Computational Linguistics and Chinese Language Processing
Vol. 17, No. 3, September 2012, pp. 29-44

聲符部件排序與形聲字發音規則探勘

Phonetic Component Ranking and Pronunciation Rules Discovery for Picto-Phonetic Chinese Characters

張嘉惠*、林書彥*、蔡孟峰*、李淑萍+、廖湘美+、黃鍔#

Chia-Hui Chang, Shu-Yen Lin, Meng-Feng Tsai, Shu-Ping Li,

Hsiang-Mei Liao, and Norden E. Huang

摘要

近年來台灣有相當多的新移民的加入，這些新移民在口語的學習上雖然有地利之便，但是在漢字的認識上則是相當弱勢。由於漢字乃是圖形文字，學習單一字的成本相對的高。如果可以讓漢字教一個字，可以學到十個字，對於漢字教學的成效應有相當的助益。本文從部件教學的概念出發，考慮聲符的發音強度、出現頻率、及筆劃數，做爲聲符部件教學順序的準則。我們利用部件發音強度(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)，以線性加總、幾合乘積、及調和平均三種方法對部件排序。根據此部件排序學習，前五個部件便可延伸學習多達 140 個相似發音的漢字。進一步，我們應用中研院文獻處理實驗室所建立的「漢字構形資料庫」，以及標記所得之形聲字，拆解形聲字組成的部件，挖掘串連漢字之間關係的形音關聯規則。我們從 600 萬條發音規則中篩選與分群出 3 組高信賴度與 5 組高支持度的規則，並藉由這些規則來輔助漢語發音的學習，提高學習效率。

關鍵詞：形聲字、聲符強度、部件教學、學習曲線、關聯規則

* 國立中央大學資訊工程所 Dept. of Computer Science and Information Engineering, National Central University, Taiwan
E-mail: chia@csie.ncu.edu.tw
The author for correspondence is Chia-Hui Chang.
+ 國立中央大學中文系 Dept. of Chinese Literature, National Central University, Taiwan
# 國立中央大學數據中心 Research Center for Adaptive Data Analysis, National Central University, Taiwan

Abstract

In recent years, there are a considerable number of new immigrants in Taiwan. Although these people are in the good position to learn Chinese, the advantages are limited to speaking and listening. Recognizing Chinese characters is a tough task since one has to memorize the shape, meaning and pronunciation at the same time. Therefore, the cost of learning a single character is relatively high compared with other languages in alphabet system. The goal of this study is to make the 80% pictophonetic characters to be organized more systematically such that the pronunciation of most pictophonetic characters can be inferred automatically. We evaluate the importance of Chinese components by considering the pronunciation strength, occurring frequency, and number of strokes using linear sum, product, and harmonic mean, respectively. Furthermore, we discover pronunciation rules by association mining with priority grouping. Three groups of high reliability rules and five groups of high support rules are demonstrated in this paper to show the effectiveness of pronunciation rule discovery.

Keywords: Picto-phonetic Character, Pronuciation Strength of Phonetic Component, Component-based Teaching Method, Learning Curve, Association Rule

1. 簡介

漢字是世界上最古老的文字之一，也是至今仍廣爲使用一種形系文字。近年來由於中國市場的興起，以華語做爲第二外語的學習也連帶地愈來愈受到重視，華語學習者的人數也倍數成長，據 China Daily 2010 的文章指出，目前全世界超過四千萬的非華裔人士正在學習華語文。由此可見未來華語文學習市場的龐大需求；再者，台灣近年來外籍與大陸配偶的人數從 2002 年的二十三萬人成長至今四十四萬人，其中外籍配偶約十四萬六千多人，已取得國籍者約九萬人，在在顯示了漢語學習的重要性。

過去學習漢語只能靠資深的中文老師的教導或是學習者慢慢累積經驗，不僅對於海外華語師資的培育緩不濟急，對於學習者而言更是一條漫長的路。然而，漢語字形讀音繁複，初學者並不易掌握學習要訣，尤其漢語的發音更是複雜多變。事實上華語作爲第二語言的學習，比起英文作爲第二語言的學習更是難上許多，因爲漢語的字形與音調相較拼音文字複雜，學習者要同時進行形、音、義三者的連結，如果沒有適當的學習方法，個別漢字的學習成本相當高。比起傳統的拼音拉丁文字，即使會說華語的海外華人對於漢字的認識也可能相當有限。其最主要的原因在於漢字是圖形文字(pictograph system)，無法像英文等拼音文字(alphabet system)一樣，一旦學會拼音方法(phonetic representation)，即有基本的閱讀能力。相較之下，一般漢字學習者讀寫的學習進展則會比較緩慢，而且必須搭配注音符號(Chinese phonetic symbols)或是其他拼音方法，才可知道每個漢字的發音。這樣的限制，對於漢字的學習相當不利，這也是爲什麼二十世紀初期中國大陸欲將

漢字拉丁化的主要原因。

漢字的構成包含象形、指事、會意、形聲、轉注、假借(總稱六書(許慎，1999))。據統計資料，7000 個現代漢語通用字中，屬於「形聲」結構的有 5631 個，約佔總字數的 80.5%，這麼多的形聲字在整字的組合上，多數採用「1+1」的方式，也就是一個意符加上一個聲符。基於這樣一個語言事實，我們可以借助部件教學，充分發揮部件的組合關係強化學習者對於漢字的識記。但如何折衷構字能力強度與發音強度，篩選或排序聲符部件則是本文主要探討的研究議題。

本篇論文中，我們應用(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)，以部件發音分佈的集中性計算聲符強度，加以部件延伸字數及筆劃數的考量，提出線性加總、幾合乘積、及調和平均三種結合方法，對部件加以排序。利用此排序做爲漢字部件教學的順序，可以幫助學習者在短時間內提高閱讀效率。我們以累計延伸字個數做爲學習成效的比較，發現有效的排序，可以在學習完前五個部件，便可藉此延伸學習多達 140 個具有高度相似發音的漢字，同時累計筆劃數也是可以接受的範圍，顯示適當排序的重要性。

