中文自動校正輔助系統
An Assistant System for Chinese Automatic Correction

劉如生　　張士蓮
Robin Liu and Shih-Lien Chang

元智大學電機暨資訊工程研究所
Institute of Electrical Engineering and Computer Engineering and Science
Yuan-Ze University

摘要

本論文提出一個考量中文輸入因素(以倉頡爲例)，由系統處理相鄰相近錯字之自動校正輔助系統。當以電腦繕打文件時，常因不慎誤觸了相鄰相近的按鍵，或將輸入字根的排列順序倒置而產生錯誤的文字。

針對上述問題，我們首先找出落單字(即錯字)，再利用中文輸入法反向查詢原理，求得原輸入的字根組合。其次將字根組依照相鄰、相近、次序、多根及少根的判斷條件，重新排列組合而成另一中文字，並以此文字進行前後文字詞的比對工作，如此訂正錯字。

關鍵字：中文錯字偵測，中文錯字訂正，斷詞，構詞，單字詞

ABSTRACT

This paper proposes a consideration on the Chinese input method - using 倉頡 as an example, and implements an automatic correction assistant system. For typing documents with computer, the wrong word may be generated due to touching adjacent/similar key stroke, or making wrong order radical incidentally.

To handle the above problems, we first find the isolated word(i.e., wrong word), then obtain the combination of radicals using reverse query of Chinese input method. Next, reconstruct the possibly related word according to the judgment on errors such as neighbor, similarity, sequent, excess or missing radical. With the substituted word and its front and rear words as new pattern, we match the pattern with dictionary and obtain the correct word.

Keywords : Chinese spelling check, spelling correction, word segmentation, lexical rules, monomorphemic words

一、緒論

中文電腦化是國人使用電腦上非常重要的一項課題，中文電腦中的字碼、字型、造字、輸入法等之研發，相關應用層面的詞庫資料庫、中文字辨識、錯字辨識與校對、語音系統、機器翻譯、資料檢索，都是中文電腦化可待探討的問題。

電腦所處理的中文字，在輸入過程中，常因錯誤的按鍵、或是人爲的會錯字，使得輸入的中文字產生錯誤，因而產生錯誤的表達與訊息。因此如果有相關的辨識系統，將有助於文章撰寫時訂正之檢查，大量降低校稿所需的人力及時間。有鑑於此，本論文將以電子中文文件之錯字辨識爲主要研究方向。

經鍵盤輸入儲存在電腦的中文文章有一個有趣之特性，那就是文章裏的錯字還是字，單靠檢查BIG-5 碼是不夠的，必需經過「詞」層次的分析才查得出來。另外，關於鍵入錯字的原因方面，可能是與正確的字同音、或音相近似(注音輸入者甚易發生，字根輸入者也有可能，因爲人的運作記憶是以語音爲基礎的)，也可能是形狀近似(字根輸入者易發生)。根據原因，才可以去推算使用者原來想打的字。

基本上，本論文運用文章辨識上的斷詞等方法，首先找出鍵盤鍵入的錯誤文字。其次，依據其中文輸入法，分析可能錯誤之原因，將錯字予以訂正。本論文所製作之自動校正輔助系統，乃提供一般文件製作來使用，使用者以中文輸入法依序鍵入文字至文件編輯器中後，可隨時啓動本辨識系統進行錯字偵測，檢查結果以及各項提示建議將展示在螢幕上供使用者訂正。

二、名詞解釋

2.1 中文輸入法

中文輸入法一般可分成兩類：一、字音輸入法：就是依照中文字的發音方式，由數個注音符號組合而成爲一個中文字。二、字形輸入法：訂定數十種可拼湊中文字的字根(類似於中文字典中的部首)，由這些固定數量的字根，組合出所有之中文字。

倉頡輸入法是一個以字根爲主的輸入法，主要是以 26 個字根組合出所有的中文字，其組合的過程中，字與字之間字根的重複性並不高，字的獨立性較強，所以此輸入法的特性比較容易掌握。反觀，注音輸入法，雖然是配合國人口語的方式輸入，但是由於中文字同音字太多，除了不易判斷之外，也極易造成混淆。

2.2 詞庫

中文文章是以詞與詞接續而構成，一個中文單字通常並不具備真正的意義，中文詞才是構成中文文章的基本單位。因此在分析處理中文資訊研究上，是以處理中文詞爲主。對於詞的判斷，又須仰賴詞庫內詞的數量，所以完整詞庫的建立是非常重要的工作。

