華語詞彙分級初探*

張莉萍 陳鳳儀
lchang@ntnu.edu.tw a0410613@ntnu.edu.tw
國立台灣師範大學國語教學中心
台灣省台北市 110 和平東路 1 段 162 號

2005/3/21 被接受、2005/10/10 修改寄出

摘要

本研究是針對華語文詞彙分級方式的一個嘗試，並實際從事詞彙整理與分級工作。主要選詞來源為中研院核心詞彙、通用詞彙、參考詞彙表、CPT 詞彙表、HSK 詞彙表，運用相對頻率、加權值的方式算出每一個詞的比重，進而依據華語文學習里程，制定出初級 1500 個詞彙、中級 3500 個詞彙、高級 3000 個詞彙。經過詞彙排序分級等原則處理後的 2005CPT 初級詞彙覆蓋了百分之九十二左右的 HSK 甲級詞，顯示兩者基本詞彙的同質性相當高。詞表內容除了詞彙本身，還標示了每個詞的等級、拼音、語法類、(英文)解釋和例句，並註明該詞在 HSK 中的等級。由於選詞來源涵蓋範圍廣，並具代表性，避免了一般語料庫偏書面語，少口語詞彙的缺失，此分級詞表將是華語文教師與外籍人士學華語的重要參考指標。

關鍵字：華語文教學、詞彙、分級、加權值、頻率、測驗

* 本論文 2005 年 4 月於廈門發表時，限於大會頁數限定，刪除部分內容，此爲修改後完整版本。

1. 研究動機

近來諸多報導指出中文已成爲世界第二強勢語言。不過，台灣教育當局對於華語教學界著力甚少，既沒有規範性的華語文教學指標，更遑論研製標準化測驗或能力檢定考試。許多外籍人士在台灣各華語中心學了好多年，除了學校發給的單科成績單之外，無法提供給公司雇主或申請入學的學校單位一份客觀而標準的評量。台灣師大國語教學中心在有限的人力、經費之下，自 2001 年開始有系統的研擬華語文聽與讀的能力測驗，制訂初步的分級方式、測驗的形式、內容、題數，經過預試、於 2003 年 12 月在台灣推出第一次華語文能力測驗(CPT)。在這過程中，詞彙量與詞彙的分級一直是我們關心的課題。

這個研究的目標是發展華語文詞彙庫及等級劃分，以做爲華語文教材編製的基本參考資料與華語文能力測驗的基本指標，將更進一步地利用做出來的成果編製兩岸華語文教學詞彙的對照手冊，標示詞彙在兩岸的能力測驗中等級所屬，這個成果將是未來華語教師與外籍人士學中文的重要參考資料。

2. 選詞來源

對於哪些詞彙是基本詞彙，是學習華語者應該先學的，至今並沒有定論。到目前爲止，葉德明[1995]是台灣華語文詞彙等級劃分主要的文獻，也是我們分級的一項重要參考。不過，因爲所蒐集語料的教材特殊性及其採用的統計方式，使得有些罕用字或詞反而比常用詞頻率高，如：合資、秋瑾、郭子儀（專有名稱）、朱、湘、玄（黏著詞素）等。所以，只由華語教材去分析詞彙、統計頻率，會產生許多不足的地方。劉英林等[1992]對漢語教學字詞的統計與分級則是做了完善、全面的研究，然宥於兩岸詞彙用法差異，以及該詞彙等級製作時間較早，有些常用詞彙如：嘴巴、支票、傳真、現金、房租、發票、房東、菜單等等出現在該詞表的丁級詞（即高級詞）並不符合現在華語教學所需。

儘管採取頻率做爲常用詞的依據爲部份人士所詬病，認爲頻率高低多取決於語料庫的文本，也有人主張語言既是用來溝通，詞彙的學習該以情境（或功能）爲主要內容。

然而後者所主張的，仍然需要機制來判定什麼是基本的表達方式，什麼是複雜的表達，以符合語言學習的原理。所以折衷的方式，我們還是採取頻率爲詞彙排序的主要依據，但爲避免一般語料庫書面語料偏多的態勢，我們在選詞的來源上做了較廣較全面的收集，如表一所述。最後，透過統計、排序、分級原則出來的結果，再邀請專家學者討論修正完成。

選詞的來源除了台灣師大華語文能力測驗詞彙分級表外[張莉萍 2002](以下簡稱 CPT 詞彙表)，我們還收集了中研院詞庫[中研院詞庫小組 1998] (以下簡稱 CKIP 詞彙表)、漢語水平詞彙等級大綱[國家對外漢辦 1992] (以下簡稱 HSK 詞彙表)，以下就所有選詞的來源及其詞數列表如下。

