形音相近的易混淆漢字的搜尋與應用

劉昭麟　　黃志斌　　翁睿妤　　莊怡軒
國立政治大學 資訊科學系
{chaolin, g9614, s9403, s9436}@cs.nccu.edu.tw

摘要

在中文裡面，漢字包含因爲發音相近或者形體相似的易混淆字，這一些易混淆字對於電腦輔助教學和語言心理學的相關研究具有相當意義。我們運用倉頡碼的設計理念和電子詞典所提供的發音資訊，配合網際網路可以得到的文字資訊，設計一個不須仰賴影像處理技術，就可以找到形音相近漢字的方法。經過實驗證明，以提交五個甚至一個建議字爲限，我們的方法所建議的形音相近字集，能夠包含一般與專業受試者所提供的常見錯別字集。

關鍵詞：漢字研究、漢字搜尋、漢字構字資訊、電腦輔助語言教學、語文認知

1. 簡介

個別漢字是構成中文的基本單位，有自己的發音、筆畫構造與所攜帶的意涵；透過個別漢字所組成的單字詞、雙字詞等詞彙，依據漢語語法組成中文句子。因此，學習漢字雖然不是學習漢語會話的必要工作，但卻是進階中文學習者一個重要的功課。同時，語言使用者如何透過語言的聲音(pronunciation)和文字的形體(grapheme)來擷取語意，更是研究語言認知歷程的學者所專注的重要議題。因此，本論文探討如何利用軟體技術找尋因爲發音和形體近似而容易混淆的漢字，以供電腦輔助教學和認知語言學的研究之用。

中文句子「今天上午我們來試場買菜」包含一個典型的錯誤；試場雖然是一個存在的詞彙，代表考試的場所，但是除非情境特殊，否則在這一例句裡面的「試場」應改爲「市場」。「經理要我構買一部計算機」這個句子也有一個錯誤：「構買」應改爲「購買」。雖然在簡體字的環境中比較多的人會寫「構買」，但是在繁體中文的使用群中，也有人把「購買」寫成「構買」。

因爲形音近似而誤用詞彙並不是中文所特有的現象，英文也有類似的問題[4]。舉例來說，“John plays an important roll in this event.” 包含一個錯誤的字；“roll” 應改爲“role”。其他像下列這一些字組，都是易混淆字的範例，principle和principal、teen和team、there和their、leak和leek、wait和weight、knows和nose以及knit和nit等等。

形音近似的漢字常被用於國民小學國語科試題的「改錯字」試題[6]。教師把一句正確的中文句子其中一個字改成另一個具有相當吸引力的錯字，以這一句帶有錯字的中文當作試題，要求受測學生找出並且更正這一錯字。這一類的試題也可以變形爲中文的

克漏詞試題(cloze) [9, 13]，克漏詞試題雖然在中文試題中比較少出現，卻是國內外英文測驗，如托福、GRE和大學指考等，幾乎是必然採用的題型。

形音近似的漢字在語言心理學的研究上也相當有用。Taft、Zhu和Peng [15] 研究部首位置對於受試者的詞彙決策(lexical decisions)與命名反應(naming responses)。Tsai等學者[16]則研究相近漢字的字數的多寡(neighborhood size)對於詞彙決策與閱讀的影響。Yeh 和 Li [17] 研究近形字對於一個熟練的中文閱讀者所執行的詞彙決策的影響。

發音相近的字可能可以藉由電子詞典所記載的資訊來判斷；相對地，形體相近的字則尚未有簡易的方法來找尋。影像處理技術雖然可能有用，但是對於爲數眾多、且近似方式繁複的漢字來說，應用影像處理技術的時效恐怕不佳。本文從應用朱邦復所設計的倉頡碼出發[2]，改變倉頡碼的原始設計，參考原本爲了補足漢字字形缺字所創造的漢字構形資訊[1]，得到一套可以爲任何漢字找尋形體近似的漢字的方法。

結合所找到音形相近的漢字字集之後，我們利用谷歌(Google)的搜尋介面所提供的資訊來排序所找到的字集的候選字，藉此排序可以限制我們所提供的近似字的字數。實驗結果顯示，不管以真人受試者或者專家意見作爲評比的標準答案，我們的系統所提供的字集都能有效協助教師編輯高品質的「改錯字」試題。

我們在第 2 節討論如何利用倉頡與構形資訊來建構一個找尋近形字的子系統。在第 3 節討論找尋漢字同音、近音字的技術問題。在第 4 節討論如何利用谷歌搜尋所得的資訊，來評比形音相近的字當中哪一些字是比較具有吸引力的錯別字。我們在第 5 節提報和分析相關的測試的結果。第 6 節則是簡單的結語。

2. 搜尋形體近似的漢字

我們在第 1 小節介紹一些近形字，在第 2 小節簡述倉頡輸入法如何將中文字編碼，在第 3 小節說明我們如何改進現有倉頡碼的編碼方式，最後第 4 小節說明我們利用關於個別漢字的資訊來找尋近形字的方法。

2.1 近形字實例

圖一、圖二和圖三包含三大類容易搞混的中文字，我們用空白將相似的中文字做分群。圖一當中的近形字，差別只在於筆劃的層次。圖二第一行各群的近形字分享同一個部件(component)而非部首。圖二第二行各群近形字則是分享同一個部件同時也是部首。圖二各組的近形字都有不同的發音。圖三爲六組分享同一部件的同音異義字。發音與內部結構相近的近形字最能造成語文學習者學習上的困擾。

士土工干千 戌戍成 田由甲申
母毋 勿匆 人入 未末 采釆 凹凸

圖一、主要差異在筆畫層次的漢字

頸勁 搆溝 陪倍 硯現 裸棵 搞篙
列刑 盆盎盂盅 因困囚 間閒閃開

圖二、形體相近的漢字

形刑型 踵種腫 購構搆 紀記計
園圓員 脛逕徑痙勁

圖三、形體與發音皆相近的漢字

要有效率地找到形體相近的漢字並不見得是一件簡單的事。藉由圖像比對方法找出

形體相似的漢字，雖然是一個可能的方法，但是卻有相當的困難。以「構」與「購」爲例，雖然以肉眼比較這兩個字的影像的時候，我們會覺得這兩個字的右側所共享的部件「冓」會重疊。實際上，經過我們測試，這樣的直覺是一個誤判。字形檔的建構，並不保證共享的部件的所有影像點(pixels)都必須能夠重疊，即便共享的部件確實有相當的影像點應該可以重疊在一起。

