The Journal of National Chiao Tung University
Vol. 2, December 1976
pp. 1-8

新中文字根集合的設計

The Design of a New Chinese Character Component Set

杜敏文 Min-Wen Du

Department of Computer Science, N. C. T. U.

(Received August 9, 1976)

ABSTRACT — A new character component set for chinese Input/output of a computer has been designed for the following two ends. First, the maximum number of components necessary to identify a Chinese character can be reduced to only 4. Second, characters composed from components may have better appearance. The first end is achieved through introducing new character components. The second purpose is approached by introducting character component variation concept.

摘要 — 本文敘述如何設計用以作電算機中文輸入輸出的一組新字根。該組字根設計目標有二。即一，最多只要用四個字根即可單一區分每個中文字。二，使由字根組合出來的字能較為美觀。我們增加了一些新字根，同時使用字根變形的方法以達到這兩個目標。

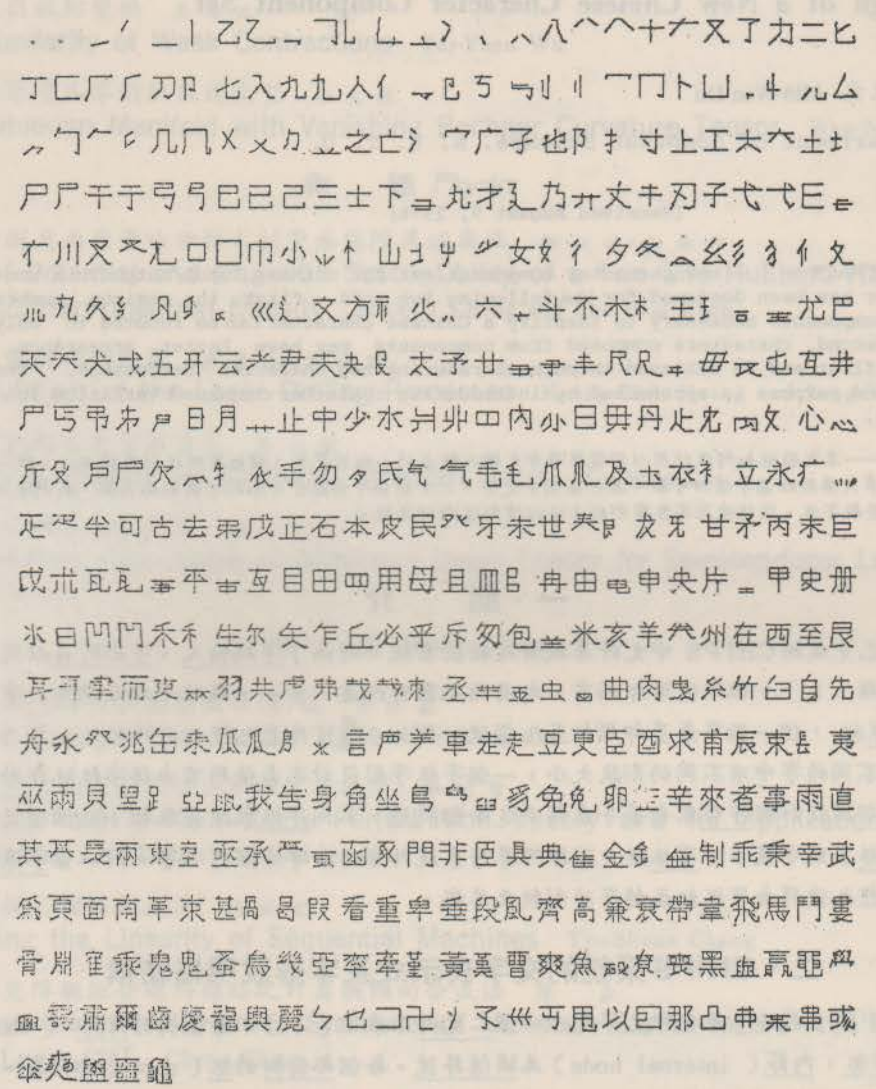
一、簡　介

在交大已完成的CHIPS 中文計算機輸入輸出系統，解決了字碼輸入、字形貯存以及內用字碼等基本問題〔 1 〕。但對使用者而言，却有兩點需作改進。第一，雖然平均每個中文字只需打 1.98 個字根鍵，但一個字最多却須打至九個鍵，例如：龘：龍匕匕龍匕匕龍匕匕。第二，由於有些字根在不同的字中有不同的形狀大小，一個字根字形設計不易使所有由該字根組合的字都很美觀。本文將從設計新字根集合着手來解決這兩個問題。不同字根數從原來的 460 個增至 532 個使得在新字根系統中最多打字根鍵 4 下即可尋得系統內 8532 字中任何一字。同一個字根給予數種字根的構想也使得由字根組成的字外形較為美觀。

