On the Formalization of Glyph In the Chinese Language

C. C. HSIEH
C. T. CHANG
Jack k. T. HUANG

A contribution to the
AFII meeting at Kyoto

Feb. 6, 1990

Content

On the Formalization of Glyph in the Chinese Language

1. Introduction

Although the word "glyph" is a newly created word, the concept of a "glyph" has been used in China for thousands of years. Traditional morphological studies of the Chinese language include numerous documents that study the "forms" of characters. The knowledge about the "forms" includes the originality, the expressional structure, the content structural phenomena, the structual law, and the functional rules of Chinese characters. The 2000 years old morphological work, LIU-SUE (六書), still provides a keynote guide-line for studying the "forms" of character. This knowledge includes how a character was formed, how a character was composed of by its components/roots, how a character appears, what is the original meaning of a character, etc. This information helps people to understand the "forms" of a character, but it doesn't help computers. Therefore, we need to formalize this related knowledge in order to let the computer handle characters, glyphs, and fonts more elegently than before.

In this paper, the authors are trying to formalize the related phenomena of glyphs for Chinese characters,according to the characteristics of the Chinese language. There will be no invention concerning language, but the formalized system will provide the computer with logical data structures and procedures so that it can carry out the basic functions needed to create glyphs and characters in Chinese language.

From now on, for simplicity's sake, the paper will not mention to any morphological works from Chinese linguistics. This assumption does not mean that the work in this paper ignored the traditional Chinese linguistical aspects. On the contrary, all the proposed mechanisms are based on the morphological nature of the Chinese language, although it will not be explicitly cited. For those who are

not familiar with the morphological characteristics, or for those who can not read Chinese, the book "Formalization of Natural Languages" by P. Kummel (ISBN 3-540-08271-9) is highly recommended as a reference.

2. Some Characteristics of Characters and Glyphs

A character, from modern linguistic view point, is a meaning bearing unit.In the Chinese language,its visualform(形),pronounciation(音),and meaning(義) can be considered as the content/attributes of a character. When people are using characters in language,different degrees of emphasis on its form, pronounciation, or meaning are made according to the environment or situation in which they are applied. For example,when people talk, no form will be used; but for identifying a person, a place, an object,or an affair in a document, the forms play a key-role. Therefore, it is not justified to say that because of some applications that we must differentiate two glyphs by assinging them as two charcters. The characters should be differentiated by their meaning,as we stated in the first sentence of this paragraph.

A character may have more than one form, or glyph (字形). For everyday use, most characters have just one form, but for some frequently used characters they have several forms/glyphs. People are used to this phenomena,and usually they select the glyph that they think properly represents that character. In CCCII, other glyphs of a character are called variants and they are systematically coded in variant layers. Other standards of character code do not have the structure for coding glyphs. Therefore, they are actually a mixture of characters and glyphs coded in the same structure and space. But, how to differentiate a character and its variants, and how variants are selected are not clearly defined in any of the CCCII publications.

A glyph may have some degree of freedom to change its appearance. In our last meeting, we called this phenomenon micro-differences.Where a font is concerned,the visual forms of a character form a matrix of glyphs and fonts, as shown in Figure 1.

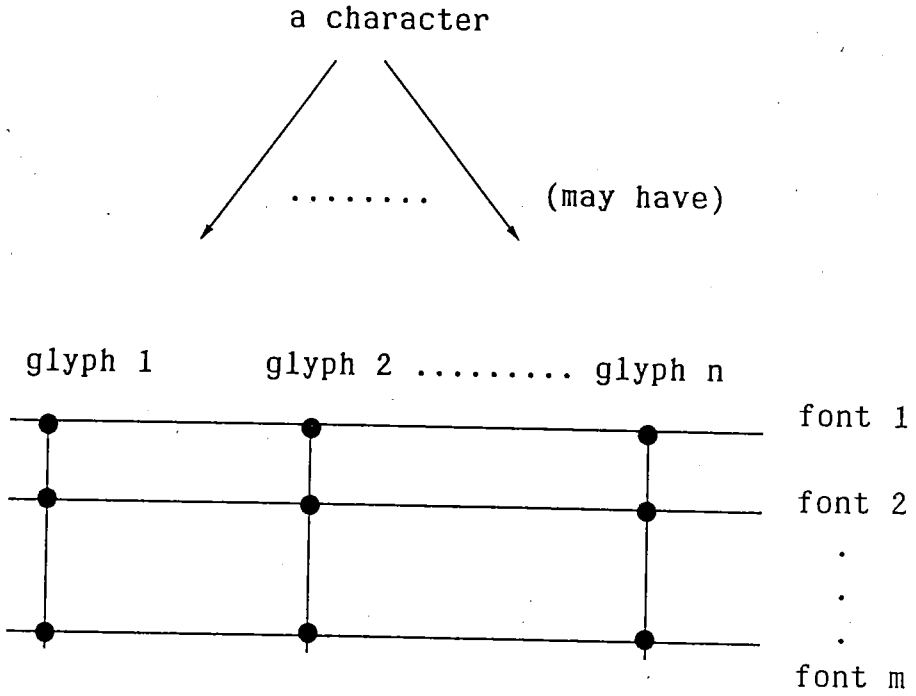


Figure 1, A character may have m x n different visual appearances, where n is the number of glyphs it has, and m is the number of different font designs. Every node of the matrix in the figure represents one visual form of the character.

