The 2014 Conference on Computational Linguistics and Speech Processing
ROCLING 2014, pp. 153-162
© The Association for Computational Linguistics and Chinese Language Processing

使用中文字筆畫構形資料庫校正字形相似之別字

Using Chinese Orthography Database to Correct Chinese Misspelling Words With Graphemic Similarity

張道行[1]　陳學志[2]　鄭健良[1]

摘要

中文別字自動偵測與校正是個相當重要的工具，許多分析別字類型的研究指出，「字音混淆」、「字形混淆」與「字義混淆」是別字產生的主要原因，因此近年來許多別字自動校正的研究也都採取分別針對字音、字形、字義造成的混淆進行探討。但對於字形相似別字的校正正確率仍不夠好，主要原因之一是因為在中文字字形結構的資訊不夠完整。本文的目的是利用中文部件組字與形構資料庫的筆畫結構資料提出一個演算法，計算兩個中文字筆畫結構序列的相似程度，並用於字形相似類別字的偵測與校正。實驗結果顯示筆畫結構用在偵測與校正字形相似別字的效能較原先以部件的方法來得有效。

關鍵字：別字偵測；別字校正；字形相似；筆畫；中文部件組字與形構資料庫；LCS 演算法

1. 緒論

中文別字自動偵測與校正是個相當重要的工具，在資訊應用中也具有相當的價值。例如許多文書處理軟體都有提供別字檢查與校正建議。但由於中文別字的偵測與校正較困難，這些軟體在中文別字校正效能上仍有不足。因此除了英文的別字校正研究[1][7]之外，一直以來也有許多研究[4][6][8][11][13][15][16]提出不同的方法試圖解決中文別字偵測與校正的問題。

早期別字自動偵測方法大多是透過建立混淆字集方式搭配語言模型運作。例如 Chang[2]將欲處理語句中的每個字視為別字，並以該字混淆字集中的字逐一替換原字，組合出多種可能的句子。再利用 bi-gram 語言模型計算所有句子的分數，選出最可能的句子。

[1] Department of Computer Science and Information Engineering, National Kaohsiung University of Applied Sciences, Kaohsiung, Taiwan
[2] Department of Educational Psychology and Counseling, National Taiwan Normal University, Taipei, Taiwan
Email: changth@gm.kuas.edu.tw

許多研究也提出了類似的方法，但改用不同的語言模型，如 tri-gram[17]、混合規則式與機率式的模型[12]。此類方法相當容易實作，也可同時適用於字音相似與字形相似兩類別字的偵測與校正。但這種方法的主要缺點為適當的混淆字集建立不易，如果廣泛蒐集所有可能的混淆字，很容易造成誤報率(false alarm)提高。但若只蒐集常見別字，則會遺漏不在混淆字集中的字。

許多分析別字類型的研究[18][19][20]指出，「字音混淆」、「字形混淆」與「字義混淆」是別字產生主要原因，因此近年來許多別字自動校正的研究也都採取分別針對字音、字形與字義造成的混淆進行分析、再加以整合的策略。這些研究在字音造成的別字部分都有相當好的校正正確率，但對於字形與字義造成的別字的校正正確率仍不夠好。對於字形造成的別字校正效能有限，主要原因之一是因為在中文字字形結構的資訊不夠完整，導致建立某些中文字字形混淆字集時容易產生相似度誤差使得字集不夠精確。

陳學志等人[22]建立的「中文部件組字與形構資料庫」是一個解決這個問題的可能線索。該資料庫於 2010 年起提供每個中文字的部件組字資料，又於 2013 年進一步擴充，將每個部件進一步拆解為筆畫結構的組合，得到每個字以筆畫為基本單位的構形資料，我們稱為筆畫結構序列。本文的目的就是利用這個中文字筆畫結構資料庫處理字形相似度問題，並用於字形相似類別字的偵測與校正。但是筆畫結構資料是描繪基本筆畫單元間的空間關係，要如何使用於計算兩字間的字形相似度並不是個容易的問題。本文的主要貢獻在於提供一個演算法可計算兩個中文字筆畫結構序列的相似程度。

本文其餘內容將組織如下。第 2 節說明字形相似別字的相關研究以及所遭遇的問題。第 3 節介紹「中文部件組字與形構資料庫」中筆畫結構資訊的表達方式。第 4 節說明本文所提的字形相似度計算方法。第 5 節說明如何將字形相似度計算結果整合在別字預測工具中。第 6 節展示應用此方法於字形錯別字的實驗結果。最後提出未來可能的研究方向。

.

