

不同類型的字詞「差很大」嗎？解析各類文本的字詞

【語文教育及編譯研究中心副研究員 李詩敏】

英國著名的企鵝分級閱讀讀本（Penguin Readers）這套書是依據廣泛閱讀（又名分級閱讀）精神編寫的英語課外讀物，每一本書都有對應的 CEFR（The Common European Framework of Reference，歐洲語言共同參考架構）等級標注，而每個等級各有對應的詞彙數量，編寫時需要嚴格控制每本書的詞彙數。

比較臺灣課外讀物的分級，政府單位、各大圖書館或坊間出版社每年公布中小學優良課外圖書推介，廣播電臺或媒體相關基金會也定期發布優質兒少節目推薦，這些圖書和媒體節目的推薦清單大多附有適合閱讀或閱聽年齡層的資訊。如果以小學國語教科書的字詞作為比較基準，上述圖書、媒體和國語教科書為不同類型的文本資料，彼此之間的字詞是否存在著差異呢？學了小學國語教科書的字詞後，可以辨識圖書、媒體、課外讀物（以上圖書加媒體資料）字詞的百分比是多少呢？以下我們從這些實際文本資料字詞之間的覆蓋率來探索語言資料背後表達的數字與意義。

首先，比較的文本類型請參考表 1。以「九年一貫國語文教科書」為例，小學三個版本共 18 冊的課本內容（包含標題、作者簡介、課文、注釋、語譯、賞析、題解等內容）：所有的國字、語詞的出現次數，加總的數量共 210,325 字（即「總字（頻）數」）、137,100 詞（即「總詞（頻）數」），不同的國字、語詞，分別共有 3,400 個字（即「相異字數」）、16,943 個詞（即「相異詞數」）。

表 1

文本類型的內容

文本 教科書	九年一貫 國語文教科書 (18 冊)	十二年國教 國語文教科書 (12 冊)	圖書	媒體	課外讀物
相異字數	3,400	2,156	4,448	3,626	4,701
相異詞數	16,943	7,208	30,241	20,220	38,703
總字（頻）數	210,325	77,605	619,663	805,794	1,425,457
總詞（頻）數	137,100	48,638	395,841	536,097	931,938

由於表 1 教科書語料總字數和其他類型語料的總字數差距過大，直接比較各類文本和教科書的字詞差異，容易造成教科書和其他文本的字詞數量差異很大、交集率很低的錯誤印象，所以我們採用計算教科書字詞對其他文本字詞的覆蓋率，不僅可以分析認識教科書字詞以後可以辨識其他文本字詞的百分比，也才能說明教科書字詞對其他文本字詞的涵蓋程度的高低以及彼此之間的字詞差距。

表 2、表 3 分別是教科書國字、語詞對各種文本國字、語詞的覆蓋率，標灰底的空格是教科書對不同階段的各類文本的最高覆蓋率。從表 2，我們可以知道，學生學了第一階段教科書國字之後，不分教科書課綱或版本，大致可以辨識第一階段圖書 78% 以上的國字，如果學習累積到第二或第三階段教科書，學生約可辨識圖書國字分別達 90%、96% 以上。整體而言，不論累積教科書國字學習數量到哪個階段，對三類文本國字的覆蓋率由高到低排序都是：課外讀物、圖書、媒體；大約到小學畢業，學生可以辨識這三類文本的百分比分別約達到 97%、96%、87%。

表 2

教科書國字對各類文本國字的覆蓋率

單位：%

教科書 \ 文本		第一階段			第一至二階段			第一至三階段		
		圖書	媒體	課外 讀物	圖書	媒體	課外 讀物	圖書	媒體	課外 讀物
九年 一貫	A 版本	78.88	69.26	82.76	92.17	83.79	94.24	96.87	87.80	97.51
	B 版本	81.24	71.61	84.96	91.71	82.60	93.66	96.71	86.91	97.41
	C 版本	80.23	71.38	83.57	91.73	83.64	93.80	96.96	87.55	97.73
十二 年國 教	A 版本	80.93	71.13	84.59	90.59	81.71	92.77			
	B 版本	80.16	70.19	83.78	90.66	81.96	92.63			
	C 版本	80.45	68.95	84.53	91.27	81.82	93.59			

表 3 則是教科書語詞對各類文本語詞的覆蓋率，整體而言，不論累積教科書語詞學習數量到哪個階段，對三類文本語詞的覆蓋率由高到低排序都是：媒體、課外讀物、圖書；大約到小學畢業，學生可以辨識這三類文本的百分比分別約是 87%、84%、80%。

表 3

教科書語詞對各類文本語詞的覆蓋率

單位：%

教科書 文本		第一階段			第一至二階段			第一至三階段		
		圖書	媒體	課外 讀物	圖書	媒體	課外 讀物	圖書	媒體	課外 讀物
九年 一貫	A 版本	63.98	69.26	67.79	74.96	83.79	80.35	80.98	87.80	84.90
	B 版本	66.19	71.61	70.11	73.97	82.60	79.24	79.74	86.91	83.87
	C 版本	65.84	71.38	69.84	74.01	83.64	79.89	79.95	87.55	84.33
十二 年國 教	A 版本	65.84	71.13	69.66	73.58	81.71	78.54			
	B 版本	64.64	70.19	68.55	73.58	81.96	78.70			
	C 版本	63.71	68.95	67.50	72.97	81.82	78.38			

從表 2 和表 3 的同版本教科書字詞或是不同版本教科書字詞對於累積到同一個學習階段的字詞的覆蓋率顯示，教科書各版本選取的課文內容或是不同課綱的字詞規範，這兩者對於教科書字詞對同一類文本的覆蓋率影響不大，反而是不同類型文本資料之間的覆蓋率差異較大。即使是課外讀物的相異字詞數量或所有字詞數量都遠多於圖書、媒體，但教科書字詞對課外讀物字詞的覆蓋率都不是最低的，顯見字詞數量多寡並非影響覆蓋率的最主要因素，文本類型才是影響覆蓋率高低差距的最重要因素。以國字而言，常用國字約 3,000 字，教科書文本有書面語及口語文本，因此教科書字詞對混有書面語及口語的媒體覆蓋率較高；就語詞而言，因為中文國字與國字即可組成語詞，不同類型文本影響了語詞的類別與常用度，因此語詞數量、內容以及文本類型影響了教科書語詞的覆蓋率，教科書語詞對語詞數量較少的媒體文本的覆蓋率較高。

針對文本資料類型不同而形成彼此之間字詞「差很大」的情形，我們建議在解析文本資料時，應該要盡量考量文本文類、文體、語式、場域、主題、媒體的平衡性，並採用字詞的頻率、交集率或覆蓋率等數值，了解字詞在文本的出現次數、被重視度以及常用度，才能讓編制產出的字詞表或比較的數值具有科學性及參考價值。

資料來源

李詩敏、蔡旻穎（2023）。小學國語文課外讀物與教材常用字詞分析比較研究。國家教育研究院整合型計畫案成果報告（NAER-2022-022-C-1-1-F1-01）。新北市：國家教育研究院。連結網址：<https://www.grb.gov.tw/search/planDetail?id=14901468>