Computational Linguistics and Chinese Language Processing
Vol. 21, No. 2, December 2016, pp. 1-18

基於詞語分布均勻度的核心詞彙選擇

A Study on Dispersion Measures for Core Vocabulary Compilation

白明弘*、吳鑑城*、簡盈妮*、黃淑齡*、林慶隆*

Ming-Hong Bai, Jian-Cheng Wu, Ying-Ni Chien,

Shu-Ling Huang and Ching-Lung Lin

摘要

核心詞彙是不受文本類型、主題、應用情境等影響，穩定使用的詞彙。在自然語言中，核心詞彙的數量相對稀少，卻構成溝通內容的主要部份，因此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，在語料庫語言學興起後，詞頻與詞彙分布均勻度統計提供了客觀的統計數據協助核心詞彙的選取。在本論文中，我們提出一個多面向均勻度整合公式，使詞語均勻度的計算能夠同時考慮到不同的分類面向。其次，我們也針對傳統公式統計結果偏差的問題，提出詞頻正規化的方法。對於實驗的評估，我們提出了一個以異源語料庫評估核心詞彙的方法，可以比較各種統計公式的優缺點與特性。在實驗結果的部份，我們實際比較了多種不同的核心詞彙表選擇公式，分析不同公式的特質，並驗證了詞頻正規化的確能夠修正傳統公式的缺點。最後，我們也驗證了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。

* 國家教育研究院編譯發展中心
Development Center for Compilation and Translation, National Academy for Educational Research
E-mail: wujc@mail.naer.edu.tw
The author for correspondence is Jian-Cheng Wu.

Abstract

Core vocabulary is a set of words that are stable used across different text types, theme, and application scenario. In natural language, the number of core vocabulary is relatively small, the core vocabulary, however, plays an important part in language learning because it constitutes a major part of communication content. The traditional core vocabulary selection method is mainly based on the expert knowledge and rule of experience. With the rise of corpus linguistics, word frequency and dispersion uniformity provide objective statistical data to assist the selection of core vocabulary. In this paper, we propose a formula that integrates multi-dimensional uniformity , so that the estimation of word uniformity can take different classification dimensions into account. Secondly, we also propose a method of word frequency normalization for the problem of deviation of the traditional method. For evaluation, a method of evaluating the core vocabulary with a heterogeneous corpus is proposed and it can compare the advantages, disadvantages, and characteristics of various statistical formulas. In the results, we actually compare the different core vocabulary selection formulas, analyzed the characteristics of different formulas, and verified the word frequency normalization can correct the shortcomings of the traditional formula. Finally, we also verified that the proposed method which integrates multi-dimensional uniformity can pick out the vocabulary with more core characteristics.

關鍵詞：語料庫語言學、核心詞彙、邊緣詞彙、分布均勻度。

Keywords: Corpus Linguistics, Core Vocabulary, Fringe Vocabulary, Dispersion Uniformity.

1. 緒論

核心詞彙(core vocabulary)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(Huang, Zhang & Yu, 2005)。這些詞彙的穩定性是多方的，除了跨類別、主題、應用情境之外，還包括跨年齡層、性別等性質(Stuart, 1991)。這些詞彙相對於非核心詞彙(邊緣詞彙, fringe vocabulary)來說數量較稀少，卻構成溝通內容的主要部份(Vanderheiden & Kelso, 1987)。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(Liu, 2012)，因此核心詞彙是語言學習中重要的一環。核心詞彙除了被應用在語言教學之外也應用在詞典編輯、輔助溝通系統、比較語言學等領域(Juilland & Chang-Rodríguez, 1964; Carroll, 1970; Juilland, Brodin & Davidovitch, 1970; Rosengren, 1971; Huang *et al*., 2005; Liu, 2012)。

傳統的核心詞彙選擇方法主要依據專家知識與經驗法則(Huang *et al*., 2005)，語料庫語言學興起後，統計式的方法逐漸取代經驗法則。然而從語料庫的觀察中可以發現，單

純使用詞頻無法分離核心詞彙與邊緣詞彙，許多高頻的詞彙只在特定的情境下高頻出現。我們在中研院平衡語料庫(後簡稱平衡語料庫)中觀察四個詞頻接近的詞在不同主題中的分布情況（如圖 1），「*網路*」在大部份主題中都屬低頻詞，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(Juilland & Chang-Rodríguez, 1964; Carroll, 1970; Juilland *et al.*, 1970; Rosengren, 1970; Huang *et al.*, 2005; Liu, 2012)，當分布程度越均勻時，該詞語的核心程度就越高。