2. 相關研究

早期研究多使用混淆字集偵測與校正別字，然而這種方法未必能將真正字形相似字完整涵括。因此有些研究針對字形相似性設計自動建立混淆字集的方法。最早的相關研究是使用與字形相關的倉頡輸入法偵測字形相似別字。如 Lin 等人[9]利用倉頡碼建立混淆字集，再採用類似 Chang[2]的方法偵測與校正。實驗結果顯示校正的成功率有明顯提升且有較少的誤報率。Liu 等人[10]也是利用倉頡碼對形構相近字提出計算相似度的方法。[10]利用倉頡輸入法將中文字編碼，透過倉頡碼 26 個基本單位的組合比較兩兩中文字的相似性。然而

倉頡輸入法為了維持每一個漢字輸入不超過五個碼的效率，因此某些部件較多或較複雜的漢字倉頡碼會被簡化，使得字與字之間的相似證據難以找出。於是[10]建構了一套「倉頡詳碼」完整地還原字的構形，將每個字以倉頡詳碼表示，可提供較精確的相似度訊息。

基本上使用倉頡碼建立混淆字集的概念是認為漢字可拆解成數個基本單元、或稱為部件(radical)，比對部件的使用可判斷字形相似性。而有一些研究[22][23]提出了更完整的中文字部件構形資料庫，例如[23]提出的「漢字構型資料庫」，以及 [22]建構的「中文部件組字與形構資料庫」。以後者為例，其蒐集了 6097 個常用字，從內拆解出 439 個基礎中文部件，並歸納出 11 種字形空間的結構關係。透過以上指標與部件跟結構關係的組合，可以更加瞭解字形的構成，進而觀察字與字之間的相似關係。與倉頡碼相比，使用中文字部件資料庫比較字形相似度是更為合理且精確的作法。

先前 Chang 等人[5]的研究也利用「中文部件組字資料庫」中的部件偵測與校正字形別字。該研究也發現使用部件偵測及校正字形相似字的確有效，但該研究也指出有些部件由於未能進一步拆解導致某些字的相似度計算有相當大的誤差。例如酉本身是一個部件，因此當他和西計算相似度時因為沒有相同部件導致相似度非常低。這個問題突顯了以部件量測字形相似性的侷限所在，但也指出改良的方向，也就是應該以更基本的字形組成單元測量相似性。

3. 中文字筆畫結構表示法

在「中文部件組字與形構資料庫」中，所有漢字可歸納出 41 個基礎筆畫與 11 個字形結構關係。筆畫中包含基本的橫{一}、豎{|}、撇{丿}等。其中部分筆畫無法透過系統 unicode 編碼呈現，便會將其透過組合的方式表達，例如「國」的第二劃為橫豎，但橫豎無法以一般 unicode 編碼呈現，因此以筆畫{口 2}(代表「口」字書寫時的第二劃，也就是橫豎)表示。有些更複雜的筆畫必須透過結構關係描述，例如「大」的第三劃為撇捺，同樣無法以 unicode 呈現，因此會以筆畫{[尺/]}表示，其中筆畫內置於「尺」右側的「/」代表右下方的筆畫，即「尺」字右下方的那一劃，也就是撇捺。

字形結構部分則包含 10 種組合及其它結構。10 種組合包含：垂直組合(例如「二」字結構是由上下垂直組合而成)、水平組合(例如「林」字是由左右水平方式組合)、封閉組合(例如「國」字是由四面包圍方式組成)等。這些組合以符號表示，例如垂直組合為「-」、水平組合為「|」、封閉組合為「0」等。大多數字都可以用這 10 種組合結構表示，但有些字的結構不屬於任何一種組合，故這些字歸類為無字形組合之結構，例如「一」、「乙」字