All visual forms of a character are composed of graphemes. In some papers, they are also called "roots", "components", and even sometimes incorrectly called "radicals". In this paper, we will use the term components (字根).** Components form a finite set. It is a closed set of approxiamately 1200 finite elements. The rules of composition of a character is very complicated from a traditional linguistical viewpoint, but when only the visual expression is concerned, the number of basic components can be reduced to about 300 to 600 . A recommend set of components, the Chiao-tung set, is shown in Figure 2. There are 446 components which can be used to generate more than 48700 characters. The composition rules of the Chiao-tung set is simple. There are only three operations, horizontal combine, vertical combine and containing. They are illustrated in Figure 3.

字根依出現頻率排列，置入CCCII之第85面，自第5段至第9段，共446字

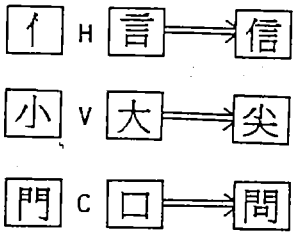
752521 口	752522 亻	752523 日	752524 白	752525 儿	752526 辶	752527 門	752528 木	752529 一	75252A 言	75252B 氵	75252C 女
75252D 月	75252E 宀	75252F 人	752530 文	752531 子	752532 也	752533 口	752534 艹	752535 阝	752536 疋	752537 扌	752538 不
752539 彳	75253A 𠂉	75253B 走	75253C 小	75253D 𠂇	75253E 寸	75253F 攵	752540 糸	752541 我	752542 又	752543 夕	752544 貝
752545 了	752546 目	752547 十	752548 田	752549 禾	75254A 夂	75254B 刂	75254C 心	75254D 大	75254E 力	75254F 广	752550 匕
752551 八	752552 可	752553 厶	752554 立	752555 上	752556 方	752557 王	752558 巾	752559 在	75255A 𠂆	75255B 西	75255C 斤
75255D 竹	75255E 皿	75255F 冖	752560 來	752561 𠀉	752562 頁	752563 古	752564 礻	752565 里	752566 主	752567 二	752568 隹
752569 殳	75256A 土	75256B 至	75256C 工	75256D 灬	75256E 卜	75256F 止	752570 車	752571 生	752572 羊	752573 虫	752574 去
752575 礻	752576 亠	752577 尸	752578 與	752579 金	75257A 艮	75257B 冂	75257C 尔	75257D 火	75257E 戶	752621 尤	752622 中
752623 勹	752624 巴	752625 矢	752626 干	752627 乍	752628 天	752629 欠	75262A 厂	75262B 巛	75262C 弓	75262D 用	75262E 犬
75262F 于	752630 馬	752631 已	752632 几	752633 戈	752634 丷	752635 刁	752636 𠆢	752637 者	752638 山	752639 耳	75263A 臼
75263B 幺	75263C 自	75263D 戊	75263E 三	75263F 豆	752640 士	752641 為	752642 米	752643 足	752644 下	752645 母	752646 彡
752647 少	752648 牛	752649 五	75264A 乂	75264B 事	75264C 正	75264D 罒	75264E 雨	75264F 看	752650 重	752651 水	752652 灬
752653 堇	752654 其	752655 弔	752656 片	752657 永	752658 石	752659 𠂤	75265A 仪	75265B 发	75265C 开	75265D 云	75265E 且
75265F 手	752660 直	752661 長	752662 本	752663 丁	752664 更	752665 皮	752666 勿	752667 之	752668 聿	752669 面	75266A 先
75266B 门	75266C 冫	75266D 屮	75266E 亠	75266F 犭	752670 刀	752671 卩	752672 血	752673 民	752674 亥	752675 非	752676 𠂉
752677 彐	752678 亡	752679 夬	75267A 东	75267B 夕	75267C 吕	75267D 卜	75267E 歹	752721 幽	752722 仆	752723 氏	752724 七
752725 冉	752726 兩	752727 凹	752728 尹	752729 九	75272A 羊	75272B 而	75272C 東	75272D 夫	75272E 由	75272F 央	752730 亜
752731 無	752732 己	752733 告	752734 艮	752735 臣	752736 𠂢	752737 气	752738 マ	752739 身	75273A 疒	75273B 垄	75273C 丸
75273D 高	75273E 舟	75273F 牙	752740 未	752741 黃	752742 及	752743 才	752744 头	752745 匚	752746 夂	752747 电	752748 世
752749 叉	75274A 必	75274B 吏	75274C 酉	75274D 南	75274E 尧	75274F 乂	752750 太	752751 求	752752 豕	752753 巳	752754 乃

Figure 2: The Chiao-tung Component set according to the frequency of usage order(交通大學字根表, 1972)

