應用動、靜態詞典以加速鍵盤輸入中文之方法
A Dynamic-and-Static-Dictionaries Based Method for Accelerating Chinese-Character Inputting with Keyboard

古鴻炎　　楊仲捷
Hung-yan Gu　　Chung-Chieh Yang
國立台灣科技大學電機系
Department of Electrical Engineering
National Taiwan University of Science and Technology
E-mail: root@guhy.ee.ntust.edu.tw　Fax: 886-2-27376699

摘要

在中文輸入上，鍵盤是一種普遍被使用的工具。為了加快以鍵盤輸入中文的速度，本文研究了以詞為單位來輸入的觀念、並提出實踐的方法，而將舊有的以字為單位來輸入的觀念加以擴充了。關於詞輸入觀念的實踐，我們研究了兩個重要的課題，第一個是，可減少按鍵次數之多字詞按鍵規則的設計，我們以注音輸入法為例，提出了一種設計方法，可據以獲得簡單且相容的按鍵規則；第二個是，可動態地記錄專有名詞、及支援短詞組合成長詞之動態詞典的設計，我們提出了一種基於赫序表的資料結構，使得動態詞典的功能、及查尋速度能夠符合實際需求。除了研究相關的問題，我們也實際去製作出一個整合了單字注音連續轉換、及多字詞直接輸入的原型系統，而驗證了所提出方法的可行性。

1. 導言

關於中文輸入之研究，在許多方向上都有人投入心力，如光學文字辨識(Chuang 1995, Huang 1997)、線上手寫輸入(蔡 1997)、國語聽寫機(Wang 1997)、鍵盤輸入(古 1992)等，其中以鍵盤輸入之方向較容易實行，也擁有最多的使用者，但這並不表示以鍵盤爲基礎的中文輸入方法(如注音類的漢音、自然、忘形，字根類的倉頡、大易)，就無再改進之空間，或者將窮途末路而不再需要研究，相反地，基於省力性(手寫輸入易疲勞)、隱密性、錯字更正之方便性(只用語音聽寫不易操作錯字之更正)、費用、輸入速率等等因素之考慮，我們覺得近期內鍵盤仍將是中文輸入上一種普遍被使用的工具(包括滑鼠按的圖像式鍵盤)。因此，若有一種能夠改進鍵盤中文輸入的效率(輸入一個中文字所需之按鍵數)、或速率(單位時間內輸入的字數)的方法被提出，那將讓爲數眾多的中文輸入

者受益，基於此一信念，我們遂投入於改進鍵盤中文輸入之研究，如今也獲得了幾許成果而可供參考。

本文提出的中文輸入之加速方法，其所依據的基本觀念是，將過去只能以中文字爲單位來輸入的作法擴充成能夠以詞爲單位來輸入，這裡所謂的詞其含意是廣義的，如"做了再說"、"中看不中用"都可被當成是詞，且詞的長度可以從 1 到一個預先設定的最大值(如 13)，說以片語爲輸入單位應較恰當，不過本文內仍將稱之爲詞。我們所以會想到以詞爲單位來輸入，是考慮到原始的注音輸入法裡，有同音字選字的問題，而當以二字詞或更多字之詞的注音爲單位來看時，同音詞的問題就小很多了，因此，我們想到進一步去爲二字以上之多字詞來定義一些能夠減少按鍵次數的按鍵規則，以加速中文之輸入或達到節省氣力之目的。這裡所說的以詞爲單位來輸入的觀念，和以往注音輸入法裡的連續注音轉換之觀念(Chen 1987)並不相衝突，因爲本文所說的詞是包含字的，當輸入的是單字的注音時，仍可去使用連續注音轉換的機制，事實上我們已製作出一個整合多字詞直接輸入、和連續單字注音轉換的原型系統。除了注音輸入因爲有同音字而需要連續轉換之處理，其實字根類的一些簡化的輸入法也需要類似的連續轉換處理(Fan 1991)。

雖然說以詞爲單位來輸入中文的觀念並不難，並且這樣的觀念對注音類或字根類的輸入法都可應用，但是，如何去實行就不是一件容易的事了，因爲有兩個實際的問題需要解決，一個是按鍵規則的設計與實行，而另一個是動態詞典的設計與製作。輸入二字以上多字詞的按鍵規則的設計，除了要減少按鍵的次數以達到加速輸入中文的目的外，各種長度的詞的按鍵規則要能夠相互相容，即不需額外按鍵來選擇接著要輸入之詞的詞長，這樣才可讓使用者隨心所欲，想輸入 N 字詞就直接用 N 字詞的按鍵規則， N 的值可以是傳統上所用的 N=1(即以字爲單位)，或新加入的 N=2,3, ..., 13(本文設定之詞長最大值)；除了考慮按鍵規則間的相容性之外，也要考慮如何設計一組既容易理解、又幾乎不需背誦的按鍵規則，不然按鍵規則又多又複雜而令人生畏，那將引不起使用者的使用意願，也就讓所訂的按鍵規則變得沒有意義了。