都為單獨存在的形體結構，不屬於任何一種組合。

在資料庫中每一個漢字都是由筆畫、筆畫間連接關係、與字形結構三項所組合，稱為「基礎筆畫組合」。例如「二」字，其基礎筆畫組合為「-({一 s},{一})」，從中可看出垂直組合(-)是主結構，這個組合包含了兩個筆畫{一 s}與{一}，其中{一 s}代表書寫時較短的{一}。「二」字無筆畫連接關係，對於有筆畫連接關係的字也會在基礎筆畫組合以符號組合「+{a:b@c}」表示連接關係。此符號組合一般都會緊鄰在一筆畫之後，例如「{丨}+{a:b@c}」，表示會以筆畫{丨}為基準，本文稱為基本筆畫。符號組合中的 a 代表整個基礎筆畫組合中由左而右數來第 a 個筆畫和基準筆畫{丨}有連接關係，這第 a 個筆畫本文稱為連接筆畫。符號組合中的 b 代表基準筆畫在連接筆畫的 b 位置相接，「@c」則代表連接筆畫在基準筆畫的 c 位置相接。此資料庫將每一個基礎筆畫依照比例平均劃分為 10 等分，用來表示前述相接位置。符號組合中的「+」表示筆畫間的連接關係為相互交錯；若筆畫間為相接但非交錯的關係，則以「~」符號表示。筆畫間連接關係只有以上兩種。

圖 1 是一個基礎筆畫組合的範例。範例中的「十」字其基礎筆畫組合為「[{一},{丨}+(1:5@5)]」。「十」的筆畫連接關係表示為{一}在其 5 的位置(也就是正中間)與{丨}在其 5 的位置相互交錯。然而「十」的結構為單獨存在，故在筆畫組合裡面不會看到結構符號、只會有筆畫。但是兩個筆畫有連接關係，因此僅以「{丨}+(1:5@5)」表示。

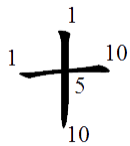

圖 1：基礎筆畫組合範例字「十」

4. 中文字筆畫結構相似度

由於前述的筆畫結構表示法不容易直接用來計算兩中文字筆畫結構的相似度，因此必須先將整體筆畫結構轉換為筆畫結構配對。轉換方法是先將每個漢字的基礎筆畫組合中以有連接關係的一組筆畫配對作為相似度計算的基本單位。例如「大」字，基礎筆畫組合為「[{一},{月 1}+(1:5@3),{[尺/]}~(1:5@0)~(2:3@0)]」，將其透過兩兩筆畫連接關係配對成({一},{月 1})、({一},{[尺/]})及({月 1},{[尺/]})三個筆畫配對，並以漢字書寫時的筆畫順序將配對排序，形成一個配對序列。

接著便可使用最長公共子序列(Longest Common Subsequences，LCS)演算法[7][14]計算兩個配對序列的相似度。LCS 演算法可以經由比對得到兩序列中最長相同的子序列，

因此廣泛被使用於計算兩字串或序列之間的相似度。其計算方法是將序列中的所有子序列依照順序一個個與另一序列之子序列做比對，若相同則比較下一個子序列，不同則記錄比較至目前最長的相同子序列，接著再從頭比對另一序列的下一個子序列，重複上述動作直到所有子序列比較完成即可取得最長相似的子序列組合。例如比較兩字串「Look」與「Books」，會先以「Look」的 L 與「Books」的 B 比較，不同則比較「Books」下一個子字元 o，重複以 L 比對完「Books」所有字元後，再以「Look」的第二字元 o 重新比對「Books」。藉由不斷地循環比對便能得到最長的子字串為「ook」，長度為 3。

由於 LCS 演算法具備「擷取序列必須順序相同」的特性，符合漢字是依固定筆畫順序書寫的規則，因此我們採用 LCS 演算法作為兩字間筆畫配對序列相似計算的方法。以「大」和「太」兩字之相似度比較為例，其配對序列分別為({一},{月 1})、({一},{[尺/]})、({月 1},{[尺/]})與({一},{月 1})、({一},{[尺/]})、({月 1},{[尺/]})、({、})，兩字間除了配對({、})不一樣之外，其餘三個配對均相同，且其配對經由 LCS 演算法計算後，最長連續相似配對的配對數為 3。

上述的配對方法雖然依照順序的特性能有效計算大多數字之間的相似度，但有兩種特殊情況必須考慮。第一種情形是部件相同但部件位置不同的鏡寫字會造成相似度不足的問題。例如「部」與「陪」兩字，其筆畫配對完全相同，但出現的順序不同，若如果利用原本配對序列經由 LCS 演算法計算並無法得到合理的相似度。因為有鏡寫障礙的學生很容易寫出這類的別字，故我們修正前述配對序列產生的方式以改善 LCS 演算法的問題。