除了考量聲符部件學習順序之外，我們也試圖分析漢字發音規則，做爲學習發音的參考。爲了要產出易懂的發音規則，讓中文的學習者可以應用形聲字的特性來推測漢字的發音，在本文中我們應用關聯規則探勘挖掘形聲字發音所存在的規則。我們應用中研院文獻處理實驗室所建立的「漢字構形資料庫」，拆解其組成的部件，挖掘串連漢字發音關係的形音關聯規則，來輔助學習者學習，讓漢字不是教一個字才學到一個字，而能搭配關聯規則「一舉數字」，發揮數位學習的優點。我們從 600 萬條發音規則中篩選與分群出 8 條高信賴度與兩組各約 10 條高支持度的規則，並藉由這些規則來輔助漢語發音的學習效率。

2. 相關研究

最早有關漢字構造的研究，應屬中央研究院資訊科學研究所文獻處理實驗室，從 1993 年開始，陸續建構古今文字的源流演變、字形結構及異體字表，做爲記錄漢字形體知識的資料庫，也就是漢字構形資料庫(中研院文獻處理實驗室)。漢字構形資料庫不僅銜接古今文字以反映字形源流演，也記錄了不同歷史時期的文字結構。另外也由於開發漢字部件檢字系統，得以解決缺字問題。然而漢字構形資料庫過去的研究著重在字形知識的整理，尚未涉及字音與字義的處理；因此文獻處理實驗室近年來開始文字學入口網站建置計畫(莊德明、謝清俊，2005；莊德明、鄧賢瑛，2008)。一如其文所述：“漢字構形資料庫目前只著重在字形知識的整理，尚未涉及字音與字義；建立一個形、音、義俱備的漢字知識庫，仍是我們長遠的目標”。因此本論文的目的即是以挑戰漢字的發音規則知識庫爲出發，除了了解漢字發音規則外，也希望藉由此項研究找出一套形聲字發音轉換規則，讓華語學習者可以在聲符與規則的輔助下，順利讀出字的發音出來。

與本研究最爲相關的研究計畫是淡江大學中文系高柏園、郭經華、胡映雪等教授所

主持之“字詞教學模式與學習歷程研究”。其概念是藉由即時回饋的寫字練習(學文Easy Go!)，比較部件拆解做爲漢字教學策略成效(洪文斌，2010)，輔以線上教學平台「IWiLL Campus」（郭經華，2010），進行「以字帶詞」之詞彙學習策略（高柏園，2010）。此計畫在美國加州地區 Saratoga High School 針對 26 名修習 AP 中文課程之學生，實施四週約八堂之主題課程，用以評估漢字部件教學之學習策略對於海外華語文學習者之成效。從其國科會期中報告顯示，採用多媒體自習一組的學生在認字、書寫、及字的結構上，比傳統標示筆劃順序的習字方法呈現較佳的成果，顯示以部件拆解做爲漢字教學策略的可行性。

張嘉惠等人於 2010 年提出了兩種自動化判定形聲字聲符的方法(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010) ：其一是藉由聲符構件與原字的發音相似度高於非聲符構件與原字的發音相似度的概念，與語言學專家的所制訂聲母與聲母、韻母與韻母之間發音相似度，做爲第一種形聲字聲符的方法。同時也比較採用限制性最佳化技術，求得發音相似度分數。第二種方法則爲構件發聲分佈比較法，藉由聲符構件其衍生字的發聲分佈比非聲符構件的漢字發聲分佈較爲集中的概念，來計算每個構件的發聲分佈與所有漢字的發聲分佈 KL 值，做爲構件做爲聲符的強度。實驗結果顯示，發音相似度比較法在 7340 個形聲字中的判定聲符準確率爲 93.35%，而構件發聲分佈比較法則可達到 98.66%的準確率。雖然形聲字聲符的判定只是過渡性的需求，但是構件發聲強度卻可做爲學習漢字順序的重要參考準則，這也是本篇論文的重點之一。

3. 部件重要性排序

首先我們從部件教學的概念出發，希望對於聲符的教學順序，提出一個考慮聲符發音強度、出現頻率、及筆劃數的排序方法，做爲聲符部件教學順序的準則。由於構件發聲分佈比較法對於判定形聲字聲符有高達九成八的準確率，因此我們此處即採用做爲聲符發音強度。根據(張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔，2010)的定義，每一個部件的聲母發音強度、韻母的發音強度、及調號的發音強度可由下列三式計算而得：

$$I(w) = KL(P_I(W) \,\|\, P_I(A)) \tag{1}$$

$$F(w) = KL(P_F(W) \,\|\, P_F(A)) \tag{2}$$

$$T(w) = KL(P_T(W) \,\|\, P_T(A)) \tag{3}$$

其中 A 表示所有漢字所成的集合，W 則表示部件 w 所延伸的字所成的集合。函數 $P_I(A)$、$P_f(A)$、$P_T(A)$分別表示 A 集合中漢字的聲母、韻母及調的分佈機率。KL(P||Q)則代表兩個機率分佈的 KL-divergence：

$$KL(P \,\|\, Q) = \sum_i P(i) \log \frac{P(i)}{Q(i)} \tag{4}$$

對於聲符而言，由於發音集中度較高，因此 w 的聲母分佈 $P_I(W)$與所有漢字的聲母分佈 $P_I(A)$會有較大的差異。同理韻母分佈 $P_f(W)$與 $P_f(A)$差異，以及聲調分佈 $P_T(W)$與

P_T(A)差異也會較大。因此我們即可以 KL-divergence 公式對此差異值計算出其程度，換句話說我們利用公式 1, 2, 3 分別計算一個部件的聲母、韻母、及調號的 KL 值，這三種數值分別反應出此部件的聲母、韻母、及調號的發音強度。

除了部件的發音強度，在部件學習排序上，我們也必須考慮部件的頻率。因爲對於漢字學習者來說，發音強的部件，也要有一定的出現頻率，才能發揮其做爲聲符的功能。因此若單純以發音強度來決定教學順序，並不是非常適當的選擇。再者，對於學習者來說，漢字的筆畫數多寡也會影響學習的效率。因此如何將三者同時考慮於部件教學的順序，是此處最主要的挑戰。常見的結合方式是以線性加總，然而在此處並非最佳的結合方法，如圖一部件發音強度與頻率散佈圖顯示，若以線性加總發音強度與部件的頻率（部件頻率定義爲包含部件 w 的形聲字字數|W|除以全部字數），可能先找到的是頻率高但發音強度較弱的部件，或是發音強的部件但是頻率較低的部件，而非同時據有高頻及高發音強度的部件。

