

# 對應於歐洲共同架構的華語詞彙量\*

張莉萍

國立臺灣師範大學國語教學中心

## 摘要

本文先從外語詞彙學習文獻入手，檢視英語詞彙量的研究，再從文本覆蓋率、教學與認知學習等方向來探討漢語詞彙量和歐洲共同架構(CEFR)之間的關係。筆者建構了一個學習者語料庫--內容取自以 CEFR 為架構的電腦華語寫作能力考試，分析現階段 113 萬字語料，根據語料覆蓋率概念，並參考華語教材與學習之間的關係，初步建議對應於 CEFR 等級的漢語詞彙量：A2 約 1000 個詞；B1 是 2300-3000 個詞；B2 是 4500-5000 個詞。至於 C 級詞彙量，則限於考試寫作語料庫 C 級語料不足，加入外籍學生手寫作文語料統計，估計 C 級詞彙量介於 8000 到 10000 個詞之間。

**關鍵詞：**詞彙量，詞表，學習者語料庫，覆蓋率，歐洲共同架構

## 1. 前言

「歐洲語言共同參考架構：學習、教學、評量」(以下簡稱 CEFR)(Council of Europe 2001)自 2001 年出版以來，受到外語教學、測驗評量領域廣泛的注目，這個共同架構之所以得到高度的關注，究其原因，除了受到歐洲語言學習的龐大人口影響之外，三等六級(A1/A2; B1/B2; C1/C2)的清晰架構、兼具科學性驗證與簡明的能力描述風格也是得到外語教學與考試機構青睞的主要原因。就語言測試的角度來看，不同的測試如果有一個共同的架構為平台，那麼不同的測試所報告的分數或等級就可以在一共同標準上互相參照，如此可以減輕學習者、教學機構對於不同測試分數或成績理解解釋上的負擔，也有利於不同國家或地區相互承認各種語言能力證書。近年來，不少外

---

\* 本研究得到國科會計畫(NSC 99-2631-S-003-012, 100-2631-S-003-004)以及教育部邁向頂尖大學計畫部分經費補助，特此感謝。本文初稿發表於 2011 年 10 月 22-23 日於國立台灣師範大學所舉辦的紀念梁實秋先生國際學術研討會上，感謝與會專家學者及本期刊兩位匿名審查人給予之寶貴建議。

語考試已經完成對應於 CEFR 架構，做出對應於該架構的分數或等級說明。例如，IELTS 考試 6.5 級分對應到 C1 等級 (Taylor 2004)；TOEFL 紙筆考試 560 分對應到 C1 等級 (Tannenbaum & Wylie 2005)；德語考試、法語考試也都有相對應的說明。中華民國教育部則在 2005 年明訂推動英語能力檢定測驗時，應參考這個共同架構 (臺社 (一) 字第 0940075287C 號令)；另於 2007 年發函國家華語測驗推動測驗委員會 (以下簡稱華測會) 評估華語文能力測驗 (TOCFL, 前 TOP) 與 CEFR 接軌的可行性 (臺華字第 0960077767 號)。華測會自 2007 年開始依據相關文獻 (Council of Europe 2003; Figueras, North, Takala, Verhelst, & Van Avermaet 2005) 進行對應工作 (高詩涵、陳俐蘋、藍珮君 2007; 藍珮君 2007)。雖然將測驗對應到 CEFR 這個架構上，仍有不少模糊難界定之處 (Weir 2005)，但透過測驗研發人員共同的理解，初步的對應後，顯示等級對應如下表。

表 1、TOCFL 對應 CEFR 一覽

TOCFL	基礎級	初等/進階級	中等/高階級	高等/流利級
CEFR	A2	B1	B2	C1

為了提供考生報考訊息，華測會建議各級適用對象為：基礎級適合在臺灣學習半年 (約 240 小時) 或具有 800 個詞彙量的學習者；初等是適合 360-480 小時或具備 1500 個詞彙者；中等 480-960 小時或具備 5000 個詞彙者；高等適合 960 小時以上或 8000 個詞彙量的學習者 (張莉萍 2007)。而 2010 年初，中國國家漢辦推出新版漢語水平考試 (HSK)，該測驗亦宣稱對應於 CEFR，然所建議詞彙量與 TOCFL 及其它外語差異甚鉅 (請參見表 2)，可見雖然大家有共同的平台，但由於 CEFR 並沒有針對個別語言建立充分而詳細的語法、詞彙等語言知識內容，以致於即使描述同一個語言的內在知識，各家說法不同。姑不論其它外語詞彙量的估算，不同學者或單位對於漢語詞彙量的推估差異即不小，從表 2 中可以看出，華測會所公布的詞彙量與德語區漢語教學協會所公布的詞彙量，差距最小。漢辦所公布的詞彙量和各家差距最大，各等級詞彙量明顯較其它系統低許多。

表 2、對應於 CEFR 的外語詞彙量一覽

	德語 <sup>1</sup>	俄語 <sup>2</sup>	漢語（華測會） <sup>3</sup>	漢語（漢辦） <sup>4</sup>	漢語（德語區漢語教學協會） <sup>5</sup>	漢語（蔡雅薰） <sup>6</sup>
A1	500		500	150	600	300
A2	1000	2100	800	300	1200	1000
B1	2000	4400	1500	600	2500	2000
B2	4000-5000	10000	5000	1200	5000	3500
C1		12000	8000	2500		5500
C2				5000		8000

本文在這個動機下，試圖探討漢語做為外語，學習者在每一個等級所需具備的詞彙量。也就是說，究竟多少詞彙量可以做到 CEFR 能力指標所描述的任務或活動。希望藉由這個研究可以提供給教學者、學習者和評量者具體參考。