表一　選詞來源一覽表

來源	類別	詞數	內容說明
CPT2002 詞彙表	初等	1,507	依台灣初級教材各種情境編輯
	中等	3,684	依台灣中級教材各種情境編輯
	高等	4,964	依台灣高級教材編輯，以新聞題材居多
CKIP 詞彙表	核心詞彙庫	15,595	收錄在五本辭典[1]的詞項且出現在中央研究院平衡語料庫 10 次以上的詞。
	通用詞彙庫	13,371	1. 選詞原則：收錄在三本辭典以上的詞項且出現在中央研究院平衡語料庫 4 次以上。 2. 詞數 13,371（已除去核心詞彙庫的詞）。
	參考詞彙庫	6,865	1. 選詞原則：(1)收錄在三本辭典以上的詞項；(2)收錄在所參考的任一本辭典中且出現在中央研究院平衡語料庫 10 次以上的詞。 2. 詞數 6,865（不包含前兩個詞彙庫的詞）。
HSK1992 詞彙表	甲級	1,033	相當於初級
	乙級	2,018	相當於中級
	丙級	2,202	相當於高級
	丁級	3,569	相當於優級

[1]這五本辭典二部來自台灣（中研院詞庫、教育部國語會辭典）、二部大陸（現代漢語辭典、信息處理用現代漢語常用詞詞表）、一部爲華語學習者所編的字典（ABC漢英字典）。

3. 研究方法

由於本研究對於基本辭彙的認定或哪一個詞應該在哪一級完全是以華語爲第二語言學習爲目標，因此 CPT 詞表和 HSK 詞表是我們最主要的參考來源。究竟這兩個詞表的交集或是覆蓋的情況如何？可以從表二看出端倪。

表二　CPT2002 vs. HSK 等級詞彙的交集

HSK ∩ CPT2002		CPT 詞彙表（10,155 個）		
		初(1,507)	中(3,684)	高(4,964)
HSK 詞彙表（8,767 個）	甲(1027)	683	231	17
	乙(2003)	330	805	265
	丙(2188)	141	486	474
	丁(3569)	54	406	632

HSK 甲、乙、丙、丁各級詞數和華語文能力測驗初、中、高三級詞彙表所收集的各級詞數分列於表二。以 HSK 的甲級詞和 CPT 的各級詞彙爲例，同時出現在 HSK 甲級和 CPT 初級的詞彙共有 683 目詞，出現在中級的詞彙有 231 目詞，而同時出現在 HSK 甲級和 CPT 高級的詞只有 17 個詞，以此類推。可以推算出，HSK 的甲級詞覆蓋 CPT 初級詞的比率爲 67%（683/1027）；而 CPT 的初級詞彙覆蓋 HSK 甲級詞的比率爲 45%(683/1507)，也就是 CPT 初級詞出現在 HSK 甲級詞中的比率也有 45%，依此類推。

3.1 詞彙的重量

綜合表一十個選詞來源，經過篩檢、合併以後，共得 40,402 目詞，然後給每一個詞加權值和頻率值，算出總值以作爲排序的依據。加權值的給法如下：

1. CKIP 詞庫分成核心詞彙、通用詞彙和參考詞彙，我們分別指派 4,*3*,*2* 的值。
2. HSK 甲級詞、乙級詞、丙級詞和丁級詞彙等也分別給予 *4*,*3*,*2*,和 *1* 的值。
3. CPT 初級詞、中級詞和高級詞分別給予 4,*3*,*2* 的值。

頻率值則以中研院 500 萬詞平衡語料庫的頻率做爲排序的最主要依據。不過我們不

採用詞的絕對頻率，而採用這些詞的相對頻率[2]。相對頻率是以每個詞在語料庫出現的次數除以所有詞出現次數的總和。以「喜歡」爲例，其在平衡語料庫出現的次數爲 1993 次，所以，「喜歡」的絕對頻率就是 1993，而其相對頻率爲 0.000406(1993/4908867)。

雖然頻率的高低一般是我們在定義詞的等級時最重要的依據，但是因爲目前所根據的語料庫仍以書面語居多，爲了彌補此缺憾，我們進一步將平衡語料庫中「口語」形式的文章（包括演講、電視談話性節目及大學生日常的對話內容）、國小課本等語料抽出，分別統計詞頻產生二個詞表，並算出其相對頻率[3]。