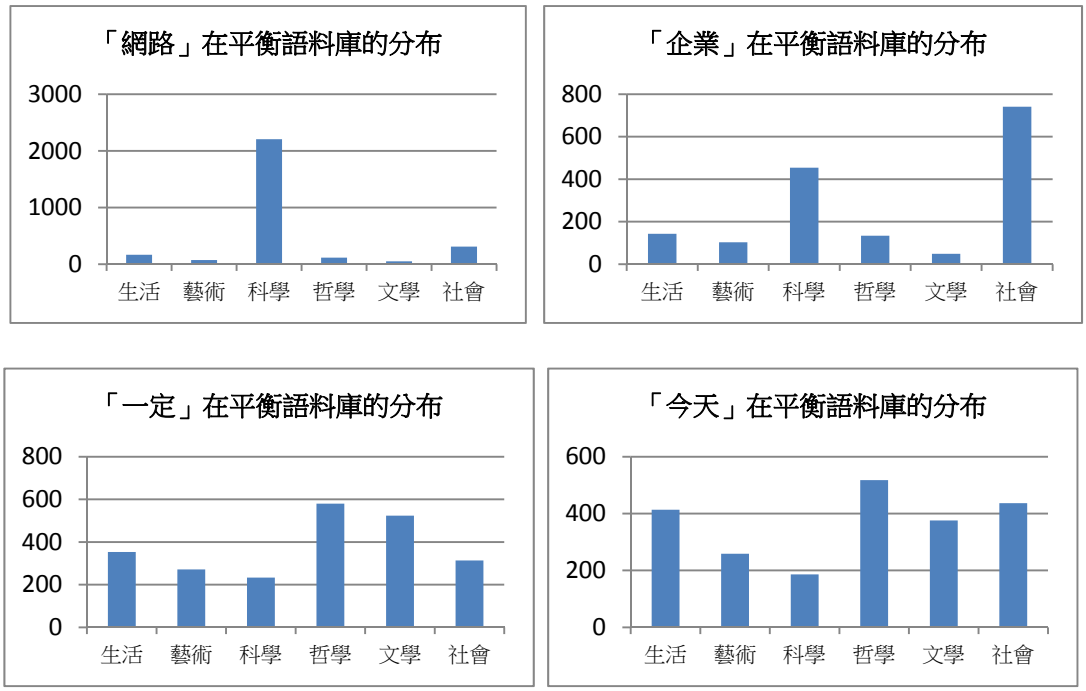

圖 1. 觀察四個詞頻接近的詞語在平衡語料庫不同主題中的分布（單位：次數/每百萬詞)
[Figure 1. The Observation of the distribution of four words with similar frequency in different topics in Sinica corpus (unit: words/per million words)]

詞語分布均勻度確實是衡量詞語核心程度的有效指標，然而，均勻度的計算具有多重分類面向。以上述的例子（圖 1）而言，均勻度的計算面向為主題，統計所得到的均勻度，也僅限於在主題中分布的均勻程度，無法保證在其他面向(如時間、語式、文體、媒體等)的均勻度。Huang *et al.* (2005)曾提到從不同分類面向計算詞語的均勻度的概念，不過該文並未提出整合不同面向均勻度的具體方法。

在本論文中，我們首先提出一個整合多面向均勻度的方法，使詞語均勻度的衡量能夠同時考慮不同的分類面向。其次，我們提出詞頻正規化的方法來修正傳統公式遇到分類區塊大小不一造成統計結果偏差的問題。最後，我們提出了一個以異源語料庫評估核心詞彙庫的方法，以比較各種均勻度公式的優缺點與特性。

以下為本論文的結構說明：我們將在第 2 節中介紹詞語分布均勻度的計算方法；在第 3 節中提出核心詞彙庫的驗證方法；第 4 節將說明實驗的設計、實驗的結果與數據分析；在第 5 節中我們將討論使用均勻度公式在選擇詞彙的一些特質；在第 6 節我們將為本文做結論。

2. 詞語分布均勻度的計算

2.1 語料區塊的切分

計算詞語分布均勻度前必須先將語料庫切分成幾個區塊，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。

語料區塊切分的方法大致可分成隨機切分、依篇章、時間或主題類別切分等。隨機切分的優點是方法簡單、不需要倚賴額外的訊息，且仍有不錯的效果。不過有幾個問題必須注意：首先，當語料本身收錄的文章主題不平衡時，收錄篇數較多的主題可能分布在較多區塊中，因而造成此主題的高頻詞均勻度被高估的問題。其次，在切分區塊時必須選擇適當的區塊大小，因為過小的區塊及過大的區塊都會造成均勻度計算的結果偏差。以極端的例子來說，當區塊極小到以句子為單位時，則均勻度序列將趨近於頻率序列，亦即，以均勻度選擇的詞彙庫將近似於以詞頻選擇的詞彙庫。此問題的原因在於，單一詞語通常在單一句子中很少重覆出現，因此，當一個詞的頻率越高時，必然出現在越多句子之中，因而計算出來的均勻度也越高。另一個極端例子是當區塊極大時，例如只切割成 5 個區塊，假設 *a* 詞在 5 個區塊中均勻地出現數萬次，而 *b* 詞在各區塊中恰巧都低頻出現過。依均勻度來看 *a* 詞和 *b* 詞可能有相近的均勻度，這種現象在區塊數極少時發生的機率很高。

如果依篇章、時間或主題分類來切分區塊，則必需考慮到幾個問題。首先，語料中每個區塊的大小差異可能很大。以平衡語料庫為例，社會主題所收錄的文章數量大約是藝術主題的 4 倍之多。然而，過去研究所提出的均勻度公式大多只考慮詞語在區塊間的詞頻差異，而沒有考慮每個區塊的相對大小。忽略區塊的大小差異的問題將造成詞頻分布均勻度的偏差，亦即原本分布均勻的詞語，卻因為區塊大小不一，而造成均勻度低估的現象。我們將在 2.3 節提出詞頻正規化的方法來解決這個問題。其次，語料的分類方式非常多種，以平衡語料庫來說，目前的分類方式就有主題、文類、媒體、語式、文體、子主題、時間等不同的切割區塊方法，以不同分類切分語料代表以不同面向來觀察詞語的均勻度。我們將在 2.4 節中提出整合多重均勻度的計算方式。