除了以上所描述的「非完美重疊部件」的問題之外，漢字之間的相似關係還有別的類別。以「員」和「圓」爲例，不管我們把這兩個字的影像如何平移，所得的最大交集的影像點的數量可能都不容易讓我們認定這兩個漢字的相似性。所謂「相似」，其實有其主觀的因素存在，雖然不一定每一個人都會認爲「員」和「圓」相似，但是大多數的人應該都會接受這樣的看法。在某一些可能是有一些極端的應用之中，我們或許還會希望我們的程式可以找到「貝」和「圓」的相似處，這時「貝」甚至只是「圓」的內部構件的一小部分。又請看圖三中第二行右手邊的字群，他們共同分享的部件出現在不同的位置。這時候影像處理技術雖非毫無用武之地，但是所須進行的計算量可能就不小，除了平移還須要考慮放大（或者縮小）的問題。不管是平移或者是放大，都須要決定平移量、平移方向和放大的比例，這一些決策都會使得計算變得相當地複雜。而即便引入其他更加複雜的演算法，例如紋路分析(texture analysis)，計算速度也是很難提供即時快速的服務。

上述的討論，還侷限在兩個漢字的直接比對上。如果考慮到漢字的數量龐大，計算的功夫就可能更加耗時費力。中文擁有超過 22000 個漢字[11]，所以直接用影像比對字的相似度須要很大的計算量；如果欠缺一些有效資訊支援，直接比較任意兩個漢字的話，就必須處理超過 4.8 億種組合。如果只有考慮我國教育部所提出的 5401 個中文常用字[3]，則大約會有 2900 萬種組合。

詞典編纂者利用中文字的部首(radicals)，將中文字在字典中有組織地進行分段，因此部首訊息是有用處的。在圖二中的第二行，我們舉了一些例子。這些字群中擁有的共同部件，皆爲這些中文字的部首，所以我們可以在中文字典中的某一段落，找到同屬這一個字群的中文字。然而光靠詞典編纂者定義的中文部首資訊是不夠的。在圖二中第一行的中文字群，有著共同的部件。然而這些部件並非中文字的部首，舉例說明：「頸」及「勁」在字典中分屬於兩個不同的部首。

2.2 倉頡原始碼

倉頡輸入法以 25 個字作爲基本單位，創造出一套分解漢字的方法；透過這 25 個字的組合，就能把漢字輸入到電腦中。倉頡輸入法分解漢字的方法，雖然不是非常完美，但是這一個分解個別漢字爲基本單位的出發點，跟我們尋找近形字的需求是相接近的。

表一分成三個主要部分，由左而右分別列出圖一到圖三部分漢字的倉頡碼。在一部有安裝倉頡輸入法的電腦上，可以用倉頡碼輸入中文字，例如輸入「一一一月金」的話，就可以得到「頸」（註：「一一一月金」是英文鍵盤上的 MMMBC）。在倉頡輸入法中，每個漢字都被分解成爲一個有序的元素；簡而言之，我們可以發現其中的子序列能組合成一個字的主要部件。很顯然地，透過計算個別漢字所分享的倉頡碼的數目，是一個可

以決定相似字的方式。舉例來說，我們可以說「搞」和「篙」是相似的，因爲他們的倉頡碼裡都有代表「高」這個部件的「卜口月」。我們也可以輕易發現「踵」、「種」和「腫」分享了「重」這一個部件，因爲他們的倉頡碼都包含了「竹十土」這一個子序列。

表一、 一些漢字的倉頡（原始）碼

漢字	倉頡碼	漢字	倉頡碼	漢字	倉頡碼
士	十一	頸	一一一月金	踵	口一竹十土
土	土	勁	一一大尸	種	竹木竹十土
工	一中一	硯	一口月山山	腫	月竹十土
干	一十	現	一土月山山	購	月金廿廿月
勿	心竹竹	搞	手卜口月	構	木廿廿月
匆	竹田心	篙	竹卜口月	圓	田口月金
未	十木	列	一弓中弓	員	口月山金
末	木十	刑	一廿中弓	脛	月一女一
		因	田大	逕	卜一女一
		困	田木	徑	竹人一女一
		間	日弓日	痙	大一女一
		閒	日弓月		

然而，某些形狀有微妙變化的漢字，倉頡碼似乎無法提供出它們相似的證據；例如「士土工干」和其他列在表一最左邊欄位內的字。這些字是依據特殊的分解規則解構的，這種特殊的規則使得我們無法輕易利用倉頡碼的相似度來找尋近形字。

爲了維持輸入一個漢字不須要敲擊超過五個鍵的輸入效率，倉頡輸入法蓄意簡化某些部件較多或者較複雜的漢字的倉頡碼。例如，在「脛」和「徑」的倉頡碼裡，「一女一」代表了「巠」這個部件，但是在「頸」和「勁」的倉頡碼裡，「巠」這個部件卻被簡化成「一一」。而「員」的「口月山金」在「圓」的裡面只剩下「口月金」。

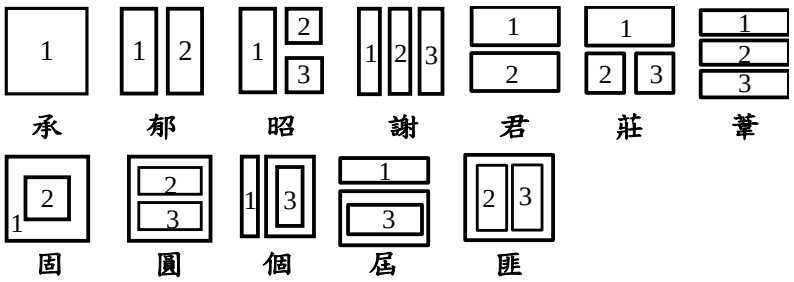
以輸入效率作爲設計要件，倉頡輸入法簡化用來代表個別漢字的內碼序列這一作法是可以理解的。然而，這樣的簡化程序使我們難以依照真實的倉頡碼來比對字的相似度。因此，我們並沒有直接採用倉頡碼做爲比較漢字相似度的基礎。爲了自己的需求，我們恢復了被倉頡輸入法所簡化的內碼，以鍵盤能夠用倉頡輸入法打出的最小構件做爲恢復原則。例如不管「巠」這個部件怎麼樣被簡化，如果我們用倉頡輸入法打「一女女一」能夠得到「巠」的話，那麼就將「巠」被簡化後的倉頡碼一律恢復爲「一女女一」。爲了在稱呼上有所區分，我們稱呼原始的倉頡碼爲***倉頡原始碼***，而我們所建構的倉頡碼爲***倉頡詳碼***。

