一個速度更快的中文輸入法─師大大師輸入法

A Faster Chinese Input Method — The NTNU-Master Input Method

林順喜
國立臺灣師範大學資訊工程學系
linss@csie.ntnu.edu.tw

魏仲良
國立臺灣師範大學資訊工程學系
idreamer@tp.edu.tw

摘要—科技的進步，讓電腦成為工作上最重要的工具，因此選擇一個好學、快速的中文輸入法能增進資料處理的效率。

本研究的目標是設計出一個簡單、好學且快速的輸入法。我們先研究倉頡以及大新倉頡輸入法的特點之後，確定以縮短最大取碼數為 3 與省略不必要的空白鍵為主要的改進方向，再搭配重新編排的字根表，以期達到預期的目標。中文字屬於圖形系統，由於每一個字的形狀裡包含著文字結構的資訊，因此在編排字根表的過程中，利用文字結構的資訊設計了一個快速產生碼表的程式，並從中獲得重要的統計數據，並以此做為字根抽換以及字根與鍵盤對應的參考依據。

分析及實測結果顯示，師大大師輸入法確實降低了平均編碼長度及平均按鍵次數，即使不背誦簡碼，也可以取代大新倉頡(需背誦大量簡碼才會快)成為目前最快速的輸入法。若本輸入法再加入 1、2 碼簡碼，速度將逼近理論上的最佳解。

關鍵詞—中文輸入法、倉頡、師大大師、快速編碼工具、編碼長度

Abstract – Because of the improvement of information technology, computer has become the most important tool for work so that an easy-to-learn and fast Chinese input method can advance the efficiency of processing data.

The goal of this study is to develop a better input method. After looking into the characteristics of ChangJei and DaSinChangJei input methods, shortening the maximum encoded length to 3 and reducing the times of typing the space key are set to be the main aspects of improvement. We conduct a new and rapid strategy utilizing the structural information of Chinese characters to produce a mapping table which maps the Chinese characters to their encoding strings during the process of rearranging mapping tables. The statistics gathering from the developed program are used to decide how to adjust the Chinese character roots and how to rearrange their positions on the keyboard.

The analytic and experimental results show that NTNU-Master Chinese input method, derived from this study, have reduced the maximum average encoding length and the average number of key strokes. Theoretically, even without using any brevity code, it already has the potential to take DaSinChangJei's place to be the fastest Chinese input method at the present time. Furthermore, when it is integrated with carefully-defined brevity codes, it can achieve the near-optimal speed of the Chinese input problem.

Keywords–Chinese input method, ChangJei, NTNU-Master, rapid encoding utility, encoding length

一、研究背景及文獻探討

科技的進步，引發了資訊科技的革命，在近二、三十年間，因為個人電腦的普及，使得原本需依賴大量人工處理的資料，漸漸藉由電腦強大的運算能力來提昇處理的效率。再加上

網路科技的進展，人們溝通以及交換資訊的方式也跟著「e 化」了，不管是電子郵件、會議資料、營運報表等，皆需將內容輸入至電腦再進行後續的處理。因此，如何快速地將資料輸入電腦，便關係到資料處理的效率。

電腦在設計之初，僅考慮西方社會的使用，因此在中文資料的處理上，相對較為不便；再者中文字圖形系統亦較西方拼音系統來得複雜許多，也因此導致中文輸入遠較英文輸入複雜且緩慢。所幸在國內有識先進的努力之下，利用中文字的各種特性(讀音、筆畫、字形)，各式各樣的中文輸入法應運而生。

(一) 中文輸入法之發展

現今中文輸入法的形式均屬於「委選式」輸入法，由使用者輸入中文字的屬性資訊篩選出符合條件的中文字來進行文字輸入。一般將中文字的屬性資訊粗分為兩大類：

(1) 字音：輸入讀音篩選出同音字，再由使用者選出正確的字，例如：注音、漢語拼音等輸入法為其中的代表。

(2) 字形：輸入代表字形結構的碼(字根)來篩選出同碼字，再由使用者選擇，例如：倉頡、三角號碼、嘸蝦米、行列等輸入法。

為了提升輸入速度，一些增強型的輸入法則在原本輸入法的架構之下，透過詞庫的使用，能夠在輸入的過程中，依據上下文自動挑選可能的字，如自然注音、新酷音輸入法等屬於字音類的增強型輸入法，而新倉頡、自然倉頡等則為字形類增強型輸入法的代表。然而這些智慧型選字的系統在一些文章(如一般人名)輸入時誤判率極高，回頭修修改改反而拖慢速度，並造成打字者疲累。

除了上述透過鍵盤進行輸入的中文輸入法之外，由於小型隨身行動裝置(如個人數位助理 PDA、智慧型手機 Smart Phone 等)的出現，受限於機體的體積，手寫以及語音辨識等也成為中文輸入的方式。

在挑選和評估輸入法的時候，大致可依照以下向度來考量：

(1) 學習門檻：學習門檻愈低，表示愈容易學，且學習時不需花費太多時間。

(2) 輸入速度：輸入速度愈快，表示輸入的效率愈好。一般說來，每字所敲的平均按鍵數愈少、同碼字愈少(選字愈容易)，輸入的速度愈快。

(3) 使用鍵數：若用來輸入的鍵數少於 30 個，則可將使用的按鍵安排於鍵盤上的 26 個英文字母鍵加上逗號(,)、句號(.)、分號(；)以及斜線(/)等 3 排共 30 個按鍵所組成的主要輸入區之內，則可以減少打字時手指移動的距離，增加輸入的速度以及降低因長距離移動手指按錯的機會。