考慮以詞爲單位來處理使用者的按鍵輸入時，由於本文所謂的詞也含蓋了傳統上的字的單位，因此，在一般注音輸入法裡提供的連續注音轉換之功能，可說是基本且必備的功能，而此功能內所用到的一個支援注音到詞語轉換、查尋的詞典，我們也可再拿來

利用，以支援作多字詞之輸入按鍵轉換成詞語的查詢動作，這樣一個預先準備好的、用以儲存一般常用詞語的詞典，本文稱之爲靜態詞典(static dictionary)。若僅使用靜態詞典來查詢多字詞按鍵輸入可能對應到的語詞，將會面臨如下所說的困難，第一個是，對於靜態詞典裡未收錄的詞彙，如專業術語、專有名詞、及個人之慣用語，如何在輸入過一次後就能操作多字詞的按鍵規則來輸入？若規定使用者要自己去爲每一個新詞彙操作一次登錄的程序，很明顯地會造成使用者的不便，一者是費力氣，二者是要停頓思慮去想想是否需將碰到的新詞登錄到詞典；第二個困難是，一般的靜態詞典並不支援短詞自動組合成長詞的功能，使得一個長詞如"後天免疫不全症候群"，即使已經被輸入過數次，下次要輸入時，還是得依序逐一輸入"後天"、"免疫"、"不"、"全"、"症候群"等詞語(假設未操作過登錄的程序)，而不能夠直接(不用操作登錄之程序)就去操作 9 字詞之按鍵規則。請不必擔心 9 字詞的按鍵規則是否很複雜，其實我們設計的按鍵規則簡單得一看就記住了，且不必理會欲輸入的長詞(字數大於 3 者)究竟有幾個字。

雖然過去已有不少人研究了連續注音轉換成中文字之問題(Lin 1987, Gu 1991, Hsu 1994, Kuo 1995, Ho 1997)，但是對於前面提到的兩個關於靜態詞典的缺點，則尚未看到有人提出實際且有效的解決方法。因此，我們便思考此一問題，而想到引進動態詞典(dynamic dictionary)的觀念，並去研究動態詞典功能的實行方法，以便用動態詞典來自動記錄使者最近輸入進來的文句或詞彙，這樣，使用者就完全不需要去操作新詞登錄的程序，而且任何一個新詞彙不管其長度多長(在最大長度的限定內)，只要被輸入過一次，就確定可在動態詞典裡尋找到，如此當要再次輸入該新詞時，就可直接操作多字詞之按鍵規則了。例如輸入過一次"台北市基隆路四段"後，以後想要輸入"台北市基隆路"之詞時，就可直接操作 6 字詞之按鍵規則。

由前面的說明可知，本文倡議在原有的以字爲單位的輸入法上，再擴充支援以詞爲單位的中文輸入方式，藉由訂定減少按鍵次數的多字詞輸入之按鍵規則來加速鍵盤中文之輸入，這樣的想法可被應用到許多現有的、不管是注音類或是字根類的中文輸入法上，雖然本文裡的說明都拿注音輸入法作例子，但這並不表示只有注音輸入法才適用。要實踐以詞爲單位來輸入中文，動態詞典之觀念及製作是很重要的，所以在第二節就先說明我們對於動態詞典製作之研究成果；關於多字詞之按鍵規則的設計，在第三節裡我

們以注音輸入法爲例提出了一種相容、且簡單可行的設計；在第四節裡，我們則對同鍵詞(具有相同輸入按鍵的詞語)的問題加以分析；第五節裡則說明原型系統製作時會面臨的問題，及我們採取的作法；最後一節是結語。

2. 動態詞典設計

2.1 動態詞典的功用

回顧過去，可知動態詞典的觀念很早就被資料壓縮領域所應用(Bell 1990)，即調適型(adaptive)詞典式編碼方法，此外，建造統計式靜態詞典的問題也有不錯的方法可用。動態詞典的觀念簡單說來是，利用一個貯列(queue)之資料結構來儲存最近被處理過之資料，當新的一筆被處理過之資料要放進去時，就固定從貯列的一端放入，而當空間不夠時，就從貯列的另一端將最舊的資料移出。動態詞典之所以能夠發揮功用是因爲，一序列等待被處理的資料中，常常有反覆出現的資料片段存在，使得處理之軟體常常可從動態詞典中找到對應的已被處理過的、與等待被處理的相同資料片段，而達成特定的目標。