2.3 倉頡詳碼與漢字構形資訊

雖然經過簡化所得的倉頡碼可以提升輸入效率，卻也造成我們在比對相似度的困難。因此，我們選擇使用完整的倉頡碼來建構我們的漢字資料庫。舉例來說，在我們的資料庫裡，「巠」、「脛」、「徑」、「頸」和「勁」的倉頡碼分別對應到 「一女女一」、「月一女女一」、「竹人一女女一」、「一女女一一月山金」和「一女女一大尸」。

除了不因輸入效率而簡化倉頡碼之外，我們還可以利用漢字的構形資訊來提高我們找尋近形字的效果。中研院早在將近十年前就開始了漢字構形的相關研究[1]，這一研究方向發展出了一套可以建構各種漢字字形的技術[11]。漢字的構造特殊，一般都認爲是方塊字，在這四四方方的空間裡面，我們又可以把一個漢字切割成幾個小部分，每一部分都是一個子結構；這一些子結構雖然可能正好是詞典所列的部首，有些則不然。目前我們採用Lee切割漢字的方式[12]。這一方式是以倉頡碼的「連體字」、「字首」、「字身」、「次字首」和「次字身」的觀念出發。儘管中研院所提出的切割方式並沒有完全侷

限在倉頡碼的切割觀點[1]，但是因爲漢字本身幾乎欲蓋彌彰的構造，使得我們目前所採用的切割方式與中研院的看法多所類似。



圖四、以倉頡碼爲基準的漢字基本構形

以圖二所列的漢字爲例，有些字可以從垂直方向分解成兩個部分，像是「盅」可分解成上面的「中」和下面的「皿」；有些字可以被水平分解，像是「現」是由「王」和「見」這兩個部件所組成的；而有些字則是一個包圍的結構，譬如說「囚」這個字是「人」部件被「口」部件包圍起來。因此，在決定漢字的相似度時，我們可以同時考慮構形資訊和所分享的相同部件這兩個因素。

雖然漢字有其看似自然的構形方式，但是並非每一個學者或者母語使用者都認同某一切割方式。圖四是Lee切割漢字的構形方式[12]，每個構形方式下方有一個漢字作爲範例。爲了記錄構形方式，我們賦予構形方式編號(以下稱爲***構形編號***)，在圖四裡面，由左到右、由上到下依序從 1 號開始編號。構形中的一個小方塊表示一個子結構，一個漢字最多可以有三個子結構，而且我們用數字替這些部件編號(以下稱爲***構件編號***)。依照倉頡區分子結構的方式，圖四上列最左邊的字是***連體字***，其他的字都有子結構，所以稱爲***分體字***。在分體字裡面，構件編號 1 號的子結構稱爲***字首***；如果一個分體字總共只有兩個子結構的話，構件編號 2 號的子結構就是***字身***。如果一個分體字有三個子結構的話，則構件編號 1、2 和 3 號的子結構分別是***字首***、***次字首***和***次字身***。倉頡輸入法的漢字切割方式，十分看重最左側的子結構和最上方的子結構，因此，構形編號 2、3、4 和 10 號的字的字首都是最左側的子結構；構形編號 5、6、7 和 11 號的字首都是上方的子結構。有外框的構形的字則都是以外框子結構當作是字首。

表二、一些漢字的倉頡詳碼

漢字	構形編號	構件 1	構件 2	構件 3	漢字	構形編號	構件 1	構件 2	構件 3
承	1	弓弓手人			頸	2	一女女一	一月山金	
郁	2	大月	弓中		徑	3	竹人	一女女一	
昭	3	日	尸竹	口	脛	3	月	一女女一	
謝	4	卜一一口	竹難竹	木戈	痙	7	大	一女女一	
君	5	尸大	口		員	5	口	月山金	
森	6	木	木	木	圓	9	田	口	月山金
葦	7	廿	木一	手	相	2	木	月山	
因	8	田	大		想	5	木月山	心	
國	9	田	戈	口一	箱	6	竹	木	月山
個	10	人	田	十口					
屆	11	尸	山	土					
匪	12	尸	中一	卜卜卜					

這一些構形資訊，除了影響倉頡碼簡化的規則之外，對於我們的系統提供了有利的資訊。如果把構形資訊也放進個別漢字的資料裡面的話，我們就可以知道兩個漢字的外型是否相似。也可以知道兩個漢字如果有相同的子結構，那這一相同的子結構分別出現在方塊字的那一個子空間中。這樣的資訊有助於我們更加精確地掌握兩個相似的程度。

當所有簡化的倉頡碼都回復成倉頡詳碼之後，我們可以把每個字都加上所屬的構形編號，並且將其倉頡碼依照構形中的構件編號分類。因此，我們所建構的字的紀錄包含構形標籤和一到三組倉頡碼序列。表二所列的紀錄包含圖四中的字的倉頡詳碼，並加上一些我們之後會討論的字的紀錄。在原本倉頡碼裡被省略的元素以方框框起來表示；例如，「徑」的倉頡原始碼是「竹人一女一」，同時「徑」的倉頡詳碼是「竹人一女女一」。

將原本倉頡碼中省略的元素重新加進去，以及把倉頡碼的子序列分成三部分都有助於我們辨別相似的漢字。現在我們可以輕易的分辨出由「一女女一」和「一月山金」所組成的「頸」，與「竹人」和「一女女一」所組成的「徑」是兩個相似的漢字。我們的系統也可以推測得知「員」及「圓」是相近的字。如果我們只採用倉頡原始碼的話，這將是一件不容易辦到的工作。

構形資訊則提供更加精確的近似度的訊息，「頸」、「徑」、「脛」和「痙」的倉頡詳碼都包含了「一女女一」這個子序列，但是因爲「徑」和「脛」屬於同一構形，而且所共享的子序列屬於同一子結構，所以「徑」和「脛」的相似度要高於「徑」和「頸」及「徑」和「痙」的相似度。實際上，我們還可以導入一個簡單的機制，在尋找某一特定構形編號的字的時候，特別偏好某一特定構形編號的字。例如，當我們在搜尋本身屬於構形編號 2 的字的近形字的時候，比較偏好構形編號是 2、3、4 和 10 的字，因爲這一些屬於這一些構形的字的輪廓線是類似的。

2.4 比對的方法和計算的效率

我們已經以人工建立了五千多個漢字的倉頡詳碼的資料庫[7]。從這五千多字之中，找尋一組形體相近的漢字，只須要用最簡單的字串比對即可。在最糟糕的情況下，比對兩個都擁有三個子結構的字的時候，我們必須比較他們各自的倉頡碼子序列九次。實施字串比對的步驟是很簡單的，因此計算量極小。