我們將欲比較的兩序列其中之一複製後連結在原先序列的後面，形成一個長度為原先兩倍的新序列，稱為複製倍增序列，再將複製倍增序列與另一序列做比較。以「abcdef」與「defabc」兩序列為例，由於兩序列僅「abc」與「def」順序顛倒，若使用原先 LCS 演算法計算，其最長相似長度為 3。因此我們將序列「abcdef」倍增為「abcdefabcdef」後，再與另一序列「defabc」比較，可得最長共同序列為「defabc」，長度為 6。但由於這種做法會使一個字與自己的最長共同序列的相似度會和它與鏡寫字間的相同，因此當兩字為鏡寫字時，必須在使用最長共同序列長度作為相似度評估時略做調整，才能有較合理的結果。

第二種情形則是配對序列省略了字形結構資訊導致某些字的相似性計算不合理。例如「昌」與「旧」兩字，其筆畫配對序列完全相同，若以 LCS 方法比較兩字，兩序列完全相同，但兩字並不是相同字，是屬於主字形結構不同的相異字。因此相似度公式必須考慮比較兩字之間的主結構關係，根據研究經驗給予不同的字形主結構的結構相似係數。在 LCS 計算相似性時須另外再乘上結構相似係數，作為兩字最後的字形相似分數。例如「昌」

與「田」字，其主字形結構分別屬於「垂直組合(-)」與「水平組合(|)」，這兩個結構關係差異較大，較容易分辨，因此兩結構的相似係數為 0.4。。

另外，本文使用筆畫評估字形相似性的原始動機之一，是因為有些相似的部件無法被有效判斷相似性。雖然用筆畫能將部件拆解成更小的比較單位，但筆畫與筆畫之間仍有相似的問題，例如筆畫{一}和筆畫{刁 2}、{乚 2}及{冖}相當相似。為解決這個問題，我們使用[22]所提供的筆畫相似度資訊，[22]將所有筆畫分類為 21 個筆畫相似集。因此本文在計算筆畫配對序列相似度時，對於兩個筆畫配對，若對應的筆畫在同一個筆畫相似集中，即使筆畫不同仍視為相同筆畫配對。

綜合以上討論，對於兩個中文字 C_1 與 C_2，其筆畫配對序列分別為 S_1 與 S_2，則兩字的字形相似度公式如下:

$$sim(C_1, C_2) = sms(C_1, C_2) \times \frac{2 \times lcs(rep(S_1), S_2) - mirror}{num(S_1) + num(S_2)} \quad (1)$$

其中 *rep* 函數可產生輸入序列的複製倍增序列；*lcs* 函數計算兩序列的最長共同子序列的長度；*num* 函數計算該序列的配對數量；*sms* 函數計算兩字主字形結構的相似性；參數 *mirror* 在 *lcs* 函數值分別等於兩個 num 函數值且兩字相異時，其值為 1，否則為 0。

5. 別字校正模型

本文使用的別字訓練模型是以 Chang 等人[5]所提的方法為基礎。[5]利用訓練資料中別字所產生的候選詞的四個參數(字音相似度、字形相似度、字頻機率比值、詞性機率比值)建立一個線性迴歸公式作為預測模型，用來判別候選詞是否可以對原句的可疑字組進行替換校正。本文與[5]的差別就在於字形相似度的計算方法不同。為了建立預測模型，本文準備 650 筆含有別字的資料，在計算有別字的可疑字組與有正確字的候選詞間的四個參數後，成為一個 positive 樣本集合;另外準備 650 筆句子包含無別字句子與含有別字句子，在計算可疑字組與無正確字的候選詞間的四個參數後成為一個 negative 樣本。使用這兩個樣本集合共 1300 句可以建立線性迴歸公式。這個線性迴歸公式如下：

$$y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 \quad (2)$$

其中X_i為候選詞與可疑字組間的四個參數，β_i為四個參數的迴歸係數，將候選詞對應

可疑字組的四個參數及各項迴歸係數代入上述公式可得到 y 值。當候選詞代入公式所得之 y 值越大時，其作為應替換的正確詞彙的可能性越高。相對地，若 y 值越小其作為可替換詞的機會也相對減少，因此可以利用 y 值作為候選詞應替換原字組可能性的依據。

