TR-82-011

中文字辨認系統研究計畫

第一期報告

主持人：陳克健

協同研究人員：梁德昭、鄭守祥

中華民國七十一年七月

一、導論

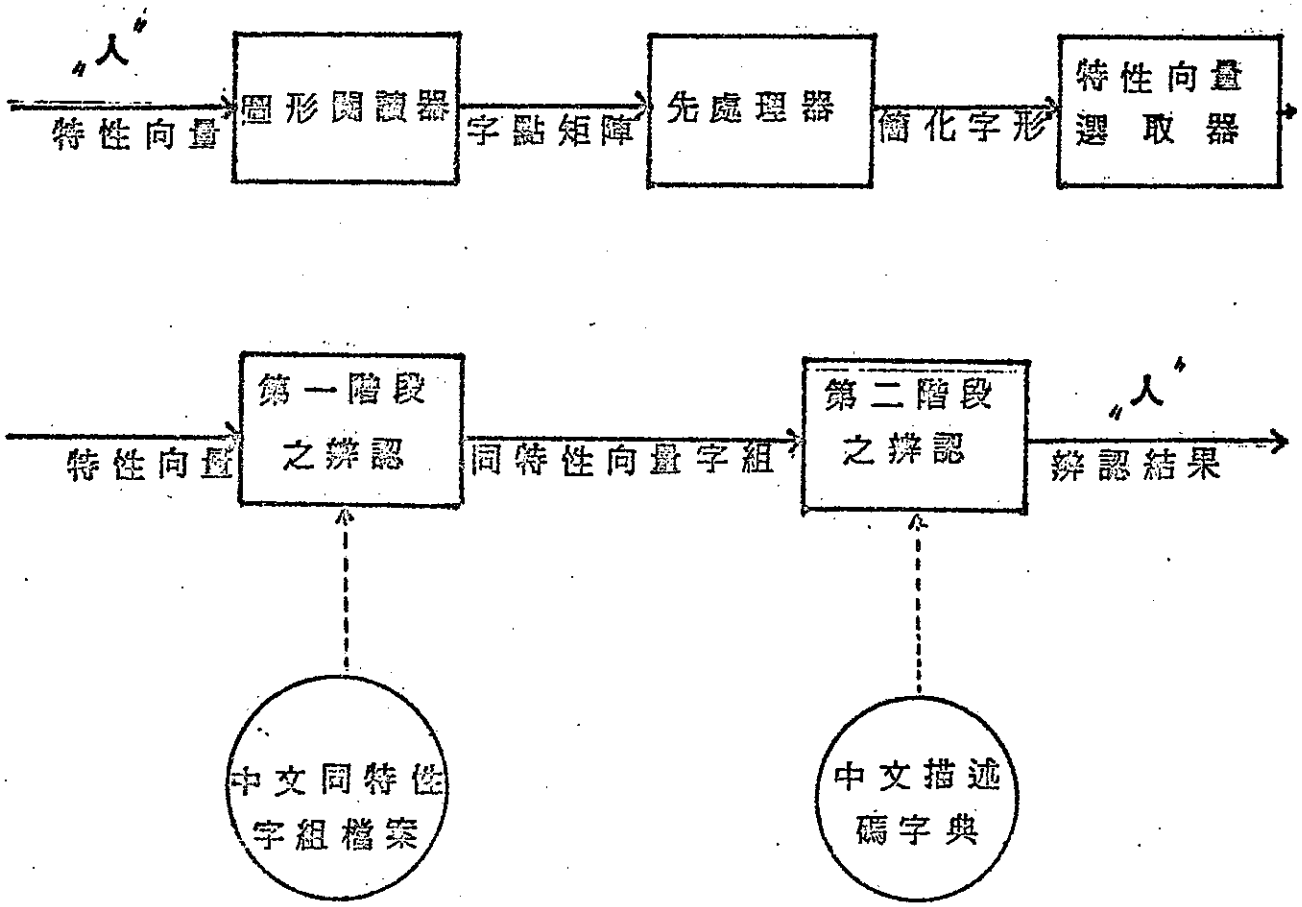
中文資訊的輸入是中文電腦處理上的一個瓶頸[1,2]輸入中文的方式大致上可以分爲三種，第一種是鍵盤輸入，第二種是語音輸入，第三種是圖形輸入。每一種輸入的方式都有其困難的地方。鍵盤輸入方面的研究最多[2,3,4,5,6,7]成就也不少，不過至今皆未有一個完美的方法。語音及圖形輸入方式，兩者皆牽涉及圖形的辨認[1,2]其中比較有希望且具有挑戰性的要屬於圖形輸入。圖形輸入的方式是利用圖形閱讀機，輸入中文圖形，經過圖形辨認程式，決定輸入的中文圖形是什麼字。在這一方面日本已經有十五年以上的研究[8]且有相當的成果，對標準的印刷字體，可以每秒鐘輸入100個字，比人爲鍵盤輸入快一百倍以上[9]且正確的比例高達百分之九十九以上，我實在應該急起直追。然日人用的中文字不過兩千字左右，而我們平日所用的字，根據教育部訂中文常用字[10]將近有五千字。中文圖形輸入的工作更形艱鉅。