二、CHIPS系統的文字表示式及字形說明檔結構

在CHIPS 系統中，每個中文字是以一個二元樹來表示〔 1 〕。每個樹的外節 (External node) 為字根，內節 (internal node) 為關係符號。每個部份樹的根 (root) 都代表一個字的部份字體。關係符號則表示部份字體的相對關係。在系統中共使用三種關係符號，即橫連△△，直連△和包含▲。字根數則共 460 個（不包括同字根異形體，如木，木，圖一），可表 48,713 個中文字〔 1 〕。

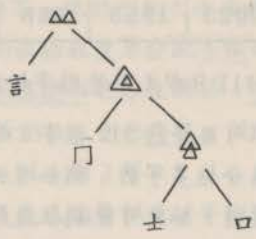
為了節省貯存空間，使字形說明檔內每個單元定長，以及尋字方便，CHIPS 中將不同數目字根的字分別建表貯存。同時，每個字都用抽出式波式法 (Extracted Polish Form) 表〔 1 〕。每個文字表示式分成三個部份，一個字根序列，一個關係符號序列及一個字根關係符號向量。根



圖一　CHIPS系統所使用之中文字根

據這三個部份就可把文字從字根組合起來。

例一：調的二元結構樹如圖二。



圖二 “調”字的二元結構樹

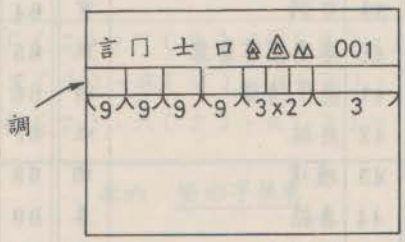
其抽出式表示式為

字根序列　關係符號序列　字根關係符號向量

字根序列及關係符號序列排列順序各為二元結構樹以波氏法（Polish Form）表時的相對順序，字根關係符號向量中0表字根，1表關係符號，此向量表二元結構樹以波氏法表時之字根與關係符號排列位置。

由於以抽出式波氏法表示，字根序列與關係符號是分離的，將此種文字表示式列表建檔後字根序列即可當做尋找鍵尋找文字表示式。若將字形說明檔依照字根大小順序排列，則可使用二元搜尋法（Binary search）有效率地找到為某一字根序列的字。

例二、調有4個字根，所以應於4個字根的字形說明檔中尋找。圖三為4個字根字的字形說明檔。CHIPS系統中共用499個字根，三種關係符號，所以每個字根用9筆（bit）表，每個關係符號用2筆表。4個字根的字其字根關係符號向量長度可省為3筆〔1〕。



圖三　4個字根字的字形說明檔

三、減少尋字所需字根鍵的數目

表四為在CHIPS系統中，所有8532個字使用字根數目的統計。由這個統計我們可看出超出5個字根的字共1208個。這對使用者是一大負擔。是否可能不用全部的字根序列即可單一地去找到一個字呢？我們嘗試的方法是用字根序列前面（左起）的幾個字根來單一區分所有的8532個字。經過程式測試的結果，發現若是只用每個字最前三個字根來尋字，將有933個字不能區分。若是用最前4個字根來尋，則僅有213個字不能區分。所以我們選擇用最前4個字根為尋找鍵的方法。不能分的213個字則設計新字根使能單一分出。

字根數	1	2	3	4	5	6	7	8	9
字數	271	2178	2923	1953	888	221	79	18	1

表四 CHIPS中文字使用字根數目統計

表五為CHIPS 中用4個字根不可區分的213個字，以及新增用以區分此213個字的字根。增加字根的原則有二，一為選擇可區分較多字的，例如增加㇕則許多字皆可由不可分而成可分。其二為由使用者看來該字體成為一整體，如此可便利往後的由字分字根工作。