本文從訓練的 1300 句 positive 及 negative 樣本集合裡候選詞與可疑字組間四個參數代入公式所得之 y 值找出最佳分類正確率，並以此 y 值作為預測模型的門檻值。若候選詞是應替換原字組的可能性分數大於門檻值即會被判定為可替換之正確詞彙而進行校正，反之若一個候選詞是應替換原字組的可能性分數小於門檻值，該候選詞則會被判定為無法替換詞彙而被捨棄。

6. 實驗

本文採用 SIGHAN-7 Bake-off 2013: Chinese Spelling Check[15]所提供的資料集作為實驗的測試資料。資料集內包含 SubTask1 與 SubTask2 各 1000 份文本。其中 SubTask1 只提供含有別字的文本與別字位置的資料，因此本文僅採用 SubTask2 資料作為量測中文別字偵測與校正效能使用。SubTask2 的 1000 份文本是來自於蒐集學生寫作常見的錯誤類型，每一份文本內至少含有一個以上的別字及數句未包含別字的句子，全部共包含 1265 個別字。本文保留含有別字的句子作為測試資料，合計共 1219 筆句子。由於我們是測試字形相似度的效能，因此只採用字形相似以及形音相似實驗的別字。另外，實驗使用的別字偵測工具只能擷取 WECAn[3]斷詞系統將詞彙中別字分成單字詞者，因此我們只使用符合上述兩項條件的 451 個別字作為測試資料。

本文別字校正實驗將以第 5 節所提方法設計兩個預測工具，一個是使用 Chang 等人[5]所提出的部件相似度計算方法，另一個則是使用本文所提筆畫結構相似度計算方法。兩個工具分別針對 451 筆測試資料進行測試。另外由於測試的別字有些為字音與字形皆相似的別字，為了評估字形相似計算之結果差異，在系統提出可替換候選字時會淘汰僅字音相似的正字詞彙，只比較字形混淆與形音同時混淆的別字。結果於表 1 所示。

表 1、使用部件相似與筆畫相似方法之校正結果

	校正字數	Recall	Precision	False Alarm
部件相似法	250	55.43%	63.29%	0.93%
筆畫相似法	417	92.46%	79.58%	0.60%

由表 1 的結果可知，在只比較形別字與形音別字的情況下，使用本文所提之筆畫相似方法能有效提升校正成功率。從本文所提方法的三項評估指標均遠優於使用組字部件的方法，可以得知筆畫計算字形相似的方法確實能改善部件相似性不足的問題。

7. 結論

本文提出了一個使用中文字筆畫結構資料的字形相似評估方法，並將其與現有別字偵測方法整合。從實驗結果可以證實本文所提方法用於字形相似別字的偵測與校正效能較原先以部件相似的方法來的好，其原因可以推論是本文所提方法讓字形相似度計算變得更加精確所導致。但受限於實驗材料沒有明顯區分字形別字以及形音均相似別字，因此無法證實本文所提方法在次分類別字上的效果，後續研究可以進一步加以證實。

另外，字形相似度計算的正確性評估可以考慮先建立真人相似度評估的標準，用以檢驗各種方法所得字形相似度與人類認知結果的一致性。我們目前正在進行這項工作。若是這項工作得以確認本文所提方法在計算字形相似度的正確性，本文所提方法可廣泛應用於漢字教學以及提供漢字習得理論研究相當有效的工具。

誌謝

本文作者感謝科技部計畫編號 NSC 102-2511-S-151-002 及 NSC 103-2911-I-003-301 的支持，同時也感謝教育部及國立台灣師範大學「邁向頂尖大學計畫」的支持。

參考文獻

[1] Bressan, S. (2004). Morphologic non-word error detection. Proceedings of IEEE 15th International Workshop on Database and Expert Systems Applications. 31-35

[2] Chang, C. H. (1995). A New Approach for Automatic Chinese Spelling Correction. Proceedings of Natural Language Processing Pacific Rim Symposium’95, Seoul, Korea. 278-283.

[3] Chang, T. H., Sung, Y. T., & Lee, Y. T. (2012). A Chinese word segmentation and POS tagging system for readability research. Paper presented at 42nd Annual Meeting of the Society for Computers in Psychology, Minneapolis, MN, USA.