## 2. 文獻探討

大概沒有人會否認詞彙在外語學習上的重要性，甚至有人說詞彙是人與人溝通必要的元素，沒有詞彙就沒有辦法傳達意義（Wilkins 1972: 111）。因此在外語教學大綱或評量上，詞彙量的擬定都佔了重要的地位。也有大量研究顯示詞彙量與各項語言技能之間的關係，例如，詞彙量與閱讀理解之間的關係（Laufer 1992; Qian 1999; Stæhr 2008）；與寫作能力之間的關係（Astika 1993; Laufer 1998; Stæhr 2008）；與聽力之間的關係（Milton, Wade, & Hopkins 2010; Stæhr 2008; Zimmerman 2004）；與口語流利之間的關係（Milton et al. 2010; Zimmerman 2004），顯著相關性基本上都介於 0.6 到 0.8 之間。華測會（2011）分析 98-99 年度考生詞彙得分和總分之間的相關性，結果是各等級

<sup>1</sup> 詞彙量的訊息是來自與德國顧安達教授(Prof. A. Guder) 電子郵件交流（2010 年 9 月 9 日）。

<sup>2</sup> 資料出自彭桂英（2007: 157）。其中針對 B2 和 C1 的詞彙量，另加註「其中需能自由使用 7000 字」。

<sup>3</sup> 資料來源 <http://www.sc-top.org.tw/>

<sup>4</sup> 資料來源 [http://english.hanban.org/node\\_8002.htm#nod](http://english.hanban.org/node_8002.htm#nod)

<sup>5</sup> 資料來源 [http://www.fachverband-chinesisch.de/fachverbandchinesishev/thesen-papiereundresolutionen/FaCh2010\\_ErklaerungHSK.pdf](http://www.fachverband-chinesisch.de/fachverbandchinesishev/thesen-papiereundresolutionen/FaCh2010_ErklaerungHSK.pdf)

<sup>6</sup> 資料來源 <http://140.138.144.150/~s912250/1004slide.pdf> (page 14)

顯著相關在 0.629-0.860 之間。顯示詞彙能力越高所展現的語言溝通能力越佳。

關於二語學習者詞彙研究，以英語文獻居多。研究大致有兩個方向，一是詞彙廣度（breadth），一是詞彙深度（depth），前者就是詞彙量（size），後者是指詞彙使用能力。Vermeer(2001)曾論證過這兩者其實是同一個向度，深度是詞彙量其中的一個功能。兩者很難劃分開來，在測試詞彙量的同時，往往也包括了詞彙深度的測試。另外，在討論詞彙量問題時，還有一個概念常常被提起，那就是產出性（productive）/接收性（receptive）的詞彙能力，究竟能不能區別二者，二者是不是有顯著差異。關於這個問題，學界爭辯許久，但至今也還無定論，可以參考 Melka（1997）一文。近年來，似乎學者傾向於不區分二者差別，主要理由是詞彙量的定義很清楚，也可以由心理語言學實驗驗證，但是其它一不論是詞彙深度或是產出/接收詞彙能力的定義則無法得到普遍的共識，也就無從驗證（Milton 2010）。本文也只能先忽略這些問題，直接討論詞彙量和 CEFR 等級關係。

那麼究竟詞彙量與 CEFR 之間是否有關聯？是不是可能訂出每個等級學習者所需要具備的詞彙量？關於第一個問題的答案應該是肯定的，因為 CEFR 的指標最早就是從 Van Ek（1975）*The Theshold Level* 一書的概念（功能—意念大綱）而來，書中列有詞表，大約 2000 個詞彙，而 *Theshold* 相當於後來 CEFR 的 B1 等級；*Waystage*（相當於 CEFR A2 等級）則列了 1000 個詞彙（Van Ek 1980）。只是 CEFR 不是針對特定語言設定，因此後來書中也就沒有提及詞彙量的問題，但從近來許多學者的研究可以發現，愈高等級的學習者詞彙量愈高的這個結果，CEFR 與詞彙量之間的關聯無庸置疑。例如，Meara & Milton（2003）以參加劍橋英語考試通過者為受測對象，利用詞彙測驗 *XLex*（常用 5000 詞）來估計學習者的詞彙量，結果如表 3 所示。顯示訂定每個等級詞彙量是可行的作法，可以讓 CEFR 這個共同平台更具有說服性。不過，表 3 所呈現的僅是學習者針對最常用的 5000 詞所做的統計，實際上 B2 以上學習者所需要的詞彙量一定超過這個估計。Milton & Alexiou（2009: 208）指出劍橋系列考試中相當於 B2 等級的 FCE 考試架構基礎是採用 Hindmarsh（1980）的詞表，這個詞表所列的詞語則有 4500 個。如果表 3 中所列 3250-3750 個常用詞，另加上某些特殊情境所需詞語及非屬於前 5000 高頻詞的詞語，對 B2 學習者而言，4500 個詞彙量應該是合理的推測。

表 3、詞彙量與 CEFR 的關係（摘錄自 Meara & Milton, 2003, p.8）

CEFR Levels	Cambridge exams	XLex (5000 max)
A1	Starters, Movers and Flyers	<1500
A2	Kernel English Test	1500 - 2500
B1	Preliminary English Test	2750 - 3250
B2	First Certificate in English	3250 - 3750
C1	Cambridge Advanced English	3750 - 4500
C2	Cambridge Proficiency in English	4500 - 5000

另外，Milton & Alexiou（2009）更利用 *XLex* 測驗的三個語言版本來分別測試英語、希臘語、法語學習者，希望得知不同語言所需的詞彙量是否不同。得到的結果是學習希臘語需要較大的詞彙量；而法語需要的詞彙量比英語少。這個結果顯示，不同外語學習者在每個等級所需要的詞彙量或許不同。Milton（2010）對這個現象做了一番討論，大致提出 4 個方向來思考，第一是詞彙形成方式不同，可能導致頻率計算結果很不一樣。例如，在英語頻率極高的詞類是代名詞或介詞，但不是每個語言都是這樣的表現。像匈牙利語、芬蘭語、土耳其語這些黏著性語言（*agglutinative language*），則是在動詞或名詞後面以加上詞綴的方式來表示，可能這些語言需要的詞彙就比英語要多出個 1000 詞左右。第二、由於歷史的因素，英語在某些類別的詞彙較多，例如，農場動物與肉類是完全不同的詞形，如，*sheep*（羊）和 *mutton*（羊肉），要理解這些概念的詞彙量可能比其它語言來得多。第三、個別語言造詞的方式不同，例如，德語傾向於用已存在的詞來組合或複合成新詞，但是有些語言傾向於創造新詞或衍生新詞。另外，有的語言詞界不清，如漢語，不同的分詞方式，則會造成不同數量詞彙結果。最後，不同語言因為語域（*register*）不同，詞彙表現方式也不同，例如，法語的高頻 2000 詞，除了用在平常生活所需外，也適用在正式和學術語域；英語則需要特別的學術詞彙。這也呼應了 Milton & Alexiou（2009）的研究結果，法語所需的詞彙比英語少。