中文圖形輸入法困難的地方不只是字數衆多，而且(1)中文字的結構複雜(2)相類似的字很多(3)有各式各樣不同的字體、字形及大小(4)印刷字體的影像不一定很清晰。爲了要克服上述的困難，同時兼顧系統的辨認速度。我們的系統將採用兩個階段的辨認方式。在第一個階段，我們先把輸入的字選出一些特性[11]這些特性包含總共筆畫的長度和週邊的形狀，利用這些選出的特性，可以找出一組字，它們都具有這些相同的特性，希望正確的字也在這一組字裏面。如果這一組字只有二十個字，則第一階段的辨認，已經把搜尋的對象從五千字減低到二百五十分之一。在第二個階段，我們將利用一個預先建好的字典，這個字典中存有五千中文字的筆畫描寫語法，把從第一階段裏挑出的字的描寫語法[11, 12]和輸入字的圖形比較，找到一個最相似的字當做輸入的字。

二、系統的規畫

此一中文字辨認系統的規畫，分爲兩部分，㈠中文描述碼字典及同特性中文字組據實的建立。㈡辨認系統的建立。同特性中文字組及中文描述碼字典分別辨認的第一及第二階段提供辨認所需的資料。辨認系統的辨認步驟順序爲(1)先處理 (Preprossing)[13,14,15]，先處理所作的工作是，把所要辨認的字從光學讀圖機，讀入字的點矩陣，這個點矩陣再經過筆畫寬度的收縮，及決定筆畫類別，得到筆畫的簡單形式，(2)特性的選取[16,17,18,19,20,21]:從先處理得

到的結果，可以求得總共筆畫的長度及週邊的形狀。我們把筆畫的長度分爲十類，週邊的形狀也分爲十類，共有一千種不同的筆畫長，上週邊及下週邊形狀組合特性向量(feature vector)。有相同特性的字，就被分在同一組，希望每一組不超過二十個字。同特性中文字組的建立就是根據這些特性來決定同一組的字。(3)第一階段的辨認：從輸入字的特性向量，在同特性中文字組資料表找到有相同特性向量的字組。(4)第二階段的辨認[8,19,22]:把第一階段裏得到的同特性字組中的字與輸入字作圖形對比(Pattern Matching)。圖形對比的方式是從中文描述碼字典中找出同特性字組裏字的筆畫描述語碼，這個描述語碼經過分析以後，可以知道字的每一筆、每一畫的長度及相對位置，從這個筆畫的資訊和輸入字的圖形做一個對比，給予一個相似程度的分數。其中相似程度最高的一個字即爲辨認的字，整個辨認系統可以用下面的圖形表示。



擬定之工作程序如下，第一期六個月完成：1.組字語法的訂定。2.字根表的建立及字根的描述碼。3.五千常用中文字的描述碼。第二期一年完成：1.同特性字檔案的建立。2.中文字典的建立。3.辨認系統的完成。

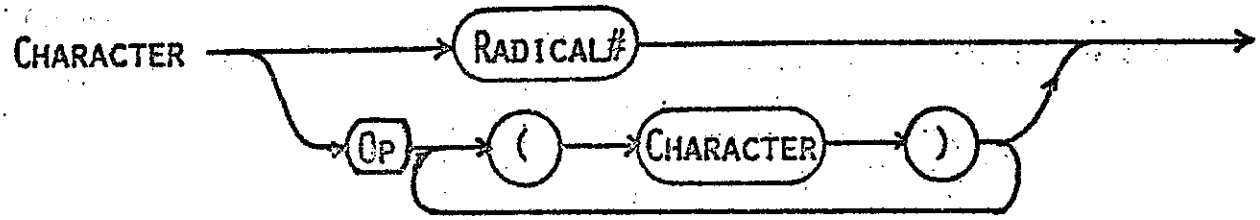
三、中文組字語法的分析 [11, 12]