[4] Chang, T. H., Su, S. Y., & Chen, H. C. (2012). Automatic Correction for Graphemic Chinese Misspelled Words. Proceedings of the 24th Conference on Computational Linguistics and Speech Processing, Taoyuan, Taiwan. 125-140.

[5] Chang, T. H., Chen, H. C., Tseng, Y. H., & Zheng, J. L. (2013). Automatic Detection and Correction for Chinese Misspelled Words Using Phonological and Orthographic Similarities. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan. 97-101.

[6] Chiu, H. W., Wu, J. C., & Chang, J. S. (2013). Chinese Spelling Checker Based on Statistical Machine Translation. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan. 49-53.
[7] Friedman, C., & Sideli, R. (1992). Tolerating spelling errors during patient validation. Journal of Computers and Biomedical Research, 25(5), 486–509.
[8] Huang, C. M., Wu, M. C., & Chang, C. C. (2007). Error Detection and Correction Based on Chinese Phonemic Alphabet in Chinese Text. Proceedings of the Fourth Conference on Modeling Decisions for Artificial Intelligence(MDAIIV). 463-476.
[9] Lin, Y. J., Huang, F. L., & Yu, M. S. (2002). A Chinese Spelling Error Correction System. Proceedings of the Seventh Conference on Artificial Intelligence and Applications. 207-212.
[10] Liu, C. L., Lai, M. H., Tien, K. W., Chuang, Y. H., Wu, S. H., & Lee C. Y. (2011). Visually and Phonologically Similar Characters in Incorrect Chinese Words: Analyses, Identification, and Applications. Journal of ACM Transaction on Asian Language Information Processing, 10(2), 10:1-39.
[11] Liu, X., Cheng, F., Luo, Y., Duh, K., & Matsumoto, Y. (2013). A Hybrid Chinese Spelling Correction Using Language Model and Statistical Machine Translation with Reranking. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan. 54-58.
[12] Ren, F., Shi, H., & Zhou, Q. (2001). A hybrid approach to automatic Chinese text checking and error correction. Proceedings of 2001 IEEE International Conference on Systems, Man, and Cybernetics. 1693-1698.
[13] Varol, C. (2011). Pattern and Phonetic Based Street Name Misspelling Correction. Proceedings of 2011 IEEE Conference on ITNG, Las Vegas, NV. 553-558.
[14] Vachharajani, V. (2012). Effective Label Matching for Automatic Evaluation of Use – Case Diagrams. Proceedings of 2012 IEEE Fourth International Conference on Technology for Education, Hyderabad. 172-175.
[15] Wu, S. H., Lin, C. L., & Lee, L. H. (2013). Chinese Spelling Check Evaluation. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan. 35–42.
[16] Yeh, J. F., Li, S. F., Wu, M. R., Chen, W. Y., & Su, M. C. (2013). Chinese Word Spelling Correction Based on N-gram Ranked Inverted Index List. Proceedings of the Seventh SIGHAN Workshop on Chinese Language Processing (SIGHAN-7), Nagoya, Japan. 43-48.
[17] Zhang, L., Huang, C., Zhou, M., & Pan, H. (2000). Automatic detecting/correcting errors in Chinese text by an approximate word-matching algorithm. In: Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics, 248–254.
[18] 丘慶鈴 (民 92)。避免小學生寫錯別字之教學策略 (碩士論文)。國立新竹師範學院，新竹市。
[19] 陳思彧 (民97)。國民小學高年級別字矯誤教學研究 (碩士論文)。臺北市立教育大學，臺北市。
[20] 林佳儂 (民 98)。國中生錯別字教學策略研究 (碩士論文)。國立彰化師範大學，彰化縣。
[21] 張嘉惠、林書彥、李淑瑩、蔡孟峰、李淑萍、廖湘美、孫致文和黃鍔 (民 99)。以最佳化及機率分佈標記形聲字聲符之研究。中文計算語言學期刊，15(2)，70-84。
[22] 陳學志、張瓔匀、邱郁秀、宋曜廷、張國恩（民 100）。中文部件組字與形構資料庫之建立及其在識字教學的應用。教育心理學報。269-290。

[23] 莊德明、謝清俊（民 94）。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。