詞庫依照內容，大致上可分爲「基本詞庫」與「專業詞庫」兩類。此外，本系統亦提供使用者自建「動態詞庫」之功能。

2.3 斷詞

所謂中文斷詞是指將輸入的中文句子，依據語意及文法結構，切割成以詞爲基本單位的工作[10]。

斷詞方法一般分爲「法則式斷詞法」與「統計式斷詞法」兩類。法則式斷詞法主要是由構詞規則爲出發點，使用長詞優先的原則，強調語言現象，由於有些漢字大多只出現在詞彙的首位或末位，因此也使用了字詞的結合性規則[9]。這類型的研究偏重於歧義性分析(disambiguation)，以詞典爲根本，配合不同的方法(規則法、相互訊息、限制傳遞法)找出最佳的組合[11]。

統計式斷詞法著眼於大量資料的處理。該方法認爲語言的性質，可以從大量的語料庫，經由數學模式獲得。因此主要是用歸納之統計數據資料爲判斷憑據，並且引用機率模式作爲斷詞的依據。這類系統以 Fan,C.K.,Tsai,W.H.的鬆弛法，及 Sproat-Shih 的統計式斷詞法爲代表。

部份學者將上述兩種方法合併，首先利用中文中自由詞素(free morpheme)或附著詞素(bound morpheme)的性質簡化斷詞步驟，並用一階馬可夫機率(Markov Model)列出所有可能的結果。

目前斷詞的技術都能達到 95%以上的正確率，然而因爲方法的不同，某些需要反覆的計算而造成速度較慢，有些則須建一個龐大的統計表。因此，這些斷詞系統在實用性上都會受到一些限制。

2.4 單字詞頻表

某些字很可能因爲無法與相鄰的字合併爲一個詞而落單，形成單字詞[13]。但是並不是所有單字詞都是錯字，事實上中文有許多使用頻率很高的單字詞，如「的」、「了」、「是」..等等，因此單字詞詞頻表是判斷落單的字是否爲錯字的重要條件之一。

2.5 構詞

有某些特別詞，在詞庫中找不到，也無法以人工預建詞典，但是卻有規則可跡尋。因此我們須內建有規則的統計表，利用這些構詞規則，來補充詞庫的不足。

構詞是利用構詞法則，來處理具有衍生性的中文詞。包括數字、定量詞、姓名和其他構詞規則。例如：數量詞(三個、1997、一九九七)，AABB(高高興興、快快樂樂)，A 一 A(看一看、吃一吃、玩一玩)，A 不 A(好不好、要不要、去不去)，AAB(吃吃飯、打打球)[13]。

三、相關研究

中文校正之處理，近年來已有不少先進積極從事開發工作，如 1992 年由施得勝、王良志、陳志達、聶素芬共同發表之基於統計的中文錯字偵測法[13]，主要是提出在文章中自動偵測錯字所在位置的方法。該論文是以統計的方法，在訓練大量的文句庫後，得到單字詞詞頻表及接續強度表，再以這兩個表爲基礎，經由評分函數(Score Function)計算出被懷疑的單字詞所得之分數，若小於「門檻值」(Threshold)，則標示爲錯字。其實驗數據顯示檢錯率可達 70%以上。

張照煌博士於 1994 年提出一種自動偵測並訂正中文文書中錯別字的方法[12]。其方法爲事先整理含字形、字音、字義或輸入碼相近字所形成的「綜合近似字集」代換原文，產生候選字串，其次利用語言模型評分，找出評分最高的候選字串，即可自動偵測文書中的錯別字；但是事先準備的「綜合近似字集」會因爲蒐集不足、過量的字集會造成誤判，另外近似字集代換後，所產生的候選字串數量龐大，在中文的檢錯方面勢必造

成軟體的負擔；

中文校正方面，另有 Kin Hong Lee 和 Chin Lu 在 1995 年所發表的文章[19]。

四、系統說明

本系統之研發步驟分別爲(1)分析輸入錯字的因素(2)錯字偵測(3)錯字訂正(4)準確率的評估。

4.1 分析輸入錯字的因素

電子中文字並無拼字或書寫上的錯誤，主要是因爲中文的輸入，使用者是利用各種字形或字音的中文輸入法，將條件吻合的中文字加以輸出在螢幕畫面上。因於能夠輸入的中文字，是確實存在且對應至電腦內的中文內碼及字型，不會有人爲誤寫的錯字字型出現在畫面上。因此，電腦中的中文字錯字的定義，是指輸入了不適當的中文字。