動態詞典爲什麼對中文之輸入有幫助？因爲我們可用大容量之動態詞典來記錄相鄰中文字的共出現(co-occurrence)關係，例如一'枝'花、一'隻'狗、一'支'筆等名詞與量詞之間的共出現關係，如此，當使用者再次輸入音節注音／ㄓ ㄍㄡˇ／時，"隻狗"之詞就會被查出；再者，動態詞典可用來記錄靜態詞典中未收錄的專有名詞、術語，如"紫杉醇"、"低鈉鹽"、"虛擬實境"，如此，當使用者要再次輸入曾被輸入過的專有名詞時，就可直接操作多字詞的按鍵規則來加速輸入；此外，還可利用動態詞典來進行短詞組合成長詞之處理，詳細作法在第五節裡說明，這樣，使用者只要輸入過一次某一長片語，以後就可直接操作多字詞的按鍵規則來快速地輸入那個長片語。由前面的說明可知，我們只使用了一個共用的動態詞典，並且爲了能夠記錄得愈多愈詳盡，詞典的大小是愈大愈好，不過，也不能大到使得搜尋速度變慢得不能接受，再者，動態詞典裡的資料在程式結束執行後也要繼續維持下去(即存檔)，這樣才能讓所記錄的資訊累積下去。

2.2 動態詞典之結構設計

由於我們要利用動態詞典來提供短詞(含單字詞)組合成長詞之功能，並且，這個功能除了支援以詞(二字以上)爲單位來輸入的方式之外，也要支援以字爲單位來輸入時用馬可夫中文模型來進行連續注音轉換之處理方式(古 1995, 古 1997)，因此，只採用簡單的貯列結構來實現動態詞典在實作上是不可行的，因爲在馬可夫模型中，一個音節若對應到10個同音字，則 5個連續音節可能組合出的五字詞就有 100,000 個，不用說6、7或更多字組成的詞了，再加上我們希望動態詞典愈大愈好，則字串比對的運算量及所需的時間是可想而知的，因此，我們就決定以赫序表(hash table)結合貯列的複合結構之設計來實現動態詞典，以赫序表的功能來加快搜尋的速度，而以貯列的功能來控制資料的存入與刪除。

爲了配合赫序表的處理，實作上就必需把一個中文句子的組成字(如: ABCDE)，拆成相鄰字組成的雙字詞序列(如: AB, BC, CD, DE)，然後將這些雙字詞(含注音資料)插入赫序表，可是如此做會伴隨產生兩個問題，第一個問題是無法迅速地確定雙字詞之間的時間次序、及存入時間，例如由兩音節/ㄓ ㄏㄨㄚ/去查出"支花"與"枝花"都在赫序表裡時，如何知道那一個是較晚存入的、及多久之前存入的，以便由存入時間來推測某一詞語再次被用到的機率，我們採用的一個解決辦法是應用郵戳(time stamp)的觀念，即每次要將一個雙字詞存入赫序表時(不管是新插入的或是更新的)，就將目前時間計數器的值取出一起存入到赫序表裡，然後將時間計數器加一。第二個問題是會發生張冠李戴的現象，假設先前輸入的語句 "一隻黑貓帶著兩支黑拐杖並撐著一支花雨傘"，已被存入動態詞典裡，則接著輸入 "一隻黑貓" 的四個音節注音時，會轉換出來的是"一支黑貓"，這是因爲 "一支" 的郵戳比 "一隻" 晚，而 "支黑" 比 "隻黑" 晚，乍看之下，也許會認爲何不依據郵戳值是否連續去判斷原先是否是接在一塊的，不過，這樣做是不可行的，例如若輸入 "一隻黑狗和二隻黑貓" 後，那是不是使得"隻黑" 與 "黑狗"不被認爲是連在一塊的，因爲"隻黑"的郵戳值會被第二次出現的"隻黑"所更新，對於這樣的問題，我們的解決方法是再增加一個赫序表，用以儲存中文句子裡相鄰三字所形成之三字詞，例如要將中文句子 ABCDE 存入動態詞典，就相當於把 AB,BC,CD,DE 等雙字詞存入第一個赫序表，而把 ABC, BCD, CDE 等三字詞存入第二個赫序表。如此，當要查尋一個四字詞 WXYZ 存不存在於動態詞典裡時，就先到第一個赫序表去看 WX,XY,YZ

是否都能比對成功，包括注音資料之比對，這樣才不會因破音字而發生如／ㄕㄤˊ ㄐㄧㄠˋ／被轉成"時間"而非"實踐"之現象；然後，再到第二個赫序表去看 WXY 與 XYZ 存不存在。如果都比對成功，才算是存在於動態詞典裡，如此，前述例子"一隻黑貓"的音節注音就不會轉換錯了。至於 5 、 6 、...等多字詞的檢查可依此類推。

3. 按鍵規則設計

3.1 原則與方法

前面我們曾提到，設計多字詞的按鍵規則時，要在兩個基本原則之下去考慮減少按鍵的次數，一個原則是按鍵規則間要兩兩相容，這樣才不需額外按鍵來選擇欲輸入之詞的詞長，另一個原則是按鍵規則要簡單易理解、而幾乎不需背誦。基於這兩個原則，我們爲普通注音鍵盤設計了一套多字詞的按鍵規則，詳細情形如下面的說明，雖然這裡只以注音輸入法爲例，提出對應的多字詞按鍵規則之設計，但我們認爲同樣的原則、同樣的設計方法也可應用到字根類中文輸入法裡，去設計對應的多字詞按鍵規則。