以「喜歡」爲例，表三簡單地表達了這個詞的總值是怎麼得來的。

表三　「喜歡」的總值說明

詞項	詞的來源	Weight/Value
喜歡	CPT 初級詞彙	4
	HSK 甲級詞	4
	中研院詞庫(核心詞彙)	4
	平衡語料庫	0.000406
	國小課本（國編版）	0.000145
	口語語料庫	0.001188
總值		**12.001739**

3.2 詞彙的分級

詞表中每個詞有了它的重量值，再依其值由大至小排序得出每個詞的排名。接下來就是分級的工作了。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爲依據，如柯華葳（2003）依照詞的累計頻率值爲 75%，85%，90%，95%，將詞彙分成 5 個等級。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。

[2]這裏的相對頻率是根據要比對的詞表中所有詞數的總和計算出來的。

[3]這二個語料的詞數：(1)「口語」形式的文章，共 9068 目詞；(2)國小課本(國編版)，共 8349 目詞。

加以每一個等級的詞彙數量參差不齊，例如前者的五級詞彙量分別是 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。

而本研究爲詞彙的分級目的很明確，分級是爲了給華語教學學程參考，同時作爲華語能力測驗的基本骨幹。因此在界定分級詞彙量時，主要以學習者爲依歸，所以訂出初級語言能力詞彙量約 1500 個，中級能力詞彙量爲 5000 個，高級能力詞彙量爲 8000 個。也就是依詞表中每個詞的總值排序，取前 1500 名的詞爲初級詞彙、第 1501-5000 名的詞彙爲中級詞彙，第 5001-8000 名的詞彙爲高級詞彙。之所以取詞彙量八千，是採取鄭錦全（1998）「詞涯八千」的概念，從計量的觀點來解釋人類的語言認知能力。

4. 詞彙處理原則

這一節主要是簡單陳述我們在處理詞表時所做的一些處理。由於每個詞表對詞的分類、詞的認定等方面或多或少有些差異，於是我們必須做一些統整。

4.1 詞彙前置處理（Pre-Processing）

首先，在進行各個詞表比對之前，先審視所有的詞表，將所有專有名詞（包括人名、地名）、由地名組成的詞彙，如：法國菜、日本料理等、和由數字組成的所有詞組，如：一百三十二，三成三，百分之二十，三點三十分，一九九九年等，自詞彙表中移除，然後再針對不同詞表來源、不同狀況做前置處理。

4.1.1 CPT 詞表處理

利用中研院詞庫小組的斷詞程式，自動將所有的詞彙標上詞類，而多重詞類的部分，則利用人工方式一一標上。另外針對一些內部結構不同的詞項，給予不同的處理方式，其處理步驟分列如下。最後，再依其出現在初級、中級、高級詞表中，給予 4,3,2 等值。

1. 詞組的處理：

因爲本詞表是直接從教材中抽取，所以收錄了許多詞組，尤其是在高級詞彙表中，例如：「金融性衍生商品」、「二十出頭」等。針對詞組的部分，首先依據斷詞原則，依類、依詞斷開。然後再依此比對這些詞是否已經在詞表中，如果沒有的話，則依其

詞組原來出現的等級來存放這些詞彙。再以高級詞彙表中的「金融性衍生商品」爲例，依詞依類將其斷開之後，變成「金融(N)性(N)衍生(V)商品(N)」，經過比對之後發現「金融」和「商品」已在中級的詞彙表中，但找不到「衍生」這個詞，因爲這個詞組是在高級詞彙中，因此，將「衍生」增至高級詞彙表中。

2. 詞項增補：

有些詞項是經由構詞的規律所產生或衍生，因此存有「基本」詞和「衍生」詞的關係，但是在本詞表中可以發現許多「衍生詞」卻找不到其相對應的「基本」詞項，爲了彌補此缺項，我們依其不同結構分別審視察看有無缺漏並將其補上，增至適當等級的詞彙表中。

(1)動補結構：在詞表中也可發現動補結構「長不大」，卻找不到「長大」這個詞，針對這種情形，我們一一將其補上。

(2)包含詞綴、接頭詞/接尾詞的詞項：在詞表中也發現不少包含詞綴、接頭/接尾詞的詞項，卻找不到其「原型」的詞，例如：中級詞彙表中有「海外部」，但是沒有「海外」，因此增加「海外」至中級詞彙表中。