電腦中的中文字其所發生的錯字，是在輸入過程時，由於人爲對於字的誤認，或是操作上不慎誤觸，因而輸入另一個不相關的中文字，而這個中文字在文章中可能產生錯誤的意義，此時我們稱此中文字爲「錯字」。

產生的文字錯誤可分爲「認知上的錯誤」與「操作上的錯誤」兩類，所謂「認知上的錯誤」就是使用者直接輸入一個誤認的文字。「操作上的錯誤」就是當使用者在輸入時，由於快速的輸入，而不自覺的誤觸了錯誤的按鍵。

本論文所討論的錯字，爲使用者按錯鍵所造成「操作上的錯誤」。依據使用者所用的輸入法之不同，產生出按鍵的錯誤也不盡相同；例如注音輸入法「ㄗ」與「ㄔ」位置相近，容易誤觸，而倉頡輸入法「大」與「十」的倉頡字根位置相近，較易誤觸。因此，應考慮使用者所使用的輸入方法與種類，分析可能產生的錯誤，進而推導出原來正確的字。

本系統之製作乃針對倉頡輸入的使用者而設計。分析使用者在輸入字的組合過程中，可能會發生以下的錯誤情形：字形相近的錯誤、字根的排列錯誤、缺字根、多字根等。

4.2 錯字偵測

圖 4-1 爲找出錯字位置的方法，主要是參考[13]，但是對於部份功能，爲配合系統的製作而有所改善，茲說明如下。

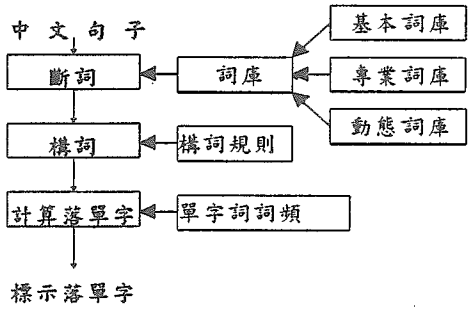

圖 4-1 錯字偵測流程圖

1.斷詞　　文章藉詞庫加以斷詞，在斷詞方面本論文將採用長詞優先方法進行文章斷詞。現行使用詞庫的應用軟體，大多以詞句的第一個字爲主要索引，進行搜尋工作。當詞句中第二個或其後的字發生錯誤時，依詞庫對應的關係，由第一個字搜尋詞庫，即可能找出原正確字。

但是，當詞句的第一個字發生錯誤時，上述方法所引導的方向，將會搜尋到截然不同的詞句，因此，比較不易找到正確的字。

有鑑於此，本系統將基本詞庫之資料結構加以改變。其做法爲利用資料庫所提供的索引檔特性，將詞庫中的所有詞，分作 4 個索引檔案。當第一個字發生錯誤，在詞庫中找不到相對應或類似的詞時，將搜尋的索引檔改爲第二個字，以第二個中文字爲首進行搜尋詞庫。依照此種多重索引，將可改善詞句中第一個字產生錯誤的情形。

2.構詞　　利用構詞規則[13]，來處理具有衍生性的中文詞。

3.單字詞詞頻表 目前所使用的單字詞詞頻來源是中文網路討論區中討論字的統計資料，單字詞詞頻爲判斷落單的字是否爲錯字的條件之一。

本系統在錯字偵測過程當中，使用長詞優先判斷處理方式，以詞庫處理長度爲 2 個字以上的詞，以預先準備的單字詞詞頻處理 1 個字長度的詞。當文章經由上述檢查後，若仍屬落單文字，則定爲可疑的錯誤字。

上述的錯字偵測，容易將人名、專有名詞等判定爲錯誤。因此，本系統設計一動態詞庫讓使用者隨時放入自訂的詞，將可有效的解決該類問題。

4.3 錯字訂正

關於錯字訂正的做法，乃是針對上節中標示落單字的原始輸入方式加以分析處理。我們假設使用者在輸入過程中，出現了操作上的錯誤，也就是在輸入文字時，不慎誤觸了相鄰或相似的按鍵，而產生了錯誤的字根，這些字根組合時雖有可對應的中文字，但卻無法與前後字組成合理的