		增根			增根			增根
1	鬲鷊	冏	27	償儻	尚	53	敝彆蹩嫳獘憋彆鱉	
2	羆罷	能	28	睿壑叡	㕡		瞥幣弊	㡀
3	棗棘	棘	29	參毿	参	54	嵩峒	峒
4	螃蠔	古	30	能態熊	能	55	從聳聳	從
5	莞蒄	元	31	毚劖	毚	56	愆衍	行
6	簽籲	僉	32	芻雛皺	芻	57	甭黼	甭
7	讀譖	買	33	渟濠	古	58	瘉癒	俞
8	麟麋	鹿	34	潮滸瀚	卓	59	歲劌翽	歲
9	鑒擘	々	35	澰瀲	僉	60	衞躗	韋
10	鄠靈	靈	36	瀅澇濴瀠濼	炏	61	僉歛	僉
11	鐓錞	亠	37	湯盪燙	易	62	侖歈觎	侖
12	鏜鐺錦	尚	38	漉瀌	鹿	63	鄉嚮響饗	鄉
13	晉鄑戩	晉	39	寅戭	寅	64	懋鬱爵鬱攣	楙
14	殄餮	歹	40	搴蹇騫搴賽塞	寒	65	椁槨	古
15	卒頓	卒	41	隳隋隋墮	隋	66	檀檁	回
16	熱墊	亠	42	搡掇	双	67	檻欖	𠂉
17	襄瓤	襄	43	壇壞	回	68	鬲鬴	冏
18	就鷲	亠	44	喜話	喜	69	景晾影顥	景
19	亶鸇氈顫皸	回	45	尋鄩	尋	70	旱旰	旱
20	鼓鼙鼗鼕鼞	鼓	46	嘵嚎	古	71	只叭	只
21	區毆鷗甌歐	區	47	啜嗪	双	72	顯歇	㬎
22	匿慝	若	48	叻另	另	73	辪櫱	𠂤
23	厭壓饜檿黶魘靨	厭	49	咢顎鶚鄂	咢	74	敬槃驚警擎	敬
24	旭晉	旭	50	單戰鄲辭獸疊	畄	75	觀歡顴勸鸛	雚
25	佼儆	交	51	吧邑	吧	76	藥藥	㗊
26	倞傹	亠	52	唄員	員			

表五 用4個字根不可區分的字

四、字根變形及組合字形之美化

一個字根在不同的字裡有可能需要不一樣的形狀大小才能使組合出來的字美觀。在〔1,2,3〕中使用定比重的辦法，根據每個字根的複雜程度來分配字根佔據的空間。如此解決了字根在字中的大小問題。同一字根不同形狀可用字根變形的方法解決。字根的變形可分為3類：