在設計注音輸入法的多字詞按鍵規則時，我們注意到注音鍵盤上的注音符號的排列方式已有許多種被提出，一種使用傳統排列方式的我們稱爲普通注音鍵盤，也就是我們要據以設計多字詞按鍵規則的鍵盤，除此之外還有如宜韻注音鍵盤(古 1992)、許氏注音鍵盤等等的排列法。基本上，注音符號的排列方式和詞爲單位來輸入的觀念並不相衝突，不過，某些排列方式的確會對詞輸入觀念之實行造成麻煩，因爲在原先字爲單位的輸入觀念裡，雖可讓二個或多個注音符號共用一個按鍵而不會有分辨不清(ambiguous)的情形，但是當考慮要設計多字詞輸入之按鍵規則時，就會發現有分辨不清的情形存在，而增加了中文輸入軟體製作上的複雜度，並增加了操作同鍵詞選詞動作的機會，因此，字單位的輸入觀念裡認爲不錯的注音鍵盤設計，不見得就一定適合在以詞爲單位來輸入的觀念裡使用，所以本文也給注音鍵盤之設計導入了一個新的考慮因素。

由於本文所謂的詞單位是包含字的單位的，所以在設計多字詞的按鍵規則時，第一個考慮到的便是如何區分使用者目前要輸入的是單字還是多字詞，爲了解決這個問題，我們就訂定如下的規則：

(R1) 多字詞的按鍵規則裡，不可用到聲調按鍵。

因爲輸入一個單字的注音時，最後一定要輸入一個聲調的按鍵，而當輸入軟體接收到一個聲調按鍵時，就可依規則(R1)確定使用者不是在輸入多字詞。這樣的規則也可應用在字根類的輸入法裡，例如倉頡輸入法裡輸入一個單字的最後一個按鍵固定是空白鍵。

能夠區分目前輸入的是單字還是多字詞之後，接著我們考慮，使用者欲輸入之多字詞的長度可能從 2 變化到 13 ，那麼輸入軟體如何知道目前被輸入之多字詞的詞長？相關的一個較基本的問題是，是否一定要知道詞長才能夠作處理？如果從軟體製作者的觀點來看，當然是希望使用者明確告知欲輸入之多字詞的長度，當反映在多字詞的按鍵規則上，可能就是要求使用者多輸入一個暗示詞長的按鍵，這樣的作法不僅會增加按鍵的次數，而使輸入速度變慢，最主要的缺點還是造成使用者的不便，因爲要記住多輸入一個暗示詞長的按鍵這件事(而輸入單字時不用)，及輸入之前還得自己先去算詞的長度。因此，我們設計多字詞的按鍵規則時，是在實作上可行的條件下儘量考慮讓使用者方便，實際上我們的規劃是，將多字詞依詞長分成三類，即二字詞、三字詞、與四字以上之多字詞，然後各別去設計各類詞的按鍵規則，以便兼顧不同詞長之詞的特性及簡化按鍵規則，其實從所設計的按鍵規則來看，可以說只有兩類詞，因爲三字詞可看成是四字詞的特例，也就是說使用者大約只需分辨欲輸入之詞是二字詞還是二字以上之詞。關於作分類的考慮因素，前面提到的一個是不同詞長之詞的特性，我們指的是，詞的長度愈長(如四字以上之詞)，則只需較小的按鍵率(按鍵次數與詞長之比率)就能夠互相區分而少有同鍵詞出現，而二字詞需要較大的按鍵率，不然發生同鍵詞的機率會很大；此外，我們把四字以上之多字詞歸爲一類，並爲它們設計通用的按鍵規則，目的是讓使用者不用去算欲輸入之詞的詞長，而這樣也讓需要記憶的按鍵規則的數目減少了。

3.2 多字詞按鍵規則

詳細說來，我們設計的雙字詞按鍵規則是:

(R2) 欲輸入雙字詞 XY 時，先按 X 的頭尾兩注音符號，再按 Y 的頭尾兩注音符號。

例如要輸入雙字詞"電腦"則按ㄉㄢㄋㄠ等四鍵，而雙字詞"實際"要按ㄕ*(Enter)*ㄐㄧ等四

鍵，也就是說注音符號不夠用時要以*(Enter)*鍵替補，以湊成四鍵。我們所以選用頭、尾而不用頭、中之兩注音符號，是因爲確定頭、尾後，中間能夠被插入的注音符號只有含不含介音ㄧㄨㄩ等四種可能，不像確定頭、中後，尾部可能插入的韻母注音符號的數量有 12 個之多，而發生更嚴重的混淆不清之情況。類似(R2)之按鍵規則，也可用在字形類的中文輸入法裡，以輸入雙字詞。接者，我們爲三字詞設計的按鍵規則是：