不同語言在每個等級所需要的詞彙量既然不同，那麼如何訂定才可行或合理？Milton & Alexiou（2009）提出利用語料的覆蓋率來觀察每個語言所需的詞彙量。關於覆蓋率的概念，是指我們如果要理解一個文本，需要文本中的多少詞彙。已經有不少學者提出閱讀理解的相關數據，例如，Laufer（1989）

和 Hu & Nation (2000) 分別提出百分之九十五覆蓋率可以達到一般理解和百分之九十八覆蓋率可以做到輕鬆理解的數字根據。Nation (2001: 147) 更進一步提出百分之八十的覆蓋率可以做到抓住篇章要旨的理解。最近的研究則是，Nation (2006) 根據英國國家 (書面) 語料庫 (BNC) 的資料統計，覆蓋百分之九十八的語料，需要 8000-9000 詞族 (word family)<sup>7</sup>。也就是說，要能沒有障礙的閱讀報章小說，需要 8000-9000 詞族。Nation (2006) 研究也指出，要能覆蓋口說語料庫內百分之九十八的語料，需要 6000-7000 詞族。這些數據也得到其他研究者的支持 (Adolphs and Schmitt 2003; Schmitt 2008)。

然而上述的 8000-9000 詞或許可以視為學習者要達到英語 C2 程度所需的詞彙量，那麼其它等級的詞彙數量又是多少呢？究竟英語文獻中經常提出的常用詞 2000, 3000, 5000 是否能對應到 CEFR 的等級？目前看來，較清楚也為較多研究者支持的是最常用的 2000 詞彙，應該是初階學習者要往上繼續學習，所需要的詞彙量門檻 (Stæhr 2008)。刻在進行的英國劍橋大學與其他合作機構所進行的英語水平計畫 (English Profile Program, EP)，也提供了一些具體的數據。Capel (2010) 指出這個計畫已經利用母語人士語料庫、學習者語料庫以及學習者教材等，提供對應於 CEFR 的詞彙內容如下：A1-601；A2 新詞-925；B1 新詞-1429；B2 新詞-1711；總計四級的詞數為 4666。她指出這個結果和 Hindmarsh (1980) 詞表 4500 個這個數字非常相近，隱含要達到英語 B2 等級的學習者大概要有 4500 詞彙量；B1 約 3000 個；A2 約 1500 個；A1 約 600 個。

### 3. 對外漢語的詞彙量

相較於外語尤其是英語詞彙量的豐富文獻，關於漢語做為第二語言或外語的詞彙量的探討則非常貧乏。臺灣多數學者探討華語詞彙分級議題，例如，葉德明 (1997)、鄭昭明 (1997)、張郁雯 (2003)、張莉萍 (2002, 2004) 等。各家對於詞彙等級的分界，大都採用詞的使用頻率和累積頻率作為依據，如葉德明 (1997: 20) 將教材收集來的詞彙，依出現在教材中的使用頻率排序，取前 8000 左右，經過專家學者的討論，訂出由易到難的五個等級，第一

---

<sup>7</sup> Nation (2006) 舉 abbreviate 說明詞族的概念，abbreviate, abbreviates, abbreviated, abbreviating, abbreviation, abbreviations 都是同一家族，也就是由詞根加屈折詞綴或衍生詞綴都算在同一個家族詞之下。

級（易）有 1,471 詞，第二級累計有 3,104 詞等等。鄭昭明（1997）也對常用字（4583）和常用詞（44908），依其使用頻率/總次數的 50%，75%，95%，99%和 100%，提供了 5 層不同難度的等級。這樣的劃分方式是否符合華語學習者的目標，以及所依據詞庫的詞頻是否客觀都值得商榷。加以每一個等級的詞彙數量參差不齊，例如張郁雯（2003: 96）依照詞的累計頻率值為 75%，85%，90%，95%，將詞彙分成 5 個等級。五級詞彙量分別是 2720、1440、2372、4093、7636，也就是說對初學者（第一級）的要求是 2720 個詞彙量，對第二級的要求只有 1440 個詞彙，似乎不符合學習原理，而且五級詞彙量總計 18261，對於語言學習者而言，似乎負擔過重。張莉萍（2002，2004）爲了華語文能力測驗所研製的詞表，將詞彙量 10000 或 8000 分爲三個等級（見表 2，B1-C1 等級），總的來說，我們看到的普遍問題是各家的詞彙量標準不清以及分級的定義不一。

鄭錦全（1998）在〈從計量理解語言認知〉這篇文章中，深入檢查歷代經書、史書、字書等用字，發現雖然字書（即現今字典）所用字種最多達 5 萬多，但其他書種、不同篇章、不同作者所用字種都不超過八千<sup>8</sup>，因此提出「詞涯八千」的概念——一個人所能掌握運用的詞素和非衍生性詞語的數目。文中也舉鄒嘉彥等人（T'sou et al. 1997）以覆蓋率所得出的數據爲例來證明，以百分之九十的覆蓋率來看，所需的詞彙量無論在新加坡、香港、臺灣三地都只需幾千詞<sup>9</sup>。鄭之研究從認知與學習的觀點出發，提供語言學習領域啓發性的訊息。張莉萍（2004）沿用 8000 詞這概念做爲華語學習者的詞彙量依據，利用語料庫詞彙頻率與教材詞彙等訊息，劃分了初、中、高三級詞彙內容與詞彙量。