在計算近似度的時候，目前我們考慮幾個因素。首先是兩個漢字的倉頡詳碼的子結構是否有完全相同的倉頡碼。如果有的話，我們會記錄這一共享的子結構的倉頡碼的個數當作分數。如果共享的子結構是同一空間位置的子結構的話，則分數還會加倍。分別比對完最多九個子結構組合之後，會把所得的分數加總，然後得到一個初步的總分。我們以這一個總分把資料庫裡面的字排序(爲了計算效率，我們很早就把得到零分的字捨棄)；如果有同分的字，則再以構形編號和目前所查詢的漢字的構形編號相同者作爲最後分數較高者，然後才是構形相似的字。

以表二的「頸」、「徑」、「脛」和「痙」爲例。如果我們是要尋找「徑」的近形字時，因爲這四個字共享「一女女一」這一個構件，所以我們可以知道「頸」、「脛」和「痙」跟「徑」有基本的相似性；相對之下，其他的字，如「員」，跟「徑」沒有相同的構件，因此不是相近的字。除了知道「頸」、「脛」和「痙」跟「徑」相近之外，我們如何知道

這三個字之中哪一個字跟「徑」比較相像？在這一個例子裡面，我們發現「脛」和「徑」的共享構件都是出現在第二個構件。相對地，「頸」的「一女女一」是在第一個構件，而「痙」的構形編號是 7，並不是構形編號 2、3、4 和 10 的字，因此視覺上跟「徑」的相似度還是比不上「脛」。

在一部使用 2.24G 的 RAM 和 Windows XP 的 Pentium 4 2.8G 機器上，從五千多個字搜尋資料庫內的字的相似字，只須要花費不到一秒鐘的時間。如果是要把這一系統當作是電腦輔助出題系統的基礎子系統的話，這樣的速率，在實務上應該符合時效。

3. 發音相近的漢字

我們可以利用一部可以提供漢字發音訊息的電子化詞典，找出任意字的同音字或者近音字。如果能夠掌握一部可靠的詞典，則找尋同音字是一件簡單的事情；例如「試場」和「市場」。然而要找尋所謂的近音字，例如「光陰」和「光影」，則須要一些語音學知識的支援。

目前我們還沒有考慮實際生活中漢語發音的變調(Sandhi)[10]現象，所以在判斷個別漢字發音的近似度的時候，還沒有考慮語境的影響，因此並不是絕對準確。變調是許多語言都有的現象，最常聽到的漢語變調規則是兩個連續的三聲字的第一個三聲字要以二聲來發音；例如，在自然發音的情境下，「選舉」和「演講」的「選」和「演」都被當作二聲字來發音。如果有多個三聲字連續出現，則還有更加複雜的發音慣例。此外，習慣上我們會把「媽媽」的第二個「媽」以輕聲來發音，這也是一種受語境影響而改變發音的例子。目前我們還沒有完成處理這一類發音變化的程式。

除此之外，我們可以蒐集一些生活中一些常見的發音問題或者透過問卷調查，以獲取母語使用者對於所謂「容易混淆的音」的資料。舉例來說，在一般國民小學的教學經驗裡面，ㄓ、ㄔ和ㄕ這三個音分別容易跟ㄗ、ㄘ和ㄙ混淆。類似的經驗如，ㄣ跟ㄥ這兩個韻母相當接近，所以「金雞」和「京畿」聽起來很相像；而「ㄏㄨㄢ」 這一個發音跟「ㄈㄢ」聽起來相當接近，所有常有人把「冒險患難」寫成「冒險犯難」；或者有人把「翻書」唸成「歡書」。這一類的「近音字」如果有好的資訊來源，很容易由程式列出所有相關資料。目前我們採用了中研院語言所李佳穎研究員所提供的一些資料[5]，作爲尋找漢字近音字的依據。

4. 排比易混淆的漢字

我們可以用第 2 節和第 3 節所闡述的技術，來找尋單一漢字的近形字、近音字和同音字。我們可以利用這一些相近的字，爲某一個詞彙找出易混淆的錯誤寫法，例如，蓄意把「冒險患難」寫成「冒險犯難」。這類的錯誤詞彙對編寫國語科的「改錯字」試題和研究閱讀認知歷程都有特定的用處。

對近形字而言，不管是以中研院的漢字構形資訊[1]做爲構形編號的依據也好，或是以Lee切割漢字的構形方式[12]做爲依據也罷，本論文所提出的找近形字技術只是一種過濾機制，試圖取出「可能」會被誤用的易混淆字群而已；我們並不急著在近形字的部分就取出最常被誤用的易混淆字，因爲最常被誤用的易混淆字也有可能是同音字或是近音字。

以「勉強」這一個詞爲例，如果目的是把「勉」由一個錯別字來取代，則可以應用找尋近形字的技術找到「免」和「兔」，再用找尋近音字的技術找到「冕」、「娩」和「緬」等其他同音字。於是我們面臨了如何把這一些近形字、同音字和近音字排序，好讓我們系統的使用者盡快找到滿意的錯別字的需求。跟一般的注音輸入法類似，愈容易混淆的漢字最好是放在建議名單的前頭，以減少使用者的搜尋時間。這就是本節所要交代的「排比」問題。

實際經驗顯示，雖然我們可以依據倉頡詳碼找尋近形字，但是光是形體相近並不見得就是很有用的錯別字；因爲即使兩個漢字真的有一些相似的地方，也不一定會讓人們感到容易混淆。以經驗上的直覺來看，「改錯字」試題比較常用同音字或者近音字來取代正確的字。形體相近的漢字可能對於語言心理學中關於語文閱讀的研究有比較大的用途。儘管如此，我們仍將以編寫「改錯字」試題爲目的，探討如何排比我們所得到的候選字（包含近形字、同音字和近音字）。