綜觀市面上的各種輸入法，一般說來，字形類的輸入法輸入速度較快，但學習門檻較高；而字音類的輸入法則相反，學習門檻低，但輸入速度較慢。至於替代性的語音及手寫輸入，學習門檻低，但需要準備額外的裝置(例如麥克風、手寫板等)且需經過辨識的過程，以現存的輸入法來看，效果還不算非常準確，而且速度極慢，因此目前中文輸入法的主流還是使用鍵盤做為輸入的媒介。而在中文輸入法檢定的競賽中幾乎皆是字形類輸入法的天下，故本文所改良及設計的輸入法亦以字形為依據。

(二) 倉頡輸入法之簡介

目前在中文輸入檢定中奪冠的輸入法為「大新倉頡」輸入法[4]，為「倉頡」輸入法[7][9]的改良版本，故於此節先簡單介紹倉頡輸入法。倉頡輸入法為朱邦復先生於 1976 年發明之「形義檢字法」於 1978 年正名而來。目前在臺灣地區大量推廣使用的版本為 1982 年推出的倉頡輸入法三代，該輸入法版本亦為多數作業系統所內建的中文輸入法之一。

朱邦復先生透過文字的整理，歸納出 24 個組成中文字的倉頡字母，其中又可分為 4 類，並加入 90 個輔助字根，茲整理如表 1-1。

表 1-1 倉頡輸入法字母與輔助字形對照表

分類	倉頡字母	英文字母	輔助字形
哲理類	日	A	罒
	月	B	冂 夕 爫 冖 目 ⺆
	金	C	八 丷 儿
	木	D	朩 寸
	水	E	⺢ 氵 氺 又
	火	F	⺌ 灬 小 ⺍ 个
	土	G	士
筆畫類	竹	H	厂 丿
	戈	I	广 丶 厶
	十	J	宀
	大	K	疒 𠂇 乂
	中	L	衤 丨 ⺺
	一	M	工 厂 ㇀ 丆
	弓	N	⺈ 亅 ㇇ 乙 ク ⺄
人體類	人	O	㇏ 亻 人 𠂉 入 ㄑ 𠂆
	心	P	勹 忄 ⺗ 匕 七 乀 乀
	手	Q	⺧ 扌 龵 ⺘
	口	R	
字形類	尸	S	⺁ コ 匚 ㇆ 卜
	廿	T	䒑 廿 非 业 共 廾 ⺍
	山	U	屮 凵 乚 屮
	女	V	乚 ㄑ 𡿨 ㇗ ㇀ ㇙
	田	W	囗 ⺟
	卜	Y	辶 卜 亠 ⺀

倉頡輸入法定義了幾個名詞如下：

(1) 字首

字形可從縱向或橫向切割者，其「最左方」或「最上方」之部份稱做字首。

(2) 字身

在可分割的字中，除字首以外的部份稱為字身，若字身還可再分解，則又可分割為次字首與次字身。

(3) 分體字

可分割為字首與字身的中文字稱為分體字。

(4) 連體字

字形筆劃相連而形成一個字，如:「我」、「凹」。字形筆劃不全部相連，但形勢上為一完整之字，如:「來」。

複合字、難字、特殊字亦視作連體字。

(5) 複合字

複合字首：以表 1-2 中的字當作「字首」或「次字首」，視為一整體，只取首尾碼。複合字：表 1-3 中的字，不論作字首或字身，或單獨出現，均只取首尾碼。

表 1-2 倉頡輸入法複合字首表

字首	麻	厭	辰	气	合	羽	薛
取碼	ID	MK	MV	ON	OR	SM	TJ
字 例	蘼	壓	褥	滊	盒	[illegible]	孽

表 1-3 倉頡輸入法複合字表

複合字	門	鬥	卩	隹	幾	𣎆	虍
編碼	AN	LN	NL	OG	VI	YN	YP
字 例	問	鬧	鄰	雄	畿	瀛	唬

(6) 難字

有些字的部份形狀過於瑣碎，取碼困難，故列為難字，對應到 X 鍵。取碼方式有以下兩種:(甲)若首、尾碼易取碼，僅中間部份難取，則取首、難、尾三碼。見表 1-4。(乙)若尾碼也不易決定，則只取首難二碼。見表 1-5。

(7) 特殊字

表 1-6 五種字形，若有其他筆劃介於其中時，先取此五種字形，再取其餘部份。

(8) 重複字

只要字碼相同之字，稱為重複字。其中使用頻率高者為本字，其餘為重複字。重複字在原編碼前方多加一個「X」，以作區別。如果加了「X」而超過五碼，則原字碼刪去最末一碼。若再有第二重碼字，則依此類推。

倉頡輸入法是將分體字先分解為字首與字身兩個部份，再分別針對其形狀依據由左而

右，由上而下，由外而內的順序進行拆解取碼。在拆解時還需遵守儘量不破壞字形結構的完整原則以及略過包覆字形內筆畫的省略原則，而連體字、分體字以及輔助字形在取碼時的規則如下：