(R3) 欲輸入三字詞時，依序按各字最前之注音符號，然後按〕鍵。

例如要輸入三字詞"王陽明"則按 ㄨㄧㄇ〕 等四鍵。規定多按一個注音符號以外的 〕鍵，是要使按鍵數湊成四，以消除 ambiguous 之情況，因爲若不加按一個〕鍵，就不能確定使用者是要輸入單字詞、二字詞、三字詞、還是四字詞，而只有當確切知道使用者要輸入幾字詞時，若從詞典中只查到一個對應的詞語，就可立刻將該詞語送出，以免除使用者需要再按一鍵來選取所要之詞語的步驟。關於四字詞的按鍵規則，我們的設計是：

(R4) 欲輸入四字詞時，依序按各字最前之注音符號。

例如要輸入四字詞"一帆風順"則按 ㄧㄈㄈㄕ 等四鍵。由規則(R4)可知，爲什麼我們說三字詞的按鍵規則可看成是四字詞規則的一個特例。此外，應該不難看出規則(R4)與規則(R2)是有衝突存在的，因爲使用者輸入一個二字詞的四個按鍵時，可能由規則(R4)找到一個對應的四字詞，例如欲輸入"距離"時，按ㄐㄩㄌㄧ等四鍵，依據規則(R4)會對應到"金玉良言"。基本上發生這種衝突的機率很小，靜態詞典裡的二字詞經由分析只發現到 15 個會有這種衝突，因此我們依長詞優先的原則來設定規則(R4)比(R2)具有較高的套用權，即當使用者按了四次注音符號按鍵後，先試圖套用規則(R4)，如果詞典中未能找到對應的詞語，再去嘗試規則(R2)。雖然我們訂定規則(R4)具有較高的套用權，但是使用者想輸入的可能是發生衝突的那個二字詞，再者還要考慮如何銜接到五字以上多字詞之輸入，因此我們不能在套用規則(R4)成功後，看到只有一個對應的詞語就立刻將該詞語送出，而必需令使用者再按一數字鍵來選取所要的詞語，此時使用者按的鍵若是特殊的*(Esc)*鍵，就可確認他想輸入的是二字詞，然後改成去套用規則(R2)。

關於五字以上至十三字之多字詞的按鍵規則，我們的設計是：

(R5) 欲輸入五字以上之多字詞時，依序按各字最前之注音符號。

例如要輸入六字詞"台北市民政局"(假設曾輸入過此片語)則按 ㄊㄅㄕㄇㄓㄐ 等六鍵。比較規則(R4)與(R5)可知，事實上兩規則並無不同，這也就證實了我們在前面所說的，雖然詞的長度有長有短且愈長愈難算清楚，但是依據本文設計的按鍵規則，使用者不需去算所輸入長詞的字數，也無需掛慮會分不清楚該用那一條規則，而只需分辨欲輸入之詞是二字詞、三字詞、還是四字以上之詞。另一方面，從實際製作的觀點來看，規則(R4)與(R5)在實行上並無衝突或不可行的地方，當使用者輸入四個注音符號後，就先去嘗試套用規則(R4)，若詞典中找不到對應的詞語，就再去嘗試套用規則(R2)，若能夠找到對應之四字詞，就將找到的詞語顯示於螢幕上，然後等待使用者的下一個按鍵，接著，將讀到的按鍵分成三類，第一類如果讀到的是*(Esc)*鍵，則表示之前鍵入的四個按鍵是要輸入二字詞的；第二類如果讀到的是注音符號之按鍵，就可確認是，使用者想輸入比四字更長的詞，因此就套用規則(R5)，此時若未能從詞典找到對應的詞語，就將之前找到的詞語中的一個送出，若能找到對應的詞語，就再將找到的詞語顯示於螢幕上，然後再等待使用者的下一個按鍵；第三類如果讀到的是數字按鍵，就表示使用者要選取目前顯示出的詞語中的一個。在普通注音鍵盤上，由於部分的數字鍵(如 1,2,5,8,9)也是注音符號之按鍵，而產生四字以上多字詞輸入時的混淆不清情況，這樣的情況可採用前述解決規則(R2)與(R4)衝突的相同作法來化解，也就是令注音符號有較高的優先權，而先嘗試從詞典找對應的增長一字的詞語，第一種情況若找不到，就將剛輸入的按鍵當成是數字鍵，然後去選取前一次顯示出的詞語；第二種情況若能找到對應的詞語，就將找到的詞語顯示於螢幕上，然後等待使用者的下一個按鍵，此時使用者可按*(Esc)*鍵，以強迫將前一次的按鍵當成數字鍵並選取前一次顯示出的詞語。

4. 同鍵詞分析

前一節裡我們提出了一套輸入多字詞的按鍵規則，初步看來應可說是簡單易懂，但相對的一個考慮是，簡單的按鍵規則會不會導致嚴重的同鍵詞問題，即依據按鍵規則來輸入一個詞時，卻從詞典裡查出許多的詞語都可對應到所輸入的按鍵，而需操作同鍵詞