延續「勉強」這一個例子，我們如何猜測「免」、「兔」、「冕」、「娩」和「緬」個別的適合程度？一位出題老師當然有他的主觀感覺，但是一個軟體程式如何能有這樣的「主觀」意識呢？我們借重谷歌(Google)所提供的搜尋服務來模擬這樣的直覺。如果我們以「免強」加上雙引號進行查詢，在所得的查詢結果中，觀看 “Results of 1-10 of about 220,000 …”（附註：如果是使用谷歌的中文介面，則會看到 “關於**免強**大約有 220,000 頁…”），可以知道大約有廿二萬筆網頁資料用到「免強」這一個詞。在查詢的時候加上雙引號，用意在於告知谷歌把查詢的詞彙當作一個連續字串、不可以分開查詢，因此會排除只包含類似「免，強」之類的網頁資料，所得的結果會比較符合我們的需求。我們可以撰寫一段簡單的程式，把所要查詢的條件傳送給谷歌，然後再從所得的回傳資料，透過簡單的資訊擷取機制得到所要的數字。這一程序所得的數字大略地反應了網路社群中使用這一個錯誤詞彙的頻率，可以詮釋爲生活中人們犯同樣錯誤的相對機會，如果數量愈大則表示人們採用那一個錯誤的詞彙的機會也相對地高。

我們以「免強」、「兔強」、「冕強」、「娩強」和「緬強」這五個詞，分別加上雙引號作爲給谷歌的查詢關鍵詞，會得到 222,000、4720、506、78 和 1510（附註：這一批數字得自於 2008 年 7 月 7 日的試驗）。因此，如果我們要從「免」、「兔」、「冕」、「娩」和「緬」之中，提交三個候選字給使用者時，我們依序提出「免」、「兔」和「緬」；如果是要提交五個候選字的話，則依序在後面加上「冕」和「娩」給使用者。

5. 實驗結果與分析

在這節裡，我們將以編寫「改錯字」試題爲目的，檢驗我們的系統是否能找出現實生活中易混淆的漢字。本節將分成四個部分來介紹我們所使用的實驗設計、分別使用一般受試者與專家意見所進行的實驗結果、最後討論使用倉頡詳碼判別近形字的缺點。

5.1 實驗設計

首先我們從*新編錯別字門診*[8]這一本書，找出 20 個包含有容易混淆的字的詞彙。表三所列的是我們所選定的詞，每一個詞彙都包含一個以底線標示的粗體字。爲了行文簡潔，以下我們以***易混淆字***來稱呼表三裡面這一些以底線標示的粗體字。這個易混淆字將會特意地被一些***錯別字***來取代。

我們以兩種資料來檢驗我們系統所提出的建議錯別字的品質。我們請真人受試者寫出他們認爲適合取代這一些易混淆字的錯別字。然後以所蒐集的這一些資料，來評比我們系統所建議的錯別字的效用和評量受試者之間的一致性。同時，我們也會利用*新編錯別字門診*這一本書所討論的常見錯字，來評比我們系統的建議字和受試者所寫的錯別字。我們以真人受試者所提供的資料來反應一般人對於這一些錯別字的選擇，而用書本所提供的錯別字來代表專家的意見。

首先，利用我們的系統將這 20 個易混淆字找出「近形」、「同音」、「近音」三個候選字表；再將這三個候選字表組合成「同音、近音」、「近形、同音、近音」兩個***建議字表***，並以建議字表裡的字逐一取替表三所列詞彙的易混淆字。以這樣程序所產生的詞彙，再利用第 4 節所描述的程序，取得個別詞彙被網頁資料採用的頻率，藉以將建議字表內的字排序，使得最前面的字爲被採用頻率最高的字。如果前一程序所得到的搜尋結果數量爲 0，再利用谷歌搜尋「包含全部的字詞」功能(即不加上雙引號直接進行查詢)所回傳的搜尋結果數量，由大到小排於前一個程序的排序之後。

本實驗希望探討我們的系統是否能找出現實生活中被用來取代易混淆字的錯別字，因此請了 21 位大學在學生（校名因匿名投稿之故，暫且不明列）擔任受試者來進行實驗。我們請這 21 位受試者針對前述 20 個詞裡的易混淆字寫下至多五個錯別字。所收集到的共 420 (=21×20)個題次的回覆中，一共包含 599 個字（包含重複的字），其中有 24 個其實不是漢字的錯字。這 24 個錯字答案分佈在 7 個真人受試者的答案裡面。如果不管字的對錯，平均每一題次，真人受試者每一題平均填寫了 1.426 個建議字；如果扣除錯誤的字的話，就只剩下 1.369 個建議字。

表三、測試詞彙的列表

編號	詞彙	編號	詞彙	編號	詞彙	編號	詞彙
1	一**剎**那	2	一**炷**香	3	眼花**撩**亂	4	相形見**絀**
5	作**踐**	6	剛**愎**自用	7	可見一**斑**	8	和**藹**可親
9	**彗**星	10	**委**靡不振	11	**穠**纖合度	12	待價而**沽**
13	獎**券**	14	意興闌**珊**	15	**罄**竹難書	16	**搔**首弄姿
17	根深**柢**固	18	**椿**萱並茂	19	煩**躁**	20	璀**璨**

我們採用資訊檢索相關研究中最常使

用的評估標準，即精確率(precision)與召回率(recall) [14]做爲評估我們系統的方式；精確率是系統所建議的字當中是標準答案的比例，召回率則是標準答案中被囊括到系統所建議的字的比例。在計算精確率和召回率的時候，我們把錯誤的答案也當作答案，實際上我們的系統是不能建議這一些根本不存在的漢字的。因此，我們目前並不會因爲真人受試者的回覆的品質不佳，而高估了我們的效用。這一些真人受試者的錯誤甚至還讓我們低估了我們系統的分數。

在目前的研究中，我們沒有採用 F 分數(F measure)。F 分數是利用召回率和精確率計算所得的單一分數，雖然可以提供基礎的比較。但是在我們的實驗跟漢字輸入法的評估相當類似，使用者所能接受的建議錯別字的數量可能極少，因此分別檢視精確率與召回率，比起只有提供 F 分數更能讓研究者看清問題的本質。

5.2 一般受試者的評估

我們以 21 位受試者爲表三的 20 個易混淆字所寫下的 20 組錯別字做爲標準答案來進行評估。表四爲「同音、近音」與「近形、同音、近音」兩個建議字表各取前五個字與前十個字的實驗結果，我們利用建議字表所提供的錯別字與另外 21 位受試者所寫下的 20 組錯別字一組一組地進行評估比對，分別得到 20 組各個易混淆字的精確率與召回率，接著把這二十組精確率與召回率做平均的計算，因此得到以這 21 位受試者的錯別字爲標準答案時建議字表的精確率與召回率。然後再計算這 21 組數據的平均，所得到的計算結果就是表四中的平均精確率與平均召回率。