2.2 均勻度公式

為了建立語言學習的參考詞表，過去有很多種均勻度公式被提出來。Juilland & Chang-Rodríguez(1964) 以及 Juilland *et al.*(1970)都以標準差的概念為基礎，提出分布均勻度(Juilland's coefficient of dispersion, JD) 定義如下：

$$JD = 1 - \frac{V}{\sqrt{n-1}} \tag{1}$$

其中的變數說明如下：

n: 語料庫切分區塊數

V: 詞語在區塊中分布的變異數：$V = \frac{\sigma}{\bar{f}}$

σ: 詞語在區塊中分布的標準差：$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(f_i - \bar{f})^2}{n-1}}$

f_i: 詞語在第 *i* 區的詞頻

$\bar{f}$: 詞語在區塊中的平均詞頻，即$\bar{f} = F/n$

F: 詞語在語料中的總詞頻

Juilland 提出分布均勻度的目的是為了適度調整從語料庫中統計的詞頻，因為高頻的詞語未必都是重要的詞語，尤其是只出現在特定主題的詞彙。所以 Juilland 認為選擇學習參考詞表，除了考量詞頻之外，同時也要考量詞語被廣泛使用在不同的主題類別。Carroll (1970) 認同 Juilland 以分布均勻度調整詞頻的觀點，但認為以標準差為基礎所設計的均勻度並不是一個好的評估方法。他提出以訊息熵(entropy)為基礎的分布均勻度公式(Carroll's coefficient of dispersion, CD) 定義如下：

$$CD = \frac{H}{log_2(n)} \tag{2}$$

其中的變數說明如下：

H： 訊息熵 $H = logP - \frac{\sum_{i=1}^{n} p_i \, log_2 p_i}{P}$

p_i：詞語在第 i 區塊中的分布比例，$p_i = f_i/s_i$

s_i：第 i 區塊總詞數

P：$\sum_{i=1}^{n} p_i$

Rosengren (1971) 則調整方均根公式，提出新的均勻度公式(Rosengren's coefficient of dispersion, RD)如下：

$$RD = \frac{\left(\sum_{i=1}^{n}\sqrt{f_i}\right)^2}{n} \cdot \frac{1}{F} \tag{3}$$

其中的變數說明如下：

n: 語料庫切分區塊數

f_i: 詞語在第 *i* 區的詞頻

F: 詞語在語料中的總詞頻

Lyne(1985) 則依據 chi-square 為基礎提出均勻度公式(Lyne's coefficient of dispersion, LD)如下：

$$LD = \frac{1-\chi^2}{4F} \tag{4}$$

Huang *et al.*(2005)為了預測與驗證 Swadesh(1952) 所提出的基本概念表(basic concepts)，提出一個均勻度為基礎的方法。他們所提出的均勻度公式 (Distributional Consistency, DC)如下：

$$DC = \left(\frac{\sum_{i=1}^{n}\sqrt{f_i}}{n}\right)^2 \cdot \frac{1}{f} \tag{5}$$

基本上 Huang 等人所提出的均勻度公式 DC 與 Rosengren 所提出的 RD 相同，不過兩個研究者的目的不同，Rosengren 的 RD 公式是作為調整詞頻的參數之一，為讓調整後的詞頻能夠同時具備高使用率與跨主題均勻分布特性。而 Huang *et al.* (2005)則直接使用 DC 值做為 Swadesh 基本概念表的預測指標。

2.3 公式的使用限制與正規化

在 2.1 節中曾提到，依篇章、主題、時間、語式等分類將語料切分成區塊，則區塊所包含的詞數差異可能很大，進而造成分布均勻度的偏差。例如：「開始」在平衡語料庫中的分布大約每百萬詞出現 600 次，而且在不同文類中，分布十分穩定。但因為平衡語料庫六個主題所收錄的文章數量差異很大，而使得「*開始*」在各主題中出現的頻率差異也很大（見表 1）。

如果我們以每百萬詞出現頻率來看就會發現，「*開始*」在每個文類中都以非常穩定的分布出現。因此，在使用均勻度公式之前，我們可以先將頻率正規化。頻率正規化(frequency normalization)的公式如下：

$$u_i = \frac{f_i \cdot 10^6}{N_i} \tag{6}$$

其中 N_i 表示區塊 i 收錄的總詞數。在計算均勻度時，可將原公式中的詞頻數 f_i 改成正規化後的詞頻 u_i ，但需注意的是，與詞頻有關的參數亦需要調整。我們以 DC 公式的正規化為例來說明，正規化後的公式(Normalized Distributional Consistency, NDC)定義如下：

$$NDC = \left(\frac{\sum_{i=1}^{n}\sqrt{u_i}}{n}\right)^2 \cdot \frac{1}{\bar{u}} \quad (7)$$

其中 $\bar{u}$ 為詞語在所有類別中的平均詞頻，也必需依正規化的詞頻重新計算，即 $\bar{u} = \frac{\sum_{i=1}^{n} u_i}{n}$。

表 1. 「開始」在平衡語料庫各主題出現的頻率與每百萬詞頻率
[Table 1. The original frequency and the frequency in per million words of "Kāi shǐ" in topics of Sinica corpus]

主題	文類大小	頻率	每百萬詞頻率
藝術	846,593	624	737.07
生活	2,243,362	1,471	655.71
文學	2,234,564	1,481	662.77
科學	1,128,083	632	560.24
哲學	1,126,081	698	619.85
社會	3,624,244	2,113	583.02
總計	1,202,927	7,019	626.53