詞句。

處理中文落單字時，主要是先找出此中文字原來的字根組合，也就是剖析中文輸入碼。其次，列出使用者對於各種字根可能誤觸的情形，並經輸入碼重組得到某些字，再將這些可能的正確字與前後文字連接，經由詞庫的比對及單字詞詞頻表的檢查，即可找出可能符合文句的正確中文字，此錯字訂正的流程圖如下。

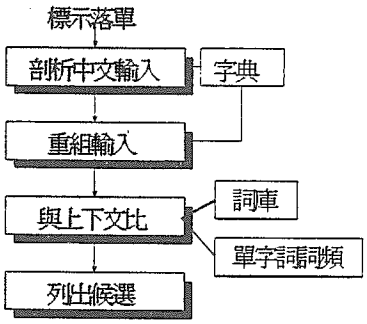

圖 4-2 錯字訂正流程圖

基本上，不同的中文輸入法會有不同的錯誤產生，例如：

1.字音輸入法　(注音、漢音輸入法等)

在字音、聲調、破音字、諧音容易產生錯誤。

2.字形輸入法　(倉頡、嘸蝦米、大易輸入法等)

在字形、部首、筆畫容易產生錯誤。

通常字音輸入法所列出的同音字數量相當多，容易造成錯字訂正的選字率過高，而降低校正效率。而字形輸入法所鍵入的錯誤，大都爲字義不同，字形卻相像的錯誤字，所以不易被文字校正人員察覺。我們將以字形輸入法(倉頡輸入法)爲主，作爲討論的方向，並製作一實務系統。

現以一實例來說明字碼剖析以及列出相關字的流程。

文章中的原意爲　「今天天氣很好！」

但是使用者卻輸入爲「今天天氮很好！」

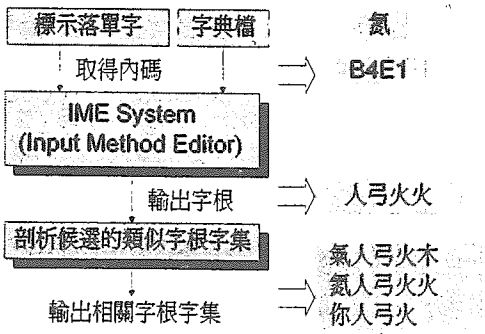

圖 4-3 反向字根查詢之流程

拆碼與重組字根應考慮以下五種情形：

1.相似字根：按鍵位置是與正確碼形狀相似的字根，拆碼時，較不容易辨識的字根。

2.相鄰字根：通常按錯的位置是正確碼周圍的 8 個按鍵，各字根可能錯誤的情形可參考倉頡輸入法鍵盤配置圖。

3.相同字根排列錯誤：當輸入字根碼順序不同時，也可能產生不同的中文字。

4.多餘字根：使用者多鍵入一個字根。

5.短少字根：使用者少鍵入一個字根。

茲就上述情形，分析「氮」字組合情形。依照反向查詢，「氮」的字根組爲「人弓火火」，而其周圍相鄰的按鍵如下所示：

「人」-> 戈、大、中、心

「弓」-> 月、竹、十、一

「火」-> 水、口、廿、木、土、金、女、月

最後一碼爲空碼，故有二十四種可能情形(因爲倉頡輸入法共有 25 個相異的字根)，所以全部組合共有 5*5*9*9*24=48600 種。在眾多組合之中只有 5 種組合能夠構成真正的中文字歸類如下：

落單字　「氮 人弓火火」

相似字根　「氣 人弓火木」

多餘字根　「你 人弓火」

多餘字根　「倓 人火火」

相鄰字根　「儋 人弓金口」

相鄰字根　「傸 人弓月金」

經由重組字根的方法，雖會有大量的組合產生，但是經由字形輸入法的特性分析，可發現約有 99.99%的組合無法找出所對應的中文字，因此我們採用 pruning 演算法來找出有效的組合字，以提升系統之速度。