字根依出現頻率排列，置入CCCII之第85面，自第5段至第9段，共446字

752755 予	752756 魚	752757 亦	752758 廾	752759 廿	75275A 巳	75275B 曲	75275C 角	75275D 声	75275E 婁	75275F 八	752760 儿
752761 内	752762 刂	752763 艮	752764 丈	752765 凵	752766 入	752767 丰	752768 小	752769 黑	75276A 斗	75276B 甫	75276C 羽
75276D 申	75276E 衣	75276F 丰	752770 乚	752771 曰	752772 丨	752773 韋	752774 犮	752775 央	752776 辰	752777 共	752778 关
752779 半	75277A 尹	75277B 刃	75277C 制	75277D 坐	75277E 子	752821 鬼	752822 咼	752823 片	752824 带	752825 丰	752826 丘
752827 亞	752828 毛	752829 束	75282A 奴	75282B 州	75282C 兆	75282D 飛	75282E 弋	75282F 办	752830 无	752831 甘	752832 为
752833 虍	752834 矛	752835 亟	752836 鼬	752837 尺	752838 卑	752839 垂	75283A 革	75283B 弗	75283C 丙	75283D 丰	75283E 缶
75283F 卜	752840 彐	752841 丸	752842 束	752843 末	752844 屮	752845 乎	752846 毋	752847 川	752848 久	752849 戋	75284A 乡
75284B 甲	75284C 曷	75284D 柬	75284E 凡	75284F 肉	752850 县	752851 門	752852 巨	752853 骨	752854 具	752855 弋	752856 匕
752857 鸟	752858 瓜	752859 耒	75285A 仄	75285B 乘	75285C 巨	75285D 羊	75285E 由	75285F 小	752860 刂	752861 戊	752862 乚
752863 承	752864 爪	752865 朮	752866 屯	752867 瓦	752868 壽	752869 乑	75286A 事	75286B 史	75286C 冥	75286D 祭	75286E く
75286F 平	752870 乙	752871 段	752872 丹	752873 劣	752874 麗	752875 夂	752876 韦	752877 兼	752878 互	752879 丞	75287A 龍
75287B 丐	75287C 丄	75287D 册	75287E 烏	752921 业	752922 曹	752923 率	752924 肅	752925 氷	752926 井	752927 歹	752928 典
752929 豸	75292A 卵	75292B 冓	75292C 土	75292D 重	75292E 尸	75292F 少	752930 厂	752931 喪	752932 夷	752933 牽	752934 斥
752935 衰	752936 甚	752937 亏	752938 七	752939 丐	75293A 爽	75293B 尤	75293C 弔	75293D 申	75293E 壬	75293F 勹	752940 互
752941 函	752942 齒	752943 巫	752944 鳳	752945 鬲	752946 两	752947 黽	752948 匆	752949 肩	75294A 包	75294B 雀	75294C 乐
75294D 一	75294E 疒	75294F 與	752950 曳	752951 毗	752952 声	752953 内	752954 幾	752955 氺	752956 丿	752957	752958 丶
752959	75295A 羊	75295B 少	75295C 毋	75295D 臣	75295E 亚	75295F 爾	752960 叉	752961 匚	752962 歪	752963 尸	752964 尢
752965 冂	752966 豕	752967 乖	752968 隶								

Figure 2 (continued)



, horizontal combination

, vertical combination

, one contains the other

Figure 3, Examples of the three composition rules that generate characters from components

All components are composed of a finite set of strokes. The number of strokes used in some high quality vector character generators is around 40. An example of a basic stroke set is shown in Figure 4, and another example of constructing a character from components and then from strokes is shown in Figure 5.