2.4 整合不同面向均勻度

在前面曾提到，將語料庫依不同主題類別切分後計算均勻度，即代表以不同面向來觀察詞語的核心程度。但是在過去的研究中，並沒有提出一套整合不同面向均勻度的方法。所以，在這一節中，我們嘗試使用模糊集合論(Zadeh, 1965; Kwakernaak, 1978)的角度來解釋核心詞彙。

假設 U 表示中文詞彙的模糊宇集合，$S_j \subset U$ 是在某一分類面向的分布均勻模糊集合。則某一詞語 x 屬於 S_j 的均勻度可以表示成 $\mu_{S_j}(x)$ 。我們以 NDC 均勻度公式當作詞語分布均勻度的評估方法，我們將 x 屬於 S_j 的程度定義如下：

$$\mu_{S_j}(x) = NDC_{S_j}(x) \quad (8)$$

其中，$NDC_{S_j}(x)$ 表示以 NDC 均勻度公式計算在 S_j 主題類別面向的均勻度。

我們要計算詞語 x 整合多個分類面向的均勻分布率，亦即是要計算 x 同時屬於多個模糊集合的程度即 $\mu_{\{\cap_{j=1}^{m} S_j\}}(x)$ ， 其中 m 表示有 m 種模糊集合，亦即有 m 種語料庫切分的方法，我們假設均勻模糊集合之間為獨立事件，則依據模糊集合論，可以表示為：

$$\mu_{\cap_{j=1}^{m} S_j}(x) = \prod_{j=1}^{m} \mu_{S_j}(x) = \prod_{j=1}^{m} NDC_{S_j}(x) \quad (9)$$

這個結果代表當要計算多面向均勻度時，在模糊集合獨立性的假設下，我們可以將多面向的均勻度(Multi-dimensional Normalized Distributional Consistency, MNDC) 定義為每個面向均勻度的乘積，亦即：

$$MNDC(x) = \mu_{\cap_{j=1}^{m} S_j}(x) = \prod_{j=1}^{m} NDC_{S_j}(x) \quad (10)$$

以實際的例子來說明，假設要計算詞語在篇章、主題、時間三個面向的均勻度，計算詞語 x 在這三個面向的均勻度可表示為：

$$MNDC(x) = NDC_{篇章}(x) \cdot NDC_{主題}(x) \cdot NDC_{時間}(x) \quad (11)$$

3. 核心詞彙表驗證

為了驗證不同核心詞彙抽取公式的效果，我們使用兩個獨立建置的語料庫來分別進行核心詞彙的抽取與驗證，一、來源語料庫：用來抽取核心詞彙的語料庫。二、驗證語料庫：用來驗證核心詞彙的核心程度。

在驗證的方法上，依照過去研究對核心詞彙的定義，大致可歸納出兩種特性，第一、跨情境特性，第二、詞彙重要性。在跨情境特性的驗證上，我們提出核心詞彙在驗證語料庫中不同主題下的再利用率定義如下：

$$Reuse\ Rate(C) = \frac{1}{n}\sum_{i=1}^{n} \frac{|C \cap L(T_i)|}{|C|} \quad (12)$$

其中 C 表示從來源語料抽取的核心詞彙表，T_i 表示驗證語料庫 T 中的主題 i 語料，$L(T_i)$ 表示構成 T_i 的詞表。公式的意義是檢驗核心詞彙表在驗證語料庫的不同情境中，是否都能保持很高的使用率。

再利用率可用來評估詞表的跨主題特性，但不保證這些詞在使用上是重要的。所以，我們以語料庫覆蓋率來驗證詞彙的重要性，定義如下：

$$Coverage(C) = \frac{\#\ of\ tokens\ in\ T\ which\ was\ covered\ by\ C}{\#\ of\ tokens\ in\ T} \quad (13)$$

當詞表在語料庫中的覆蓋率越高時，表示詞表中的詞彙涵蓋了語料庫中較為重要的詞語和概念。

4. 實驗設計與結果

4.1 核心詞彙來源語料庫

在實驗的設計上，我們使用國家教育研究院建置的華語文語料庫(Corpus of Contemporary Taiwanese Mandarin , 簡稱 COCT)中的書面語語料作為核心詞彙抽取的來源語料庫(柯華葳等人，2016)。COCT 書面語語料庫蒐集來源以圖書為主，目前共包含 1 億 1,220 萬字。這些文章在分類上包含出版年份(1986-2015 年)、書籍冊別，書籍主題(10 類)，我們利用篇章、書籍冊別、書籍主題、與出版年份四個分類面向來計算詞彙的分布均勻度。

4.2 驗證語料庫

在核心詞彙的驗證上，我們採用中央研究院平衡語料庫 4.0 版，該語料庫包含了約 1,000 萬詞，每一份文本都標示主題、文類、媒體、語式及文體等不同的後設資料。(Chinese Knowledge Information Processing Group [CKIP], 1995)

4.3 不同均勻度公式比較

在第一個實驗中，我們驗證了不同均勻度公式在核心詞彙的抽取效果。在來源資料庫切分上，我們採用隨機區塊切分法，將 COCT 書面語語料庫隨機切分成包含 3,000 字的小區塊，然後使用不同公式計算每個詞語的分布均勻度、詞頻或調整詞頻，公式說明如下：