選詞的步驟，類似於傳統注音輸入法裡的同音字選字的操作，例如輸入二字詞"美妙"的 4 個按鍵ㄇㄟˇㄇㄠ時，會查出"美妙"，"美貌"，"眉毛"等詞語。爲了檢視同鍵詞的發生機會及嚴重情形，我們就以原型系統裡所用的靜態詞典來進行分析，該詞典裡含有 33,275 個二字詞、8,083 個三字詞、10,448 個四字詞，詞語的收錄參考了中研院詞典(中研院 1994)的二至四字詞，經過分析後，我們得到表 1 至表 3 的數據，在這三個表裡，

表 1　二字詞之同鍵詞分析

同鍵範圍	1	2	3	4	5	6	7	8	9	10	11	12	13	14
詞個數	14,185	8,646	4,644	2,472	1,425	864	406	192	81	70	66	36	13	28
比率%	42.6	26.0	14.0	7.4	4.3	2.6	1.2	0.6	0.2	0.2	0.2	0.1	0.0	0.0

表 2　三字詞之同鍵詞分析

同鍵範圍	1	2	3	4	5	6	7	8	9
詞個數	3,418	2,188	1,314	616	280	114	70	40	18
比率%	42.3	27.1	16.3	7.6	3.5	1.4	0.9	0.5	0.2

表 3　四字詞之同鍵詞分析

同鍵範圍	1	2	3	4
詞個數	9,481	848	81	12
比率%	90.7	8.1	0.8	0.1

同鍵範圍表示輸入一個詞的按鍵後會查到的同鍵詞的個數，詞個數表示靜態詞典中有多少個詞語具有某一個特定的同鍵範圍值，例如從靜態詞典中找輸入按鍵爲ㄇㄢㄏㄚ的二字詞，只能找到詞語"漫畫"與"棉花"，因此這兩個詞的同鍵範圍值都定義爲 2 ，且都要算在同鍵範圍值爲 2 的詞個數之計數裡。由表 1 可知，靜態詞典裡有 42.6%之二字詞的同鍵範圍值爲 1 ，也就是沒有同鍵詞，而可在輸入四個注音按鍵後即刻送出對應之詞，另外 56.7%的二字詞(同鍵範圍值爲 2 至 9)需要多按一個數字鍵來選取所欲輸入之詞，至於同鍵範圍值到達 10 以上的二字詞有 213 個。接著，由表 2 可得知， 42.3%的三字詞沒有同鍵詞，而可在輸入四個按鍵後即刻送出對應之詞，其餘的需多按一個數字鍵來選取所欲輸入之詞。最後由表 3 可知， 90.7%的四字詞沒有同鍵詞，這個比率比起表 1 與表 2 裡的都高許多，因此我們可推測五字以上之多字詞的同鍵詞問題會更小。不過，對於四字以上之多字詞來說，即使輸入之按鍵只對應到一個詞，使用者還是要加按

一個數字鍵來選取欲輸入之詞，因爲輸入軟體依據前一節裡的按鍵規則(R5)會無法判斷使用者是否要輸入更長的詞，然而在需加按一個數字鍵的情況下，四字詞的輸入效率尙可達到 1.25 鍵每字。

由於表 1 、表 2 顯示，輸入二字詞或三字詞時，發生同鍵詞選詞的機率(1 - 42.3%)並不小，而將造成使用者需要常常注視螢幕的情況，因此，我們便思考可能的改進方法，後來想到一種可行的作法是，以聲調來分辨、選取同鍵詞群中的一個詞，例如要輸入二字詞"市長"時，按 尸(Enter)虫大 等四鍵，結果螢幕上會出現"師長"、"師丈"、"時裝"、"市長"等詞語，並聽到電腦發出"嗶"一聲來通知有同鍵詞之情況，此時，使用者可不必看螢幕，而直接在心裡想"市"、"長"的聲調分別是 4 聲與 3 聲，然後就去按聲調 43 所對應的按鍵而選到"市長"之詞語。這樣的方法需要 5*5 = 25 個表示聲調組合之按鍵，而數字鍵之外的按鍵數量多於 25 ，所以實行上並無困難。依據前述的按聲調鍵的方法，再去對靜態詞典中的二字詞、三字詞作分析，其中三字詞之聲調鍵由詞首兩字決定，結果我們得到表 4 與表 5 之數據，在此二表裡，同鍵範圍值爲 1 的詞個數表示，不用按聲調鍵就已無同鍵詞情形的詞語個數，加上按聲調鍵後才變成無同鍵詞情形的詞語個數，而其它的同鍵範圍值下的數值都是在加按聲調鍵後的統計。由表 4 可知，透過聲調鍵的篩選，在輸入二字詞時可有 82.1%的機會不用注視螢幕，並且最大的同鍵範圍值也由 14 降到 6 了，即減低了同鍵詞問題之複雜度；此外由表 5 也可看到類似的改進，透過聲調鍵的篩選後，輸入三字詞時可有 85.7%的機會不用注視螢幕，並且最大的同鍵範圍值會由 9 降到 5 。

