言模型，如生成式預訓練變換模型（Generative Pre-trained Transformer, GPT）系列，正是通過深度學習技術在大量的語料庫上訓練而成。語料庫作為「教材」，模型從中學習語言的模式、規則，並將這些知識轉化為人機互動、語言生成的實用性技能，使模型能夠預測、生成符合語法結構的文本，並在文本中理解上下文的關聯性。

國家教育研究院所建置的臺灣華語文語料庫（Corpus of Contemporary Taiwanese Mandarin, COCT）收錄了書面語、口語、華英雙語及華語中介語等各類語料。其中，正體中文的書面語語料截至 111 年底已有約 4 億 4,401 萬字，且為了便於使用者檢視及分析語料，更以英國蘭開斯特大學（Lancaster University）所研發的 CQPweb 為基礎，建置了國教院語料庫索引典（後稱系統，見圖一）讓使用者可進行靈活的查詢和分析，並藉由搭配中文語料庫，深入挖掘各種詞彙現象。

圖 1、國教院語料庫索引典

選單	COCT 書面語語料庫2020
語料庫查詢	標準查詢
標準查詢	查詢模式： 簡易查詢（不區分大小寫） Simple query language syntax 每頁的查詢結果： 1000 Match strategy： Standard 限制（檢索範圍）： None (search whole corpus) 開始查詢 重設查詢
限制查詢	
單詞查詢	
詞類列表	
關鍵詞	
分析語料庫	
分析語料庫	
Saved query data	
查詢歷史	
儲存查詢結果	
分類查詢結果	
上傳查詢指令	
建立/編輯子語料庫	

除了可直接查詢目標詞外，系統提供了多樣的強大查詢方式，當我們想觀察中文詞綴(affix)現象，例如中文常見的前綴「阿」，只需輸入「阿+」，就可以獲得所有「阿」開頭的詞語出現的例句（如圖二），還可進一步透過系統內建統計分析功能，取得各個詞語出現的頻率跟比例（如圖三），讓我們立刻可以瞭解前綴「阿」常組成像「阿嬤」、「阿姨」等親謂稱呼，也會用於名稱之中，如「阿里」、「阿拉伯」、「阿福」。

圖 2、檢索「阿+」取得所有含有前綴「阿」所組成語詞的句子

Solution 1 to 100		Page 1 / 1,930
今花蓮縣光復〕一帶的阿眉住屋，跟A萊	阿眉	的一樣。穀物不像高山蕃放在屋內，
連聲稱讚阿笛是個好孩子。這時，老公公對	阿笛	說：「孩子，竹子可以砍了，快去
好吃極了，看得阿欽直往肚子裡吞口水。	阿欽	的結論是，小孩子不准吃雞屁股，可能是
》也有一份特別的依賴：一九六三年，當	阿拉巴馬州	州長華里士（George Wallace，正試圖以擋住校門的
」阿努比斯說：「要尊敬並敬畏牠。」	阿穆特	顯然在睡覺時聽見有人叫牠的名字。
衣裳，趕到河灣去了。阿大來到河灣上，	阿二	也趕到了。圓圓的月亮，高高掛在天
外交關係協會在全球企求的特定輿論氣候。」	阿達憲	（Ken Adachi）也說：「絕大多數美國人所認為
大岩石外「墓地」哭喊：我兒啊！	阿爸	斷腸、阿母心碎。你們疼痛嗎？恐懼嗎？
摺疊椅）；一個角落擺了兩張長方形桌，	阿蘿	（Alo）占據一桌，奇亞歐（Kyaaro）在
爭來爭去時，膽小的鮎川竟大膽跟	阿房	私奔。增田因不死心而磨磨菇菇，結果被捕
何相干？不走就算了！」於是，	阿凱	兄弟就隱匿在里漏部落外的樹林內。
道：「蜂王蜂王！我想請你上天去喊	阿方	和阿珍快回來，不知行不行？」蜂王隨口

圖 3、有前綴「阿」所組成語詞的分析情形

No.	Query result	No. of occurrences	Percent
1	阿嬤	5709	2.96%
2	阿姨	5324	2.76%
3	阿公	4089	2.12%
4	阿里	2941	1.52%
5	阿拉伯	2688	1.39%
6	阿媽	2176	1.13%
7	阿福	2002	1.04%
8	阿爸	1882	0.98%

除了能夠分析語詞的結構，系統還能協助探索語詞之間的關係。以量詞「座」為例，透過系統的搭配詞功能，我們能夠迅速查找常與「座」一同出現在句子中的其他詞彙，如數詞「一」、指示代詞「這」、「那」，以及名詞「山」、「城市」、「橋」等（見圖四）。進一步深入分析這些搭配詞，有助於揭示有關「座」更多語言現象。例如，名詞「山」、「城市」、「橋」等顯示了「座」常與地理元素和建築物相關聯。除了前述的分享之外，系統還具有許多功能，歡迎大家共同來探索。

圖 4、「座」的搭配詞資訊

No.	Word	Total no. in whole corpus	Expected collocate frequency	Observed collocate frequency	In no. of texts	Log-likelihood
1	—	3,701,685	3,470.606	23,063	14801	49,603.592
2	山	54,936	51.507	3,292	2187	21,122.664
3	這	2,114,032	1,982.063	10,258	6619	17,439.645
4	城市	32,019	30.020	1,699	1250	10,474.064
5	橋	12,169	11.409	1,216	807	9,073.921
6	整	105,785	99.181	1,636	1438	6,128.334
7	那	754,138	707.061	3,637	2649	6,093.701
8	城	42,831	40.157	970	727	4,341.473

中文語料庫的存在，不僅讓語言學家能夠更系統地研究語言的變化、規律和演變，再透過強大的索引典，我們更得以窺探中文詞彙在不同時期、不同語境下的變化，從而更深入地理解中文的豐富性。無論是對於語言學者還是中文學習者而言，這樣的探索都將是一場豐富而有趣的冒險！

資料來源

林慶隆、林崇熙、白明弘、吳欣儒、連育仁（2022）。**華語文教育課程指引研發與語料庫應用推廣_111 年計畫期末報告**。國家教育研究院研究計畫成果報告（編號：NAER-2022-012-C-3-4-C1-02）。新北市：國家教育研究院。連結網址 <https://www.grb.gov.tw/search/planDetail?id=14223081>

林慶隆、柯華蕓、吳鑑城、白明弘、陳茹玲（2019）。**《建置應用語料庫及標準體系》期末研究報告**。國家教育研究院研究計畫成果報告（編號：NAER-107-12-F-1-01-00-1-11）。新北市：國家教育研究院。連結網址 <https://www.grb.gov.tw/search/planDetail?id=12562546>