(1) 連體字：4 碼(含)內依序取完，否則取首、次、三、尾共 4 碼。

(2) 分體字：若字首為倉頡字母或輔助字形，僅取 1 碼，否則先取字首的首、尾 2 碼。

若字身為連體時，3 碼內依序取碼，超過時僅取連體字身之首、次、尾 3 碼；字身為分體字時，若次字首可 1 碼取完，則次字身取首、尾 2 碼，否則次字首取首、尾 2 碼，次字身僅取尾碼。

(3) 輔助字形：單獨使用時，不可僅取單碼，視為連體字依筆畫順序取碼。

表 1-4 首、難、尾字表

難字	字碼	範例字
身	HXH	軀、鎢
慶	IXE	櫻
鷹	IXF	薦
鹿	IXP	麒、麓
弗	LXH	姊
鬧	LXL	淵
龜	NXU	鬮
黽	RXU	蠅
兼	TXC	鶼、廉

表 1-5 首、難字表

難字	字碼	範例字
臼	HX	舅
𦥑	HX	盥
𦥑	HX	輿
肅	LX	蕭、瀟
卍	NX	
齊	YX	霽、薺

表 1-6 特殊字表

字形	編碼	範例字
木	DB	柬
	DJ	末
	DL	束
	DOO	來
	DW	東
	DWF	東
火	FC	灬
	FQ	㷮
禾	HDL	秉
	HDLP	乘
大	KKKK	爽
	KMAA	奭
	XKN	夷
	KOO	夾
七	PU	屯

由以上的說明可知倉頡輸入法每個字最多取 5 碼，實際上使用於編碼的鍵數有 25 個(24 個倉頡字母鍵加上 1 個難字鍵 X)，因此編碼的空間多達 $\sum_{i=1}^{5} 25^i$ =10172525 種組合，遠大於中文字的數量，故不同文字衝碼或同碼的機率會比較低，因此花在選字上的時間少。相較於字音類的輸入法來說，輸入速度雖然快上不少，但由於輸入一個字需要的按鍵次數較多，拆碼規則又過於繁複，難字及特殊字表的處理又另有規則，因此相當難學，速度的提升也有限。

(三) 大新倉頡輸入法

大新倉頡輸入法為蘇清得先生改良傳統倉頡輸入法的取碼規則後，於 2001 年推出的中文輸入法，目前在中文輸入檢定紀錄中是速度最快的輸入法。輸入法的主體主要還是依據傳統倉頡輸入法，在字根表與取碼規則方面僅做些微的調整[10]。但改進的部份主要如下：

(1) 碼數減少

總碼數最多僅取 4 碼，分體字之字首與字身分別最多各取首、尾 2 碼；若為連體字，則最多僅取首、次、尾 3 碼。

(2) 增加使用按鍵與輔助字形

除了傳統倉頡使用的 25 個鍵之外，另外使用 Z 鍵代表「言」字根以及分號鍵「；」代表「食」與「禾」兩字根，並將使用之字根(含倉頡字母)總數由原本的 114 個擴增為 150 個。

(3) 大量使用簡碼

將使用率高的常用字另外編排成 1、2 碼簡碼，需記憶並經常使用，方可增進輸入速度。

透過以上的調整，大新倉頡自推出之後便在中文輸入檢定中取代了嘸蝦米輸入法[3][10]，成為輸入速度的冠軍，目前的紀錄甚至可達每分鐘 238 字。

二、改良與新輸入法設計

仔細觀察過曾經成為中文輸入檢定冠軍的輸入法之後發現，不管是行列輸入法、嘸蝦米輸入法乃至於大新倉頡輸入法，它們與之前的傳統倉頡輸入法以及大易輸入法等的差別，除了字根以及拆碼規則的不同之外，其共通的特點就是它們皆縮短了最大取碼數，由 5 碼減為 4 碼，另外就是大量使用需要記憶的簡碼。

使用倉頡輸入法的 25 鍵來進行編碼，若在一字最多只取 3 碼的狀況之下，編碼空間為 $\sum_{i=1}^{3}25^{i}=16275$ 個組合；若改以鍵盤主要輸入區中的 30 個按鍵來進行編碼時，編碼空間可擴增為 $\sum_{i=1}^{3}30^{i}=27930$ 種組合，因此我們認為一字 3 碼應已足夠處理臺灣地區 Big5 字集所收錄的 1 萬 3 千餘字。

決定方向之後，接下來便需慎選字根、安排字根與按鍵的對應以及設計取碼規則，再依此產生碼表(所有文字與其取碼的對應表)，然後評估結果再進行調整，然後重新產生碼表，再評估調整...，如此反覆循環，直到發現足夠好的組合為止。由於過程十分費時費力，因此需要設計一套快速產生碼表的程式以方便後續評估與調整的動作。