無獨有偶，8000 這個數量在大陸學者的研究中，也得到驗證。劉英林、宋紹周（1992）觀察 1959-1991 年之間大陸所做的詞表、詞庫後，指出前人經過多次反覆統計形成的常用詞共識，至少有兩個標界，一個是常用詞 3000；一個是常用詞 8000。3000 常用詞可以覆蓋一般語料的 86%，5000 常用詞可以覆蓋語料的 91%，8000 常用詞可以覆蓋語料的 95%。雖說，劉宋兩人分析觀察眾多語料庫與詞表所得成爲編製對岸漢語水平詞彙等級大綱的重

<sup>8</sup> 鄭文根據不同經書、史書，不同作者用字都不超過四、五千字到七、八千字（四、五千佔多數），先提出「字涯八千」的說法，再從每個漢字都能代表一個詞素，漢語的詞素大多能單獨成詞，提出「詞涯八千」概念。

<sup>9</sup> 鄒文是以 1995-1996 華人社區報章爲文本內容統計。

要依據一甲乙丙丁四級詞共 8822 個。但同樣覆蓋率的概念，如果以臺灣中研院建置的平衡語料庫統計資料來解釋（詞庫小組 1998），排名前 9100 個詞，只能覆蓋百分之八十五的語料，也就是說，認識 9100 個詞可以理解五百萬詞語料中的百分之八十五。如果要達到百分之九十五的覆蓋率，需要擁有 35,500 個詞彙。可見不同的語料庫性質、大小，以及分詞的方法，都關係到統計出來覆蓋率的數據。基本上，語料庫越大、語料同質性越小，文本覆蓋率所需的詞彙量就會越大。由於語料庫的大小、內容的平衡性以及詞彙切分、計算的方式這些面向還沒有得到妥善的解決，不同標準下得出覆蓋率的詞數也有不小的差距。而且這些覆蓋率的數字都是在本國人產出的語料庫所得出的結果。

另一方面，既然談的是學習者詞彙，也應該從教學、學習與評量的實際層面來討論華語詞彙量的問題，劉英林、宋紹周（1992: 12）報告了根據對外漢語教學群體經驗，在大學四年中，前兩年的詞彙量是 5000 詞，後兩年 3000 詞，因此認為訂定 8000 詞做為對外漢語教學專業的通用詞彙總目標是合理的。表 4 是筆者歸納臺灣多數學習者的學習狀況，包括學時和詞彙數的關係一覽表：

表 4、學時和詞彙量

教材	詞彙數	學時	CEFR 等級
視聽華語 I	449	120	A1/A2
視聽華語 II	481	130	A2
視聽華語 III	1195	140	B1
視聽華語 IV	1250	140	B1/B2
視聽華語 V	667	200	B2

從表 4 可以看出，學完視聽華語這一系列教材，約需要 730 小時<sup>10</sup>，可以學得 4042 詞，預期可以達到 B2 程度。要注意的是，這是在目標語地區的學習狀況，高階的學習者透過母語人士的交談、課室活動、網路，還可以接觸到或學習更多的華語詞彙。似乎 4500 的詞彙量對要達到 B2 等級的華語學習者而言，是合理，可能也是至少得具備的詞彙量。至於之後的學習，由於學生依興趣或專業的需求選擇不同教材，很難估計實際學得的詞彙。如果以 1 年

<sup>10</sup> 以台灣語言中心的教學制度約一年半時間。



半的時間學得 4500 詞的標準來計算，平均 1 天學得 8 個詞，在臺灣學習 3 年，可以擁有 9000 個詞彙量。

綜上所述，無論從教材教學層面或覆蓋率觀點欲探求華語學習者所需的詞彙量都有其侷限。以認知學習的觀點，鄭錦全的詞涯八千概念倒可視為詞彙學習的總目標。這個結果與英語近年來的研究相當，本文假設華語與英語所需總詞彙量近似，相對於 CEFR 各等級的詞彙量也與英語近似，也就是說，以 A1-600, A2-1500, B1-3000, B2-4500 這樣的數量來檢視表 2 所蒐集到的漢語詞彙量數據，摒除 C 級詞彙不看，德語區漢語教學協會所擬的詞彙量較接近英語詞彙量。中國漢辦所擬數量則遠遠不足。漢語所需詞彙量真的較英語或其它外語來得少嗎？究竟不同程度的學習者能產出什麼樣的詞彙、能運用多少詞彙，目前為止，並沒有系統性的觀察或報告。在下一節，本文利用一個符合 CEFR 分級的學習者語料庫，來觀察不同語言水平的學習者實際產出的詞彙內容與詞彙量。

#### 4. 研究方法

語料庫語言學興起於 1980 年代，臺灣中央研究院則是從 1990 年代開始蒐集語料，於 1995 年完成第一版兩百萬詞的漢語語料庫（請參考 <http://dbo.sinica.edu.tw/ftms-bin/kiwil/mkiwi.sh>），限於早期收集的方式與限制，不論臺灣或對岸所收集的文本都以書面語體（尤其報章雜誌）居多，對於對外漢語教學者或學習者而言，功效有限。本文擬利用學習者語料庫，觀察漢語做為第二語言學習者實際產出的語料。關於華語學習者語料庫，大陸已建置完成的有北京語言大學的漢語中介語語料庫、南京師範大學留學生語料庫、暨南大學留學生漢語中介語語料庫、哈薩克族學生漢語中介語語料庫系統、HSK 動態作文語料庫、中山大學中介語語料庫，但只有 HSK 語料庫是唯一對外公開的漢語學習者語料庫（<http://202.112.195.192:8060/hsk/login.asp>）。在臺灣，鄧守信教授於民國 86 年間與 94 年間所蒐集的中介語料是最早的華語學習者語料庫，但學習者背景多針對英語為母語的人士，語料來源比較多是課後習作或測驗，目前在臺師大華研所網站上可以下載語料。為切合這個研究需要，我們希望觀察的學習者語言是在完成溝通任務下所產生的語料，而且符合歐洲共同架構的等級劃分。因此本文利用近年來臺灣華測會開發的華語文能力電腦寫作考試的語料來建構一個學習者語料庫。這個語料庫（以下簡稱 TOCFL 語料庫）的特色如下：