表 4　加聲調鍵之二字詞同鍵詞分析

同鍵範圍	1	2	3	4	5	6
詞個數	27,329	4,882	804	188	65	6
比率%	82.1	14.7	2.4	0.6	0.2	0.0

表 5　加聲調鍵之三字詞同鍵詞分析

同鍵範圍	1	2	3	4	5
詞個數	6,924	982	135	24	5
比率%	85.7	12.1	1.7	0.3	0.1

5. 原型系統製作

爲了驗證本文提出的多字詞輸入之按鍵規則的可行性，我們遂實際去製作一個可以詞爲單位來輸入中文的原型軟體系統，由於這個系統也必需能夠接受字爲單位的注音輸入，因此我們決定拿過去建立的字爲單位來輸入、且以馬可夫語言模型來進行連續注音轉換之中文輸入系統[14]來作基礎，然後加以擴充使它能夠處理多字詞之按鍵輸入。

由第 3 節裡的按鍵規則可知，多字詞的輸入按鍵只提供了音節注音的部分資訊，不像單字的輸入按鍵提供了整個音節的注音資訊，因此，到詞典去查多字詞輸入按鍵可能對應的詞語時，注音資料的比對就必需改變成一種非精確比對的方式，實作上我們以空白符號(即 ASCII 碼 32)來填充缺少的注音資料，然後令空白符號可和任何注音符號比對成功。此外，原型系統裡的詞典是分成靜態與動態兩種詞典的，靜態詞典裡的詞語長度爲 1 至 4 ，所以只能支援按鍵規則(R2)、(R3)、(R4)，用以查詢一至四字詞的注音輸入會對應到的詞語，並且不能用以查詢新增的、未登錄過的詞語；相反地，動態詞典支援的按鍵規則除了(R2)、(R3)、(R4)之外，還增加了(R5)，即接受查詢的詞語長度可由 2 變化到 13 ，並且可用以查出先前曾輸入過的專有名詞、慣用語(在不需主動登錄的條件下)。這樣的靜、動態詞典的功能差異，意味著在原型系統裡製作這兩種詞典時，需要採取不同的資料結構、不同的處理方式。

關於動態詞典的一個重要的、尚未說明的功用是，支援短詞組合成長詞的處理。例如，使用者曾各別輸入“電腦”與“文盲”兩詞語來串成“電腦文盲”之詞，然後按*(Enter)*鍵以送出該詞並將它存入到動態詞典裡，則下次使用者依按鍵規則(R4)來按 ㄉㄋㄨㄇ 等四鍵時，輸入軟體要能夠從動態詞典中找出“電腦文盲”之詞。依據第二節裡的動態詞典的結構設計，我們達成短詞組合成長詞的作法是，先用前三字的部分注音(如前述的ㄉㄋㄨ)到第二個赫序表去作非精確比對，以查出可能和它對應的三字詞 $U_nV_nW_n$, n = 1, 2,...,N ，再用後三字的部分注音(如前述的ㄋㄨㄇ)去作非精確比對，以查出可能對應的三字詞 $X_kY_kZ_k$, k=1,2, ...,K ，然後對所有可能的 n, k 組合下的 V_nW_n 與 X_kY_k 作比對，以找出當 V_nW_n 相同於 X_kY_k 時所組合出的四字詞 $U_nV_nW_nZ_k$ 。

此外，若使用者要依按鍵規則(R5)來輸入六字詞的注音按鍵 $P_1P_2P_3$ $P_4P_5P_6$ ，則

當他輸入到 P_4 時，就先以 $P_1P_2P_3P_4$ 去第二個赫序表查出可能對應的四字詞 $A_n B_n C_n D_n$ ，而當他輸入到 P_5 時，再以 $P_3P_4P_5$ 去查出可能對應的三字詞 $E_kF_kG_k$ ，然後對所有可能的 n, k 組合下的 C_nD_n 與 E_kF_k 作比對，以找出當兩者相同時所組合出的五字詞 $A_n B_n C_n D_n G_k$ ，令找出的五字詞重新排定下標後以 $A'_j B'_j C'_j D'_j G'_j$, j=1,2,...,J, 表示；接著當輸入 P_6 後，就以 $P_4P_5P_6$ 去查詢可能對應的三字詞 $Q_iR_iS_i$ ，然後對所有可能的j, i 組合下的 $D'_j G'_j$ 與 Q_iR_i 作比對，以找出當兩者相同時所組合出的六字詞 $A'_j B'_j C'_j D'_j G'_j S_i$ 。對於其它字數(大於六字)的多字詞按鍵，只需將前述的作法延續下去即可。不過，檢視前述的作法可看出，此種組合成長詞的作法有時會組合出未曾輸入過的詞語，例如使用者曾輸入過"小獅子"與"獅子頭"之詞，則依按鍵規則(R4)按ㄒㄕㄗㄊ等四鍵時，會找出未曾輸入過的"小獅子頭"之詞，然而這種情形在很多時候都是可接受的。