(一) 編碼工具的設計

研究初期(約 2006 年)之時，碼表是透過人工檢視編輯直接產生而成的，因此任何改變，即使只是簡單的字根與鍵盤的對應的調整，皆需將 1 萬 3 千餘字從頭到尾檢查一遍，一一挑出進行修改，所需耗費的時間過長。後來我們改為兩階段的處理，將人工編輯分成兩個部份，其一為產生每個中文字與套用拆、取碼原則所取用的字根的表格(每一個字在最後取碼時的第 1、2、3 碼分別取用哪一個字根)，另一個表格則記錄著每個字根與鍵盤的對應關係，最後再由程式產生碼表。但這種簡單改良的編碼法也僅簡化了調整字根鍵位的步驟，僅需修改第二個表格後重新執行程式即可，但是對於字根的抽換或者取碼規則的改變並無多大助益。因此必須重新設計一個更有效率的方法，不管是對於字根鍵位的調整、字根的抽換、編碼或取碼規則的更動，都能夠較方便且快速地進行修改。由於中文字屬於圖形系統，每一個字的形狀裡包含著文字結構的資訊，因此若能善用文字的結構，便能透過程式的協助來產生碼表。

文字結構的資訊參考自中央研究院文獻處理實驗室所釋出之漢字構形資料庫[1]，我們接著撰寫程式找出底層部件(該部件不可再以其他部件組合而成)，再指定底層部件與字根的組合關係，以及字根與按鍵的對應，這幾個步驟類似在定義 context-free grammar 的規則，只是每個非終端符號(non-terminal character)右方的替換目標是唯一的，如圖 2-1 所示。最後透過程式逐步替換即可產生碼表。

如此一來，一些對應關係的改變，不管是字根與鍵盤的對應，或是字根的抽換等，只要改寫少數幾條相關的規則即可；至於取碼規則

的改變，則需要實作在程式中。善用文字結構的資訊，除了可以快速產生碼表之外，更可減少人工逐字編碼可能產生前後不一致的情況。

劗	贊/刂
劘	靡/刂
劗	鑫/刂
功	工/力
加	力/口
劣	少/力

圖 2-1 替換規則的範例

每次編完碼表之後，我們利用程式統計一些資訊來當成評估效果的指標，再依這些指標資訊進行調整：

(1) 總編碼數：指所有文字的取碼方式的總數，愈大愈好。

(2) 重碼/同碼字的數量：同一個取碼方式對應到的文字要愈少愈好，表示選字的機會愈低。

(3) 字根與被取碼位置的關係：每個字根被取成第 1 碼、第 2 碼與第 3 碼的數量，由這個指標來決定字根與鍵盤按鍵的對應，此外也可以評估是否要加入或抽換字根，基本上要讓每個按鍵被當第 1 碼、第 2 碼與第 3 碼的次數愈接近、愈平衡愈好。

圖 2-2 的範例表示 13060 個中文字在某種字根鍵盤對應以及取碼規則之下，總共有 9717 種不同的取碼法，而這 9717 種取碼法之中，有 7122 種對應到單一個字，不與別的字同碼，而 1975 組則是 2 個字有相同取碼法，509 組則是 3 個字同碼，而一個取碼法最多會對應到 6 個字，而這樣的狀況只有 3 組。

在圖 2-3 的範例中，e 鍵(即「水」鍵)被 815 字取為第 1 碼，被 352 字取為第 2 碼，有 450 字將它取為第 3 碼。而對應到 e 鍵的字根當中，三點水(氵)開頭的字有 774 個。

這個編碼方法的概念很簡單，程式也相當容易撰寫，執行效率也不錯，在本研究中利用 P3-1GHz 的舊電腦，約莫 12 秒可以產生碼表並且完成所有的統計工作。

```
Total lines:   9717
7122   1
1975   2
 509   3
  89   4
  19   5
   3   6
```

圖 2-2 總編碼數與同碼字數量統計範例

```
=== e ==================
三    1  16  52
丶   33 237 292
水    7  20  49
氵  774  75   0
氺    0   3  57
氷    0   1   0
     --------------
     815 352 450
=== f ==================
一   17  75  20
宀  170 273   0
火  238  81  77
灬    0  29 212
鳥   17   9 209
亠   10  21   0
小    0   9  34
⺌    0   0  11
     --------------
     452 497 563
```

圖 2-3 字根與取碼位置的統計範例

(二) 師大大師輸入法

本研究的產出物為「師大大師輸入法」，其原始概念於 2005 年夏天開始發想，2006 年由師大資工系碩士班學生江漢昇進行初次改良嚐試[8]，當時僅修改了傳統倉頡輸入法的取碼方法，將最大取碼數由 5 碼縮減為 3 碼，並未更動字根以及鍵盤對應。其效率已經比大新倉頡好了，然而同碼字的數量甚多，選字較為不易。而我們在後續的研究中，透過前一節編碼工具的輔助，本研究大幅更換選用的字根以及字根與鍵盤的對應，以期達到更好的效果。

縮短最大取碼數會使得輸入時非同碼字的按鍵次數減少而提升打字速度，但它也會造成編碼空間下降，使得不同文字衝碼/同碼的機會提高，需要選字的機會也變多，因而拖慢同碼字輸入的速度。雖然縮短最大取碼數為改良輸