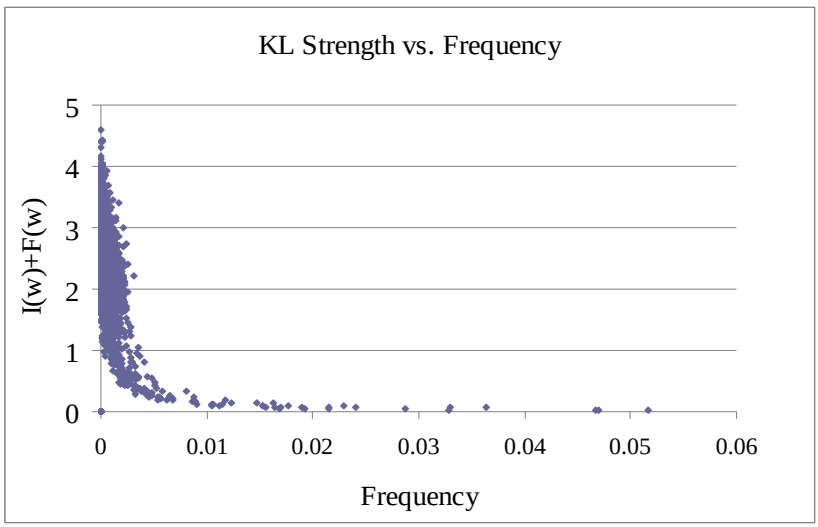


圖 1. 部件發音強度與頻率散佈圖

爲了找出頻率高且發音強度強的部件，且同時也希望能將筆劃數較少的部件優先排序。我們提出三種排序部件的依據：

1. 線性加總：$ScoreA(w)=a*Freq(w)+ I(w)+ F(w) + b*Strokes(w)$
2. 幾何乘積：$ScoreG(w)= Freq(w)*(I(w)+F(w))/\sqrt{Strokes(w)}$
3. 調和平均：$ScoreH(w)=ScoreG(w)/ScoreA(w)$

其中 *Freq(w)*代表部件 *w* 的頻率，*Strokes(w)*爲部件 *w* 的筆畫數；*a* 與 *b* 則是線性加總的權重。由圖 1 可知發音強度約爲頻率的 a=90 倍，同理，我們求得筆畫數的權重 *b=0.01*，可使線性加總的三個因素間取得平衡。第二種結合方法則是三個因素的幾何乘積，最後調和平均則是取線性加總與幾何乘積的調和平均做爲部件排序的評估。加法與乘法是結合不同因數最直接的方法，而調和平均則是取兩者的結合。

3.1 實驗評估

爲了評估三個部件排序是否能有效率地提昇學習效率，我們繪製出以幾何乘積做爲部件排序，與其累積延伸字數的關係[1]。如圖 2 所示，橫軸表示排序過的部件，從左而右依序是：分令丁方干包…等字，縱軸淺色代表累積延伸字的個數 Y_1，縱軸深色則代表聲符能正確預測聲母個數與韻母個數的總和 Y_2，兩者分別定義如下：

$$Y_1 = \Sigma_i |W_i|, \quad (5)$$

$$Y_2 = \Sigma_i \, (\mathrm{Imatch}(w_i, W_i) + \mathrm{Fmatch}(w_i, W_i)) \quad (6)$$

其中 $Imatch(w_i, W_i)$代表部件 w_i 延伸字集合 W_i 中與部件 w_i 具有相同聲母的字數，$Fmatch(w_i, W_i)$代表部件 w_i 延伸字集合 W_i 中與部件 w_i 具有相同韻母的字數。舉例來說若聲符 w_i 爲包(ㄅㄠ)，若其延伸字集 W_i 爲{炮(ㄆㄠ)、胞(ㄅㄠ)、苞(ㄅㄠ)}，那麼$|W_i|$=3，而 W_i 中與 w_i 有相同聲母的字爲{胞、苞}，因此 $Imatch(w_i, W_i)=2$；而 W_i 中與 w_i 有相同韻母爲的字有{炮、胞、苞}，因此 $Fmatch(w_i, W_i)=3$。因此兩者相加後可得正確預測聲母個數與韻母個數的總和=5。

正確預測聲母個數與韻母個數的總和（Y_2）愈接近兩倍累積延伸字的個數（$2Y_1$），表示預測正確的準確率愈高，將上述兩值相除，可得準確發音比例。從圖 2 可以看出排序在前面的字即有相當多的延伸字，同時準確發音的比例也相當的高。表 1 列出排序前十個部件及其可延伸學習的形聲字，如表 1 所示，這些部件都具有延伸字發音高度相似、出現頻率高、筆數少的特性，益於先行學習。