Freq: 詞語在語料庫中的頻率。

Entropy: 以訊息熵公式計算均勻度值。

DC: 以 Distributional Consistency 計算詞語的均勻度值。

JD: Juilland 的分布均勻度。

JU: Juilland 的詞頻調整法。

CD: Carroll 的分布均勻度。

CU: Carroll 的詞頻調整法。

LD: Lyne 的詞頻調整法。

RAF: Rosengren 的詞頻調整法。

在實驗中，我們分別使用上列的公式計算來源語料庫中詞彙的分布均勻度、詞頻或調整詞頻，然後將每個詞表依照公式評估值由高而低排序，最後，在每個詞表中取出前 10,000 個詞做為核心詞彙表抽取的結果。

在核心詞彙抽取效果的比較上，我們使用每個公式統計出來的結果各取前 10,000 詞當作詞彙表，並以平衡語料庫的六個主題來驗證詞彙表的再利用率。表 2 為 9 個公式抽取的詞彙表的再利用率驗證結果。從結果來看，我們可以發現 Entropy , DC, CD 三個方法所抽取的詞彙表，在平衡語料庫中都有較高的再利用率，這表示這三個方法所抽出來的詞彙，具有高度跨主題性質。

為進一步觀察這些詞的重要性，我們評估這些詞彙表在平衡語料庫中的覆蓋率(如表 3)。在此表中我們發現 Freq, DC, JU, CU, RAF 這五個方法的覆蓋率都很高。其中 RAF 和 CU 能夠最有效覆蓋驗證語料庫，由此可知使用均勻度來調整詞頻的確可以更精確地找出重要性高的詞彙。不過，使用 RAF 及 CU 所抽出詞彙表的再利用率則不及 DC，這表示 RAF 及 CU 的結果包含了較多的非核心詞彙。綜合來說，從再利用率的角度觀察，Entropy、DC、CD 三種公式在核心詞彙的抽取上有最佳的核心性質，而從覆蓋率的

角度觀察，這三個公式的語料庫覆蓋率相較於詞頻及調整詞頻的公式並無太大的差異，亦即這三個公式在詞彙的選擇上能夠兼顧到詞彙的核心性與重要性。

表 2. 各核心詞彙表在平衡語料庫的不同主題中的再利用率
[Table 2. The re-utilization rate of each core vocabulary in different topics in Sinica corpus]

	生活	藝術	科學	哲學	文學	社會	平均
Freq	0.969	0.920	0.897	0.927	0.972	0.973	0.943
Entropy	**0.992**	0.946	**0.924**	**0.967**	**0.999**	**0.996**	0.971
DC	**0.992**	**0.947**	**0.924**	0.966	0.998	**0.996**	0.970
JD	0.991	0.940	0.916	0.964	0.998	0.995	0.967
JU	0.972	0.924	0.900	0.930	0.975	0.975	0.946
CD	**0.992**	0.946	**0.924**	**0.967**	**0.999**	**0.996**	0.971
CU	0.978	0.931	0.906	0.939	0.980	0.981	0.952
LD	0.991	0.939	0.915	0.964	0.998	0.994	0.967
RAF	0.988	0.943	0.918	0.956	0.993	0.992	0.965

表 3. 不同均勻度公式在驗證語料庫中的覆蓋率
[Table 3. The coverage of different uniformity formulas in verification corpus]

	生活	藝術	科學	哲學	文學	社會	整體覆蓋率
Freq	0.809	0.828	**0.833**	0.893	0.853	0.814	**0.832**
Entropy	0.806	0.826	0.826	0.893	0.855	0.809	0.829
DC	0.807	0.828	0.829	0.894	**0.856**	0.811	**0.831**
SD	0.788	0.806	0.818	0.872	0.829	0.792	0.810
JD	0.798	0.819	0.816	0.887	0.852	0.799	0.821
JU	0.809	0.829	**0.833**	0.894	0.854	0.814	**0.832**
CD	0.806	0.826	0.826	0.893	0.855	0.809	0.829
CU	**0.810**	0.830	**0.833**	0.894	0.855	**0.815**	**0.833**
LD	0.798	0.818	0.815	0.887	0.852	0.797	0.820
RAF	**0.810**	**0.831**	**0.833**	**0.895**	**0.856**	**0.815**	**0.833**

4.4 正規化的實驗

在第二個實驗中，我們以 DC 及 Entropy 兩個公式為例，比較詞頻正規化前計算均勻度及正規化後計算均勻度的效果。在來源資料庫的切分上，我們採用書籍與主題分類兩種切分方式，因為這兩種切分方式的區塊中包含的詞數大小較不一致。書籍切分法即是以每一冊書籍視為一個區塊將 COCT 書面語語料庫切分成區塊；而主題分類切分法依 10 個主題類別[1] 將 COCT 書面語語料庫切分成 10 個區塊，再使用不同公式計算每個詞語跨區塊的分布均勻度。在均勻度的計算上，DC 及 Entropy 兩個公式都分別計算詞頻正規化前的均勻度與詞頻正規化後的均勻度(Normalized DC 與 Normalized Entropy)。實驗結果同樣於排序後取前 10,000 個詞作為詞彙表，並驗證詞彙表的再利用率。