表四所列的數據顯示，我們所提出的方法相當有效地捕捉到受試者的偏好。以「近形、同音、近音」資料所建構的建議表的實驗來說，平均精確率看起來雖然不高，在提交五個建議字和十個建議字的時候，平均的精確率分別略低於 0.2 和 0.1。不過這意味著不管是提交五個字或者十個字，我們的系統都能夠大約提供出受試者所寫的錯別字。依據前一小節的分析，平均而言，針對每一題次，受試者只有寫出 1.369 個實際上存在的漢字作爲錯別字。因此，我們的系統能夠捕捉到這一些特定的字並不是一件絕對簡單的任務。從這一個觀點看我們的系統，便能看出它的可用性。如果仔細比較一下，提交五個建議字的時候，平均有 0.88 個可用字；提交十個建議的話，平均就有 0.95 個可用字，命中率不可謂不高。

表四、兩組系統建議字表所達成的平均精確率與平均召回率

系統建議字表 / 效果評估	「同音、近音」		「近形、同音、近音」	
	取前五個字	取前十個字	取前五個字	取前十個字
平均精確率	0.166	0.094	0.176	0.095
平均召回率	0.618	0.672	0.649	0.680

這一組實驗，同時也讓我們看到近形字對於改錯字試題的編輯工作的貢獻度似乎不大。比較表四的左半側和右半側的實驗數據，我們發現加入近形字之後所建構的建議表雖然效果都有所提升，但是提升的幅度並不顯著。這一實驗結果暗示著中文錯別字跟發音（同音字或者近音字）的關係，可能比跟字的形體的關係要密切。這樣的觀察當然可能是跟表三裡面我們所選擇的詞彙有關。在表三裡面，常見的錯別字只有第四題是跟字形比較相關；第 15 題勉強也算跟字形相關。其他的試題則都是明顯的跟字的發音有比較高的關連。再以「不虛此行」爲例，除非是以同音字作爲搜尋的要件，否則很難找到

「不需此行」這一個誤用的形式。

表五、受試者之間的平均精確率與平均召回率

受試者代號	平均精確率	平均召回率	受試者代號	平均精確率	平均召回率
A	0.569	0.473	L	0.458	0.368
B	0.553	0.508	M	0.365	0.455
C	0.408	0.635	N	0.520	0.491
D	0.495	0.468	O	0.448	0.546
E	0.497	0.520	P	0.558	0.481
F	0.489	0.479	Q	0.370	0.513
G	0.580	0.462	R	0.379	0.559
H	0.408	0.304	S	0.441	0.444
I	0.628	0.509	T	0.435	0.543
J	0.539	0.431	U	0.451	0.491
K	0.531	0.443			

中文錯別字跟發音的關係是不是真的比跟字形的關係要來得密切呢？雖然我們是透過一個隨機的程序挑選出表三所列的 20 道題，因此也認爲我們的實驗數據應該是暗示了這一現象，但是須要一個更大規模檢驗才能更進一步地驗證這一直覺。

爲了能夠更客觀地評估系統的效果，我們又做了另一組實驗。從這 21 位受試者當中輪流取出一人所寫下的 20 組錯別字當做系統建議字表所提供的錯別字，並以其他 20 位受試者的錯別字當作標準答案，來評估這一個暫時被挑出來的受試者答案的品質。我們用英文字母 A 到 U 來代表這 21 位受試者，表五列出這一些受試者輪流被當作被評估對象時所得的分數。

在這一組新的實驗中，我們取出每位受試者所寫下的 20 組錯別字作爲假想的建議字表，再以假想之建議字表所提供的錯別字與另外 20 位受試者個別寫下的 20 組錯別字一組一組地進行評估。評估的過程中我們會計算 20 組各個錯別字的精確率與召回率，接著計算這 20 組精確率與召回率的平均，得到以個別受試者所提供的錯別字爲標準答案時的精確率與召回率。接著再計算這 20 組數據的平均，最後所得到的計算結果就是表五中的平均精確率與平均召回率。

以整個表五的數據來評估所有受試者的平均表現，分別把所有的精確率加總，然後除以人數，平均的精確率和平均召回率分別是 0.48200 和 0.48205，兩者幾乎相等。受試者之間的共識度雖然表面上看起來不高，但是這一些受試者在接受我們測試之前並未事先相互交換意見，同時是獨立回答問卷，所以這一個平均的精確率和召回率應該算是相當高。如果拿這一個總平均數跟表四的平均數相比的話，我們發現我們系統的精確率雖然比不上人類受試者，但是召回率卻能高於人類受試者間的召回率。這一項比較顯示我們的系統可以提供有用的服務，在容許系統提供五個建議字或者十個建議字的情形下，我們系統比人類受試者更能提供確實有用的建議字。

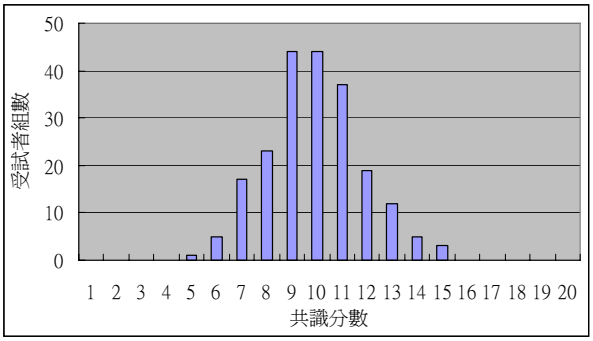
另外，我們對於這 21 位學生所認定的易混淆字是否有清楚的共識也有很大的興趣，於是我們取出每個學生所寫下的 20 組易混淆字當中的第一個字，並以此 20 個字爲該學生的***第一印象字***，故共有 21 組第一印象字，每組 20 字。接著以組爲單位，兩兩比較在相同位置上的字，若是一樣的話，則給予一分的權重；因此若是一模一樣的兩組來比較的話，則可以獲得滿分 20 分。我們將此認知共識的分析結果以方陣表示，並由以