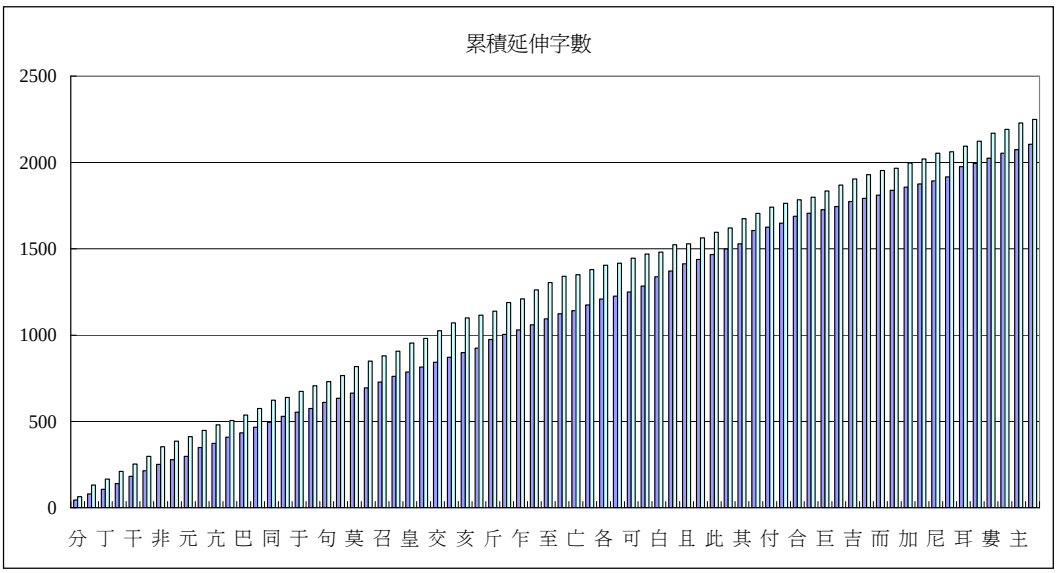


圖 2. 幾何乘積排序與累積延伸字關係

[1] 所有漢字相關資料來源則是使用中研院所開發的漢字構形資料庫。

表 1. 幾何乘積排序之部件

部件 w_i	延伸字 $\|W_i\|$	Y_1	Y_2	準確發音比例	筆劃數	累積筆劃數	延伸字
分	45	45	64	0.71	4	4	份坋坌弇吩玢枌楞棻…
令	35	80	132	0.83	5	9	伶冷坽呤囹岭狢昤泠…
丁	27	107	167	0.78	2	11	仃亭玎可叮盯宁寧玎…
方	33	140	211	0.75	4	15	仿坊彷妨枋舫放昉防…
干	42	182	253	0.70	3	18	刊平幹杆犴旰旱汗扞…
包	32	214	298	0.70	5	23	抱胞炮砲刨匏咆庖怉…
非	38	252	353	0.70	8	31	菲啡扉緋婓腓翡徘排…
屯	26	278	386	0.69	4	35	沌盹囤鈍坉佇炖飩怋…
元	20	298	412	0.69	4	39	刓岏完妧玩杬沅忨芫…
工	51	349	448	0.64	3	42	巨仜功左巧巫差式攻…

接著我們比較三種排序方法的學習曲線如圖 3，同樣地橫軸爲部件排序，縱軸爲正確預測聲母個數與韻母個數的總和。從圖 3 中可看出幾何乘積排序較線性加總法來的有效，在學到 1000 字以前幾何乘積排序呈現大幅度的成長，也就是說若我們依照乘積排序的部件順序來學習，一開始便能達到快速學習到大量的延伸字。調和平均排序採用幾何乘積與線性加總算數平均法的調和，不過其走勢幾乎與幾何乘積排序相同，這點也顯示出幾何乘積排序明顯優於線性加總。

最後我們以累積筆畫數的學習曲線來看(圖 4)，幾何乘積排序的累積筆畫數學習曲線也較線性加總排序所得來的優異。圖 3 的收斂點與圖 4 的筆劃數大增的轉折點也顯示了在學習了 2200 個部件後，累積延伸字數已呈飽和狀態，顯示接續其後的部件已是複合部件。另外圖 4 顯示 2200 部件之後筆畫數增加速度較快，可判斷排序大於 2200 後的漢字多是較複雜的字，並不是迫切的學習對象。

4. 形聲字發音規則探勘

本文第二個重點在於形聲字發音規則的探勘，藉由已標記的形聲字聲符，找出聲符與延伸的形聲字之間是否有常見的發音規則。爲了要產出易懂的發音規則，讓漢字的學習可以應用形聲字的特性來推測字的發音，在本文中我們將應用關聯規則探勘 Apriori 演算法做爲探勘形聲字發音規則的方法。每一條關聯規則必須符合最小支持度(support)及最小信賴度(confidence)，對於學習者才算有用。以下我們首先介紹如何準備形聲字成爲關聯規則探勘所需要的交易資料，以及規則的篩選與分群，以及最終所得的發音規則。

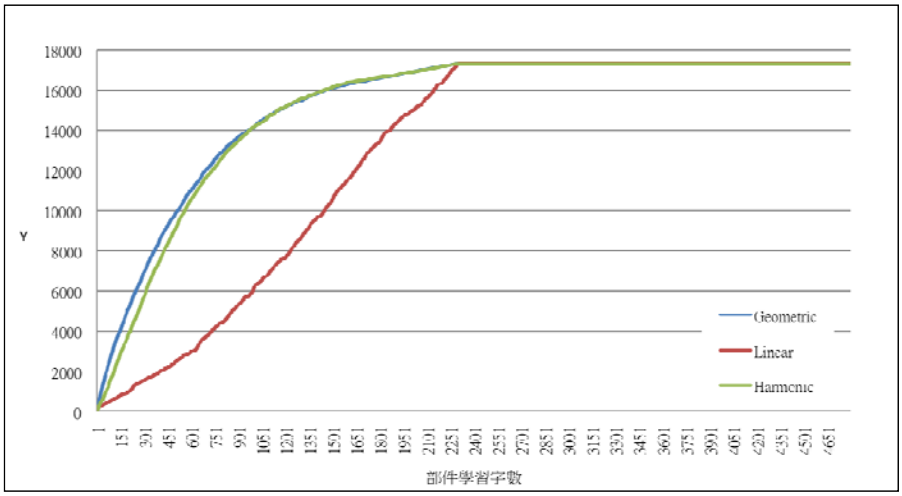


圖 3. 部件排序學習曲線比較圖

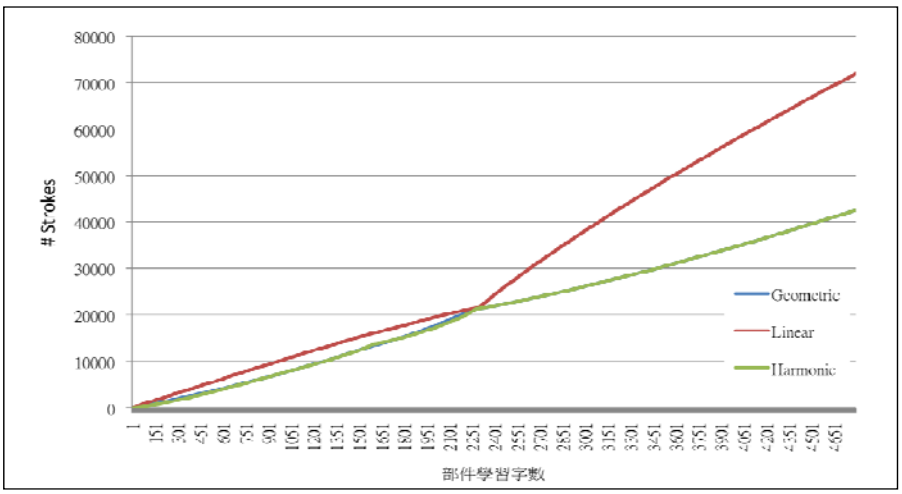


圖 4. 部件排序與筆畫數學習曲線比較圖

4.1 形聲字交易資料

關聯規則探勘原本的目的是從超市購買交易記錄的資料庫中，找出產品之間被購買的關聯程度，其主要依據爲支持度(support)及信賴度(confidence)。其中支持度代表一個規則的涵蓋率（全部交易資料中有多少百分比讓規則爲真），而信賴度則代表一個規則的準確率（前提爲真的情況下，有多少百分比資料讓結果也同時爲真）。關聯規則探勘是資料探勘領域最爲廣泛使用的工具，許多資料探勘軟體都提供此項功能，Weka[2]即是眾多資料探勘軟體其中之一。爲了推測發音規則，我們以常用字中的 3000 個形聲字準備成 3000 筆交易資料。