- (1) 這個寫作語料的來源完全是學生或考生直接以注音或拼音輸入法輸入，也就是並非透過第二人之手輸入，免去人為輸入錯誤的問題。
- (2) TOCFL 測驗採用分等分級的考試方式，運用歐洲共同語文參考架構 (CEFR) 的分級方式來命題，分為六級；A1、A2 屬於初級，B1、B2 是中級，C1、C2 是高級。因此我們收集到的語料可以得到學生精確的程度訊息，也就是他們在該等級的表現，可能在該等級通過或沒通過的訊息，以及所得分數，該測驗的分數為級分制，分為 0-5 級分。
- (3) 此寫作考試是以任務為導向的命題方式，因此不同等級有不同的寫作文體與功能要求，參加這個考試的考生需要完成兩個任務，A2 是完成實用性的便條和看圖說故事的文體；B1 是書信和一般記敘文；B2 則是應用文（特定功能的書信）和論說文（表達對特定事件的看法）；C1 是報告類（圖表說明）和論說文（提出理由支撐觀點）。

筆者主要蒐集從 2006 年至 2010 年的電腦寫作考試文本<sup>11</sup>，樣本涵蓋來自 39 種不同母語背景的學習者，60 個不同主題，總計 1,131,057 字，3,405 篇，681,859 詞數 (word token)，16,622 詞種 (word type)，也就是學習者使用了 16,622 個不同詞彙。蒐集的語料包括 A2、B1、B2、C1 各等級，每一級語料篇數及字數統計請見表 5。

表 5、TOCFL 語料庫語料分佈一覽

級別	主題數	篇數	字數	百分比
A2	19	1125	164,172	14.52%
B1	21	1423	528,335	46.71%
B2	15	650	341,734	30.21%
C1	5	207	96,816	8.56%
總計	60	3405	1,131,057	100.00%

從表 5 可以看出，中等程度學習者的語料佔多數，約百分之七十七；高等程度的語料偏少，不到十萬字。如果以百分之九十八覆蓋率為標準來觀察

<sup>11</sup> B2 因為語料篇數太少（僅 386 篇），另外加了 2011 年 264 篇。

這個學習者語料庫，所需要的詞彙量是 6481 個。如果以百分之九十五的覆蓋率來觀察，只需要 3028 個。以百分之九十八覆蓋率來看，漢語所需詞彙量表面上較英語少。當然這個結果最大的問題可能在於語料庫不夠大，除此之外，可能還存在幾個問題，例如，語料的平衡性不足，如上所述，C1 語料只佔了所有語料的百分之九左右。或是統計結果牽涉詞的認定方式，目前詞表所蒐集的「詞」是以中研院中文詞知識庫小組斷詞程式分析得來（詞庫小組 1996）。

爲了彌補 C 級語料的不足，筆者另蒐集 2010-2011 臺師大國語中心（MTC）選讀 6 級教材以上的外籍學生寫作文本（相當於 CEFR 的 C1-C2 等級），共 525 篇（59 類主題）。加入統計後，C 級語料達到 487,442 字，總詞數約 90 萬。要達到百分之九十八覆蓋率所需要的詞彙量是 9,922 個；如果是百分之九十五的覆蓋率，則需要 4,452 個。

表 6、每個級別詞數、詞種、覆蓋率、詞長訊息<sup>12</sup>

	詞數 word token	詞種 word type	95% 覆 蓋 率 coverage	98% 覆 蓋 率 coverage	詞長 char./token
A2	100,538	3,808	992	1940	1.6330
B1	324,303	10,092	2263	4654	1.6300
B2	203,021	9,179	2802	5201	1.6832
C	273,993	17,178	6411	11699	1.7790
小計	901,855	-	-	-	-

表 6 則是每一等級語料的詞種、詞數、詞長、百分之九十五、九十八覆蓋率所需詞彙量的統計資料。由於 B1 語料最多，所以詞數也最多。C 級語料則因爲加入 MTC 作文，主題類別較多元，詞種數量最多。從詞長的訊息，可以大致看出 B2 以上學習者有詞長增長現象，具有書面正式語體的特徵<sup>13</sup>。以百分之九十五和百分之九十八的覆蓋率來計算每一個等級的詞彙量（表 6 第 4-5 欄），可以看出每個等級所需的詞彙量逐漸上升，顯示要理解越高階的文本，需要越多的詞彙量。不過，可能因爲 B2 的語料數在比例上遠低於 B1、C 級語料，篇數少加上主題不那麼豐富（雖然語料量有二十萬詞以上），目前顯現與 B1 所需詞彙量的差距並不高。A2 和 B1 之間的差距則明顯可見；

<sup>12</sup> C 級資料是加入 MTC 手寫作文語料的統計數字。

<sup>13</sup> 關於書面語體特徵，有興趣讀者請參考馮勝利、胡文澤主編（2005）。

C 級和其它等級之間的差距也很顯著。

筆者以一般理解的覆蓋率（95%）大膽假設，A2 詞彙量可以設定在 1000；B1 詞彙量設定在 2300。這個數字與 A2/B1 教學、教材詞彙量相去不遠（見表 4）--屬於 A1/A2 詞彙有 930 個，屬於 B1 教材的詞彙量則在 2125-3375 之間（390-530 學時）。B2 詞彙量的設定，則需要以輕鬆理解的覆蓋率（98%）為準，因為 CEFR 區別 B2 學習者和 B1 學習者之間最大的不同是能掌握細節，而非單純的理解大意。這樣一來，從文本覆蓋率的觀點來看，B2 需要 5000 詞左右，C 級學習者則需要 12000 詞左右。但由於 C 級詞表中，排序在 9106 個詞之後的詞僅在語料庫中出現一次，不具代表性。筆者建議採取四級總詞表的 98% 覆蓋率為準，也就是近 1 萬詞；排序在總詞表中的前 1 萬個詞，在語料庫中都出現了 3 次以上。