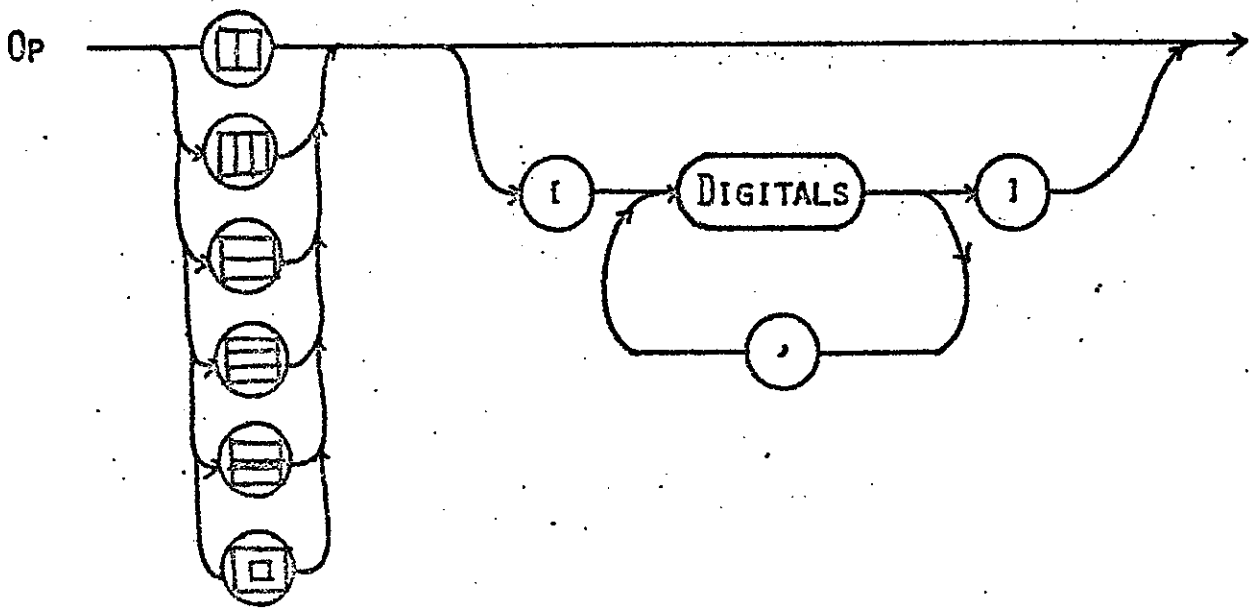
中文字的組成有一定的形式，筆畫與筆畫之間的相對位置關係決定了一個中文字，所以告訴計算機一個中文字的方法，就是如何把每一個中文字的筆畫大小、及它們的相對位置及連接關係用一種計算機可以分析的語法去描述這些關係，計算機可以利用字的描述語碼，辨認輸入的中文字。所謂一種計算機可以分析的語言不外乎是Context-free 語言或者是regular 語言，然而regular 語言太簡單了，無法描述一個二維平面上的中文字，最佳的選擇應該就是一種簡單形式的Context-free 語言。

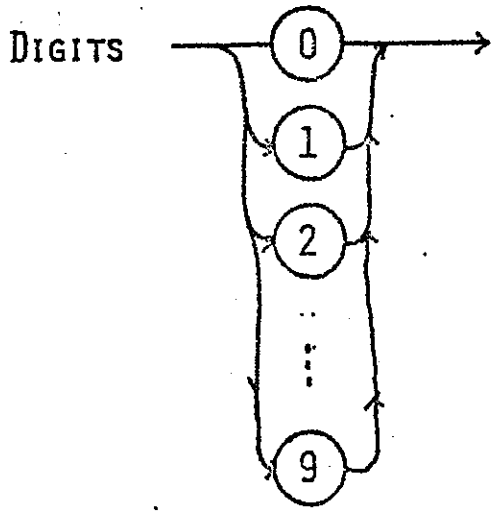
我們所定義出來的中文描述方式是這樣的，中文字由一些字根經過一些運算子的安排位置而獲得，不是直接由筆畫組合中文字[1, 23]。字根才是由一些基本筆畫所構成，所有的中文字根不到一千個，利用這些不到一千個的中文字根，可以組成的任何中文字。字根是由不到二十種的不同基本筆畫所構成，加上筆畫的長度，就可以利用運算子組成中文字根。這種層次式的結構，不僅簡化了組字的語碼，而且相對的減少了很多的計算機處理時間，及儲存空間。