[2] http://www.cs.waikato.ac.nz/ml/weka/

形聲字的發音分成三個部份:聲母、韻母、以及調號,分別將其記爲INITIAL、FINAL、TONE。另將形聲字的聲符(Phonetic component)，以及聲符的發音以 PC_INITIAL、PC_FINAL、PC_TONE 三個屬性標記。其次漢字的部首(Radical component)、形聲字排列方式(單體字、左右連接、上下連接、包圍式、其他)、形聲字筆劃(Stroke)、聲符筆劃(PC_Stroke)、兩者差值(diff_STROKE)等特徵都列爲表達發音規則的探勘項目之一。最後，形聲字的發音若與其聲符的發音相同，則標記成聲母發音不變(IU)、韻母不變(FU)、音調不變(TU)等項目，做爲交易資料的一部份。

表 2. 漢字特徵對照表及”炮”的交易範例

符號	意義	數值範圍	範例:炮
INITIAL	聲母發音	{ø ,ㄅ,ㄆ,…,ㄙ}	ㄆ
FINAL	韻母發音	{ø,一,ㄨ,…,儿}	ㄠ
TONE	調號	{1,2,3,4,5}	4
CONNECT	形聲字的連接方法	{單體字,左右連接,上下連接,包圍式,其他}	左右
PC	聲符	形聲字	包
PC_LOCATION	聲符所在形聲字之位置	{左,右,上,下,內,其他}	右
PC_INITIAL	聲符的聲母	{ø,ㄅ,ㄆ,…,ㄙ}	ㄅ
PC_FINAL	聲符的韻母	{ø,一,ㄨ,…,儿}	ㄠ
PC_TONE	聲符的調號	{1,2,3,4,5}	1
STROKE	形聲字筆劃數 L16 表示>=16 b12-15 表示介於 12 與 15 s11 表<=11	{s11, b12-15, L16}	s11
PC_STROKE	聲符筆劃數	{s11, b12-15, L16 }	s11
Diff_STROKE	形聲字與其聲符筆劃差值	{s3, b4-5, L6}	4-5
INITIAL_UNCHANGED(IU)	形聲字聲母發音不變	{false , true}	IU=false
FINAL_UNCHANGED(FU)	形聲字韻母發音不變	{false , true}	FU=true
TONE_UNCHANGED(TU)	形聲字聲調不變	{false , true}	TU=false

值得一提的部份是，由於筆劃數乃數值性屬性，爲能運用關聯資料探勘技術，我們統計了漢字構形資料庫中所有的漢字的筆劃數將其平分爲三類，分別是筆劃小於等於 11、介於 12-15、大於等於 16。同時形聲字與其聲符筆劃差值，也就是部首的筆劃數也分爲三類，分別是筆劃小於等於 3、介於 4-5、大於等於 6。每筆形聲字交易資料所包含的項目屬性如表 2 所示。

我們取最小支持度 0.3%、 最小信賴度 60%來進行形聲字發音規則探勘。針對最小支持度取 0.3%、0.5%與 1%對應各種不同的最小信賴度 60%~100%，進行 Apriori 運算後，得到不同數量的規則數如表 3。在常見 3000 筆形聲字中，支持度 0.3% 相當於符合 9 個形聲字。更小支持度的規則由於使用率不高，因此最小信賴度設爲 60%。雖然在最小支持度 1%及最小信賴度 100%時，即可探勘出 50,054 條發音規則，然許多高支持度的規則並不具有高信賴度，爲避免錯失重要的發音規則，以上各項參數設定中，我們取最多規則數的參數組合(最小支持度 0.3%，最小信賴度 60%情形下)，共 6,625,518 條規則，做爲進一步的篩選過濾。

表 3. 關聯規則探勘後規則數

sup \ conf	60%	70%	80%	90%	100%
0.3%	6,625,518	5,144,742	3,879,619	2,809,951	1,810,585
0.5%	1,573,613	1,149,779	802,029	500,708	314,523
1%	304,330	217,346	143,301	87,324	50,054

4.2 規則篩選

每條關聯規則皆是由“左邊條件[左支持度]➔右邊結果[右支持度,信賴度]”組成。雖然關聯規則探勘可以取得爲數不少的發音規則，但其中有許多是不符合我們預期的規則。舉例來說：

PC_LOCATION=右 (sup=2054) ➔ CONNECT =左右 (sup=2054, conf=1)

上述這條規則表示“若聲符位置在右，則形聲字連接方式爲左右連接”。像這樣的規則對發音的推測其實並沒有幫助。又如以下規則：“若形聲字聲母發音爲ㄅ，則其聲符聲母發音爲ㄅ”，像這樣的規則也無助於推測發音。由於我們的本意是讓學習者在具備基礎聲符的閱讀能力下，利用對聲符的相關認知，來推測出更多尚未認識的形聲字發音。因此合法的規則應該具備:“聲符條件或形聲字筆劃數”➔“形聲字發音或形聲字發音之變化”。根據此一篩選原則，我們統計出最小支持度與最小信賴度不同參數下合法的規則數如表 4。最後我們存入 368,810 條規則於資料庫中。

表 4. 篩選後規則數

sup \ conf	60%	70%	80%	90%	100%
0.3%	368,810	272,957	195,735	152,152	106,740
0.5%	61,171	32,089	15,243	7,561	5,190
1%	13,470	6,340	1,889	505	42

4.3 規則分群

雖然在最小支持度 0.3%，最小信賴度 60%情形下，規則篩選已將的規則數減少至 368,810 筆規則，但由於規則中有許多同質性的規則散佈在資料庫中，我們需要有系統地將它們分群。以圖 5 條件集爲例，可以發現 1、2、3 具有相同條件「聲符的聲母=ㄌ」，且這些規則均具有相近的支持度。仔細深入查看符合這些條件的字後發現，支持這些規則的字組也相當程度的重疊（如「老」、「呂」、「里」等聲符的延伸字），所以聲符的聲母條件可以是分群的重要參考因素。同理聲符的韻母也多涉及相同性質的規則，因此規則中若有指定相同的聲符韻母，也是我們分群的依據之一。