表 4 為以書籍切分法抽取詞彙表的結果，為了觀察，我們將驗證語料庫切分成生活、藝術、科學、哲學、文學、社會六個類別。由表 4 我們可以發現四個詞彙表在生活、文學及社會三類文本中的再利用率都很高，所以正規化前後沒有差別，而藝術、科學及哲學類再利用率較低，正規化後有微幅的改善。整體來說，正規化之後有微幅的改善，但差距不甚明顯。這代表以單冊書籍為切分區塊不致產生太大的偏差。

表 4. 詞頻正規化前後的再利用率比較，以書籍為語料庫切分區塊
[Table 4. The comparison of re-utilization rate before and after word frequency normalization(using book as a block unit)]

	生活	藝術	科學	哲學	文學	社會	平均
Entropy	0.991	0.928	0.901	0.962	0.999	0.994	0.962
Normalized Entropy	**0.992**	**0.933**	**0.906**	**0.965**	0.999	0.994	**0.965**
DC	0.994	0.943	0.917	0.970	0.999	0.996	0.970
Normalized DC	0.994	**0.945**	**0.919**	0.970	0.999	0.996	**0.971**

表 5 為以主題類別切分法抽取詞彙表的結果，由結果可以發現以主題分類當作切分區塊來計算均勻度，所抽取的詞彙再利用率較書籍區塊切割來得差。經過正規後，Normalized Entropy 在六個類別中的再利用率都比未正規化的 Entropy 有明顯的提高，平均改進達到 5%。而 Normalized DC 的再利用率也比未正規化的 DC 提高 3-4%。這結果一方面顯示了切分區塊過少時，較難選出具核心特性的詞彙；另一方面，以主題分類作為區塊切割，將造成區塊大小的落差較大，所以不論是 Entropy 或是 DC 公式，先將詞頻正規化，結果都能得到大幅的改善。

由此結果發現以下情形： 一、Entropy 比 DC 更容易受到語料庫切分不均的影響，所以 Entropy 正規化後效果提昇比 DC 來得明顯。二、切分區塊數極少的時候，區塊大小不均的落差較大，所造成的影響程度較為嚴重，所以此時正規化的效果在兩個方法中都極為明顯。三、生活、文學及社會類型文本對核心詞彙的再利用率較藝術、科學及哲

[1]參考中文圖書分類法並經調整後分成 10 類：總類，商業及金融類，哲學及宗教類，史地類，藝術類，語言文學類，社會科學類，應用科學類，科學類，休閒類。

學來得高。

表 5. 詞頻正規化前後的再利用率比較，以主題分類為語料庫切割區塊
[Table 5. The comparison of re-utilization rate before and after word frequency normalization(using topic as a block unit)]

	生活	藝術	科學	哲學	文學	社會	平均
Entropy	0.919	0.812	0.810	0.798	0.905	0.936	0.863
Normalized Entropy	**0.967**	**0.873**	**0.839**	**0.893**	**0.975**	**0.973**	**0.920**
DC	0.945	0.858	0.848	0.841	0.937	0.957	0.898
Normalized DC	**0.974**	**0.894**	**0.862**	**0.909**	**0.980**	**0.980**	**0.933**

4.5 多面向整合的效果

在第三個實驗中，本研究比較單一面向均勻度所抽出的詞表和整合多面向均勻度所抽出詞表的效果差異。在來源資料庫的切分上，採用了四種切分法，分別計算 NDC 均勻度：

Text NDC: 隨機切成 3,000 字的區塊，計算 NDC 均勻度。

Book NDC: 以書籍為單位，切分成 850 個區塊，計算 NDC 均勻度。

Class NDC: 以主題分類切分成 10 個區塊，計算 NDC 均勻度。

Year NDC: 以出版年份分成 29 個區塊，計算 NDC 均勻度。

MNDC: 將上列四種均勻度值以 MNDC 公式計算出整合均勻度。

在詞表的選取上，對於每一個方法，我們同樣依均勻度值將詞語排序，並抽取前 10,000 個詞作為詞彙表，同樣以平衡語料庫來驗證這 5 個詞彙表。

表 6 以再利用率評估詞表抽取的結果，從結果中我們可以發現，MNDC 公式所得到的詞表有最佳的再利用率，這表示整合多個面向的均勻度(MNDC 公式)的確可以選出核心程度最高的詞彙。

表 7 則是以語料庫覆蓋率來評估整合公式，從結果可以發現, MNDC 公式依然有很高的語料庫覆蓋率，但 Text NDC 所選出來的詞表，在語料庫覆蓋率上則高於 MNDC。這代表多面向均勻度(MNDC)在詞彙的選擇上，比較傾向於高核心程度的選擇，而放棄一些頻率高而核心程度較低的詞。

表 6. 多面向均勻度整合的詞表再利用率
[Table 6. The re-utilization rate of multi-dimensional uniformity integrated word list]

	生活	藝術	科學	哲學	文學	社會	平均
Text NDC	0.992	0.946	0.924	0.965	0.998	0.996	0.970
Book NDC	0.994	0.945	0.919	0.970	0.999	0.996	0.971
Class NDC	0.974	0.894	0.862	0.909	0.980	0.980	0.933
Year NDC	0.992	0.937	0.914	0.966	0.996	0.996	0.967
MNDC	**0.996**	**0.951**	**0.928**	**0.973**	**0.999**	**0.997**	**0.974**

表 7. 多面向均勻度整合的語料庫覆蓋率
[Table 7. The corpus coverage rate of multi-dimensional uniformity integrated word list]