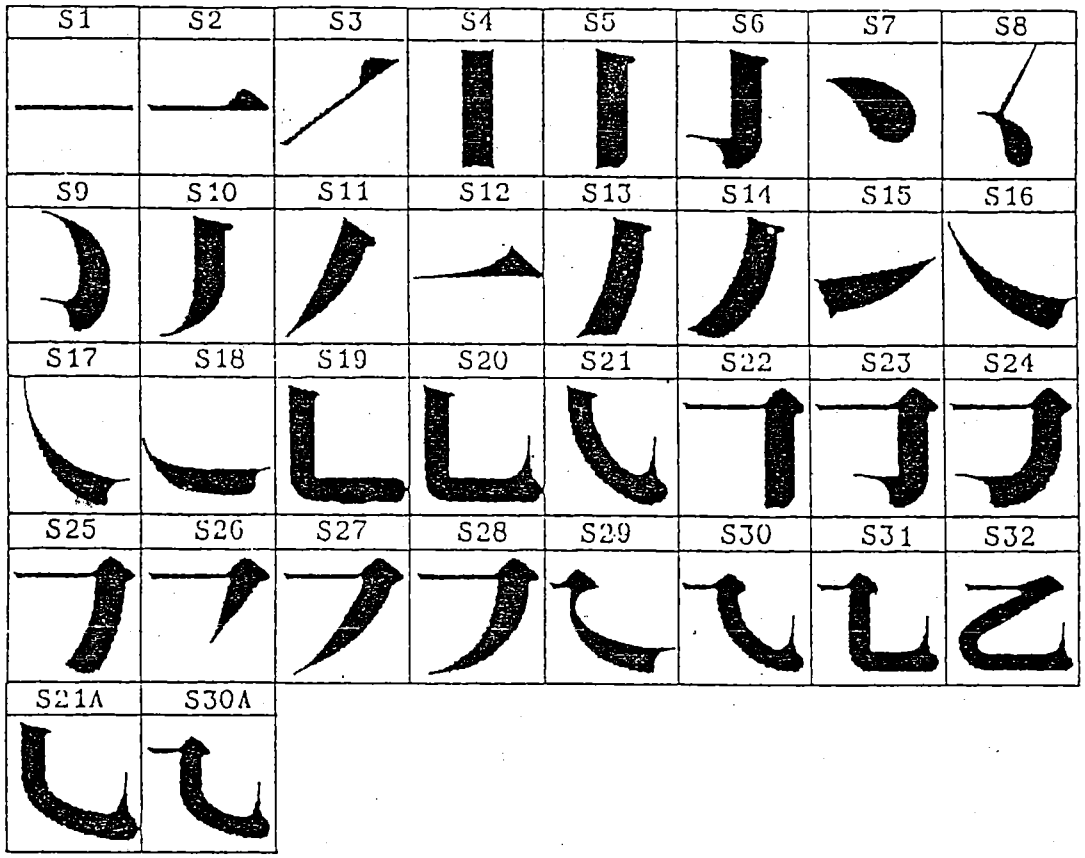


Figure 4: An example of Basic Stroke set (1984)

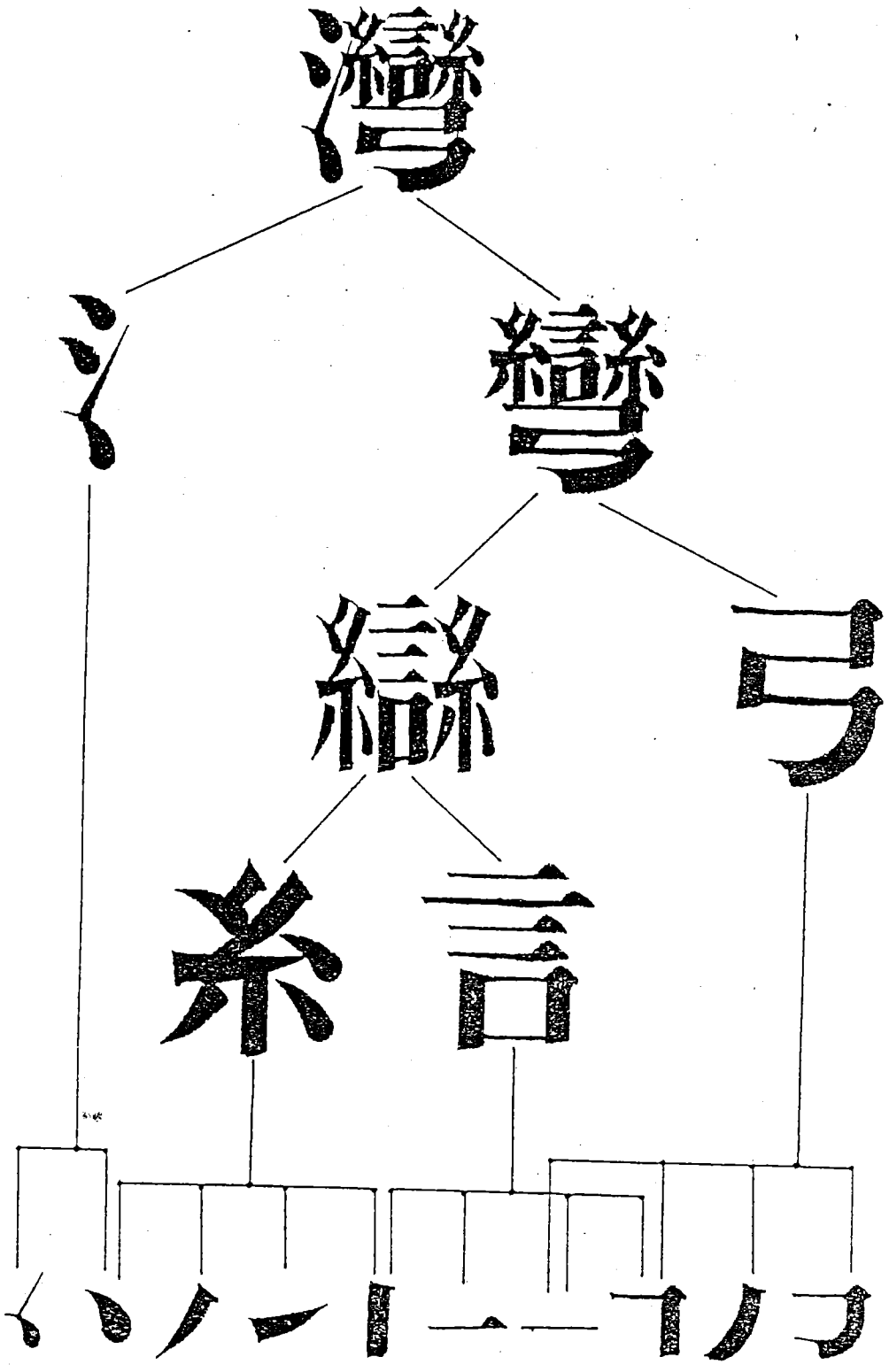


Figure 5: An example that shows the construction of character 灣 from its components and then in terms of basic strokes.

According to the above illustration, if we have the vector forms of a set of strokes and a procedure to put the strokes together, then we will have the vector forms of all components needed to generate the visual expressions for characters. Therefore, we can give each glyph a formal definition accordingly in the next section.