表六、受試者之間的共識的分數方陣

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	甲	乙	丙
A	20	13	10	13	9	9	9	8	12	10	12	9	8	13	8	12	7	7	9	10	8	9	9	11
B	13	20	9	10	9	9	8	8	11	9	12	8	8	9	8	12	6	7	9	9	8	7	7	13
C	10	9	20	7	11	8	7	5	8	8	10	6	7	7	8	11	7	8	9	9	9	4	4	7
D	13	10	7	20	9	10	11	6	11	10	10	8	8	14	8	9	6	7	8	9	9	11	11	12
E	9	9	11	9	20	10	9	6	10	10	11	8	7	9	6	9	8	10	11	9	8	7	7	10
F	9	9	8	10	10	20	14	9	11	10	9	9	9	10	7	12	7	9	8	10	10	8	8	12
G	9	8	7	11	9	14	20	8	12	10	10	10	10	14	8	11	8	9	8	10	13	10	10	14
H	8	8	5	6	6	9	8	20	8	8	7	5	7	5	5	10	6	6	6	9	4	4	5	7
I	12	11	8	11	10	11	12	8	20	12	12	11	8	13	7	10	8	10	10	11	12	7	8	13
J	10	9	8	10	10	10	10	8	12	20	12	10	9	10	6	11	8	8	10	10	9	8	9	12
K	12	12	10	10	11	9	10	7	12	12	20	9	9	10	6	10	8	8	10	9	11	7	7	11
L	9	8	6	8	8	9	10	5	11	10	9	20	7	11	5	7	6	7	7	9	9	8	8	11
M	8	8	7	8	7	9	10	7	8	9	9	7	20	8	6	11	6	9	6	9	8	5	5	8
N	13	9	7	14	9	10	14	5	13	10	10	11	8	20	8	8	6	8	8	10	11	11	11	14
O	8	8	8	8	6	7	8	5	7	6	6	5	6	8	20	9	7	8	6	7	7	7	7	9
P	12	12	11	9	9	12	11	10	10	11	10	7	11	8	9	20	9	8	10	11	11	6	7	10
Q	7	6	7	6	8	7	8	6	8	8	8	6	6	6	7	9	20	11	10	9	7	7	8	8
R	7	7	8	7	10	9	9	6	10	8	8	7	9	8	8	8	11	20	12	9	8	5	5	9
S	9	9	9	8	11	8	8	6	10	10	10	7	6	8	6	10	10	12	20	10	9	6	7	11
T	10	9	9	9	9	10	10	9	11	10	9	9	9	10	7	11	9	9	10	20	10	7	8	12
U	8	8	9	9	8	10	13	4	12	9	11	9	8	11	7	11	7	8	9	10	20	7	7	10
甲	9	7	4	11	7	8	10	4	7	8	7	8	5	11	7	6	7	5	6	7	7	20	16	12
乙	9	7	4	11	7	8	10	5	8	9	7	8	5	11	7	7	8	5	7	8	7	16	20	13
丙	11	13	7	12	10	12	14	7	13	12	11	11	8	14	9	10	8	9	11	12	10	12	13	20

上的論述推導可知，此方陣必是對角線爲 20 之對稱方陣，如表六所示（表六中的「甲」、「乙」和「丙」爲標頭的三個欄和三個列的資料，將在下一小節說明）。

表六這一組原始數據可以讓我們看到個別受試者之間共識程度的變化，但是不容易看到一般的趨勢。我們可以分析以左上到右下的對角線爲準的右上三角形的數字，在不考慮對角線上的完美分數 20 分的情形下，計算有多少組受試者的共識是 19 分、18 分、…、0 分，把得到這一些分數的組數記錄下來，然後繪製一個受試者共識的趨勢圖。圖五是這一程序的產物，橫軸是分數，縱軸是得到橫軸所列分數的受試者組數。從圖形可以看得出來，受試者的共識分數像是常態分佈。經過簡單的計算，由 21 位受試者所形成的 210 個分數的平均數是 8.905。如果以受試之間的共識分數當作分子，完美的共識分數 20 當作分母計算，則這一平均分數的百分分數只有 44.5%。



圖五、21 位受試者的共識分數的分佈

5.3 專業意見的評估

*新編錯別字門診*的作者除了表列包含有易混淆字的詞彙之外，也提供了所列詞彙最常見的錯誤寫法。我們可以把這一些錯別字當作是專業意見，以這一些專業意見來評估我們的建議字表的效用和前面一般受試者所寫的錯別字的品質。表七的內容是把表三的易混淆字改成專家所提供的錯別字，第 7 和第 13 題有兩個可能的錯別字，第二順位的錯別

字放在括號裡面。

表七、專業意見管道所列的錯別字[8]

編號	詞彙	編號	詞彙	編號	詞彙	編號	詞彙
1	一霎那	2	一柱香	3	眼花瞭亂	4	相形見拙
5	作賤	6	剛復自用	7	可見一般(班)	8	和靄可親
9	慧星	10	萎靡不振	11	濃纖合度	12	待價而估
13	獎卷(券)	14	意興闌珊	15	磬竹難書	16	騷首弄姿
17	根深底固	18	樗萱並茂	19	煩燥	20	璀燦

我們以這一專家意見來評估我們系統的成效，重複前一小節建立表四的同一實驗，得到表八的數據。結果顯示，當我們以專家意見作爲標準答案的時候，我們系統所得的平均精確率和平均召回率，都還比用一般受試者的意見爲標準答案時要高。其中一部份的原因，應該是跟專家意見裡面不會有不存在的漢字，所以召回率明顯提高有關。

表八、依據專家意見爲標準所計算的平均精確率與平均召回率

系統建議字表 / 效果評估	「同音、近音」		「近形、同音、近音」	
	取前五個字	取前十個字	取前五個字	取前十個字
平均精確率	0.170	0.085	0.190	0.095
平均召回率	0.775	0.775	0.875	0.875

表九、以專家意見爲標準答案，21 位受試者的答案的品質

受試者代號	平均精確率	平均召回率	受試者代號	平均精確率	平均召回率
A	0.550	0.550	L	0.550	0.525
B	0.650	0.725	M	0.317	0.500
C	0.371	0.675	N	0.667	0.725
D	0.575	0.625	O	0.533	0.700
E	0.504	0.625	P	0.550	0.550
F	0.600	0.650	Q	0.329	0.550
G	0.750	0.700	R	0.327	0.600
H	0.400	0.375	S	0.458	0.525
I	0.675	0.650	T	0.467	0.675
J	0.575	0.575	U	0.458	0.575
K	0.525	0.500			

表九是以專家意見的錯別字爲標準答案的時候，21 位一般受試者所提供的錯別字的品質。計算表九 21 組數據的平均，可以得到一般受試者的平均精確率和平均召回率分別是 0.51576 和 0.59881。比起表五的平均值 0.48200 和 0.48205 要明顯高很多。這一項結果可以有兩種詮釋方式；直覺上的看法是：一般受試者的意見與專家意見有比較高的一致性；而另一個詮釋則是，書本的作者確實掌握到了一般讀者能夠想到的錯別字。