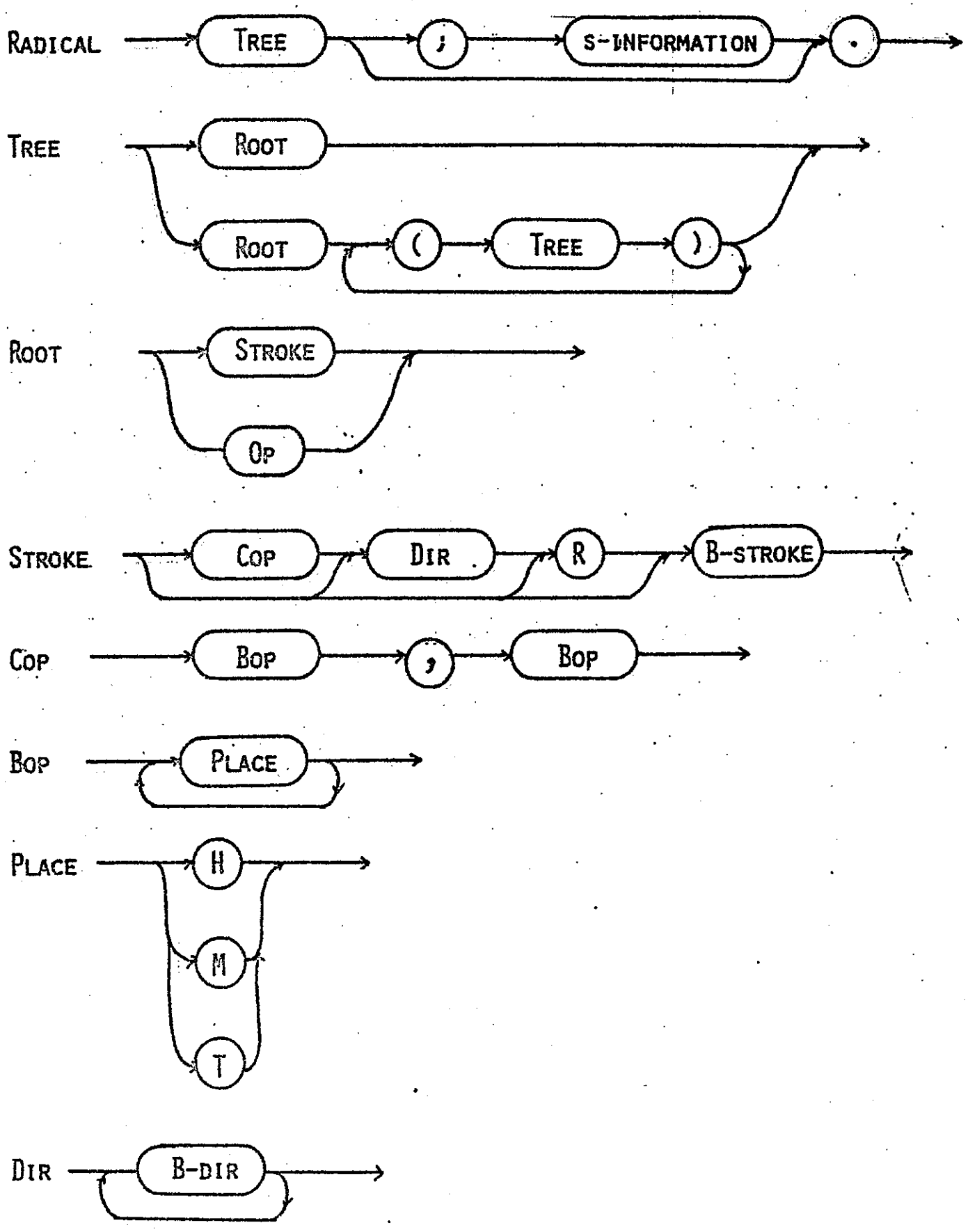
以下是我們所採用的中文字組字語法[12]。

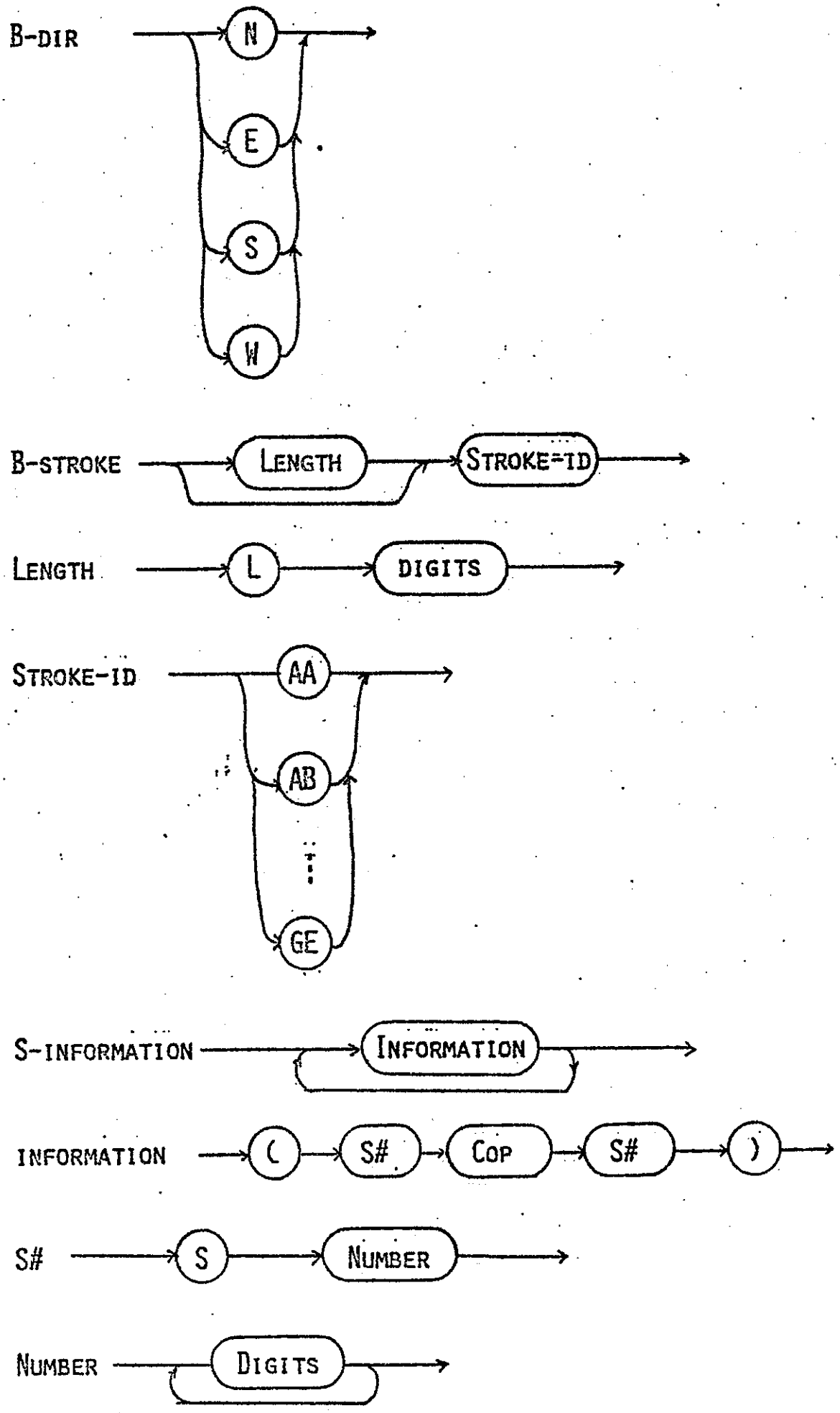
GRAMMAR RULES FOR DESCRIBING THE RADICALS OR THE CHARACTERS :








RADICAL
TREE
;
S-INFORMATION
.
TREE
ROOT
ROOT
(
TREE
)
ROOT
STROKE
OP
STROKE
COP
DIR
R
B-STROKE
COP
BOP
,
BOP
BOP
PLACE
PLACE
H
M
T
DIR
B-DIR


B-DIR
N
E
S
W
B-STROKE
LENGTH
STROKE-ID
LENGTH
L
DIGITS
STROKE-ID
AA
AB
GE
S-INFORMATION
INFORMATION
INFORMATION
(
S#
COP
S#
)
S#
S
NUMBER
NUMBER
DIGITS

OPERATORS :

01 : 

02 : 

03 : 

04 : 

05 : 

06 : 

01(日)(月) = 明

02(亻)(貝)(小) = 側

03(日)(十) = 早

04(立)(口)(羊) = 韋

