

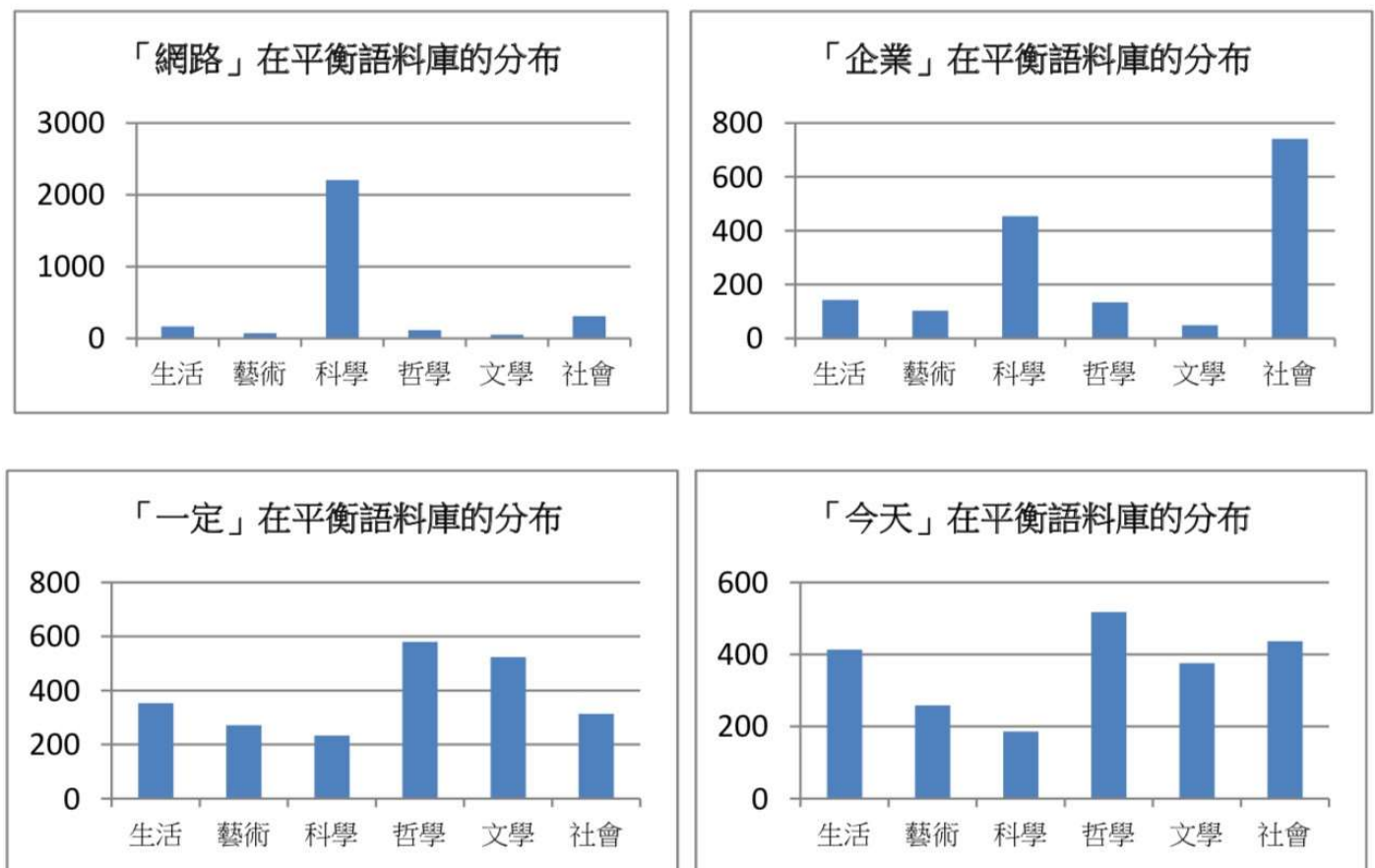


圖 1 四個詞頻接近的詞語在平衡語料庫不同主題中的分布情形（單位：次數/ 每百萬詞）

【語文教育及編譯研究中心助理研究員 吳鑑城】

核心詞彙是指一組不受文本類型、主題、應用情境等影響，穩定使用的詞彙。核心詞彙相對於非核心詞彙（邊緣詞彙）來說數量較稀少，卻構成溝通內容的主要部份。在語言的使用上，當一個句子缺乏邊緣詞彙時，雖難以確切指稱物品，但仍足以傳達說話者的主要意涵，因此核心詞彙是語言學習中重要的一環。

傳統的核心詞彙選擇方法主要依據專家知識與經驗法則，語料庫語言學興起後，統計式的方法逐漸取代經驗法則。然而單純使用詞頻無法分離核心詞彙與邊緣詞彙，例如，從中研院平衡語料庫中觀察四個詞頻接近的詞在不同主題中的分布情況（見圖一），「網路」只有在科學主題之下才大量出現。「企業」則在社會及科學主題中大量出現。相對而言「今天」和「一定」在各類主題中出現的次數較為平均。在此例中，前兩個詞語屬邊緣詞彙，後兩個詞語屬核心詞彙。由此例的觀察可以發現詞語的核心程度與分布均勻度有高度的相關，因此許多研究者提出以分布均勻度來衡量詞語的核心程度。而計算詞語分布均勻度前必須先將語料庫切分成數個區塊，然後再計算詞語在區塊中的分布是否均勻。在這樣的計算程序中，每個區塊代表一個語言使用情境的實例，當分布均勻度越高時，即表示詞語受情境的影響越小。因此語料區塊的切分方法將關係到核心詞彙選擇的結果。

本研究嘗試使用模糊集合論的角度來解釋核心詞彙，並且提出一個整合多面向均勻度的計算方法，使詞語均勻度的衡量能夠同時考慮不同的分類面向，更全面地評估詞彙的核心程度。此外，更提出詞頻正規化的方法來修正傳統均勻度公式遇到切分區塊大小不一致時，造成統計均勻度偏差的問題。在評估方面，本研究使用國家教育研究院建置的華語文語料庫中的書面語語料作為核心詞彙抽取的來源語料庫，再以採用中央研究院平衡語料庫 4.0 版作為驗證語料庫，此一運用異源語料庫評估核心詞彙庫的方法，可以準確地比較及分析各種均勻度公式所選取詞表的優缺點與特性。

研究團隊以實驗證實，正規化後的均勻度公式的確可以有效改善分布均勻度的評估，而整合多面向均勻度的計算方法，確實可以選擇到更具核心特質的詞彙。在語言教學的應用上，過去許多研究者認為均勻度公式偏好選擇功能詞，所以本研究也探討了以詞頻及分布均勻度作為詞彙選取方法的差異。其結果發現，在初階詞彙表的選擇上，無論是頻率法或是均勻度法排序，序位最高的詞彙當中，功能詞所佔的比例都非常高，對學習者來說較不適宜。所以本研究建議必須經過收詞比例的調校以增加實詞的比例，較能達到教學時的實際應用所需。

資料來源：

白明弘、吳鑑城、簡盈妮、黃淑齡、林慶隆。（2016）。基於詞語分布均勻度的核心詞彙選擇。中文計算語言學期刊，21（2），1-17。