## 5. 結語與未來工作

本文從文本覆蓋率、教學與認知學習等方向來探討漢語詞彙量和 CEFR 之間的關係，初步建議對應於 CEFR 等級的漢語詞彙量為：

A2：1000

B1：2300-3000

B2：4500-5000

C：8000-10000

至於每個等級的詞彙是哪些，要回歸到學習者應該具備哪些詞彙才可以做到 CEFR 能力指標所描述的任務或活動，然而大家都知道一個語言功能，可以運用各種不同的語言表達形式。因此，實際學習者所產出的（語料）、所接收的材料（指課堂活動、教材輸入）都是研究者應考量的來源。現階段已經由學習者語料庫蒐集了在任務溝通下學習者所產出的語料，建議未來可參考教材詞表或目標語母語人士詞頻等訊息，編輯常用 1000 詞、3000 詞、5000 詞等詞表，以供學習者、教學、測驗評量機構參考使用，也可輔助多數 A1-B2 能力學習者的詞彙學習。

據筆者初步觀察學習者語料庫發現，華測會所提供的 A2、B1 詞彙內容，學習者使用頻率極高，詳見表 7。華測會 A2 詞表所列的 800 詞，在學習者語料庫中使用了 763 個，比例高達 95.38%；B1 的 1500 詞為學習者使用的比例也高達 94.57%，可見 800 基礎（A2）詞彙或 1500 進階（B1）詞彙的內容大

致符合學習者溝通所需。另外，本研究也參考了信世昌等人（2010）所編華語詞庫中所列之核心詞彙 900 個與基礎詞彙 2000 個，比對了學習者使用狀況，如表 7 所示。核心詞彙（900 詞）在各等級使用率不低，意外的是，在基礎學習者（A2）使用比例相對較低，基礎詞庫（約 2000 詞）的現象亦同。猜測可能是編纂詞彙時，受當初情境需求分析的樣本或問卷所致，不過，由於沒有詳細資料可得知該詞庫情境需求分析的考察內容，無法進一步判斷。

表 7、學習者使用詞彙情況<sup>14</sup>

寫作考試語料庫	華測會各級詞表	核心詞庫（900 詞）	基礎詞庫（2000 詞）
A2	95.38%（800 詞）	81.13%	56.50%
B1	94.57%（1500 詞）	94.08%	84.33%
B2	72.60%（5000 詞）	90.63%	73.11%

最後，筆者也觀察到不少學習者詞彙並不落在華測會華語八千詞詞表的等級中<sup>15</sup>，例如「夜市、而且、其實、演講、觀眾、主角」等詞出現在學習者 B1 等級的語料庫中，而且都各出現 50 次以上。但前五個詞並不列在華測會 B1 詞表中，而在高階級（B2）詞表，「主角」這個詞則是出現在流利級（C1）詞表。可見詞表內容還有調整空間，建議未來可參考學習者實際產出的語料，召開專家諮詢會議，調校等級詞彙量及內容，提供學習者、教學者、評量者精確的詞彙訊息。當然，本研究方法也因為使用的工具而有所侷限，例如，採用機器自動斷詞的系統，不一定與一般使用者或教學者對漢語「詞」的認知相同，可能使得在計算詞彙或比對統計時產生不一致；另外研究採用的語料庫缺乏 A1、C2 使用者，自然也無法經由學習者端實際產出來觀察這兩個等級的詞彙內容或建議詞彙量，這些問題有待未來進一步探究。

## 引用文獻

Adolph, Svenja, and Norbert Schmitt. 2003. Lexical coverage of spoken discourse. *Applied Linguistics* 24.4: 425-438.

<sup>14</sup> 由於各家詞表對詞或詞類的分析不一致，在計算時，可能不同認知者間會有些微差距。

<sup>15</sup> 資料來源：<http://www.sc-top.org.tw/download/L1-L5vocabulary%20list20111208.xls>

- Astika, Gusti Gede. 1993. Analytical assessment of foreign students' writing. *RELC Journal* 24.1: 61-72.
- Capel, Annette. 2010. A1-B2 vocabulary: insights and issues arising from the English Profile Wordlists project. *English Profile Journal*, 1, e3 doi: 10.1017/S2041536210000048
- Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge University Press.
- Council of Europe. 2003. *Relating language examinations to the common European framework of reference for languages: Learning, Teaching, Assessment (CEF)(Manual: Preliminary Pilot Version)*. DGIV/EDU/LANG 2003, 5. Strasbourg: Language Policy Division.
- Figueras, Neus, Brian North, Sauli Takala, Norman Verhest, and Piet Van Avermaet. 2005. Relating examinations to the Common European Framework: A manual. *Language Testing* 22.3: 261-279.
- Hindmarsh, Roland. 1980. *Cambridge English lexicon*. Cambridge: CUP.
- Hu, Marcella, and Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 23.1: 403-430.
- Laufer, Batia. 1989. What percentage of text-lexis is essential for comprehension? *Special Language: From Humans Thinking to Thinking Machines*, eds. by Christer Lauren & Mar Nordman. Clevedon: Multilingual Matters.
- Laufer, Batia. 1992. How much lexis is necessary for reading comprehension? *Vocabulary and applied linguistics*, eds. by Henri Bejoint & Pierre J. Arnaud, 126-132. London: Macmillan.
- Laufer, Batia. 1998. The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics* 19: 255-271.
- Meara, Paul, and James Milton. 2003. *X\_Lex, the Swansea Levels Test*. Newbury: Exress.
- Melka, Francine. 1997. Receptive vs. productive aspects of vocabulary. *Vocabulary: Description, Acquisition and Pedagogy*, eds. by Norbert Schmitt & Michael McCarthy, 84-102. Cambridge: Cambridge University Press.