05(立)(日) = 音

06(口)(丨) = 中

MODIFIED OPERATORS :

01 [6] (奠)(阝) = 鄭

02 [4,4](序)(亻)(呆) = 衍保

03 [4](一)(立) = 立

04 [2,5](亠)(口)(冏) = 高

05 [5,6](人)(良) = 食

06 [5,1,3,3](門)(口) = 問

06 [0,7,6,0](大)(丶) = 犬

06 [0,2,2,2] (凵)(乂) = 凶

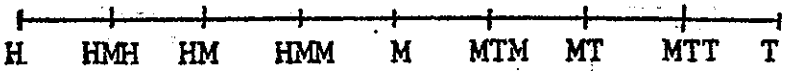
06 [3,3,2,3](勹)(口) = 句

06 [0,1,4,0](辶)(斤) = 近

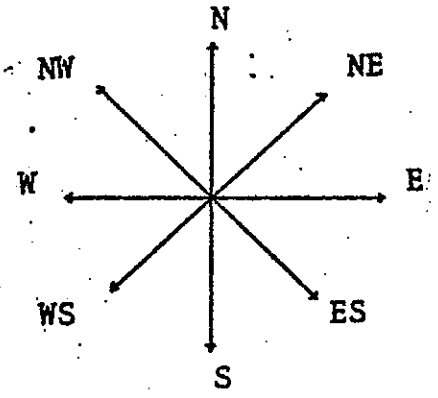
以下對描述語法做一個簡單的說明。

Place: H,M,T: 分表筆劃之頂端、中端、尾端。

如下圖所示：



B-Dir: N,E,S,W: 分表北、東、南、西。如下圖所示。



Stroke一.Id: 筆劃名稱一律以三個字母來代表，第一個字母表長度 A,B,C,D,E （長、中長、中、中短、短），第二個字母表筆劃之種類 A—G（橫、豎、點、撇、捺、鈎、挑），第三個字母表筆劃種類下之特性，如前述如 AAB, AAA,... （表長橫的第一種;長橫的第二種……）。