	生活	藝術	科學	哲學	文學	社會	整體覆蓋率
Text NDC	**0.807**	**0.828**	**0.829**	**0.894**	**0.856**	**0.811**	**0.830**
Book NDC	0.795	0.818	0.808	0.888	0.853	0.794	0.818
Class NDC	0.734	0.752	0.735	0.836	0.795	0.731	0.757
Year NDC	0.782	0.799	0.799	0.881	0.836	0.786	0.807
MNDC	0.803	0.824	0.821	0.893	**0.856**	0.805	0.825

由於詞表再利用率與語料庫覆蓋率兩個評估值存在取捨的關係，所以從表 6 及表 7 中並無法明顯確認MNDC 的效果。為了觀察詞表再利用率與語料庫覆蓋率的變化關係，本研究觀察當詞表的取詞數從 1,000 逐漸增加到 40,000 詞，詞表再利用率與語料庫覆蓋率的變化關係。從圖 2 可以發現在相同的語料庫覆蓋率基準之下，MNDC 比 Text NDC 及其他單一面向均勻度有穩定的再利用率。

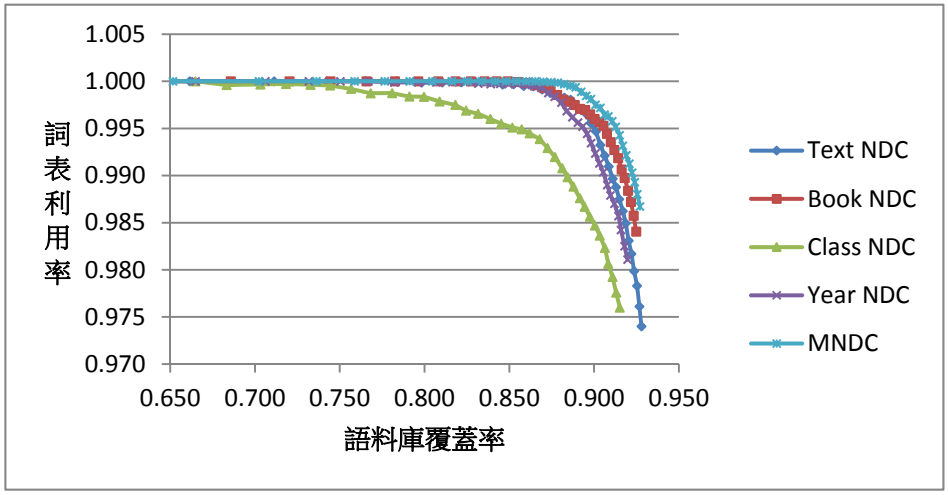

圖 2. 語料庫覆蓋率與詞表利用率關係圖
[Figure 2. The relationship between corpus coverage and word list utilization rate]

5. 分析與討論

過去的研究普遍認為以均勻度抽取詞表，會收錄較多的功能詞。本節中，我們將探討均勻度對詞彙選擇的實際影響。

首先，我們分別觀察以詞頻與均勻度公式收取 10,000 詞的詞表時，兩個詞表的詞類比例差異比較，結果如表 8。從表中我們可以發現，均勻度公式所收錄的名詞比詞頻法大量減少了約 1,000 詞，因為大部份的名詞多少都具有使用情境的特質，分布較不均勻。因為名詞收詞量的減少，所以其他詞類的收錄稍有增加，這是遞補的現象。只有副詞(ADV)、及物動詞(Vt)與不及物動詞(Vi)有較為明顯的增加。從這個結果中，並未顯示均勻度的選詞比詞頻選詞更偏好功能詞。

其次，我們把詞表的收詞數減少為 500 詞來觀察依均勻度收詞的詞類分布(表 9)，我們發現收詞數很少時，副詞(ADV)、 對等連接詞(C), 連接詞(POST)、時態標記(ASP)、介詞(P) 等詞類的收詞比例高很多。相對來說，名詞(N)、及物動詞(Vt)、不及物動詞(Vi)的收詞比例少很多。從表 9 的結果我們可以發現，當收詞量偏少時，均勻度對功能詞的偏好現象十分明顯。這一方面是因為功能詞在分布上的確是比較不受語境的影響，另一方面，功能詞的詞頻也比一般詞來得高。

在語言教學參考詞彙表的編輯上，以均勻度選擇詞彙表的確可以收錄比較穩定的跨領域核心詞彙。但是對於初階語言教學來說，因為初階所收錄的詞彙量較少，統計所得的詞彙表將包含較大比例的功能詞。對初階學習者來說，功能詞的語義模糊而抽象、歧義性高、語法規則複雜(Klammer, Schulz & Volpe, 2009)，所以不宜在初階詞表中收錄過多的功能詞。以華測會所編輯的華語八千詞表(Steering Committee for the Test of Proficiency - Huayu[SC-TOP], 2016) 為例，入門級所收錄的 500 詞中，名詞佔了極高的比例(約 60%)。而在基礎級中，名詞收錄比例降到約 50%左右。進階、高階及流利三級的名詞收錄則維持在約 40% ，可見初階語言教學的確需要包含較多的名詞。因此，不論是均勻度或是詞頻所選出來的詞表，在初階教學使用上都還需要經過收詞比例的調校以增加實詞的比例。

表 8. 前 10,000 詞收詞詞類比較
[Table 8. The comparison of Parts of Speech of the top 10,000 words]