3. A formal definition of glyphs

Let G be a visual appearance of a glyph, then G can be defined as follows:

$$G = R1\ (P1, S1) + R2\ (P2, S2) + \ldots + Rn\ (Pn, Sn) \ldots (1)$$

Where R stands for a Component of G, and P and S stand for the position of R and the size of R, respectively. The + is used to indicate the operation of ORing the image bit patterns of RS.

Equation (1) can be rewritten as in (2) for simplicity.

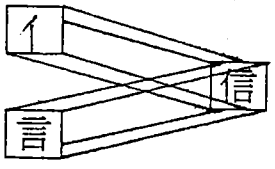
$$G = \sum R\ (P, S) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (2)$$

In the same way, each R can be expressed as a composition of strokes, T, as in (3)

$$R = \sum T\ (P, S) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

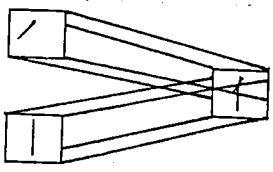
The above formalism is illustrated in the following examples.

[Example 1]:



信 = 亻 (P1, S1) + 言 (P2, S2)

[Example 2]:



亻 = 丿 (P1, S1) + 丨 (P2, S2)

In example 1 and example 2, there are no horizontal, vertical and containing operations because they are replaced by the vector operations of defining the position and size of each component and strokes, respectively.

According to the above structure, the code of a glyph can be assigned as follows:

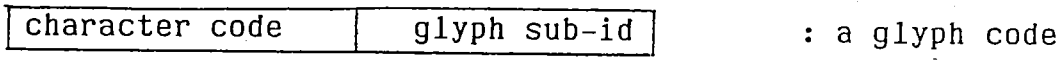


Figure 6, A structure of the glyph code base on its character code.

The code shown in Figure 6 can be used easily with every existing character code. But for a CJK union glyph set, the author's recommendation is to use the EACC or CCCII code, because they are the existing multi-lingual coding system and they have enough well-structured room for glyphs. If a new coding scheme is proposed, a in-depth comparison between the new one and the EACC code should be carefully evaluated.

4. A Study on the Variations of Character forms

Variations of character forms can be defined by transformation function on the strokes/components of a character. A study of the variations of the visual appearances of Chinese characters shows that there are three categories of transformation functions as listed below.

A. Stroke level transformation functions

On this level,a only a stroke change is made. All the changes made on this level will not cause a character change, such as change from one character to another, or a component change embedded in a character. All the functions on this level are further divided into 8 groups as follows.

A1(x): Change the relative position of a stroke, where x denotes the stroke that changes its position. In this group, the stroke-count of the character remains unchanged,the components of the character also remain unchanged. Examples of functions in this group are given in Table A1.

TABLE A1: Examples of A1 Functions

Function	Illustration at Component level	Related Characters
$A_1(\cdot)$	斥 → 斥 丸 → 丸 凡 → 凡 刃 → 刃 夕 → 夕 爿 → 爿	拆 坼 … 執 …. 恐 飆 … 仞 皴 (all character of radical 爿)
$A_1(-)$	彐 → 彐 丩 → 丩 幸 → 幸 片 → 片 爿 → 爿 鼎 → 鼎 毋 → 田	事 雪 尋 婦 …. 叫 報 倖 …. 版 … 將 … 虜
$A_1(\backslash)$	疒 → 疒 羊 → 羊 丰 → 丰 丂 → 丂 聿 → 聿 串 → 串 牛 → 丬 电 → 电	存 … 恙 害 … 愕 號 … 畫 … 數 … 告 … 奄 …
$A_1(/)$	匕 → 匕 匕 → 匕	匙

TABLE A1 (continued)

	艮→艮 女→女 厂→亠 隙→隙	(all characters have 艮) (all characters have 女) 碳
A_1 (丶)	文→文	俊凌愛夏後瓊.....

A2(X): A hook (鉤) is added at the end of a stroke. Stroke-count and component are not changed. Examples are given in TABLE-A2

TABLE A2: Examples of A2 Functions

Function	Illustration at Character/Component Level	Related Characters
A_2(1)	不→不 小→小 木→木 月→月	丕 少 尖 示... (all characters with radical 木) 胃
A_2(一)	也→也	他 弛 馳....
A_2(L)	七→七 电→电 匕→匕	奄 電.... 匕, 乖, ...

A3(x1,s2): Stroke substitution function, where stroke x1 is replaced by stroke x2. Stroke-count and component remain unchanged. Examples are given in Table-A3

TABLE A3: Examples of A3 Functions

Function	Illustration at Character/Component Level	Related Characters
A_3(丶, 丨) A_3(丶, 一) A_3(丶, 乀) A_3(丶, 丿)	母→母 言→言 刃→刃, 氷→氷 ；→氵	all characters with radical 言 伉…康… 冬寒…
A_3(一, 丿)	匕→匕 丟→丢, 刊→刊, 舌→舌, 反→反, 屯→屯, 壬→壬	尼老比化…
A_3(一, 丶)	今→今 二→亠 氏→氏 𠂉→𠂇	今倉… 佞 低抵… 監
A_3(丨, 丿)	孔→孔 刊→刊	迅訊蛔…
A_3(丿, 丶)	戶→户 ⺗→忄 辶→氵	(all characters have radical 户) 采爭… (飞→飞)
A_3(丿, 一)	壬→壬 夭→天 主→主 系→系	聖閏… 喬… 呈… 係孫…
A_3(丿, 丶)	夕→夕	敘务祭…

A4(x): One stroke is added to character. Stroke-count increases one, where x denotes the added stroke. Examples are given in TABLE A4.