入法的主要方向，但仍需搭配其他的配套措施才能有效提高速度。

表 2-1 師大大師輸入法字根表

中文符號	英文字母	字根
日	A	日曰
月	B	月冂日 冂冃
金	C	金
木	D	木
水	E	水
火	F	火
土	G	土
竹	H	竹
戈	I	戈
十	J	十
大	K	大
中	L	中
一	M	一
弓	N	弓
人	O	人
心	P	心
手	Q	手
口	R	口
尸	S	尸
廿	T	廿
山	U	山
女	V	女
田	W	田
難	X	
卜	Y	卜
貪	Z	冬
縪	,	
設	.	
痢	/	
祤	;	衤 羽 衣

為了減緩縮短最大取碼數造成重碼率上升的後遺症，我們採用了主要輸入區中的 30 個按鍵來進行編碼，如表 2-1 所示，共有 187 個字根，除了沿用原本傳統倉頡所使用的 25 個按鍵及代表的中文字母(符號)之外，另外加上了「Z」、「,」、「.」、「/」、「;」，並分別指定代表該鍵的中文符號為「貪」、「縪」、「設」、「痢」、「祤」。另外為了方便記憶，在安排字根與按鍵之對應關係時秉持著「根藏鍵中」的原則：大多數的字根可由其按鍵所代表的中文符號中拆解出來，並將衍生且相關的字根一同置於相同按鍵上。

例如：「貪」的字根有「冬」、「𠔃」、「𠂉」、「一」、「頁」、「貝」、「目」7 個，均是「貪」字內部的一個部份字形。

在拆碼時依照由上而下、由左而右、由外而內的順序，另外更明確做了規範，拆碼需要遵守以下原則：

(1) 碼少原則：拆解碼數需最少

要涵蓋整個中文字，拆解之字根總數需為最少。如：

永 → 𠂇水，不要拆成 丶一水

王 → 一土，不要拆成 一十一

求 → 丶十水，不要拆成 丶寸氺

(2) 「剪刀式裁剪」優於「疊合交錯式」拆碼，而剪裁時，裁於「空白處或線段邊緣」又優於裁於「線段中間」，如：

者 → 十𠂉日，不要拆成 土丿日

丈 → 十乂，不要拆成 𠂉乀

天 → 一大，不要拆成 二人

干 → 一十，不要拆成 工丨

哉 → 十戈口，不要拆成 土戈口

(3) 大碼優先原則：

即前碼優先取較大碼，且前碼佔用面積愈大愈好，例：

爭 → 爫 ，不要拆成 爫

亍 → 二 ，不要拆成 一丁

乇 → 千乚，不要拆成 𠂆七

以上字根拆解原則的口訣可歸納如下：

特定>碼少>裁邊>裁中>疊合>大碼

由於此輸入法最多只取 3 碼，取碼時的原則如下：

(1) 總碼數未超過 3 碼：依序全取。

(2) 超過 3 碼之連體字：取首、次、尾 3 碼。

(3) 超過 3 碼之分體字：取字首之首碼與字身之首、尾碼；若字身僅單碼，則字首取首、尾碼後再取字身。

(4) 口尾碼換碼原則：依以上 3 個原則最多取 3 碼後，若末碼為「口」且其前一碼未被取用，則將末碼改取為其前一碼，如：

喜 → 士口 艹　　嘉 → 士口力

除了編碼方面的改進之外，針對操作模式也做了重大的調整。觀察市面上大多數的中文輸入法，即使按鍵次數已達最大取碼數也要補按空白來做確認，若一個字在編碼時的按鍵數為 4，實際在進行輸入時包含做確認的空白鍵，共需按 5 次按鍵。因此，本研究的另一個改進的方向為減少不必要的空白鍵，當輸入按滿 3 鍵時，自動送出第一個候選字，而編碼長度不到 3 的字，則仍需按空白鍵表示該次輸入結束，此時仍自動送出第一個候選字。

在同碼字的處理上，將同一編碼的文字依照字頻由大至小排列，因此自動送出的第一候選字一定是該組同碼字中的字頻最高的最常用字。目前字頻的資訊來自教育部網站上取得之「八十七年常用字字頻總表」[5]，未來若有更新之統計資料可再更換。若要打的字非最高頻字，可以用右手小指按熱鍵「'」依次切換下一個候選字，也可以直接按數字鍵選擇並確定要打的字。圖 2-4 顯示操作的示範。

三、評估與比較

會影響一個輸入法速度的因素，最主要的是打出一個字所需的按鍵次數，而與按鍵次數直接相關的就是每個字的編碼長度。以下針對臺灣地區 Big5 字集所收錄的 13068 個中文字進行探討。

(一) 理論最少平均按鍵次數的推導

假設在編碼時使用了 k 個鍵，而除了編碼使用的按鍵之外，每一組編碼最多有 s 個字可供選用，將所有文字的編碼對應關係畫成樹狀結構，如圖 3-1 所示。根節點(root)表示準備輸入一個新字的狀態，第 1 層的節點表示按下第 1 個編碼鍵，第 2 層表示連續按下第 2 個編碼鍵，餘此類推。除了根節點之外的每個節點最多允許編入 s 個中文字，在根節點以下第 1 層節點上的字僅編成 1 碼，第二層的字編為 2 碼...，第 n 層節點上的字編碼長度為 n，且該層上的中文字最多可有 $s*k^n$ 個。