	依詞頻收錄詞數	依均勻度收錄詞數	依詞頻收錄之百分比	依均勻度收錄之百分比	收錄差異百分比
ADV/副詞	634	755	6.34%	7.55%	19.09%
A/(非謂)形容詞	96	92	0.96%	0.92%	-4.17%
C/連接詞	111	119	1.11%	1.19%	7.21%
ASP/時態標記	10	11	0.10%	0.11%	10.00%
DET/定詞	191	222	1.91%	2.22%	16.23%

M/量詞	169	168	1.69%	1.68%	-0.59%
N/名詞	4,752	3,749	47.52%	37.49%	-21.11%
P/介詞	102	108	1.02%	1.08%	5.88%
Vt/及物動詞	2,355	2,811	23.55%	28.11%	19.36%
T/語助詞	54	48	0.54%	0.48%	-11.11%
Vi/不及物動詞	1,479	1,869	14.79%	18.69%	26.37%
POST/後置詞	47	48	0.47%	0.48%	2.13%

表 9. 依均勻度收錄前 500 詞與收錄 10,000 詞的詞類分布比較
[Table 9. The comparison of the distribution of the Parts of Speech of top 500 and 10,000 words according to uniformity]

	收錄 10,000 詞詞類分布	**收錄 500 詞詞類分布**
ADV/副詞	7.55%	21.60%
A/（非謂）形容詞	0.92%	0.60%
C/連接詞	1.19%	7.00%
ASP/時態標記	0.11%	1.00%
DET/定詞	2.22%	8.80%
M/量詞	1.68%	4.00%
N/名詞	37.49%	16.80%
P/介詞	1.08%	5.60%
Vt/及物動詞	28.11%	21.40%
T/語助詞	0.48%	2.20%
Vi/不及物動詞	18.69%	7.80%
POST/後置詞	0.48%	3.20%

6. 結論

在本研究中，我們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。其次，我們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。最後，我們提出了一個以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的優缺點與特性。最後，我們以實驗證實，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。所以我們在本文中也探討了以詞頻及分布均勻度作為詞彙選取方法的差異。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。所以我們建議必須經過收詞比例的調校以增加實詞的比例。

參考文獻 References

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behaviour*, *3*(2), 61-65.

Chinese Knowledge Information Processing Group [CKIP]. (1995). *A Description to the Sinica Corpus*. Technical Report 95-02, Academia Sinica, Taipei.

Huang, C.-R., Zhang, H. & Yu, S.-W. (2005). On predicting and verifying a basic lexicon: proposals inspired by distributional consistency. In *POLA forever: festschrift in honor of Professor William S-Y. Wang on his 70th birthday*, Taipei: Language and Linguistics, Academia Sinica, 57-69.

Juilland, A. G. & Chang-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. The Hague: Mouton.

Juilland, A. G., Brodin, D. R. & Davidovitch, C. (1970). *Frequency dictionary of French words*. Paris, French: Mouton.

Klammer , T., Schulz, M. R. & Volpe A. D. (2009). *Analyzing English Grammar* (6th ed). Harlow, England: Longman.

Kwakernaak, H. (1978). Fuzzy random variables - I: Definitions and theorems. *Information Sciences*, 15, 1-29.

Liu, C-.P. (2012). *The effects of theme-narrative instruction with core vocabulary on oral narrative ability in elementary students with severe hearing impairment.* (Master's thesis, National University of Taiwan).

Lyne, A. A. (1985). Dispersion. *In The vocabulary of French business correspondence*(101-124). Paris, French: Slatkine-Champion.

Rosengren, I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)*, *1*, 103-127.

Steering Committee for the Test of Proficiency - Huayu[SC-TOP]. (2016). *8000 Chinese Words*. Steering Committee for the Test Of Proficiency-Huayu, http://www.sc-top.org.tw/english/download.php.

Stuart, S. L. (1991). Topic and vocabulary use patterns of elderly men and women of two age cohorts. *ETD collection for University of Nebraska - Lincoln.* Paper AAI9208116.

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts. In *Proceedings of the American Philosophical Society 96*, 152-63.

Vanderheiden, G. C. & Kelso, D. (1987). Comparative analysis of fixed-vocabulary communication acceleration techniques. *AAC Augmentative and Alternative Communication*, *3*, 196-206.

Zadeh, L. A. (1965). Fuzzy sets. *Information and Control*, *8*, 338-353.

柯華葳、林慶隆、張俊盛、陳浩然、高照明、蔡雅薰、張郁雯、陳柏熹、張莉萍、吳鑑城、白明弘、陳茹玲、李詩敏、黃淑齡、劉寶琦、丁彥平、簡盈妮、張玳維、余昱瑩（2016）。*華語文八年計畫「建置應用語料庫及標準體系」105 年工作計畫期中報告*。教育部補助之研究計畫。臺北市：國家教育研究院。[Ko, Hwa-Wei, Ching-Lung Lin, Jason S. Chang, Hao-Jan Howard Chen, Zhao-Ming Gao, Ya-Hsun Tsai, Yu-Wen Chang, Po-Hsi Chen, Li-Ping Chang, Jian-Cheng Wu, Ming-Hong Bai, Ju-Ling Chen, Shih-Min Li, Shu-Ling Huang, Pao-Chi Liu, Yen-Ping Ting, Ying-Ni Chien, Tai-Wei Chang, Yu-Ying Yu. (2016). *『The 8-year project of construction and application of Mandarin Chinese corpus and standard systems 』105-year work plan interim report.* Taipei: National Academy for Educational Research.]