

啥！電腦也會挑單字？口語語料庫對英語學習的啟示

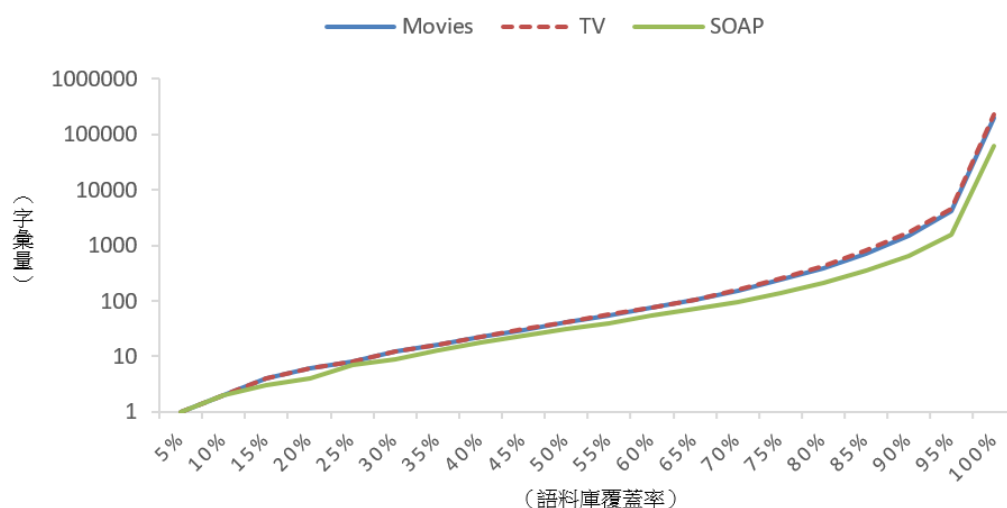
【語文教育及編譯研究中心助理研究員 周一銘】

茫茫字海，找到對的那一個，像極了愛情

英語，是當今國際往來最重要的通用語（*lingua franca*），亦是作為世界公民不可或缺的技能。字彙能力則是培養英語「聽、說、讀、寫、譯」等能力之關鍵，學習者的字彙量與語言能力關係密切。一般而言，一個語言的語法規則數量有限，但字詞量龐大。有關研究亦指出，英語系國家成年母語人士字彙量約 2-3 萬。語言教育學者普遍認為，優先學習常用字彙有助於學習者提高語言學習效率。隨著資訊科技的發展，語料庫（Corpus）巨量資料結合自然語言處理技術（NLP）成為建置字彙表的利器。篇章語料經過數位化、斷句分詞、詞形還原（*lemmatization*）、標記（*tag*）等流程後，藉由機器演算法，可獲得詞頻、標準化頻率、覆蓋率、分布率和相互資訊（*Mutual Information*）等數據。

透過語料庫，我們觀察到一個重要的事實，即每個字詞的覆蓋率並不一致，甚至可以說十分懸殊。如圖 1 所示，想覆蓋 86-92% 的語料庫內容僅需要詞頻最高的前 1,000 字左右，但要達到 95-98% 覆蓋率則需要 5,000 字。也就是，額外多學的 4,000 字僅能提高覆蓋率 6-9%。故，我們建議初學者應把有限的時間、心力放在覆蓋率較佳的字彙上。同時，在規劃課程時，應把語料庫的客觀證據作為我們安排字彙內容數量和時程先後的重要參考。