(3)重疊結構：同樣地，我們也可以發現只有重疊形式的詞，卻找不到相對應的「基本詞」，例如：在高級詞彙表中可發現「急急忙忙」，但無「急忙」，因此將其增至詞表中。

3. 等級的調整：

針對學習順序的合理性來調整詞彙的等級，例如：「成長率」和「成長」，前者出現在中級詞彙，而後者反而出現在高級詞彙中，在學習的過程中，應該是先學習較基本的詞彙，再學習具有構詞或衍生性的詞彙，因此，將「成長」從高級詞彙表中前移至中級詞彙。「單親」（高級詞彙）和「單親家庭」（中級詞彙）的情況亦同，將「單親」移至中級詞彙表。

4.1.2 HSK 詞表處理

雖然大陸漢語等級詞彙大綱(HSK)已經將詞分成四個不同等級的詞表，並標上詞類，但是因爲要同時比較不同的詞表，所以必須先將本詞表詞類的標示與其他詞表的詞類標示一致，例如，本詞表中的（形容詞）可對應至中研院詞庫的（狀態動詞），因此將其（形容詞）的標示改爲（狀態動詞）。另外，HSK 的詞表中有些項目是以固定形式出現，如：「一面...一面」，我們也另加「一面」這個詞於詞表中。最後，依其甲乙、丙、丁級詞表分別給予 4,3,2,1 等值。

4.1.3　中研院詞彙庫

中研院的詞彙庫依其選詞原則訂出核心詞彙庫、通用詞彙庫和參考詞彙庫。選詞標準是選取核心、通用詞彙庫全部的詞。以及參考詞彙庫中頻率 10 以上的詞項。所有抽取出來的詞，依其詞彙庫的不同分別給予 4,3,2 的值至核心、通用和參考詞彙庫中。經過篩選、合併之後，核心詞計有 15595 目，通用詞計有 13371 目，參考詞有 6865 目。

4.2 詞彙後置處理（post-processing）

除了依處理步驟自動比對、抽取常用詞項、分級以外，在初步的詞表產生之後，我們還針對詞項的不同特色來訂出一些選詞或過濾詞項的操作性原則，包含詞項的等級調整、合併原則和詞項的增刪原則。舉例如下：

1. 當以一般原則自動抽取出的詞項之間彼此競爭時，以功能詞、口語用語、白話文用語等優先選擇。

2. 詞項合併和等級調整方式：詞項的合併和等級調整，遇到同類詞語不一致處，通常將低頻詞併入較高頻詞。

5. 研究成果

經過前述選詞、詞彙排序分級原則等處理，CPT 詞彙分級表與 HSK 詞彙等級大綱的交集分布如表四所示。

表四、CPT2005 vs. HSK 等級詞彙的交集分佈

HSK ∩ CPT 2005		CPT 2005 詞彙表（8,000 個）		
		初(1,500)	中(3,500)	高(3,000)
HSK 詞彙表（8,822 個）	甲(1,033)	959	153	9
	乙(2,018)	354	1399	191
	丙(2,202)	32	899	854
	丁(3,569)	13	438	686

以 HSK 的甲級詞和 CPT 的各級詞彙爲例，同時出現在 HSK 甲級和 CPT 初級的詞彙共有 959 目詞，出現在中級的詞彙有 153 目詞，而同時出現在 HSK 甲級和 CPT 高級的詞只有 9 個詞，以此類推。

此研究成果包含一個分級的詞表，在這個詞表中，除了詞項，我們還標示了每個詞的等級、拼音、語法類、英文解釋和例句，並對照 HSK 中的等級，以「喜歡」爲例，其形式如表五：

表五、詞表範例「喜歡」

喜歡	xi3huan1	初級	like, love, be fond of, be happy/elated	SV (狀態動詞)	學生們都非常喜歡這位老師。	HSK 甲級

因爲每個詞皆依他們的用法來計算其使用次數、頻率，因此所有抽取出來的多義詞，也會依照其不同用法而有不同的分類及標示，以「打」爲例：

表六、詞表範例「打」

打	da3	初級	strike, hit, fight, construct, forge, mix	V (動作動詞)	外面下起大雨，雨水不停地**打**在窗戶上。 他每個月打一次電話回家。 請問這件衣服是打 5 折還是 6 折。	HSK 甲級
打	da2	中級	dozen	M （量詞）	爲了做茶葉蛋，媽媽一口氣買了三打雞蛋。	
打	da3	中級	from	Prep （介詞）	當冬天的太陽緩緩上升時，也讓人**打**心裡頭暖和起來。	HSK 丙級