1. 一般變形：所組成字多而本身筆劃數又少的字根，須較多的變形方能適合各種不同組合的字。這類字根如丶（點），丿（撇），八（捌）等。

2. 偏旁變形：一些字根單獨成字及成為偏旁部首時字形不一樣，例如木，扌；金，釒；竹，⺮，等等。

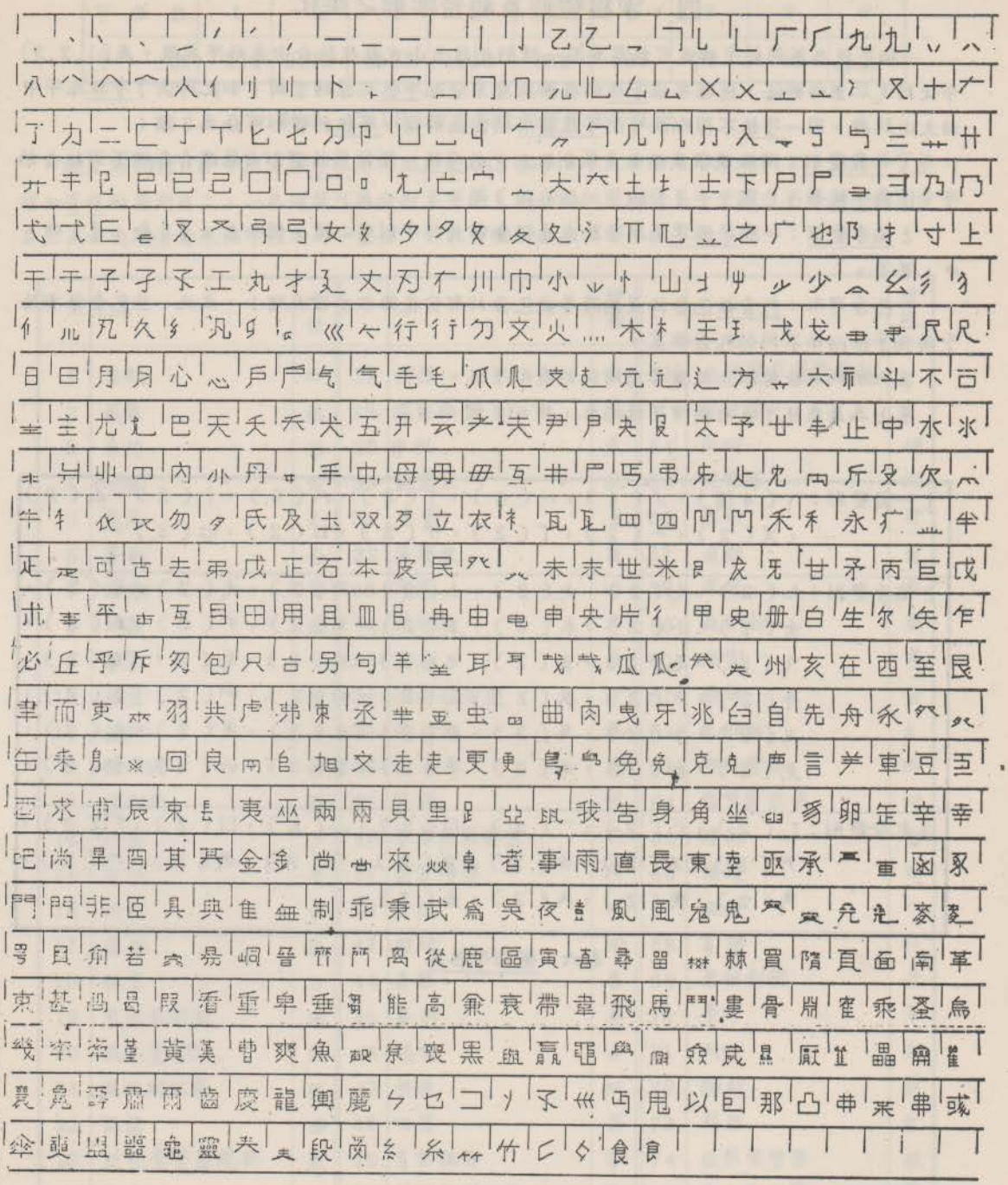
3. 包含變形：包含根包含位置應隨着被包含的部份複雜程度擴大縮小。為此一些包含根都給予兩個字根以示不同的包含位置。

表六所列為根據以上三種考慮所設的變形字根。

表七為原系統字根與新增字根總表，共616個。

1. 一般變形：	丶（4個），丿（4），一（4），丨（2），八（6），刂（2），厶（2），乂（2），亠（2），丁（2），廾（3），口（2），心（2）。
2. 偏旁變形：	九（2），儿（2），大（2），土（2），乃（2），弋（2），弓（2），女（2），小（2），木（2），王（2），戈（2），尹（2），尺（2），气（2），毛（2），支（2），元（2），六（2），尤（2），夭（2），牛（2），仌（2），衣（2），瓦（2），四（2），罒（2），禾（2），疋（2），爪（2），耳（2），瓜（2），走（2），更（2），免（2），克（2），金（2），尚（2），奓（2），糸（2），竹（2），食（2）。
3. 包含變形：	乙（2），厂（2），冖（2），冂（2），凵（2），囗（2），宀（2），尸（2），叉（2），夂（2），戶（2），癶（2），戕（2），夊（2），鳥（2），齊（2），夫（2），夫（2）。