圖 1、語料庫覆蓋率與字彙量閾值



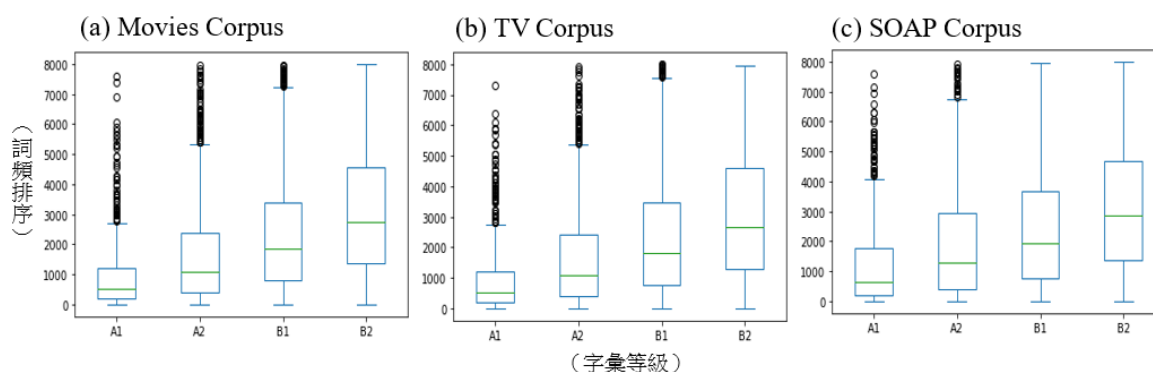
「斯斯」有兩種，語言形式也有兩種

語式 (mode)，是傳遞思想訊息的載體形態，主要可分為口語和書面形式，兩者使用的場域和風格有所差異，例如：口語常選用流行語 (buzzword)、俚語、歇後語、嘆詞、短語動詞 (phrasal verb) 等，讓言談交際生動活潑、淺顯易懂，而書面語則少用或不用。另外，口語使用動詞和連接詞的頻率較書面語高。因此，針對言談交際不同目的、功能或形式，蒐集相應的語料 (如書面語、口語、學術、或者中介語)，建置語料庫並加以分析，就顯得格外重要。

看點一：欸！別忘了「分級」

並不是所有高頻詞都適合入門初學者！圖 2 表示歐洲共同語言架構 (CEFR) 各等級字彙在口語語料庫詞頻排序的分布情形。詞頻排序越前面，使用頻率越高、越常使用。箱型圖盒狀底端表示第 25 個百分位數，頂端表示第 75 個百分位數，而綠線則表示中位數。整體而言，字彙難度越高，對應的詞頻排序中位數就越高，也可以說字越簡單、使用頻率越高，如入門級 (A1) 字彙約有 75% 的詞頻排序分布在前 1,000，初級 (A2) 字彙則是前 2,500。值得注意的是，各種難度字彙的詞頻排序為漸進式分布。以排序前 3,000 高頻詞為例，除入門 (A1) 和初級 (A2) 字彙外，尚且涵蓋中級 (B1) 或中高級 (B2) 以上字彙。不同頻帶 (frequency band) 也往往包含不同難易等級字彙，因此有必要進一步細究字彙的難易度和適用性，以符合不同程度學習者的需要。