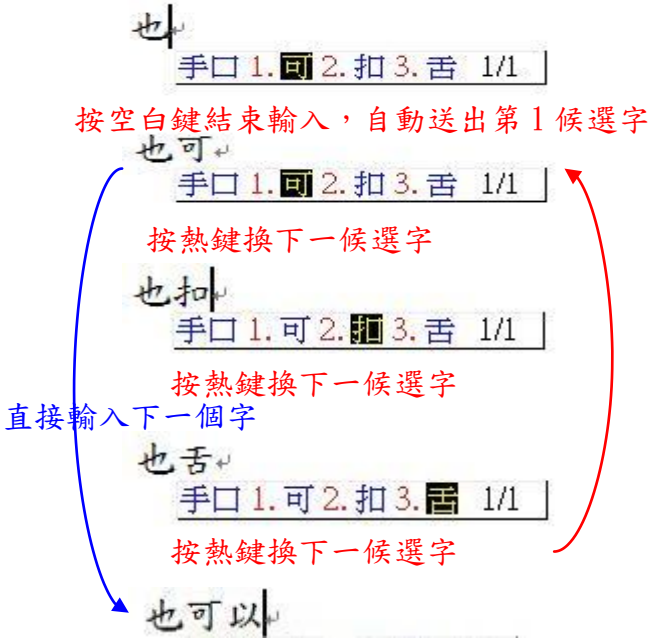

圖 2-4 以熱鍵「'」來換字的操作示範

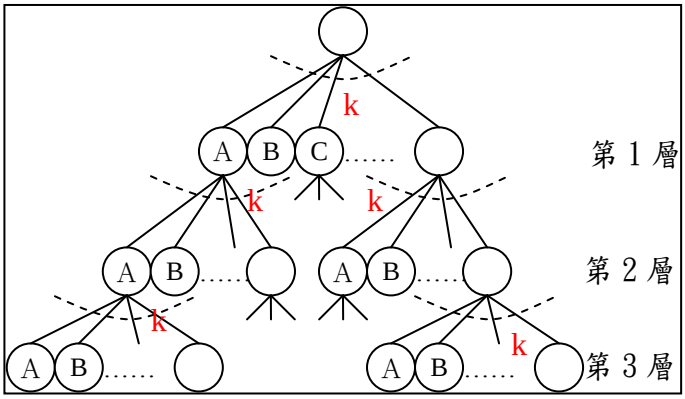

圖 3-1 編碼樹示意圖

在不考慮字頻的狀況之下，最少平均編碼長度可由以下的公式得出：

$$\frac{ks+2k^2s+3(13068-ks-k^2s)}{13068}$$

其中 13068 為本研究中討論的中文字總數，且由之前的說明，若使用 25 個鍵來編碼，在最多取 3 碼的狀況之下，編碼空間可達 16275，可編入 16275*s 個中文字，已可編出所有 Big5 的 13068 字，因此上述的公式僅考慮了 3 層的狀況。

由於在編碼樹中上層的字的編碼可為其下層字編碼的前置字串(prefix string)，因此在實際打字時無法判斷目前已輸入的按鍵組合是否已完結。所以在實際打字的情境之下，要輸入前 2 層的字時，需在最後補上一個確認的符號，一般的輸入法皆是以編碼鍵之外的其他按鍵當成結尾標記來選字(例如按空白鍵或 1～9 鍵)。故第 1 層的字需按 2 鍵(字碼 1 個+選字)，第 2 層按 3 鍵(字碼 2 個+選字)，為了省下第 3 層的選字時間，第 3 層的每個節點僅能編入 1 個中文字，所以要輸入第 3 層的字只需要 3 鍵(僅字碼)，在此我們可將系統設計成立刻上字，不需再按空白鍵。因此在不考慮字頻影響的狀況之下，每個字的重要性是一樣的，則最少平均按鍵次數可由以下公式得出：

$$\frac{2ks+3k^2s+3(13068-ks-k^2s)}{13068}=3-\frac{ks}{13068}$$

表 3-1 編碼鍵數與理論平均按鍵次數下限

編碼鍵數 k	平均按鍵次數下限	
	s=1	s=10
25	2.998087	2.980869
27	2.997934	2.979339
30	2.997704	2.977043

如表 3-1 所示，若要完全不重碼的話，在使用 25 個鍵來編碼時，平均一個字要按 2.998087 鍵，若允許每一組編碼能有 10 個同碼字的情形下，則平均只要按 2.980869 鍵就能打出一個中文字。

若要將字頻的影響考慮進來，則必須將愈常用(字頻愈高)的字編愈短的碼，所以在編碼樹上必須依照愈高頻的字安排在愈上層的原則將中文字依序填入各個節點。若將中文字依字頻由高至低排序，第 1 個字的字頻為 w_1，第 i 個字的字頻為 w_i。字頻的統計資訊來民國 83 年[11][12]、87 年[5]與 88 年[6]的統計數據(如表 3-2)。但由於每份資料在濾掉無法辨識的字後所得的合法統計字數並不一樣多，為方便計算，將每份統計數據中未載明的中文字的出現次數設為 0.5 次，並將次數換算為該字在該份數據中被使用的百分機率(也就是 w_i)，故可透過以下公式算出在字頻影響下的平均最少按鍵次數：