前面所列之圖表，在此做一比較整體性的說明：

Character （字）：可以是字根（Radical）或者是一些字根經分離運算子作用堆砌而成。

Discrete Op（分離運算子）：

橫向關係：

：如「從」是由「彳」和「𠂤」組成。

 ：如「側」是由「亻」和「貝」和「刂」組成。

縱向關係：

：如「昌」是由「日」和「曰」組成。

 ：如「章」是由「立」和「口」和「十」組成。

包含關係：

：如「犬」是由「大」包含「丶」而成。

重疊關係：

 （左右重疊）和 （上下重疊）：一些字讓其字根間有稍許的接觸會顯得更爲好看。

Digit（數字）：字根組字時，表示字根應佔此方塊比重，分爲十等分。

Radical（字根）：是由樹狀圖（Tree）以及可加上補充說明訊息（S--Information）而成。

Tree（樹狀圖）：可以是筆劃（Stroke），也可是一些筆劃之相互連接關係或具有分離關係之樹狀圖。

Stroke（筆畫）：可以是基本筆劃，或是具有旋轉(Rotate)抑若鏡射(Mirror)之基本筆劃。

Cop（連接關係）：表示兩筆劃之相互相接關係。

至於 Place（位置）、Dir（方向）、Stroke—Id（筆劃名稱）已在前面說明過。

本文之文法表示式是採由下往上的方式，由筆劃組字根，再由字根組字循次漸進。這是一種非常自然的表示方法。

為了方便說明起見，我們以例子做說明：

例：鳥

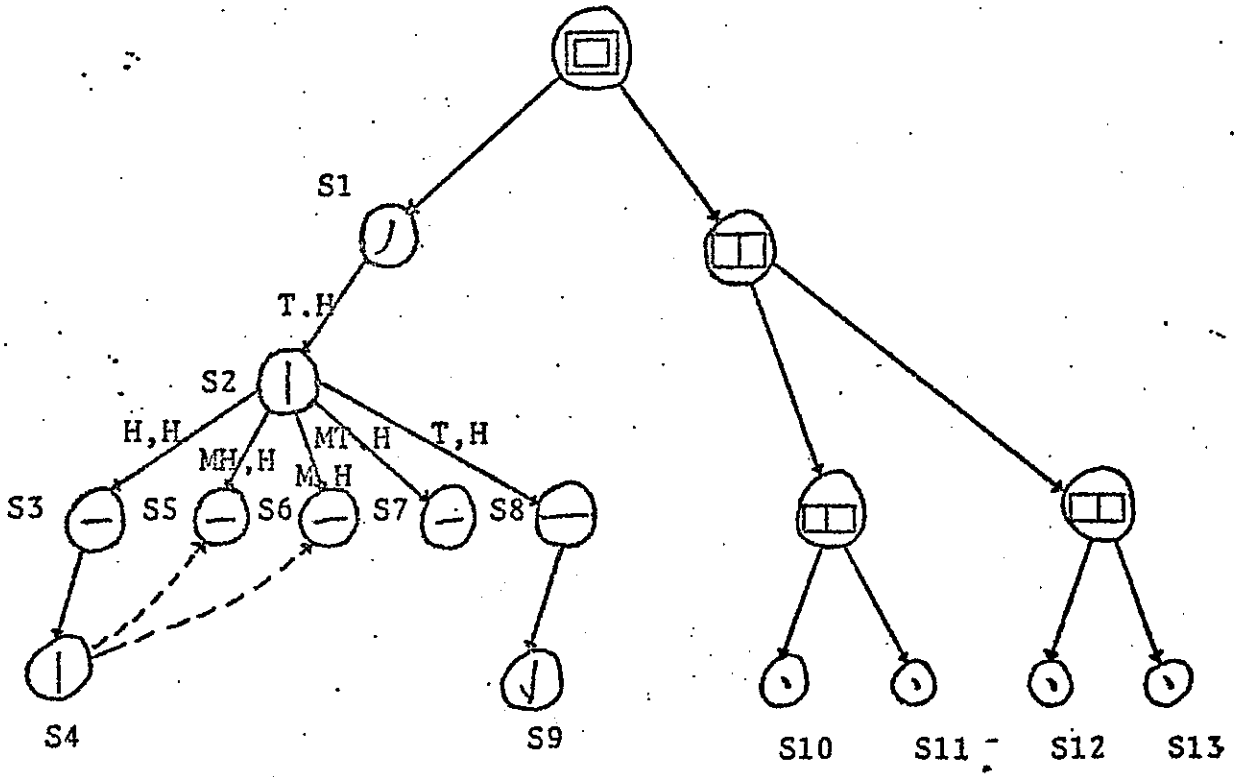


圖 3.5

筆劃名稱以中文表示，如圖3.5 所示。

文法表示式：

□ [7,1,0,2](短撇(T,H 中豎(H,H 短橫(T,H 短豎))(MH,H短橫)(M,H 短橫)(MT,H短橫)(T,H 長橫(T,H 中鉤)))) (□□ (□□(中點)(中點)) (□□ (中點)(中點))) : (S4 M,T S5) (S4 T,T S6).

