行政院國家科學委員會專題研究計畫成果報告

建立常用繁體漢字部件表音功能資料庫

A Tabulation of Phonetic and Phonological Properties in Chinese Character Components

計畫編號：NSC 89-2413-H-002-007

執行期限：民國 88 年 08 月 01 日至 89 年 07 月 31 日

主　持　人：黃榮村　　台灣大學心理系
共同主持人：邱耀初　　東吳大學心理系
計畫參與人員：許鶴鐘，蔡明家　　東吳大學心理系
王傳華　　台灣大學心理系

一、中文摘要

本研究在中研院漢字常用字集及字頻基礎上，建構出一主資料庫：漢字部件與字音資料庫，及一衍生資料庫：部件表音功能值資料庫。此為一使用者介面型的資料庫，可提供部件字音屬性分類及檢索功能。資料庫主要特色為： 1.在使用上為一開放式系統，可供使用者自行研發新的系統。 2.較過去資料庫所提供的訊息更為豐富、仔細，就評估過去漢字辨識研究實驗材料之目的而言，有助於澄清刺激字中的混淆變項，並在此基礎上檢討實驗結果何以歧異之理由。

關鍵詞：漢字字頻資料庫、部件表音功能、開放式系統

Abstract

A tabulation of phonetic and phonological properties in Chinese character components was performed to construct the main database and a generated one. This is a user-friendly interface so that phonetic and phonological attributes can be easily classified and indexed. The construction can obviously help obliterate the confoundings that might have been existed in running Chinese character recognition experiments. Looking up to this tabulation carefully can also help reanalyze and resolve the inconsistencies that puzzled separate research groups on interpreting disparate directions of the same targeted experimental effects.

Keywords: frequency-count database of Chinese characters, phonetic and phonological manifestations of Chinese character components, open system

二、緣由與目的

就文獻中部分資料庫來分析　韓布新(1993)提供的部件資料庫為簡體字，陳學志（1996）雖是採繁體字建立部件資料庫，但僅計算部件與組合結構頻率，並未計算本身即可構成字的部件組合頻率　及邊際頻率（賴惠德與黃榮村，1997）。另一採繁體字建立的部件資料庫為賴惠德與黃榮村（1997）所提出之「漢字部件、部件組合與部件位置頻率之統計分析與比較」資料庫，此資料庫兼採上述兩個資料庫的一些特性，且擴充一些關於部件的功能(如：計算部件邊際頻率），其所採用的漢字常用字集則來自中研院中文知識庫小組（1995）。

但上述資料庫卻存在以下三項缺點：1.缺乏對部件表音功能的整理。2.部件分類與部件組合分類尚未落實於部件之頻率計算中。3.非使用者介面，設計缺乏彈性屬於封閉系統，僅提供分析結果，使用者無法根據自身需求對資料庫進行各式分析，以獲取適當實驗材料。

本報告所建構常用漢字部件之表音功能資料庫，則以中研院中文知識庫小組（1995）漢字常用字集為基準，建立一使用者介面型資料庫，提供部件字音屬性分類及檢索功能。

三、結果與討論

本計畫在中研院漢字常用字集及字頻基礎上，分別建構出一個主資料庫－漢字部件與字音資料庫（範例見表一）及一個衍生的部件表音功能值資料庫（範例見表二）。

（一）主資料庫－漢字部件與字音資料庫：

1. 增添字音:主要先為 5655 個常用字增添字音（聲介韻調）訊息。
2. 部件拆解：將漢字依部件組成方式分為：上下、上下中、左右、左右間、內外、獨，共六類。並標示出每個部件的所在位置(上下中左右間內外獨)，同時計算出所有漢字筆畫。
3. 以部件為主要之索引，羅列出使用同一部件的漢字，並標示所在位置。

（二）衍生資料庫－部件表音功能值資料庫：

表音功能值在本計畫中是指：看到元件 A 時（如申），激發出 B 音（如ㄕㄣ）的強度，是部件表音功能高低的指標，本資料庫的表音功能數值計算方式主要有二：

1. 以字數為計算單位：

$$\frac{\text{具有相同元件，且發音相同的漢字總數}}{\text{具備同元件的漢字總數}}$$

2.以邊際頻率為計算單位：

$$\frac{\text{具有相同元件，且發音相同的漢字總字頻}}{\text{具備同元件的漢字總字頻}}$$

字音同異的認定寬嚴有別，漢字字音一般被界定為由三部分組成：聲、韻、調，嚴格標準下的「同音字」，係指「聲韻調」三者皆同，寬鬆標準下的「同音字」則只需「聲」、「韻」其一相同即可，一般則僅以聲韻相同作為同音與否的指標。

若以聲韻相同作為同音與否的指標，以部件「申」為例，根據「漢字部件與字音資料庫」可知漢字中含「申」部件字有 8，其中有 6 個發ㄕㄣ音；其表音功能值以字數計算為 6／8＝0.75；以邊際頻率計算為 11634/13317=0.87；又有一個發ㄎㄨㄣ音，其表音功能值以字數計算為 1／8＝0.13；以邊際頻率計算為 1058/13317=0.08；有 1 個發ㄔㄨㄣ音，其表音功能值以字數計算為 1／8 ＝ 0.13 ；以邊際頻率計算為 625/13317=0.05。