此詞表特色：

1.涵蓋華語學習者和母語使用者的常用詞彙：

選用台灣華語文能力測驗詞彙分級表和大陸漢語水平辭彙等級大綱、以及中研院平衡語料庫和台灣、大陸以及針對母語爲英語的華語學習者所制定的辭典中所抽取而得的辭彙表，因此，本詞表的選詞來源同時兼顧了華語學習者和母語使用者的常用詞彙。

2.兼顧口語和書面用語：

選詞來源涵蓋範圍廣，並具代表性，避免了一般語料庫偏書面語，少口語表現的詞彙。

3 兼顧詞彙的廣度與深度

除了詞彙的「形式」頻率統計以外，每個詞再依他們不同的用法來分別計算其使用次數、頻率。因此不但以詞彙的數量計量，部分詞彙不同的語法表現皆一併納入計算。並依其不同的詞重(weight)分別放入不同的等級。以「口」這個量詞爲例，當它計量

跟嘴巴有關的動作時，出現在初級，如：吃一口飯，喝了一大口水。當它計量言語或牙齒時，分佈在中級，如：說了一口標準的法語，一口漂亮的牙齒。計量井、刀劍、鐘或牲畜、人口時，出現在高級，如一口井，兩口豬，一家六口。

致謝

本研究部份承國科會專題研究計畫補助，特此致謝（NSC-92-2411-H-003-045）。 本論文於 2005 年 4 月於廈門發表，承國立台灣師範大學學術活動部分經費補助（T9407000249），僅此致謝。

參考書目

Jenkins, J.R. & Dixon, R., 1983, Vocabulary learning, *Contemporary Educational Psychology*, 18: 237-60

Stubbs, Michael, 1986, Language Development, Lexical Competence and Nuclear Vocabulary, In Kevin Durkin, ed. *Language Development in the School Years*. Croom Helm.

中研院詞庫小組 (1998a)，*詞頻辭典*，中央研究院中文詞知識庫小組技術報告 CKIP-98-01。台北南港：中央研究院資訊所。

中研院詞庫小組 (1998b)，*詞頻統計*，中央研究院中文詞知識庫小組技術報告 CKIP-98-02。台北南港：中央研究院資訊所。

中研院詞庫小組 (1998c)，*中央研究院平衡語料庫的內容與說明*，中央研究院中文詞知識庫小組技術報告 CKIP-98-04，台北南港：中央研究院資訊所。

杜爾文(Dew, James Erwin) (1999)，*六千個詞：中文詞彙頻率手冊* (A Vocabulary Frequency Handbook for Chinese Language Teachers and Students)。台北：南天書局。

柯華葳等（2003），*華語文能力測驗編製：研究與實務*。台北：遠流出版社。

國家對外漢語教學領導小組辦公室漢語水平考試部（1992），*漢語水平考詞彙與漢字等級大綱*。北京：北京語言學院出版社。

張莉萍（2002），*華語文能力測驗理論與實務*。台北：師大書苑。

葉德明 (1995) *華語文常用詞彙頻率等級統整研究*（附錄一），行政院國家科學委員會專題研究計畫成果報告，民國 84 年。

劉英林、宋紹周（1992），論漢字教學字詞的統計與分級，*漢語水平考詞彙與漢字等級大綱*。北京：北京語言學院出版社。

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，86 年 9 月，42-47 頁。

鄭錦全(1998)，從計量理解語言認知，In Benjamin K. T'sou et al. eds. *漢語計量與計算研究*，15-30 頁，香港城市大學。

A preliminary approach to grading vocabulary of Chinese as a second language

Li-ping Chang　　Feng-yi Chen
lchang@ntnu.edu.tw　　a0410613@ntnu.edu.tw
Mandarin Training Center, National Taiwan Normal University
162, Sec.1, Hoping E. Rd., Taipei 106, Taiwan

Abstract

This research aims to develop as well as rank the contents of vocabulary of Chinese as a second language. First, we collect the vocabulary mainly from CKIP word bank, CPT word bank and HSK word bank. Then, we use the frequency value and the weighting method to rank order of each word. And finally set up the first 1500 words for the basic level for teaching Chinese as a second language, the next 3500 words for the intermediate and the next 3000 words for the advanced learners. The output result is not only listing the word itself but also giving its English translation, ranking record, pronunciation, part of speech and one or two sentence examples for the word. The most important in the result is comparing the rank with HSK. Since the sources of words collection are various as well as representative, the output seems to be balanced instead of being short of spoken language when using the large corpus in Taiwan. We believe the word list will become the basic reference for editing Chinese teaching materials and implication of Chinese proficiency test in the future.

Key Words: teaching Chinese as a second language, weight, grading, CPT, HSK, CKIP