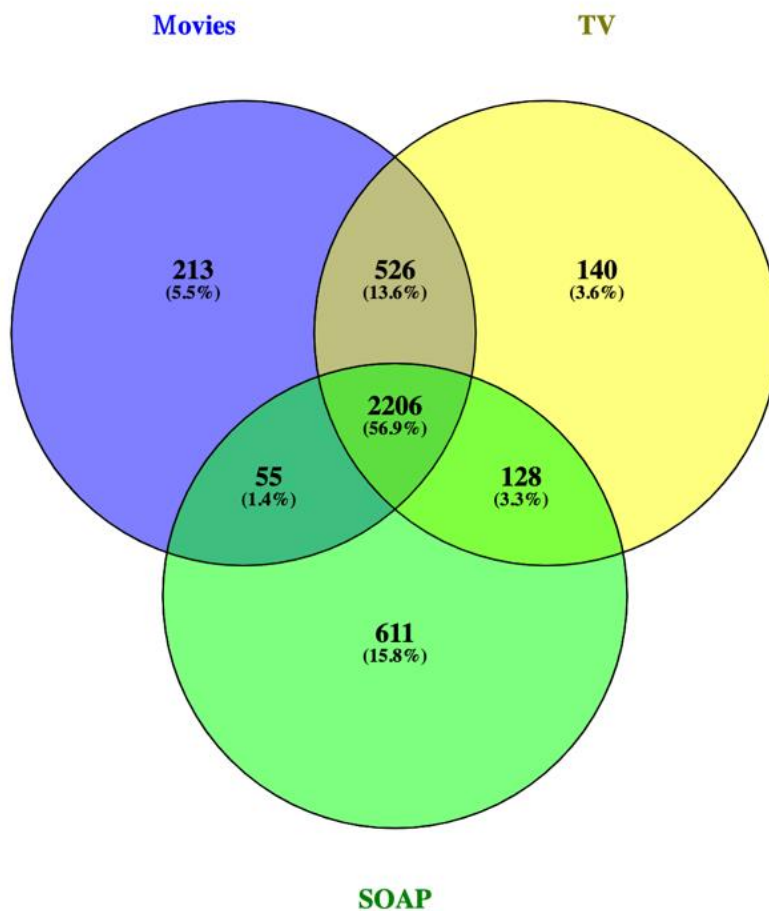
圖 2、口語語料庫中各等級字彙詞頻排序分布概況



看點二：震撼彈！沒有頭腦、沒有女朋友、還沒有...

大家可能會很好奇，108 課綱字表的「口語性」如何呢？沒有收錄哪些口語高頻詞呢？首先，三個語料庫所收錄的字彙總量分別約 3.25 億（TV）、2 億（Movies）和 1 億（SOAP），現行課綱參考字表（2,000 字）對三個口語語料庫覆蓋率約 84-87%。其次，若取各語料庫詞頻前 3,000 的字彙交集，可得到 2,206 個高頻口語字彙（見圖 3）。比對這些字彙和課綱參考字表差異情形，包含 brain, girlfriend, exactly, instead, amazing, anymore, anyway, normal 等是語料庫獨有的入門級或初級高頻口語字彙，可作為未來課綱收詞參考。不過，字表建置是門專業嚴謹的學問，仍須考量年齡經驗和文化差異等因素，並結合經驗數據（實證研究、學習時數、學生程度）和語料庫字彙統計資料，方能畢其功於一役。

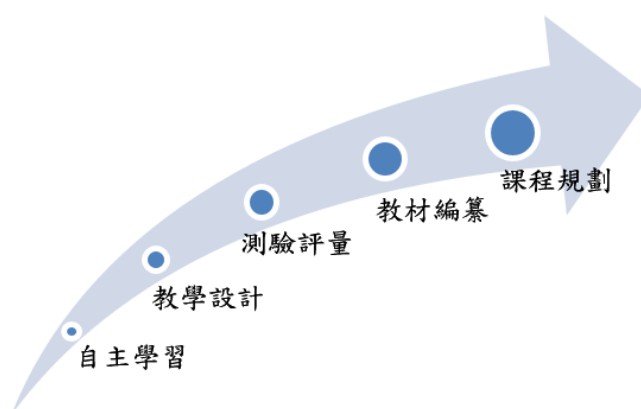
圖 3、口語語料庫高頻詞交集情形（註：非按比例繪製）



有一種溝通，叫 in English

學習語言以溝通為目的，其中又以口語表達最為自然直接。而口語高頻字彙取得，可以透過口語語料庫協助。口語語料庫收錄了語言使用者真實的言談內容，分析語料庫中常見字彙的詞頻、意義、用法、出現條件和搭配情形，可以讓英語字表收詞更貼近實際溝通及言談交際所需。藉由語料庫客觀數據，可以作為課程規劃、教學設計和教材編纂等重要參考，讓字彙學習發揮最大的效益（見圖 4）。

圖 4、口語語料庫應用示意圖



資料來源

周一銘（執行中）。現有英語口語語料庫與課綱 1,200 字彙分析比較研究。國家教育研究院計畫成果報告（NAER-2021-10-C-1-1-B6-03）。新北市：國家教育研究院。連結網址：<https://rh.naer.edu.tw/handle/w4587>