另外我們也發現相同部首的規則具有相近的支持度及相同的延伸字集，因此可再結集成群。而形聲字的連接方法在具有相同部首的狀況下，通常也會有特定的連接方法如上規則{4、5}；{6、7}，因此相同部首及形聲字的連接方法也在分群條件之一。完整分群條件優先權如下：聲符、聲符聲母、聲符韻母、部首、形聲字的連接方法。根據這些分群優先條件，可將相同性質規則分爲同群。

1. 聲符的聲母=ㄌ，聲符的調=2，聲符所在位置=右，形聲字筆劃數=12-15 (sup=17)
2. 聲符的聲母=ㄌ，聲符的筆劃數=L16，漢字與其聲符筆劃差值=4-5 (sup=16)
3. 聲符的聲母=ㄌ，聲符的調=3，漢字與其聲符筆劃差值=s3 (sup=16)
4. 部首=艸，形聲字的連接方法=上下連接，support=22
5. 部首=艸，聲符所在位置=下，support=22
6. 部首=女，形聲字的連接方法=左右連接，support=15
7. 部首=女，聲符所在位置=右，support=15

圖 5. 發音規則條件範例

針對聲母不變(IU)及韻母不變(FU)的條件下，我們將查詢所得規則經過分群之後的所得結果呈現於表 5。如表所示，分群之後，發音關聯規則即可大幅減少，有助於規則的觀察與了解。

表 5. 符合聲母不變或韻母不變的規則數及分群後規則數

Confidence / Condition	60%	70%	80%	90%	100%
IU, FU	3454/332	2004/225	1097/139	597/73	264/39
IU	9002/486	5383/383	3067/262	1758/161	809/91
FU	12171/690	8373/608	4855/470	2673/325	1392/189

4.4 關聯規則查詢介面

關聯規則的查詢介面主要是爲了教材編排者所設計，得讓使用者能根據不同條件快速篩選發音規則。「形聲字發音規則查詢系統」的設計，是採用動態呈現條件選單內容，因此第一次載入網頁時等待時間較長（約 20 秒），而後選擇條件時系統會透過 Ajax 的方式傳送搜尋條件至伺服器端擷取相關規則做爲呈現動態分群結果。查詢過程中，左下角的「下載中…」字樣會表示資料正在回傳中，右上角兩個選項則是開啓「形聲字標記系統」及「構件發聲強度列表」的連結。

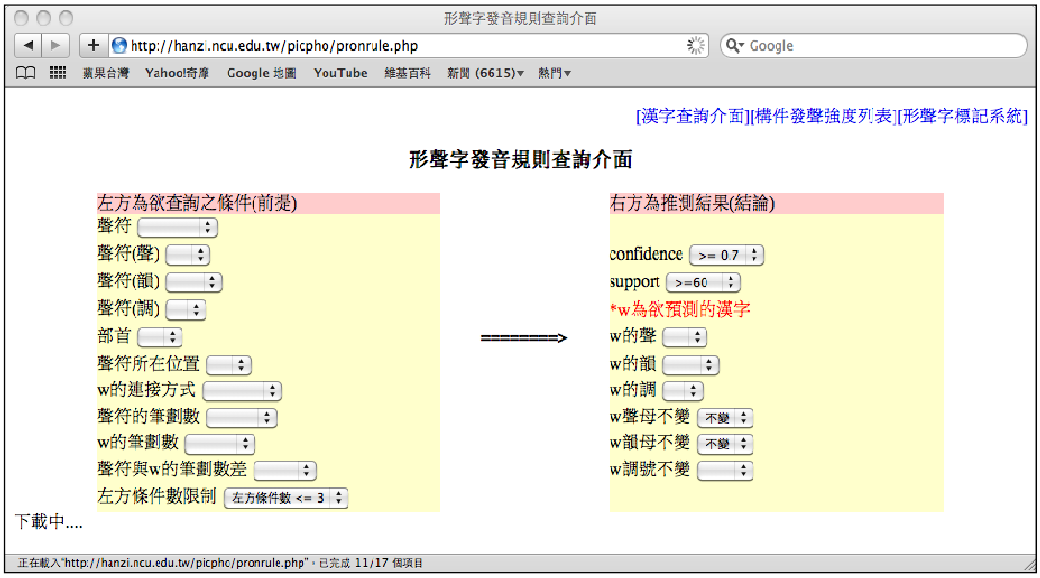


圖 6. 形聲字發音規則查詢介面(http://hanzi.ncu.edu.tw/picpho/pronrule.php)

查詢系統主要依據前述「形聲字發音規則探勘項目集」的特徵爲查詢條件，左邊是已知條件，右邊是推測結果，w 代表欲推測發音之形聲字。「左方條件限制」的功能則可篩選過長的規則。由於太長的規則通常不利於人們記誦，因此本系統預設條件數小於等於三。其他預設查詢條件爲「信賴度 >= 70%」，「支持度 >= 60」。當「下載完成」出現後，所有符合條件篩選的規則，經由分群後會顯示在介面的最下方，不同性質的規則會用不同底色做區隔（如圖 7）。每條規則中都有可供細項查詢的連結。當我們選擇[查看]連結，則可顯示滿足整條規則（包含左方前提及右方結果）的形聲字；除此之外，[例外字]則可查出究竟有哪些形聲字是符合左方前提，但是不符合右方結果的形聲字。其他連結則可顯示符合單一條件的形聲字，如[部首=木]連結，可顯示出所有部首爲木的形聲字，[聲符的聲母=ㄌ]可找出所有聲符聲母是ㄌ的形聲字。