$$\frac{2*\sum_{i=1}^{ks} w_i+3*\sum_{i=ks+1}^{ks(k+1)} w_i+3*\sum_{i=ks(k+1)+1}^{13068} w_i}{100\%}$$

$$=\frac{2*\sum_{i=1}^{ks} w_i+3*\sum_{i=ks+1}^{13068} w_i}{100\%}$$

為了方便之後與其他輸入法比較，以下以一般輸入法每一頁最多所能顯示的候選字數量 10 為 s 的值，表 3-3 為依據上述公式，套用不同字頻統計數據算出來的結果。

(二) 與其他輸入法的比較

為了比較師大大師輸入法是否有達到預期

目標，挑選了倉頡 3 代、倉頡 5 代以及大新倉頡 7.0 版進行分析，各輸入法碼表的來源來自「gcin 同好會」的網站[1]。在統計時僅將 Big5 字集收錄的 13068 字列入計算，當一個字有多種編碼時，以最短編碼為計算依據(有編簡碼時以簡碼計算)。

表 3-2 各統計數據相關資訊

使用鍵數	統計字數	總頻次	數據出處
83 年數據	13060	171894604	黃世昆、蔡志浩
87 年數據	5023	1576492	教育部
89 年數據	3563	253496	教育部

表 3-3 考慮字頻時的編碼鍵數與理論平均按鍵次數下限(假設 s=10)

編碼鍵數 k	理論平均按鍵次數下限		
	83 年數據	87 年數據	88 年數據
25	2.35719	2.43626	2.48148
27	2.33995	2.41741	2.46104
30	2.3166	2.39162	2.43318

由表 3-2 與表 3-3 的結果可知，若字頻統計資訊的愈完整，算出來的下限愈低愈精準，故以下列表的字頻資訊以 83 年的統計數據為主。另外 s 的值同樣固定為 10。

表 3-4 不考慮字頻時各輸入法平均按鍵次數

輸入法	平均按鍵次數	理論下限	備註
倉頡 3 代	5.11394	2.980869	k=25
倉頡 5 代	5.11348		
大新倉頡	4.24105	2.979339	k=27
師大大師	3.23975	2.977043	k=30

由統計數據來看，相較於其他輸入法，本研究產出之師大大師輸入法在未編簡碼的情況之下，不論考不考慮字頻的影響，平均按鍵次數皆低於其他輸入法，而其中大新倉頡的碼表中已包含了為大量的常用字指定了額外的簡碼，而由於計算時是以每一個字的最短編碼為依據，故大新倉頡在這部份並非以原始編碼長度統計。因此，若要使用大新倉頡輸入法達到極速，必須死背大量的簡碼替代原本的編碼，以提升打字速度，因此大新倉頡的打字員若不背誦簡碼的話，則打字速度將急遽降低。表 3-6 為不使用簡碼的範例，大新倉頡平均按鍵數為 3.892857 次，本法只要 2.964286 次(這和表 3-5 中純機率分析的 2.94058 次接近)，很明顯地，我們快了很多。

表 3-5 考慮字頻時各輸入法平均按鍵次數

輸入法	平均按鍵次數	理論下限	備註
倉頡 3 代 (未用簡碼)	4.45	2.35719	k=25
倉頡 5 代 (未用簡碼)	4.45009		
大新倉頡 (使用簡碼)	3.04488	2.33995	k=27
師大大師 (未用簡碼)	2.94058	2.3166	k=30

師大大師輸入法在未編列簡碼的情況之下，理論上的數據表現已勝過目前的王者大新倉頡，證明先前我們所構思的改進方向是正確的，而且利用程式及漢字構形資料庫的協助、自 2006 年 12 月至 2009 年 8 月，歷經兩年多逐步改良的字根表及拆碼規則極具成效。但若要再進一步改良，在 2006 年本實驗室江漢昇同學的碩士論文所做的嚐試中已提及，若能再額外編列 1 碼快碼(共 30 個最高頻的中文字)供師大大師輸入法使用，估計平均按鍵次數約可再降 0.1 次。在此我們亦可約略估算若再將一些有重碼的、常用的高頻字額外編列 1、2 碼簡碼給師大大師輸入法使用的結果：由於絕大多數的常用字均可按 2 鍵(字碼 1 個+空白鍵)或按 3 鍵(字碼 2 個+空白鍵、或者字碼 3 個直接上字)，因此只剩極少數不常用(頻率極低)的中文字需按 4 鍵(字碼 3 個+選字鍵)，這樣的平均按鍵次數將很接近表 3-2 的分析，也就是接近 2.39162 次；換言之，師大大師輸入法的整體效能可以逼近中文輸入法的最佳速度(near-optimal)。