就組成字根而言，大致可分成連結與分離兩部份。例如〝鳥〞就是由"鳥"與"灬"所組成的，而少部份，如〝黃〞則由"艹"與"𠂤"所組成，二部份均是連結的，也有僅分離部份的如"彡"。

現在我們就來說明如何技巧的來寫文法表示式，而不致於混淆不清。

還是以〝鳥〞做爲例子：

首先由根節(root node)開始，根節是□(7,1,0,2)先寫好，再看〝回〞有二個兒子，我們就在後面先寫兩對左、右括弧：□(7,1,0,2)(　　)(　　)，再就左邊那對括弧寫起，是 S1(Stroke1) 短撇，先寫好，再看有一個兒子，再寫一對括弧：

□(7,1,0,2)(短撇(　　))(　　)，

這時候我們發現 S1與S2有連結，因此要賦予連結之訊息，再將 S2之名稱給寫上去：

□(7,1,0,2)(短撇(T,H中豎　))(　　)，

現在我們發現S2有五個兒子，就要寫五對括弧了。

□(7,1,0,2)(短撇(T,H中豎(　　)(　　)(　　)(　　)(　　)))(　　).

還是從最左邊之括弧寫起，S2與 S3有連結：

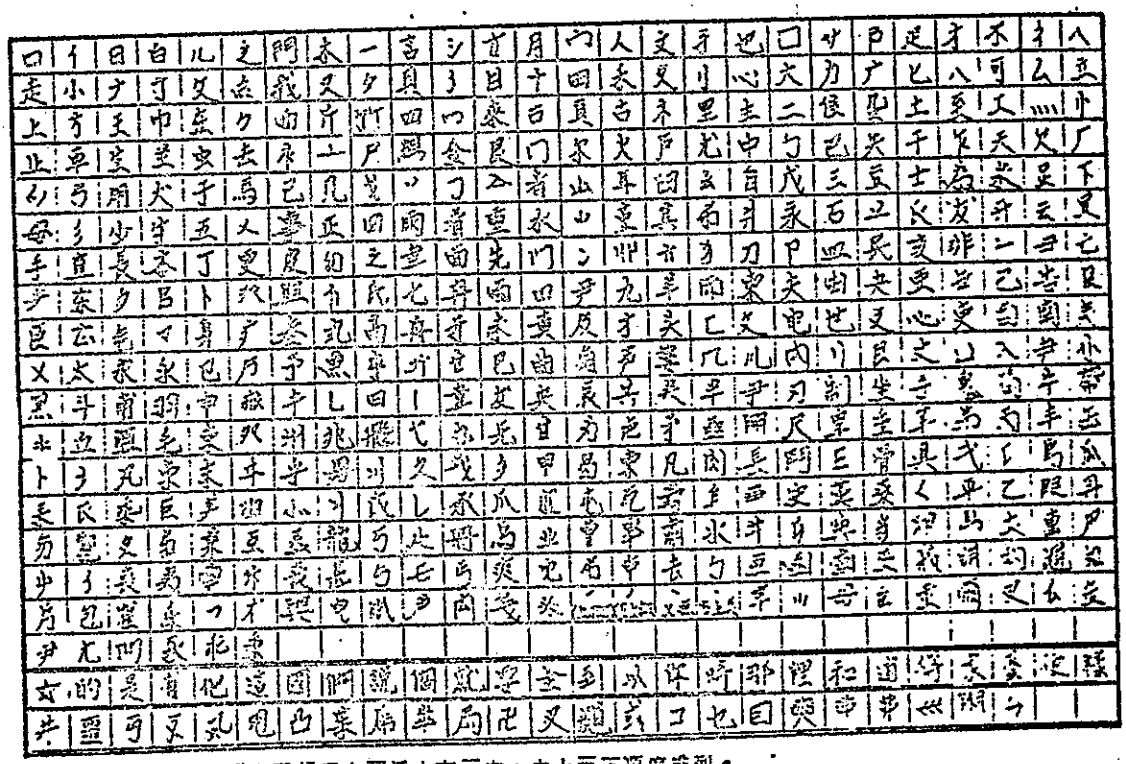
□(7,1,0,2)(短撇(T,H中豎(H,H短橫(　　)(　　)(　　)(　　)))(　　).

現在我們發現S3只有一個兒子，又得寫一對括弧了。

□(7,1,0,2)(短撇(T,H中豎(H,H短橫(　　))(　　)(　　)(　　)))(　　).