- Milton, James. 2010. The development of vocabulary breadth across the CEFR levels. *Communicative proficiency and linguistic development: intersections between SLA and language testing research: Eurosla Monographs Series, vol. 1*, eds. by Inge Bartning, Martin Maisa & Ineke Vedder, 211-232. Available online at: <http://eurosla.org/monographs/EM01/EM01tot.pdf>
- Milton, James, and Thomai Alexiou. 2009. Vocabulary size and the Common European Framework of Reference for Languages. *Vocabulary studies in first and second language acquisition*, eds. by Brian Richards, H. Michael Daller, David D. Malvern, Paul Meara, James Milton & Jeanine Treffers-Daller, 194-211. Basingstoke: Palgrave Macmillan.
- Milton, James, Jo Wade, and Nicola Hopkins. 2010. Aural word recognition and oral competence in a foreign language. *Insights into nonnative vocabulary teaching and learning*, eds. by Rubén Chacón-Beltrán, Cristián Abello-Contesse & María del Mar Torreblanca-López, 83-97. Bristol: Multilingual Matters.
- Nation, Paul. 2001. *Learning vocabulary in another language*. Cambridge: Cambridge University Press.
- Nation, Paul. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63: 59-82.
- Qian, David D. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *The Canadian Modern Language Review* 56: 283-307.
- Schmitt, Norbert. 2008. Instructed second language vocabulary learning. *Language Teaching Research* 12.3: 329-363.
- Stæhr, Lars Stenius. 2008. Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal* 36: 139-152.
- Tannenbaum, Richard J., and E. Caroline Wylie. 2005. *Mapping English Language Proficiency Test Scores Onto the Common European Framework* (TOEFL Research Reports, No. 80, RR-05-18). ETS, Princeton, NJ.
- Taylor, Lynda. 2004. IELTS, Cambridge ESOL examination and the Common European Framework. *Research Notes* 18: 2-3. University of Cambridge, ESOL Examinations.

- T'sou, Benjamin K., Hing-Lung Lin, Godfrey Liu, Terence Chan, Jerome Hu, Ching-hai Chew, and John Kwock-Ping Tse. 1997. A synchronous Chinese language corpus from different speech communities: Construction and applications. *Computational Linguistics and Chinese Language Processing* 2.1: 91-104.
- Van Ek, Jan Ate. 1975. *The Threshold Level*. Strasbourg: The Council of Europe.
- Van Ek, Jan Ate, and John Leslie Melville Trim. 1980. *Waystage English*. London: Pergamon Press.
- Vermeer, Anne. 2001. Breadth and depth of vocabulary in relation to L1/L2 acquisition and frequency of input. *Applied Psycholinguistics* 22: 217-234.
- Weir, Cyril J. 2005. Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing* 22.3: 281-300.
- Wilkins, David A. 1972. *Linguistics in Language Teaching*. London: Arnold.
- Zimmerman, Kevin J. 2004. *The role of vocabulary size in assessing second language proficiency*. Utah, USA: Brigham Young University Unpublished MA thesis.
- 中研院詞庫小組. 1996. 《「搜」文解字--中文詞界研究與資訊用分詞標準》。技術報告 NO.96-01。臺北南港：中央研究院資訊所。[CKIP (Chinese Knowledge and Information Processing) Group. 1996. *Segmentation Standard for Chinese Natural language Processing*. (Tech. Rep. No. 96-01). Taipei: the Association for Computation Linguistics and Chinese Language Processing (ACLCLP).]
- 中研院詞庫小組. 1998. 《詞頻辭典，中央研究院中文詞知識庫小組技術報告 CKIP-98-02》。臺北南港：中央研究院資訊所。[CKIP. (1998). *Accumulated Word Frequency in CKIP Corpus*. (Tech. Rep. No. 98-02). Taipei: ACLCLP.]
- 信世昌、鄧守信、李明懿（編）。2010. 《華語基礎詞庫 1.0 版》。臺北：文鶴出版有限公司。[Hsin, Shih-chang, Shou-hsin Teng, and Ming-yi Li (Eds.). 2010. *Basic Wordlist for Teaching Chinese as a Second Language Version 1.0*. Taipei: The Crane Publishing Co., Ltd.]



- 高詩涵、陳俐蘋、藍珮君. 2007. 〈TOP 與 CEFR 初步對應結果〉(未出版手稿)。臺北市：華測會。[Gao, Shi-han, Li-ping Chen, and Pei-jun Lan. 2007. The preliminary study of relating TOP to CEFR (Unpublished manuscript). Taipei: SC-TOP.]
- 馮勝利、胡文澤(主編). 2005. 《對外漢語書面語教學與研究的最新發展》。北京：北京語言大學出版社。[Feng, Sheng-li, and Wen-ze Hu (Eds.). 2005. *New Development on Written Language Teaching for Teaching Chinese as a Foreign Language*. Beijing: Beijing Language College Press.]
- 葉德明. 1997. 〈華語文常用詞彙頻率等級統整研究〉，《華文世界》，第85期，14-21。[Yeh, Teh-ming. 1997. An integrated study plan for frequency categorization of a Mandarin Chinese lexicon. *World Chinese Language*, 85, 14-21. Taipei: World Chinese Language Association.]
- 張郁雯. 2003. 〈詞彙分級研究〉，《華語文能力測驗編製：研究與實務》，柯華葳主編，83-102。臺北：遠流出版社。[Chang, Yu-wen. 2003. Research of Chinese vocabulary levels. *Chinese Proficiency Test: Research and Practice*, ed. by Hwa-wei Ko, 83-102. Taipei: Yuan-Liou Publishing Co., Ltd.]
- 張莉萍. 2002. 《華語文能力測驗理論與實務》。臺北：師大書苑。[Chang, Li-ping. 2002. *Theoretical and Practical Relevant of Chinese Proficiency Test*. Taipei: Lucky Bookstore.]
- 張莉萍. 2004. 〈華語文詞彙與句型分級方式初探 I〉，國科會成果報告(NSC-92-2411-H-003-045)。[Chang, Li-ping. 2004. A preliminary approach to grading vocabulary and patterns of Chinese as a second language. *National Science Council Research Report (NS-92-2411-H-003 -045)*.]
- 張莉萍. 2007. 〈華語文能力測驗發展現況〉，《外語能力測驗之動向與展望國際研討會論文集》，185-195。[Chang, Li-ping. 2007. The development and status of Test of Proficiency-Huayu. *Proceedings of the International Conference of Foreign Language Proficiency Tests Trends*, 185-195.]
- 張莉萍. 2011. 〈對應於歐洲共同架構的對外漢語學時建議〉。第一屆東亞華語教學研究生論壇(2011.1.15-16)。臺北：臺灣師範大學。[Chang, Li-ping. 2011. Suggested CSL learning hours based on the CEFR scale. Paper presented at the First East Asian Forum for Graduate Students of Teaching