圖 7. 形聲字發音規則分群結果

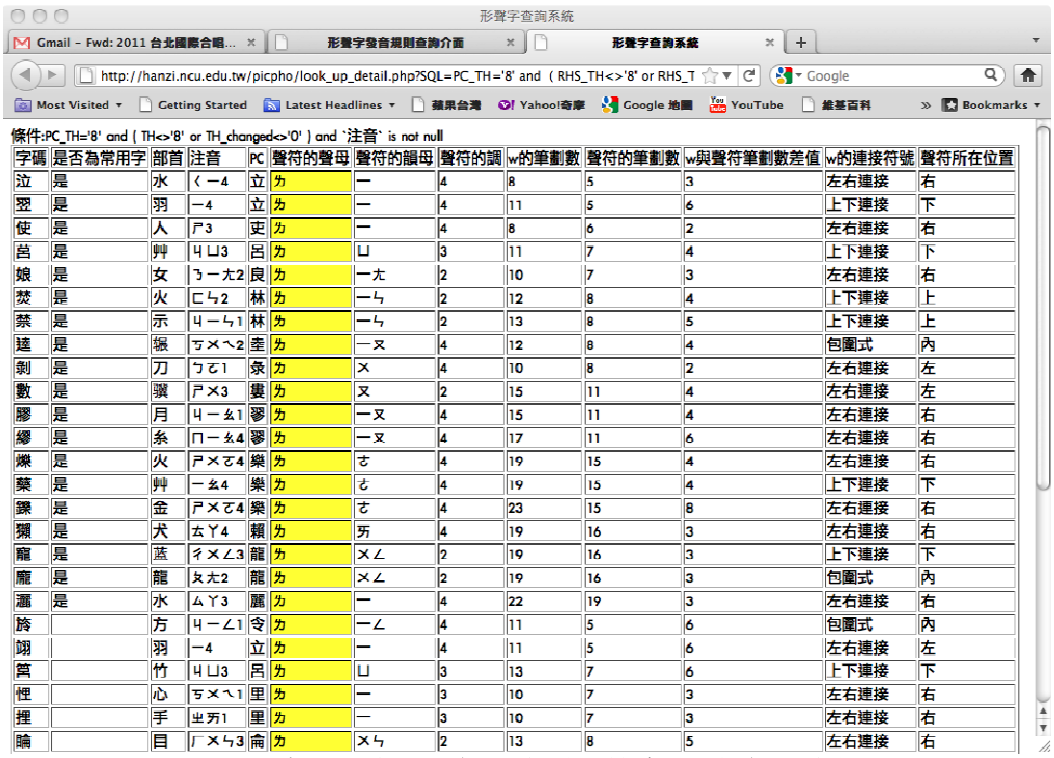
有了以上形聲字發音規則查詢系統，我們即可設定所需條件，找出相關發音規則。舉例來說，高支持度 3%、信賴度 80%且聲母發音不變的條件下的規則共有 15 條，共分成 3 組，如 R1-R3 所示。

(R1) 聲符的聲母=ㄉ (supp:197) ➔聲母發音=不變 (supp:178, conf:0.9)

(R2) 聲符的聲母=ㄇ (supp:128) ➔聲母發音=不變 (supp:105, conf:0.82)

(R3) w 的筆劃數=L16 ， 聲符的筆劃數=L16 (supp:123) ➔聲母發音=不變 (supp:98, conf: 0.8)

規則一(R1)說明聲符的聲母若爲ㄉ的前提下，形聲字的聲母發音將維持ㄉ（聲符例字：力令立列老利呂良里來侖兩戾拉林垒麥枲刺郎栗留婁累連勞量廉虜雷豊劉閭厲慮樂閭魯畾歷盧賴龍闌羅麗蘭覽）；規則二(R2)顯示聲符的聲母若爲ㄇ的前提下，形聲字的聲母發音將維持 ㄇ（聲符例字：木末毛母民冊目矛名牟米免每孟明門冒某眉眇美苗面冥迷莽莫麻悶買閔滿蒙貌麼磨瞢彌）;規則三(R3)則敍述聲符的筆劃數若大或等於16以上，則形聲字的發音也多維持原本聲符的聲母發音（聲符例字：冀贏歷燕盧磨穌縣羲翰蕭謁賴頻龍瞢裹嬰彌毚爵襄闌隱霜鮮鐵瞿聶轉離魏薩羅藝贊顛麗嚴藺蘇覺矍籣覽霸彎），不過第三個規則，由於筆劃數高，對於初學者來說幫助不大。除了查看例字之外，使用者也可查看例外字，了解符合前提(聲符的聲母=ㄉ)但是聲母發音卻改變的形聲字〔見圖 8〕。



字碼	是否為常用字	部首	注音	PC	聲符的聲母	聲符的韻母	聲符的調	w的筆劃數	聲符的筆劃數	w與聲符筆劃數差值	w的連接符號	聲符所在位置
泣	是	水	ㄑㄧ4	立	ㄌ	ㄧ	4	8	5	3	左右連接	右
翌	是	羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	上下連接	下
使	是	人	ㄕ3	吏	ㄌ	ㄧ	4	8	6	2	左右連接	右
呂	是	艸	ㄐㄩ3	呂	ㄌ	ㄩ	3	11	7	4	上下連接	下
娘	是	女	ㄋㄧㄤ2	良	ㄌ	ㄧㄤ	2	10	7	3	左右連接	右
焚	是	火	ㄈㄣ2	林	ㄌ	ㄧㄣ	2	12	8	4	上下連接	上
禁	是	示	ㄐㄧㄣ1	林	ㄌ	ㄧㄣ	2	13	8	5	上下連接	上
遠	是	辵	ㄩㄢˇ2	袁	ㄌ	ㄧㄢ	4	12	8	4	包圍式	內
剝	是	刀	ㄅㄛ1	录	ㄌ	ㄨ	4	10	8	2	左右連接	左
數	是	攴	ㄕˇ3	婁	ㄌ	ㄡ	2	15	11	4	左右連接	左
膠	是	月	ㄐㄧㄠ1	翏	ㄌ	ㄧㄡ	4	15	11	4	左右連接	右
繆	是	糸	ㄇㄧㄠ4	翏	ㄌ	ㄧㄡ	4	17	11	6	左右連接	右
爍	是	火	ㄕˇㄛ4	樂	ㄌ	ㄛ	4	19	15	4	左右連接	右
藥	是	艸	ㄧㄠ4	樂	ㄌ	ㄠ	4	19	15	4	上下連接	下
鑠	是	金	ㄕˇㄛ4	樂	ㄌ	ㄛ	4	23	15	8	左右連接	右
獺	是	犬	ㄊㄚ4	賴	ㄌ	ㄞ	4	19	16	3	左右連接	右
龐	是	龍	ㄔㄨㄥ3	龍	ㄌ	ㄨㄥ	2	19	16	3	上下連接	下
龐	是	龍	ㄤㄊ2	龍	ㄌ	ㄨㄥ	2	19	16	3	包圍式	內
灑	是	水	ㄙㄚ3	麗	ㄌ	ㄧ	4	22	19	3	左右連接	右
旖		方	ㄐㄧㄥ1	令	ㄌ	ㄧㄥ	4	11	5	6	包圍式	內
翊		羽	ㄧ4	立	ㄌ	ㄧ	4	11	5	6	左右連接	左
筥		竹	ㄐㄩ3	呂	ㄌ	ㄩ	3	13	7	6	上下連接	下
悝		心	ㄎㄨㄟ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
裡		手	ㄓㄞ1	里	ㄌ	ㄧ	3	10	7	3	左右連接	右
瞗		目	ㄏㄨㄣ3	侖	ㄌ	ㄨㄣ	2	13	8	5	左右連接	右