四、字根表的建立

字根表的組成，是參照交通大學林樹先生的中文基本字根表 [23,24]，補上在組字過程中發現不足的部分。林樹先生的中文字根表，共含字根 443 字，加上25 個常用字及23 個罕用字根，共計 496 個，如下表。



說明：1. 本表依字根出現頻率之高低由左而右，由上而下逐序排列。
2. 古碼的留"常"用字，"共"爲罕用字根。
3. 本表計收字根 443 個，酌留常用字 25 個，罕用字根 23 個，共計 496 個。

唯此一字根表尚不能滿足需要，爲了使字形的描述更精確，所以我們加入了一些新的中文字根。

五、五千常用中文字的描述

常用中文字表係依據教育部公布「常用國字標準字體表」完成，計收四千八百零七字。以後中文辨認系統的完成，並不一定要根據此一中文字表，系統上的中文字表可以擴張。初步的辨認系統是以此四千八百多字爲辨認對象，希望能在常用字的辨認上有相當好的結果。以後系統不修改，只要改變字碼檔及字群分類檔即可擴充中文字辨認的範圍。

六、參考資料

1. Proceedings of the First International Symposium on Computers and Chinese Input/Output Systems , Taiwan , Aug. 1973.
2. K.S.Fu and S.Y.Lu " Applicability of Pattern Handling Methods to Chinese Language Processing " Tech. Report , Academia Sinica, 1980 .
3. 朱邦復「倉頡一號」中文輸入方式，台北，台灣，1979.
4. "SINOTERM , an Efficient System for Composing , Displaying and Printing Chinese Characters ", Publication of Transtech Inter. Co., MA , U.S.A. ,1980 .
5. 胡立人、張源渭、黃克東，中文「三角編號法」訓練手册，系統出版社，1979.
6. " 中文簡易輸入法 " 神通實業公司，台北，台灣。
7. 陳舜齊," 首次尾三碼法" 康大資訊公司，台北，台灣。
8. K.Mori & I Masuda ," Advances in Recorgnition of Chinese Characters ", Proceedings of Pattern Recorgnition and Image Processing , 1980 IEEE .
9. S.Hirai and K.Sakai , "Development of a High Performance Chinese Character Reader ", Proceeding of Pattern Recognition and Image Processing , 1980 IEEE .
10. 常用國字標準字體表，教育部印，民國六十八年。
11. R.C.Gonzalez & M.G.Thomason , Syntactic Pattern Recognition : An Introduction , 儒林 , 1978 .
12. K.Y.Cheng and K.J.Chen , " ACCFONT - An Automatic Chinese Character Generating System ", Computer Science and Technology Conference , Bosten ,1982 .
13. E.R.Davies and A.P.Plummer , " Thinning Algorithms : a Critique and a New Methodology ", Pattern Recognition , Vol.14 pp.53-63 , 1981 .
14. W.D.Wasson , " A Pre-processor for Handprinted Character Recognition ", Proceedings of Pattern Recognition and Image Processing , 1980 , IEEE .
15. H.Ogawa and K.Taniguchi , " Preprocessing for Chinese Character Recognition and Global Classification of Handwritten Chinese Characters ", Pattern Recognition Vol.11 pp.1-7 ,1979 .

16. E.Yamamoto , N.Fujii ,T.Fujita ,C.Ito ,and J.Tanahashi , " Handwritten Kanji Character Recognition Using the Feature Extracted from Mutiple Standpoints ", Proceedings of Pattern Recognition and Image Processing ,1981 IEEE .
17. S.I.Hanaki and T.Yamazaki , " On-line Recognition of Handprinted Kanji Characters ", Pattern Recognition Vol.12 ,pp.421-429 ,1980 .
18. K.Yamamoto , "Recognition of Handprinted Characters by Convex and Concave Features ", Proceedings of Pattern Recognition and Image Processing ,1980 IEEE .
19. B.Duerr , W.Haettich, H.Tropf and G.Winkler , "A Combination of Statistical and Syntactical Pattern Recognition Applied to Classification of Unconstrained Handwritten Numerals ", Pattern Recognition Vol.12 pp.189-199 , 1980 .
20. M.Lai and C.Y.Suen , " Automatic Recognition of Characters by Fourier Descriptors and Boundary Line Encodings ", Pattern Recognition Vol.14 pp.383-393 , 1981 .
21. K.Yamamoto and S.Mori , " Recognition of Handprinted Characters by an Outermost Point Method ", Pattern Recognition Vol.12 , pp.229-236 ,1980.
22. D.J.Burr , "Design a Handwriting Reader ", Proceedings of Pattern Recognition and Image Processing , 1980 IEEE .
23. 謝清俊、黃永文、林樹," 中文字根之分析" 交大學刊,Vol IV , No.1, 1973.
24. 林樹," 中文電腦基本用字研究" 交通大學工學院，科技研究報告，CC-601 號,1972.