TABLE A4: Examples of A4 Functions

Function	Illustration	Related Characters
$A_4(\cdot)$	者→耆 少→少	堵奢都…… 步涉頻歲……

A5(x): One stroke less. Stroke-count decreases one, where x denotes the eliminated stroke. Examples are given in Table A5.

TABLE A5: Examples of A5 Functions

Function	Illustration	Related Characters
$A_5(\cdot)$	寬→寬 广→厂 次→次	廈,廚…… 盜,羨……
$A_5(-)$	德→德	
$A_5(/)$	衆→眾	

A6(x1, x2): One stroke is substituted by two strokes, where x1 denotes the stroke to be substituted and x2 denotes the two strokes that replace x1. Stroke-count increases one. Examples are given in TABLE A6.

TABLE A6: Examples of A6 Functions

Function	Illustration	Related Characters
A6(工, 工)	匹 → 匹	區 匾 臥
A6(牛, 羊)	韋 → 韋	偉 圍 …… 傑, 降, 舞 ……
	牙 → 牙	呀 …
A6(厶, 厶)	爻 → 爻	矣, 牟, 駿 ……
A6(匕, 上)	匕 → 上	比, 衛, 即 ……
	十 → 土	切 …
	小 → 小	以 似 …
	巳 → 巳	印 ……
	氏 → 氏	氐 牴 ……
	民 → 民	氓 ……
	瓦 → 瓦	瓷 瓶 ……
	瓜 → 瓜	搞 弧 孤 ……

A7(x1, x2): Two strokes are substituted by one stroke, where x1 denotes the two strokes to be substituted and x2 denotes the stroke that replaces x1. Stroke - count decreases one. Examples are given in Table A7.

TABLE A7: Examples of A7 Functions

Function	Illustration	Related Characters
A7(匕, 匕)	拒, 渠, 矩 ……	巨 → 巨
A7(内, 匕)	禺, 禽, ……	内 → 内
A7(止, 止)	些, 延, …	此 → 此
A7(了, 了)	極, ……	亟 → 亟

A8(x1, x2): Substitute a group of strokes x1 by another group of strokes x2. Examples are given in TABLE A8.

TABLE A8: Examples of A8 Functions

Function	Illustration	Related Characters
A8(儿, 川)	四 → 四	西, 要, 甄 ….
A8(木, 木)	呆 → 呆	保, 傑 … (all characters with 木)
A8(米, 米)	咪 → 咪	(all characters with component 米)
A8(氷, 冫)	雨 → 雨	雯 ….
	犀 → 犀	屬 …
A8(业, 业)	並 → 並	普, 虛 ….
A8(二, 冫)	匀 → 匀	鈞, 次, ….
A8(艮, 目)	即 → 目	(all characters with component 艮)
A8(八, 丷)	兌 → 兑	說, ….
A8(丷, 八)	益 → 益	溢, ….
A8(丷, 儿)	高 → 高	隔, …
A8(八, 人)	公 → 公	
A8(入, 人)	全 → 全	拴, 詮 ….
A8(支, 攴)		
A8(云, 去)	育 → 育	
	充 → 充	
	棄 → 棄	
	流 → 流	

B. Component level transformation functions

On this level, stroke-change causes component change, but does not cause character change.Some of the functions on this level can be classified as micro-differences if the change has been used and adopted for a period of time. The functions on this level are further divided into the following groups.

B1(y1, y2): Substitute component y1 by component y2. The difference between y1 and y2 can be recognized as simple stroke variations just as functions listed in stroke-level,although this change causes component changes in a character. Emamples are given in TABLE A9.

TABLE A9: Examples of B1 Functions

Function	Illustration	Related Characters
B_1(士,土)	吉→吉 寺→寺 壯→壯，志→志	桔，嘉，… 侍，時，等，…
B_1(土,士)	屆→屆 壬→壬 報→報	廷，望，… 幸，執，…
B_1(土,丰) B_1(土,工)	周→周 毀→毀	彫，…
B_1(人,入) B_1(入,人)	肉→肉 內→內，兩→兩	
B_1(儿,几) B_1(人,儿)	亮→亮 奐→奐	沿，虎，凳，微，… 換，煥，…
B_1(×,又) B_1(×,丷)	拔→拔 彥→彥 產→產	髮→髮 顏… 鏟，…

TABLE A9 (continued)

B_1(刀,力)	券→券, 拐→拐, 寡→寡	
B_1(力,力)	劫→劫	
B_1(勹,力)	别→别	
B_1(勹,刀)	包→包	免
B_1(刀,勹)	負→負,	召, 絕
B_1(刂,亻)	俞→俞	
B_1(冂,卄)	鼻→鼻	
B_1(月,月)	育→育	all characters with 月
B_1(牛,牛)	件→件	羊→羊, 犀→犀
B_1(朱,木)	殺→殺	剎, 悚, 術, 麻,…
B_1(冃,曰)	冒→冒	曼, 最, 冕,…
B_1(艹,艹)	敬→敬	薈, 夢, 覓, 蔺, 舊,…
B_1(艹,艹)	花→花	all characters with radical 艹.
B_1(丂,丂)	考→考	
B_1(夂,几)	殳→殳	沒, 殁,…
B_1(巳,己)	圮→圮	Note: the components 己, 巳, and 己 are sometimes interchangable in every day usage.