針對此 5 個字，本系統再代入原文，依據前後字重新進行詞庫檢查，查證是否爲符合正確意義的中文詞。

4.4 準確率概估

準確率的評估，乃依照文章的總字數、檢測出的錯字、系統訂正能力、標示的選字作爲考量，在此先定義各項說明的代號：

T=輸入文書的中文字總字數

C=標示之落單字

W=真實錯字字數

L=系統列出訂正詞數

S=使用者採用系統訂正之詞數

標示率　　M-rate = C/T

錯字偵出率　　D-rate = (C∩W)/W

訂正率　　C-rate = S/L

當標示率 M-rate 過高，表示文章可疑的文字過多，造成之原因可能爲詞庫內容不足或專有名詞、人名等過多，單字詞頻內容，以及各項表格資料無法負荷，乃致於過多文字無法處理。

當錯字偵出率 D-rate 過低，表示系統標示錯字的準確率過低。

當訂正率 C-rate 過低，表示依照此校正系統所列出的候選詞多不符合原來正確的文詞。

以下爲本系統處理的一個例子以及整體準確率概算之平均值。

實例：Microsoft outlook 是一個革命性的桌上資訊管理系統，它可以幫助您組織各種資訊，並且與其他的使用者相互溝通分享資料。您可以使用 outlook 來管理理您的電子郵件約會連絡人指派的工作或任務各式各樣的活動。

改正後的文字：Microsoft outlook 是一個革命性的桌上資訊管理系統，它可以幫助您組織各種資訊，並且與其他的使用者相互溝通分享資料。您可以使用 outlook 來管理理您的電子郵件約會連絡人指派的工作或任務各式各樣的活動。

標示率　　M-rate = 7%

錯字偵出率　　D-rate = 80%

訂正率　　C-rate = 86%

五、實作概述

本系統在 Windows 95 作業系統下製作，程式是以 VB 語言設計，並使用 Microsoft Access 做爲資料庫的處理。此外，本系統是由「錯字偵測」、「錯字比對」、「列出訂正提示詞」三個子系統所組成，而作業流程簡圖如下：

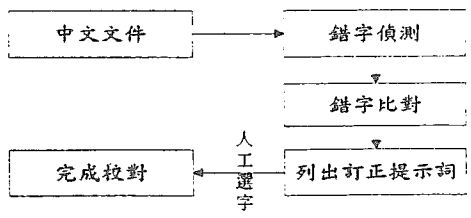

圖 5-1 作業流程簡圖

系統實作時，中文自動校正輔助系統可由主選單「工具」的「中文拼字檢查」開始啓動系統，啓動後，對文章中符合斷詞、構詞規則及單字詞的正確詞將予以跳過，表示無誤。當遇到可疑的詞時，系統出現以下「選擇提示」視窗畫面，使用者可以選擇系統提供的「選詞」項目加以改正，或是自行將這個"要在"當成新的詞，加入動態詞庫當中，或是選擇「放棄」皆可。使用者選擇後二項時，代表其認爲該詞無誤。

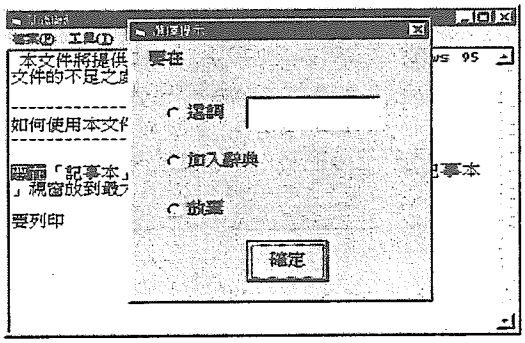

圖 5-2 選擇提示畫面

本中文自動校正系統亦爲一輔助系統，因爲本系統雖由電腦自動偵測出錯別字，並自動找出訂正詞，唯其所列出的正確字提示詞，並不直接覆蓋文章原有的詞句，而是由使用者選定所列的提示詞後再予修改。當使用者完成文章所有的拼字檢測後，將檔案儲存起來，即完成校對工作。

六、結論

在中文校正的研究發展上，多數學者著重於改善或增加大型詞庫，部分研究者則利用建立各式各樣的「字庫」、「詞庫」、「詞性分類」、「句型文法」、「近似字集」等方式予以處理。

本論文主要討論的方向爲提出一個考量中文輸入因素(以倉頡爲例)，由系統自動處理相鄰相近錯字之自動校正輔助系統。本系統之特色爲：(1)針對使用者在按鍵操作上之錯誤加以更正處理。(2)在中文詞的比對方面爲多重比對，換言之，可處理詞中的首項字發生錯誤的情形。(3)系統列出更正的字需經使用者確認，避免正確字反而被誤改。(4)提供動態詞庫，使用者可將本身認可之詞加入詞庫，以利後續工作。