6. 結語

我們認爲在近期內，鍵盤仍將是一種普遍被使用的中文輸入工具，而本文倡議的加快鍵盤輸入中文之方法，包含：(1)詞爲單位之輸入方式、(2)多字詞按鍵規則、(3)支援短詞組合成長詞並可動態記錄新詞之動態詞典等功能，如果能夠推廣至原已存在的中文輸入法上，深信可帶給爲數眾多的中文輸入者一定程度的助益。

爲了實踐以詞爲單位來輸入的觀念，本文研究了兩個重要的課題：動態詞典結構設計與多字詞按鍵規則設計。在動態詞典方面，由於動態詞典不僅要支援以詞爲單位來輸入之方式，也要支援以字爲單位來輸入時的連續轉換處理，因此我們提出一種包含兩個赫序表和一個貯列的複合結構設計，以滿足二者之快速查尋的需求；在多字詞按鍵規則方面，我們以注音輸入法爲例，設計了一組簡單易記、且相容的多字詞按鍵規則，而同樣的設計原則可被用來爲其它的中文輸入法設計多字詞按鍵規則。依據所設計的按鍵規則去對靜態詞典裡的二、三、四字詞作同鍵詞分析，我們發現各有 42.6%, 42.3%, 90.7% 的詞語沒有同鍵詞，而如果再以聲調來篩選二、三字詞，則可各別讓高達 82.1%與 85.7% 的詞語沒有同鍵詞，而減少了需要注視螢幕的機會。此外，我們也實際地去製作出一個整合了單字注音連續轉換、及多字詞直接輸入的原型系統，而驗證了前述方法的可行

性。

參考文獻

Bell, T. C., J. G. Cleary and I. H. Witten, Text Compression, Prentice-Hall Inc., Englewood Cliffs, New Jersey, 1990.

Chen, S. I., et al., "The Continuous Conversion Algorithm of Chinese Character's Phonetic Symbols to Chinese Character", Proceedings of National Computer Symposium (Taipei), pp. 437-442, 1987.

Chuang, C. T. and L. Y. Tseng, "A Heuristic Algorithm for the Recognition of Printed Chinese Characters", IEEE trans. on Systems, Man, and Cybernetics, Vol. 25, No. 4, pp. 710-718, April 1995.

Fan, C. K. and W. H. Tsai, "Reduction of Key Stroke Numbers in Chinese Input by Relaxation Based Word Identification", Proc. of Int. Conf. on Computer Processing of Chinese and Oriental Languages(Taipei), pp. 37-44, 1991.

Gu, H. Y., C. Y. Tseng and L. S. Lee, "Markov Modeling of Mandarin Chinese for Decoding the Phonetic Sequence into Chinese Characters", Computer Speech and Language, Vol. 5, No. 4, pp. 363-377, 1991.

Ho, T. H., el al., "Integrating Long-Distance Language Modeling to Phoneme-to-text Convertion", Proc. of Int. Conf. on Computer Processing of Oriental Languages (Taipei), pp. 287-299, 1997.

Hsu , W. L., "Chinese Parsing in a Phoneme-to-Character Conversion System Based on Semantic Pattern Matching", Computer Processing of Chinese and Oriental Languages, Vol. 8, No. 2, pp. 227-236, Dec. 1994.

Huang, K. Y., H. T. Yen and C. S. Han, "Neural Networks for Robust Recognition of Printed Chinese Characters", Computer Processing of Oriental Languages, Vol. 10, No. 4, pp. 425-442, April 1997.

Kuo, J. J., "Phonetic-input-to-character Conversion System for Chinese Using Syntactic Connection Table and Semantic Distance", Computer Processing of Chinese and Oriental Languages, Vol. 10, No. 2, pp. 195-210, Oct. 1995.

Lin, M. Y. and W. H. Tsai, "Removing the Ambiguity of Phonetic Chinese Input by the Relaxation Technique", Computer Processing of Chinese and Oriental Languages, pp. 1-24, 1987.

Wang, H. M., et al., "Complete Recognition of Continuous Mandarin Speech for Chinese Language with Very Large Vocabulary Using Limited Training Data", IEEE trans. Speech and Audio Processing, Vol. 5, No. 2, pp. 195-200, March 1997.

中研院中文詞知識庫小組，中文書面語頻率詞典， 1994 。

古鴻炎，"一個同時考慮鍵盤效率、人體工學原則、及符鍵對應規律性之國語注音輸入鍵盤的設計"，電工雙月刊，第 35 卷，第 2 期，第 123-132 頁， 1992 。

古鴻炎、陳志耀，"使用新式注音鍵盤及複合馬可夫語言模型之中文輸入系統"，中華民國電腦學會電腦學刊，第 7 卷，第 3 期，第 1-9 頁， 1995 年。

古鴻炎，"動態詞典及其與馬可夫中文語言模型之整合", 全國計算機會議論文集(台中)，第 D1-D7 頁， 1997 。

蔡奇峰、馬自恆，"樣本比對之中文手寫辨識系統"，全國計算機會議論文集(台中)，第 B39-B44 頁， 1997 。