B2(y1, y2): Substitute component y1 by component y2. The differences between y1 and y2 are not on stroke-level. Therefore characters with B2 variations are considered to be different glyphs. Examples are given in TABLE A10.

TABLE A10: Examples of B2 Functions

Function	Illustration	Related Characters
B2(七,亡) B2(禾,禾)	荒→荒 稟→稟	忘,慌…. 凜,懍,…
B2(囪,囱) B2(心,止)	窗→窗 恥→耻	聰,總….
B2(甩,田)	勇→勇	湧,…
B2(歹,爫)	歿→殁 猛→猛	將…. 遙…
B2(爫,不) B2(文,攵)	覓→覓 修→修	倏,務,條…
B2(口,月) B2(口,厶) B2(厶,口)	面→面 雖→雖 強→强	高,回,…
B2(艹,廿) B2(至,互) B2(廿,艹)	垂→垂 恆→恆 卉→卉	華 奔,噴…
B2(夕,夕) B2(并,幷) B2(开,幵)	条→条 併→併 妍→妍	剝,絛,…. 姘,屏,… 石开…
B2(冏,田) B2(几,乃) B2(毋,母)	點→點 朵→朵 貫→貫	會,僧,… 剎,躲,… 賓,慣,…
B2(禾,禾) B2(玉,尔) B2(九,九)	秘→秘 寶→寶 拋→拋	
B2(冏,乃) B2(米,禾) B2(乚,又)	隽→隽 糠→糠 輒→輒	攜,…
B2(耳,身)	耽→躭	

B3(y1, y2): Substitute a set of components y1 by another set of components y2. Examples are given in TABLE A11.

TABLE A11: Examples of B3 Functions

Function	Illustration
B_3(寺, 寿)	儀
B_3(采, 釆)	
B_3(臼, 由)	叟, 嫂, 瘦, …
B_3(囪, 多)	悤, 蔥
B_3(卞, 卞)	
B_3(谷, 去)	卻→却　　腳
B_3(市, 韦)	柿→柿
B_3(羊, 羊)	差→差　　嗟, …
B_3(豆, 豈)	戲→戲
B_3(豔, 曲)	豐→豐
B_3(串, 串)	衷→衷
B_3(豸, 豸)	貓→猫
B_3(豕, 豕)	豬→猪
B_3(足, 辶)	跡→迹
B_3(兔, 兎)	參→参
B_3(黃, 黄)	
B_3(鳥, 隹)	鵲→雀
B_3(面, 丐)	麵→麦
B_3(黍, 米)	黏→粘
B_3(鼉, 晏)	鼈→鼈
B_3(笁, 米)	齒→歯
B_3(員, 勻)	韻→韵
B_3(鹵, 口)	鹽→塩
B_3(火, 业)	發→発
B_3(从, 氺)	衆→衆, 聚→聚
B_3(曲, 林)	農→農

TABLE A11 (continued)

B_3(本, 夲) B_3(旦, 且) B_3(攵, 攴)	亶→亶 敦→敦	檀…
B_3(帛, 帒)	敝→敝	幣,…
B_3(匸, 凹) B_3(市, 帀) B_3(卑, 卑)	曷→曷 肺→肺	喝,…
B_3(四, 罒) B_3(水, 氺) B_3(同, 同)	罵→罵 鯨→鯨 鑾→鑾	

C. Character level transformation functions

On this level, the appearance of the whole character changed.

C1: Same components, but position changed, such as:

啟→啓
滙→匯
鄰→隣
群→羣
闊→濶
廉→廉

C2: Simplified characters, such as:

爲→爲→为
轉→転→转
韮→韭
麤→麁
鞏→巩

C3: Deformed, irregular changes, such as:

㚔→咩
裡→裏
龜→龜

5. Discriminating Glyphs

By studying the variations in section 4, a proposal is made to draw a line to discriminate glyphs, and hence defines the so-called micro-differences in previous papers.

1. All functions at stroke level are considered as micro-differences of a glyph.

2. All functions at character level are considered to cause glyph change.

3. In component level changes,
   B1 is considered as micro-difference changes because the difference is insignificant or acceptable in everyday use of the glyphs.
   B2 and B3 are considered to cause glyph change.

I must state that most of the characters in the B1 group are wrong characters from a restricted viewpoint of traditional linguistics. It is very hard to reach a consensus to differentiate glyph and micro-differences. A less argumentive proposal is to let A group functions be the micro-differences of a glyph, and let all functions in group B and C define glyphs of a character. But this proposal will increase the number of glyphs significantly. The drawback is obvious where data processing is concerned. Therefore, a model for handling the visual appearances of the character, glyph, and micro-difference is proposed in the next section.

6. A functional model for character, glyph and form derivation.

From the previous sections,a conceptual block diagram of generating the definitions for the visual appearances of a character is shown in Figure 7. In Figure 7, a character may have one or more than one glyphs. Then, each glyph has a formal definition of its visual appearance in terms of its components, and the relative positions and sizes of each component, respectively. Again,the visual appearance of each component is defined over the strokes that the component has.