本系統在未來工作方面可朝下列方向繼續進行：(1)針對使用者在中文認知上的錯誤輸入處理加以探討。(2)目前在字根的重組排列上雖使用 pruning 之演算法加以簡化，唯效率尚欠佳，應有改善之空間。(3)探討其他輸入法(如字音)產生錯誤之校正方法。

參考文獻

[1]工業技術研究院，中文辨識技術。
http://itrinews.itri.org.tw/rd/rd-5202.htm, (1996,10)

[2]許聞廉，陳克健，『自然』智慧型輸入系統的語意分析『脈絡會意法』，Proceedings of the 6th International Symposium-on Cognitive Aspects of the Chinese Language,(1993),527-540。

[3]中央研究院，自然語言研究環境之建立之附錄。
http://rocling.iis.sinica.edu.tw/ckip/CKBOOK.html#附件 1，(1996,9)。

[4]W.L. Hsu, W.K. Shih and P.H. Yeh, "Object oriented concept representation", Proceeding of ICCPOL'95,(1995).

[5]W.L. Hsu, "On physical mapping algorithms - a fault tolerant test for the consecutive ones property".

[6]Richard Sproat and Chilin Shih, "A statistical method for finding word boundaries in Chinese text", Computer processing of Chinese & oriental languages, Vol.4, No.4, March 1990.

[7]K.Y. Cheng and M.S. Hwu, "Design of a total Chinese input simulation system", Computer processing of Chinese & oriental languages, Vol.4, No.2&3, July 1989.

[8]A. Chiu and F. Wong, "An intelligent, knowledge-based Chinese input system", Computer processing of Chinese & oriental languages, Vol.3, No.1, May 1987.

[9]張俊盛，陳志達，陳舜德，限制式滿足及機率最佳化的中文斷詞方法，國立清華大學資訊所。

[10]張俊盛，彭載衍，中文辭彙歧義之研究—斷詞與詞性標示，國立清華大學資訊科學研究所。

[11]陳信希，李振昌，中文文本組織名之辨識(The Identification of Organization Names in Chinese Texts)，台灣國立台灣大學資訊工程學研究所。1994,10.

[12]張照煌(Chao-Huang Chang)，中文錯字自動訂正方法初探(A Pilot Study on Automatic Chinese Spelling Error Correction)，E000/CCL,Building 11, Industrial Technology Research Institute Chutung, Hsinchu 31015, Taiwan, R.O.C.

[13]施得勝，王良志，陳志達，聶素芬，基於統計的中文錯字偵測法，八月號 電腦與通訊雜誌，1992.8。

[14]Sun T, "A Hybrid Approach to Unknown Word Detection and Segmentation of Chinese" Chinese computing'94(ICCC94): proceedings of international conference on 1-4 June 1994, Singapore.

[15]Zhendong Dong, Chang "Negation in Chinese and English: Error Detection in Correct English ™" Chinese computing'96(ICCC96): proceedings of international conference on June 4-7 1996, Singapore.

[16]Zhendong Dong, Chang，二元接續關係及其漢語 [17] 分詞與校對中的應用 (Bi-Orderly-Neighborship and Its Application to Chinese Word-Segmentation and Proof)，Chinese computing'96 (ICCC96): proceedings of international conference on June 4-7 1996，Singapore。

[18]Sze-Sing Lam, Vincent Y. Lum, Kam-Fai Wong, "Determination of Word Sense in Chinese Full Text Using a Standard Dictionary and Thesaurus", International conference on computer processing of oriental language proceedings of 1995, Honolulu, Hawaii, November 2.

[19]Karen Kukich, "Techniques for Automatically Correcting Word in Text", ACM Computing Surveys, Vol.24, No. 4, December 1992.

[20]Kin Hong Lee & Chin Lu, "A Chinese Spelling Checker for Chinese Word Processor and its API", International conference on computer processing of oriental languages, proceedings of 1995, Honolulu, Hawaii, November 2.

[21]蔡志浩," 我們需要什麼樣的中文資訊處理技術 ", http://casper.beckman.uiuc.edu/hpp/c-tsai4/doc/comp/win95