表一：漢字部件與字音資料庫範例

字	字頻	組合	部件	位置	聲	介	韻	調	聲介韻	韻母	筆畫數
伸	790	左右	申	右	ㄕ		ㄣ		ㄕㄣ	ㄣ	7
紳	112	左右	申	右	ㄕ		ㄣ		ㄕㄣ	ㄣ	11
呻	16	左右	申	右	ㄕ		ㄣ		ㄕㄣ	ㄣ	8
砷	11	左右	申	右	ㄕ		ㄣ		ㄕㄣ	ㄣ	10
神	5566	左右	申	右	ㄕ		ㄣ	ˊ	ㄕㄣ	ㄣ	10
坤	1058	左右	申	右	ㄎ	ㄨ	ㄣ		ㄎㄨㄣ	ㄨㄣ	8
暢	625	左右	申	左	ㄔ		ㄤ	ˋ	ㄔㄤˋ	ㄔㄤ	14
申	5139	獨	申	獨	ㄕ		ㄣ		ㄕㄣ	ㄕㄣ	5

表二：部件表音功能值資料庫範例

部件	聲介韻	同音字數	總字數	字數一致	同音字頻	總字頻	字頻一致
申	ㄕㄣ	6	8	0.75	11634	13,317	0.87
申	ㄎㄨㄣ	1	8	0.13	1058	13,317	0.08
申	ㄔㄤ	1	8	0.13	625	13,317	0.05

四、計畫成果自評

以下分別從使用者介面特色、支援漢字辨識研究功能及本資料庫限制與未來改善，試說明之。

（一） 使用者介面特色：

由於本資料庫建立在高普及率軟體 Excel 上，並且是一開放式系統，故為一符合使用者介面的資料庫。以下分別詳述之：

1. 應用高普及率軟體：就作為使用者操作本資料庫方便性而言，本資料庫建立在微軟 Excel 軟體上，由於這是個相當普遍的軟體，就推廣與使用性而言相當便利，任何人只要對 Excel 有所認識，便能進行簡單地操作。
2. 開放式系統：本資料庫於設計之初即朝開放式系統的方向努力，使用者可以因自己的需要針對漢字部件的相關訊息做不同型態的組合與操作，而形成一個適合自己研究用的資料庫，本研究的衍生資料庫即是應用此開放系統的特性發展而來。

（二） 支援漢字辨識研究功能：

由於本資料庫提供可能與漢字辨識有關的訊息，較過去資料庫更為豐富、仔細，就作為重新評估過去漢字研究實驗材料之目的而言，應有助於澄清刺激字中的混淆變項。

邱耀初與許鶴鐘（2000）即根據本資料庫所提供的部件表音功能值，檢定吳瑞屯、周泰立與劉英茂（1993）的研究，發現其實驗材料潛藏一些混淆變項。並依此設計實驗，指出刺激材料篩選不當，可能為過去漢字「語音轉錄」研究爭議來源之一。未來漢字辨識研究的一些爭議，相信將可因本資料庫的完成而有所降低。

（三） 資料庫限制與未來發展：

1. 有些部件拆開後，不見於微軟的字庫系統，如果另行造字又唯恐造成轉換與計算時轉碼的問題，因之一律以代號表示，比如「牽」字被歸入上下字，上半部是以「缺 79」代之。在閱讀上雖然有所不便（不符直觀），但兩害相權取其輕。
2. 為便於資料庫之設計與使用，並符合一般漢字結構，本資料庫至多僅將漢字區隔成三部分，但是此一決定仍必須顧及以下二個問題：

(1)漢字辨識的單位是否僅止於此，這必須由相關實徵研究來回答。因之有必要在理解目前此類研究成果後，評估是否需做細部切割以擴充資料庫的效度。

(2)最初計畫原意是希望部件的拆解與文字聲韻學結合，但有部分特殊的漢字卻擁有四個以上的部件（例如：夢），又或者部件數雖然在三個以內，卻又不是以上

中下或左右間的順序構成（如：品）因之，增加代號與新部件的使用，這只是權宜之計，有待進一步的解決。

3. 目前有必要與其他資料庫做相容性的比對，作為扣除一些取樣偏誤的依據。

五、參考文獻

吳瑞屯，周泰立，劉英茂（1993）。 中文單字辨識的直接歷程。 華文世界，69期，8-16。

邱耀初與許鶴鐘（2000）。 從語音轉錄研究看漢字辨識：平行激發或序列式處理。 中華心理學會年會發表論文，台北市。

紀春興、黃居仁與陳克建（1995）。 注音檢索現代漢語字頻表。 中研院中文知識庫小組，台北市。

陳學志（1996）。 中文常用字之部件與組合結構頻次分析。 中華心理學會年會發表論文，台北。

賴惠德與黃榮村（1997）。 漢字部件、部件組合與部件位置頻率之統計分析與比較。 中國語文認知處理研討會發表論文，香港。

韓布新(1994)。 漢字部件信息資料庫的建立。 心理學報，卷 26（2），147-152。

韓布新(1995)。 部件組合-潛在的漢字結構層次。 中文信息學報，卷 9(3)，27-32。