藏在語言中的祕密：詞彙計量研究對語言教學的啟發

【語文教育及編譯研究中心副研究員　白明弘】

齊夫定律揭示詞彙使用的不均衡現象

語言學著名的齊夫定律 (Zipf's law) 描述了詞彙在真實語言使用情境中極度不均衡的現象。藉由大量文章的統計，我們得以統計出詞彙在實際語言使用的頻次表。將頻次表依出現次數由高而低排列，就可以觀察到齊夫定律的現象（如表 1）：排名序位第 20 名的詞的頻次大約是排名第 10 名的一半；而第 100 名的詞大約是第 10 名的 1/10。齊夫定律的通則就是詞頻和序位的倒數成正比（頻率 $\propto 1/$序位）。

從另一個角度來說明，齊夫定律反應了詞彙使用極度不均衡的現象：語料庫中 99% 的頻次是由最高頻的 1% 詞彙所貢獻。我們實際統計約 10 億頻次 (tokens) 的美國 COCA 語料庫，其中大約由數十萬個相異英文詞 (types) 所構成，但這數十萬詞大部分的出現頻率都極低。最高頻的前 100 詞大約就貢獻了 5 億頻次（將近 50%）。講白一點就是：只要您認識了最高頻的 100 個英文詞，就能在一般英文文章中辨認出 50% 的詞。這個驚人的不均衡現象揭示了現代語言教學中非常重要的一個概念：只要掌握最高頻的 1% 詞彙，就能讀懂文章中 99% 的詞。由此我們可以得到一個重要的啟示，語言教學應由高頻的詞彙開始學習，方能得到最佳的學習效率。

表 1、從 COCA 語料庫統計的詞彙頻次表，依出現頻次由大而小排列

序位	詞彙	頻次
1	the	47,644,615
2	be	40,310,332
3	to	24,310,041
4	a	23,738,154
5	and	23,605,940
6	of	22,467,586
7	in	15,738,070
8	I	13,977,003

9	that	12,785,937
10	have	11,714,476
……		
18	with	6,130,249
19	this	5,290,903
20	as	5,229,087
……		
98	two	1,085,199
99	first	1,085,167
100	even	1,080,589

從詞頻表到教學詞表的建置

從齊夫定律的發現，我們確立了教材的編排應該從最常用的詞彙開始，再逐漸提升詞彙的難度。然而，如果我們實際觀察詞頻表會發現，最高頻的詞以功能性的詞彙（冠詞、介詞、連詞等）居多，例如：the, be, to, a, and，它們通常沒有明確的語義內容，用於表示語法功能、連接語句或在語言結構中扮演特定的功能。對學生來說，這些詞的學習困難度較高，並不適合在初級時安排太多，而應依常用度及困難度適當的將它們安排在不同等級的教材中。所以，詞頻表通常不適合直接應用在教材的編撰上。

教學詞表則是專門為教材編撰所設計的詞表，因此，它的安排及設計必須考量到實際教學的需求。例如：依據語言能力的分級訂定各級應學習的詞彙。其中，在詞彙的順序安排上，除了依據詞頻表之外，還需將難度較高的功能詞分散到不同的能力級別。這些調整工作必須仰賴經驗豐富的語言教育專家來進行。

教材難易度的安排與評估

有了分級詞表之後，教材編撰就能依據教學大綱及教學詞表逐步提高難度級數。透過嚴格控管的情境、詞彙、文法及審慎使用的圖案，教材編撰的成果可以趨近於所設定的能力目標。然而，在編輯過程中，人工必須分心於掌握情境、詞彙、文法等要素，對編輯與審查而言都是極耗費心力的工作。透過良好的教材編輯工具，可以協助

編輯者掌握目前所使用各等級詞彙的數量，自動評估目前教材的能力等級是否超過所設定的級數等。同樣的，教材的審核者也需要類似的分析工具，才能掌握教材的內容是否和課綱所設定的能力等級相符。

結語

齊夫定律揭示了詞彙使用的不均衡現象—即少數高頻詞貢獻了大部分語言使用情境中的詞彙頻次。這提示我們在語言教學中應該優先學習高頻詞彙，以提高學習效率。而教學詞表的建置正是實踐這個目標的有效參考資料。編輯者依據教學詞表的分級，循序漸進將詞彙安排到教材的內容中，就能使教材的詞彙等級控制在合理的範圍內。最後還需要有良好的教材編輯工具輔助，使教材編輯者能專注於內容，以更有效率的方式，發展出高品質的語言學習教材。

資料來源

白明弘（2022）。**子計畫五：應用自然語言處理技術於英語教科書之分析比較研究**。國家教育研究院整合型計畫案期中報告（NAER-2022-018-C-1-1-F2-05）。新北市：國家教育研究院。