- Chinese as a Second Language(2011.1.15-16). Taipei: National Taiwan Normal University.]
- 彭桂英. 2007. 〈俄語能力測驗(TOEL)在臺灣之現況與展望〉,《外語能力測驗之動向與展望國際研討會論文集》,155-163。[Peng, Guei-Ying. 2007. The status and promotion in Taiwan of TOREL. *Proceedings of the International Conference of Foreign Language Proficiency Tests Trends*, 155-163.]
- 國家漢辦(孔子學院總部). 2010. 漢語水平考試。2010年3月15日,取自:  
[http://english.hanban.org/node\\_8002.htm#nod](http://english.hanban.org/node_8002.htm#nod) [Hanban (Confucius Institute Headquarters). 2010. HSK. Retrieved March 15, 2010, from [http://english.hanban.org/node\\_8002.htm#nod](http://english.hanban.org/node_8002.htm#nod).]
- 國家華語測驗工作推動委員會[華測會]. 2011. 〈各年度各題本詞彙得分與各變項相關分析〉,華測會技術報告,100年10月17日。[SC-TOP. 2011. The correlation between the vocabulary score and each variable. (Tech. Rep. Oct., 2011). Taipei: SC-TOP.]
- 劉英林、宋紹周. 1992. 〈論漢語教學字詞的統計與分級〉,《漢語水平考詞彙與漢字等級大綱》,國家對外漢語教學領導小組辦公室漢語水平考試部主編,1-22。北京:北京語言學院出版社。[Liu, Ying-lin, and Shao-zhou Song. 1992. Calculating and ranking of Chinese characters and words. *The Guidelines of HSK Vocabulary and Characters*, ed. by The Office of Chinese Language Council, 1-25. Beijing: Beijing Language College Press.]
- 蔡雅薰. 2011. 以CEFR為基礎之華語教學規準(投影片)。國立臺灣師範大學頂大第三次讀書會,2011年10月4日。取自:  
<http://140.138.144.150/~s912250/1004slide.pdf> [Tsai, Ya-Hsun. 2011. The CSL/CFL teaching standards based on CEFR. Retrieved October 4, 2011, from <http://140.138.144.150/~s912250/1004slide.pdf>]
- 德語區漢語教學協會. 2010. 對新漢語水平考試的幾項說明。2010年8月1日,取自  
[http://www.fachverband-chinesisch.de/fachverbandchinesishev/thesenpapiereundresolutionen/FaCh2010\\_ErklaerungHSK.pdf](http://www.fachverband-chinesisch.de/fachverbandchinesishev/thesenpapiereundresolutionen/FaCh2010_ErklaerungHSK.pdf) [Fachverband Chinesisch e.V. 2010. Statement of the Fachverband Chinesisch e.V. (Association of Chinese Teachers in German Speaking Countries) on the new HSK Chinese Proficiency Test. Retrieved August 1, 2010, from <http://www.fachverband-chinesisch.de/fachverbandchinesishev/thesenpapie>

reundresolutionen/FaCh2010\_ErklaerungHSK.pdf.]

- 藍珮君. 2007. 〈基礎華語文能力測驗與歐洲共同架構的對應關係〉, 《臺灣華語文教學》, 2007 年第二期, 39-47。臺北: 文鶴出版有限公司。 [Lan, Pei-jun. 2007. Mapping test of proficiency-Huayu for beginners onto the CEFR. *Teaching Chinese as a Second Language*, 2007: 2, 39-47. Taipei: The Crane Publishing Co., Ltd.]
- 鄭昭明. 1997. 〈漢語水平考試的定位、編製及「字彙」與「詞彙」使用的問題〉, 《華文世界》, 第 85 期, 42-47。 [Cheng, Chau-Ming. 1997. The construction of Hanyu Shuiping Kaoshi: uses of characters and vocabulary. *World Chinese Language*, 85: 42-47. Taipei: World Chinese Language Association.]
- 鄭錦全. 1998. 〈從計量理解語言認知〉, 《漢語計量與計算研究》, 鄒嘉彥、黎邦洋、陳偉光、王士元編, 15-30。香港: 香港城市大學。 [Cheng, Chin-Chuan. 1998. Quantification for understanding language cognition. *Quantitative and Computational Studies on the Chinese Language*, eds. by Benjamin T'sou, Bang-yeung Lai, Wei-guang Chan & Shi-yuan Wang, 15-30. Hong Kong: City University of Hong Kong.]

[審查: 2012.2.21 修改: 2012.3.26 接受: 2012.4.27]

張莉萍

Li-ping Chang

臺北市大安區和平東路 1 段 162 號

國立臺灣師範大學國語教學中心

162, Sec 1, Hopping E Rd

Mandarin Training Center

National Taiwan Normal University

Taipei 106, Taiwan

lchang@ntnu.edu.tw

**The study of the vocabulary size at the CEFR levels for  
CFL/CSL learners**

**Li-ping Chang**

**Mandarin Training Center, National Taiwan Normal University**

**Abstract**

The study aims to explore the relationship between the vocabulary size and learners' Chinese proficiency based on the Common European Framework of Reference (CEFR, Council of Europe, 2001) six levels. The research method is corpus-based. The source of the corpus is mainly from the computer-based writing Test of Chinese as a Foreign Language (TOCFL) which is designed according to the CEFR. So far, the learner corpus consists of more than 1 million Chinese characters from A2 to C1 levels. The paper discusses the question from the aspects of text coverage calculation, wordlists of teaching materials and cognitive learning. The results suggest the vocabulary size for A2 level is 1000; B1 level is 2300-3000; B2 level is 4500-5000; C level is 8000-10000.

**Keywords :** vocabulary size, wordlist, learner corpus, coverage, CEFR