表 3-6 不使用簡碼的比較範例：朱熹「觀書有感(其一)」，△表示空白鍵

詩文	大新倉頡			師大大師		
	正常取碼	按鍵	鍵數	正常取碼	按鍵	鍵數
半	𠂇𰀁	火手△	3	丷𰀁	金手△	3
畝	亠田𠂉乀	卜田弓人△	5	亠𠂉乀	栩弓難	3
方	亠丿冂	卜竹尸 3	4	亠丿冂	栩竹弓	3
塘	土广口	土戈口 2	4	土广申	土卜中	3
一	一	一△	2	一	一△	2
鑑	金匚丗	金尸廿△	4	金匚丗	金尸廿	3
開	日亅一廾	日弓一廿△	5	日一廾	日一難	3
天	一大	一大△	3	一大	一大△	3
光	𠂇𠂉乚	火一山△	4	𠂇𠂉乚	山大尸	3
雲	一冖一厶	一月一戈△	5	一二厶	一栩緷	3
影	日小丿丿	日火竹竹△	5	日丿丿	日竹竹	3
共	丗八	廿金△	3	廿𠂆	難廿△	3
徘	彳丨卜	山中卜△	4	彳川三	中金水	3
徊	彳口口	山田口△	4	彳口口	中田口	3
問	日亅口	日弓口△	4	日亅口	日弓口	3
渠	氵コ木	水尸木△	4	氵コ木	水尸木 2	4
那	冂𠂇一丨	尸手弓中△	5	冂𠂇阝	弓手中	3
得	彳日丶	山日戈△	4	彳日丶	中日水	3
清	氵𰀁月	水手月△	4	氵𰀁月	水手月	3
如	女口	女口△	3	女口	女口△	3
許	言𠂉十	乙人十△	4	言𠂉十	設貪十	3
為	丶𠂇灬	戈大火△	4	丶𠂇灬	水大火	3
有	𠂇月	大月△	3	𠂇月	大月△	3
源	氵厂小	水一火△	4	氵厂小	水設山	3
頭	一𠂒𠂆八	一廿一金△	5	一𠂒頁	一廿貪	3
活	氵丿口	水竹口△	4	氵千口	水手口	3
水	水	水 2	2	水	水 3	2
來	木人人	木人人 2	4	木人人	木人人	3
大新倉頡平均按鍵次數			3.892857	師大大師平均按鍵次數		2.964286

四、結論與未來方向

本研究產出之師大大師輸入法提出了下列嶄新的設計理念：

(1) 根藏鍵中：學習門檻更低

重新設計的拆碼規則比大新倉頡更簡化，而字根表的字根則盡量以「根藏鍵中」之原則進行編排，以方便記憶。

(2) 三碼上字、熱鍵換字：輸入速度更快

在未編列簡碼的情況之下，平均按鍵次數已低於大量使用 1、2 碼簡碼的大新倉頡輸入法。我們若使用 1、2 碼簡碼，則速度將接近最佳的理論極限。

(3) 僅用 30 鍵：手指移動距離短

用來輸入的按鍵 30 個按鍵，剛好位於鍵盤上 3 排主要輸入區之內，與大新倉頡的 27 鍵相仿，手指移動距離差不多。

(4) 程式協助排字根表：同碼字選字率低

如表 3-6 的範例，在輸入 28 個字時，大新倉頡有 4 個字(方、塘、水、來)需選字，而本法則只有 2 個字(渠、水)需選字。實際在打文章時亦是接近此比率。因此本法的選字率更低。

目前本文作者已在 Windows XP 系統實作了師大大師輸入法的程式，已可應付一般打字的情境，未來除可增列簡碼之外，在一些其他輔助的功能上，仍可再做一些增強，比方簡繁轉換、特殊符號輸入，或者是將 unicode 所收錄的全部中文字皆列入處理範圍。當然也可設法在其它平台上實作，以饗同好者。

現已實作好之可用於一般輸入用途之輸入法程式，將利用此論文發表時正式公開，並在網路上免費開放給大眾使用，以效法當年朱邦復先生奉獻之精神。日後將視試用之狀況，繼續改良，以提供國人一個免費、優良的輸入法而努力。

五、誌謝

本研究承蒙國科會提供研究計畫(NSC97WFA0300752)經費補助，謹誌謝忱。

六、參考文獻

[1] Gcin 同好會網站，http://cle.linux.org.tw/trac/

[2] 中央研究院文獻處理實驗室，“漢字構形資料庫”，http://cdp.sinica.edu.tw/cdphanzi/

[3] 劉重次，嘸蝦米輸入法。台北：行易，民 89 年。

[4] 大新倉頡，http://www.eztyping.com.tw/

[5] 教育部，“八十七年常用字字頻總表”，http://www.edu.tw/files/site_content/M0001/87news/page2-1.htm?open

[6] 教育部，“八十八年網頁用語資料庫”，http://www.edu.tw/files/site_content/M0001/87news/page5-1.htm?open

[7] 朱邦復工作室，http://www.cbflabs.com/index.php

[8] 江漢昇，”中文輸入法之改良研究及「師大大師輸入法」之實作”。國立臺灣師範大學資訊工程所碩士班，民 95 年 6 月。

[9] 沈紅蓮，”第五代倉頡輸入法手冊”，臺北：國榮高科技發行，民 80 年。

[10] 胡延宗，”大新倉頡與嘸蝦米輸入法之輸入績效比較研究”，大葉大學資訊管理學系碩士班，民 94 年 6 月。

[11] 蔡志浩，“Frequency and Stroke Counts of Chinese Characters”，http://technology.chtsai.org/charfreq/

[12] 黃世昆，“1994 年 Big5 中文網路討論字頻統計”，http://technology.chtsai.org/charfreq/94charfreq.html