圖 8. 查看發音規則例外字（不符合 R1 的例外字）

又如查詢高信賴度 100%、支持度 0.5%、且聲母與韻母均未改變的規則，可得 34 條符合條件的規則，分成 5 組，如 R4– R8 所示。規則左方的支持度表示滿足左方條件的常用形聲字，規則右方的支持度則爲滿足整個規則的常用形聲字。舉例來說規則七(R7)說明聲符的聲母爲ㄒ、聲調爲一聲且聲符筆劃數小於等於 11 的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「希」、「析」、「宣」、「星」、「相」、「胥」、「奚」等衍生的 16 個常用形聲字。不過使用者若是查看符合規則的形聲字，則同時可以看到其他符合條件的非常用形聲字，如「心」、「先」、「西」、「析」、「欣」、「香」、「悉」、「脩」等聲符所衍生的形聲字。規則八(R8)則說明當聲符的韻母爲ㄤ、聲調爲一聲、聲符筆劃數小於等於 11 且聲符與部首爲左右連接的時候，則衍生形聲字的聲母與韻母均不改變；符合這條規則的形聲字中包含的聲符包括「方」、「邦」、「岡」、「昌」等衍生的 17 個常用形聲字。

(R4) 聲符的聲母=ㄌ，聲符的調=3, w與聲符筆劃數差值=s3 (supp:16)

➔聲母發音=不變，韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R5) 聲符的聲母=ㄌ，聲符的調=2，聲符所在位置=右, w的筆劃數=12-15 (supp:17)

➔聲母發音=不變，韻母發音=不變 (supp:17, conf:1) [查看],[例外字]

(R6) 聲符的聲母=ㄌ, 聲符的筆劃數=L16, w與聲符筆劃數差值=4-5 (supp:16)

➔聲母發音=不變, 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R7) 聲符的聲母=ㄒ, 聲符的調=1, 聲符的筆劃數=s11, w的筆劃數=12-15 (supp:16)

➔聲母發音=不變, 韻母發音=不變 (supp:16, conf:1) [查看],[例外字]

(R8) 聲符的韻母=ㄤ, 聲符的調=1, w的連接符號=左右連接, w的筆劃數=s11 (supp:17)

➔聲母發音=不變, 韻母發音=不變 (supp:17, conf:1) [查看], [例外字]

5. 結論及未來研究

本文的研究目標係提出一套以聲符爲主的部件教學策略，將構詞能力很強的部件放在課程的前面，發揮「以簡馭繁」、「快速掌握形聲字的結構」等部件教學的優點，加強學習者利用部件線索來學習新的生字的觀念，提升其於漢字識字學習上的能力。

在本篇論文中，我們延續機率分佈比較法，考慮到發音一致性強、出現頻率高且部件筆劃數少等三種因素，我們提出三種部件排序方法，其中幾何乘積法在延伸學習字數及筆劃數曲線圖的表現上較爲出色。本論文的第二部份則是藉由形聲字的特徵，運用關聯探勘法則挖掘出許多發音規則。而發音規則經由我們歸納後可分爲，高支持度與高信賴度兩大類。藉由這兩大類的規則能幫助不同程度的初學者更易於推測未知漢字的發音。

目前有關部件發音強度的計算，以及形聲字發音的關聯規則雖已完成，但是對於輔助以聲符爲主的部件教學教材編輯，仍有不足之處。舉例來說，由於關聯規則探勘可能找到相當多的規則，而且某些規則可由其他規則涵蓋，因此如何找出一組最重要的規則涵蓋愈多的常用字及將發音規則排序，則是此處我們必須要解決的問題。再者，漢字教學步驟通常爲先教獨體字，再教簡單合體字，最後教複雜合體字。但並非每個部首和任何聲符都可組成合體字，對初學者而言，可能出現偏旁部首張冠李戴的情形。如何幫助學習者釐清這些差異，也是挑戰之一。

參考文獻

許慎撰，段玉裁注(1999)。《說文解字注》，台北藝文印書館。

莊德明、謝清俊(2005)。《漢字構形資料庫的建置與應用》，漢字與全球化國際學術研討會，台北 。

莊德明、鄧賢瑛(2008)。《文字學入口網站的規畫》，第四屆中國文字學國際學術研討會，山東煙台。

董鵬程(2007)。《台灣華語文教學的過去、現在與未來展望》，多元文化與族群和諧國際研討會，台北教育大學。 http://r9.ntue.edu.tw/activity/multiculture_conference/memoirs.html。

許聞廉、呂明蓁、胡志偉、柯華葳、辜玉旻、呂菁菁、張智凱、莊宗嚴(2009-2011)。《構建一個新移民者有機成長的多元認同平台的整合研究（期中進度報告）》。

高柏園、郭經華、胡映雪(2009-2010)。《華語文作爲第二語言之字詞教學模式與學習歷程研究》。

洪文斌(2010)。《華語文作爲第二語言之字詞教學模式與學習歷程研究－－子計畫一：中文字部件拆解教學模式與電腦輔助學習系統之研發（期中進度報告）》。

張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文、黃鍔(2010)。《以最佳化及機率分佈判斷漢字聲符之研究》，Computational Linguistic and Chinese Language Processing, 15(2), 145-160。

萬雲英，《兒童學習漢字的心理特徵與教學》，載於楊中芳、高尚仁主編，中國人、中國心－發展與教學篇，403-448。台北：遠流。

盛繼豔，《華文教學中漢語的部件教學》。

梁彥民(2004)。《漢字部件區別特徵與對外漢字教學》，語言教學與研究。

李思維、王昌茂編著(2000)。《漢字形音學》，武漢：華中師範大學出版社。

中研院文獻處理實驗室，「漢字構形資料庫」，http://cdp.sinica.edu.tw/cdphanzi/。