表六 變形字根表

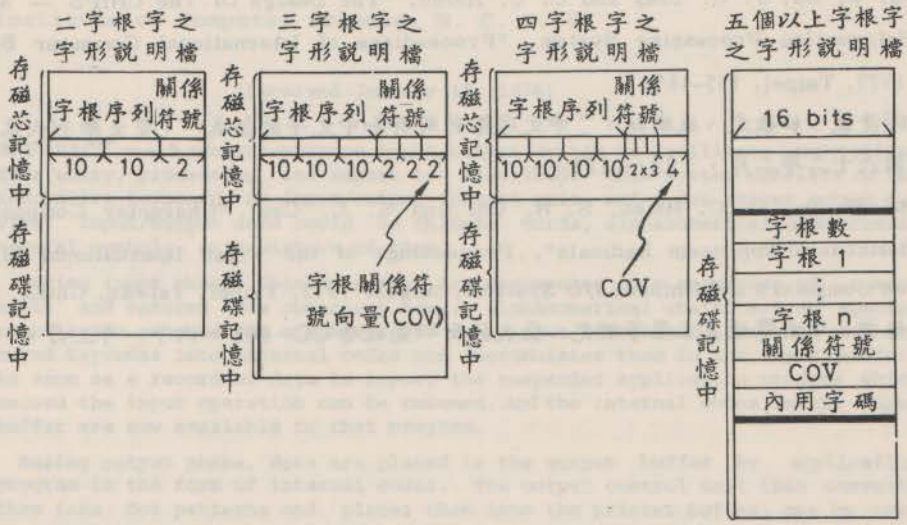


表七　新字根集合

五、字根碼之訂定，字形說明檔安排及字形說明之搜尋

由於一個字的字形說明是由字根序列作搜尋鑰來尋找，而在鍵盤上為了便利使用者，變形的字根只用一個鍵。也就是說，鍵盤上的一個鍵有可能對應系統內的幾個字根。所以字根碼的訂定，字形說明檔的安排，都必需依此設計以加速字形說明搜尋所需時間。

圖八為字形說明檔安排方法。



圖八　字形說明檔設計

一個字根的字不需字形說明檔，因各字根本身就是字。2個，3個，4個字根的字各歸併於一檔內。屬於較常用的4千字的字〔4〕存於磁慈記憶中，否則存於磁碟中。5個字根以上的字存於磁碟中。常用字存於磁慈記憶中可增進字形說明搜尋的速率。

2、3、4個字根的字形說明檔，存於磁慈記憶部份與存於磁碟部份可各別按照字根序列大小排列。5個以上字根的字形說明檔則按照前4個字根鍵排列。如此若出現的字存在磁慈記憶中，可用2元搜尋法迅速找到；若存在磁碟記憶中，可先建好一索引檔，只一次磁碟的讀入即可找到。

所謂的字根序列大小此處須加以說明。由於一個字鍵有可能對應幾個字根，字鍵輸入字碼實際上只能對應到一組變形字根的共同部份。所以變形字根的字根碼設計，以能彼此相鄰而 Hamming 距離愈短愈佳。例如點共有4個，字根碼各為 0000000000, 0000000010, 0000000001, 0000000011。其左方8個數字皆同。表七中由左至右，由上至下，字根的順序就是設計成使變形字根共同部份可最大者。

六、結　論

本文設計了一個新的中文字根集合，使〔4〕中的8532個字依照由左而右，由上而下，由

外而內的簡單規則分析，只用最前4個字根即可單一地定出任何一個字。本文提出的字根變形方法使中文字由字根組合時可有不同的選擇，因此使組合成的字較為美觀。

七、誌　謝

本文中尋找有相同4個字根字的程式由李立明先生完成，謹此誌謝。

參考文獻

〔1〕M. W. Du, J. C. Tsay and C. C. Hsieh, "The Design Of The CHIPS — A Chinese Information Processing System, "Proceedings of International Computer Symposium 1975, Taipei, 155-165.

〔2〕謝清俊、杜敏文、戚樹紅，“中文字根的貯存和中文字的合成”，交大學刊六卷一期，民國六十二年二月，122～131。

〔3〕M. W. Du, C. C. Hsieh, S. H. Chi and S. C. Chu, "Character Composition and Methods to Represent Radicals", Proceedings of the First International Symposium on Computers and Chinese I/O Systems, August 1973, Taipei, Taiwan, China, 63-78.

〔4〕林樹，中文電腦基本用字研究，交大技術研究報告CC-601，六十一年三月。