重複上一小節中第三個實驗時（只允許我們系統建議一個候選字），我們可以加入兩個建議表和專家意見。表六裡面的甲欄和甲列是「同音、近音」系統建議字表，乙欄和乙列是「近形、同音、近音」系統建議字表，丙欄和丙列是專家意見所得的分數。因爲表六是一個對稱方陣，所以同一標頭的欄與列的資料都會是一樣的。

我們計算甲欄和乙欄裡面，由上而下從 A 列到 U 列的一致性分數的平均，分別得到 7.19 和 7.52 分。也就是，這 21 位一般受試者跟我們兩種建議表的一致性大約落在七分的位置。如果計算丙欄，同樣這 21 個數字的平均的話，我們得到 10.66。專家意見跟一般受試者意見的一致性超過一半的測試題目，專家的意見還是比我們系統的建議更能捕捉到一般受試者的想法。

丙欄的甲列和乙列的分數，代表專家意見與我們兩個建議表的一致性，分別是 12 和 13 分，平均爲 12.50 分。如果拿 12.50 和 10.66 直接比較的話，專家意見給我們系統的分數還要高於給予 21 位一般受試者的平均分數。

5.4 使用倉頡碼的缺點

使用倉頡碼作爲基礎來比較字的相似度會存在一些潛在的問題。

倉頡碼在一些漢字的分類上，尤其是比較簡單的字，會採用一些模糊的規則，這讓我們在比對字的相似度時發生困難。舉例來說，「分」是採用圖四裡面的第五種構形，但是「兌」卻是採用第一種構形。此外，表一最左邊欄位所列舉出的字都很不容易被鑑定爲近形字。這一類筆畫非常簡單的字爲數不多，處理的方式可能是以人工建立資料庫，比起回頭運用影像處理技術來找近形字經濟得多。

一個字的倉頡碼的左半邊或者上半邊的字首最多只能有一個子結構。以表二的「相」、「想」和「箱」爲例，「相」這個字單獨存在時是使用構形編號 2 號；而在「箱」裡面，「相」這個子結構則是被分解成左右兩個部份。但是，在「想」這個字裡，「相」卻被當作是一個單獨的子結構，因爲它的位置處於整個字的上半邊，被當作一個字首。類似這樣的問題常常發生，例如「森」和「焚」及「恩」和「 困」。另外還有一些比較特殊的例子，像是「品」這個字使用第六種構形，但「闆」卻是使用第五種。

這一種問題的處理，暗示了我們須要重新檢視倉頡碼觀點的構形原則。我們或許應該建立自己的構形方式，這樣可以讓我們系統所找出的近形字的精確度更加提高。

6. 結語

本篇論文報告了我們如何利用倉頡碼拆解漢字的觀念來搜尋漢字的近形字。配合適當的電子詞典資料，我們的系統能夠從發音和形體兩個不同角度，找出所欲查詢的漢字的候選錯別字。我們進一步利用谷歌的查詢功能替所查到的候選錯別字排序，實驗顯示排序之後所得的建議字表確實能夠掌握一般受試者和專業意見所提供的錯別字。我們系統所產出的形音相近的字表，除了可以應用於電腦輔助試題編輯系統中的「改錯字」試題之外，也可以用於心理語言學實驗中檢驗中文使用者的閱讀行爲。

儘管現在有相當不錯的成果，但是我們也發現了以倉頡碼作爲系統設計的核心所引起的問題，我們也尚未完成對於漢語變調規則的處理程式，這一些都是正在進行的改進項目。

致謝

本研究承蒙國科會研究計畫 NSC-95-2221-E-004-013-MY2 的部分補助謹此致謝。我們感謝匿名評審對於本文初稿的各項指正與指導，雖然我們已經在從事相關的部分研究議題，不過限於篇幅因此不能在本文中全面交代相關細節。

參考文獻

[1] 中央研究院。中央研究院漢字構形資料庫。網址：http://www.sinica.edu.tw/~cdp/cdphanzi/。Last visited on 6 July 2008。

[2] 朱邦復。*第五代倉頡碼輸入法手冊*。網址：http://www.cbflabs.com/book/ocj5/ocj5/index.html。Last visited on 6 July 2008。

[3] 行政院主計處電子資料處理中心。CNS11643 中文標準交換碼。網址：http://www.cns11643.gov.tw/web/word/big5/index.html。Last visited on 6 July 2008。

[4] 任紹曾主編，張少伯譯。*同音異義詞 Homophones*，商務印書館，2000。

[5] 李佳穎研究員，中央研究員語言學研究所，私人通訊，2008。

[6] 林仁祥及劉昭麟。國小國語科測驗卷出題輔助系統，*2007 年台灣網際網路研討會論文集*，論文光碟，2007。

[7] 教育部字頻總表。網址：http://www.edu.tw/files/site_content/M0001/pin/biau1.htm?open。Last visited on 9 July 2008。

[8] 蔡有秩及蔡仲慶。*新編錯別字門診*，螢火蟲出版社，2003。

[9] J. Burstein and C. Leacock. Editors. *Proceedings of the Second Workshop on Building Educational Applications Using NLP*, ACL, 2005.

[10] M. Y. Chen. *Tone Sandhi: Patterns across Chinese Dialects*. Cambridge: Cambridge University Press, 2000.

[11] D. Juang, J.-H. Wang, C.-Y. Lai, C.-C. Hsieh, L.-F. Chien, and J.-M. Ho. Resolving the unencoded character problem for Chinese digital libraries, *Proceedings of the Fifth ACM/IEEE Joint Conference on Digital Libraries*, 311–319, 2005.

[12] H. Lee. *Cangjie Input Methods in 30 Days*, http://input.foruto.com/cjdict/Search_1.php, Foruto Company, Hong Kong. Last visited on 8 July 2008.

[13] C.-L. Liu, C.-H. Wang, and Z.-M. Gao. Using lexical constraints for enhancing computer-generated multiple-choice cloze items, *International Journal of Computational Linguistics and Chinese Language Processing*, **10**(3), 303−328. 2005.

[14] C. D. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*, The MIT Press, 1999.

[15] M. Taft, X. Zhu, and D. Peng. Positional specificity of radicals in Chinese character recognition, *Journal of Memory and Language*, **40**, 498–519, 1999.

[16] J.-L. Tsai, C.-Y. Lee, Y.-C. Lin, O. J. L. Tzeng, and D. L. Hung. Neighborhood size effects of Chinese words in lexical decision and reading, *Language and Linguistics*, **7**(3), 659−675, 2006.

[17] S.-L. Yeh and J.-L. Li. Role of structure and component in judgments of visual similarity of Chinese characters, *Journal of Experimental Psychology: Human Perception and Performance*, **28**(4), 933–947, 2002.