A character is identified by its code, and so are the glyphs. Therefore,if we construct a mechanism in the omputer that shows the mapping from a character to a glyph, then we can select a glyph for each character. This mechanism may be a table or a computable function such as in Figure 6.

The visual appearance of a glyph G is defined by assembling its components in the equation:

$$G = \sum R(p,s) \quad \ldots\ldots\ldots (2)$$

Since all the visual appearances of glyphs are defined individually, equation (2) will only define the micro-difference functions in group B1.And,the functions in the B1 group are simply a substitution of the components that cause a micro-difference change on the glyphs. In the same way, all the functions in group A are defined in the equation:

$$R = \sum T(p,s) \quad \ldots\ldots\ldots (3)$$

The above relation is clear and simple,and computable if the equation (2) of each glyph and equation(3) of each component have been established in a machine readable form in the computer.

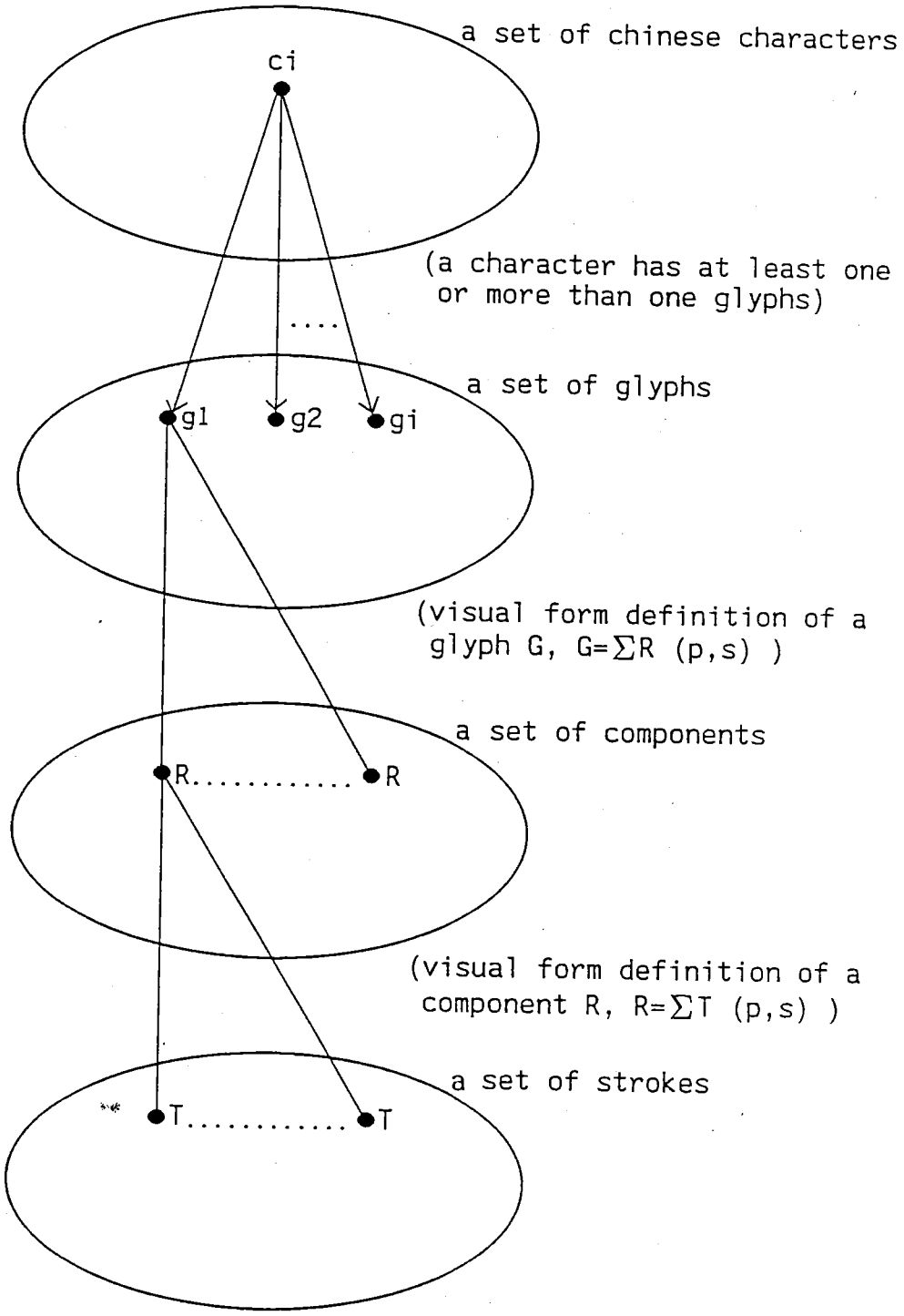


Figure 7: A conceptual diagram for generating visual form definitions of a character.

There are many ways to implement the set of equation(2) and the set of equation (3) in the computer. One way of implementation is to utilize the partical ordering relation between a glyph and a component, as well as between a component and a stroke. So, we can establish a tree data structure for processing the relations given in equation (2) and equation (3).

An example is given in Figure 8.

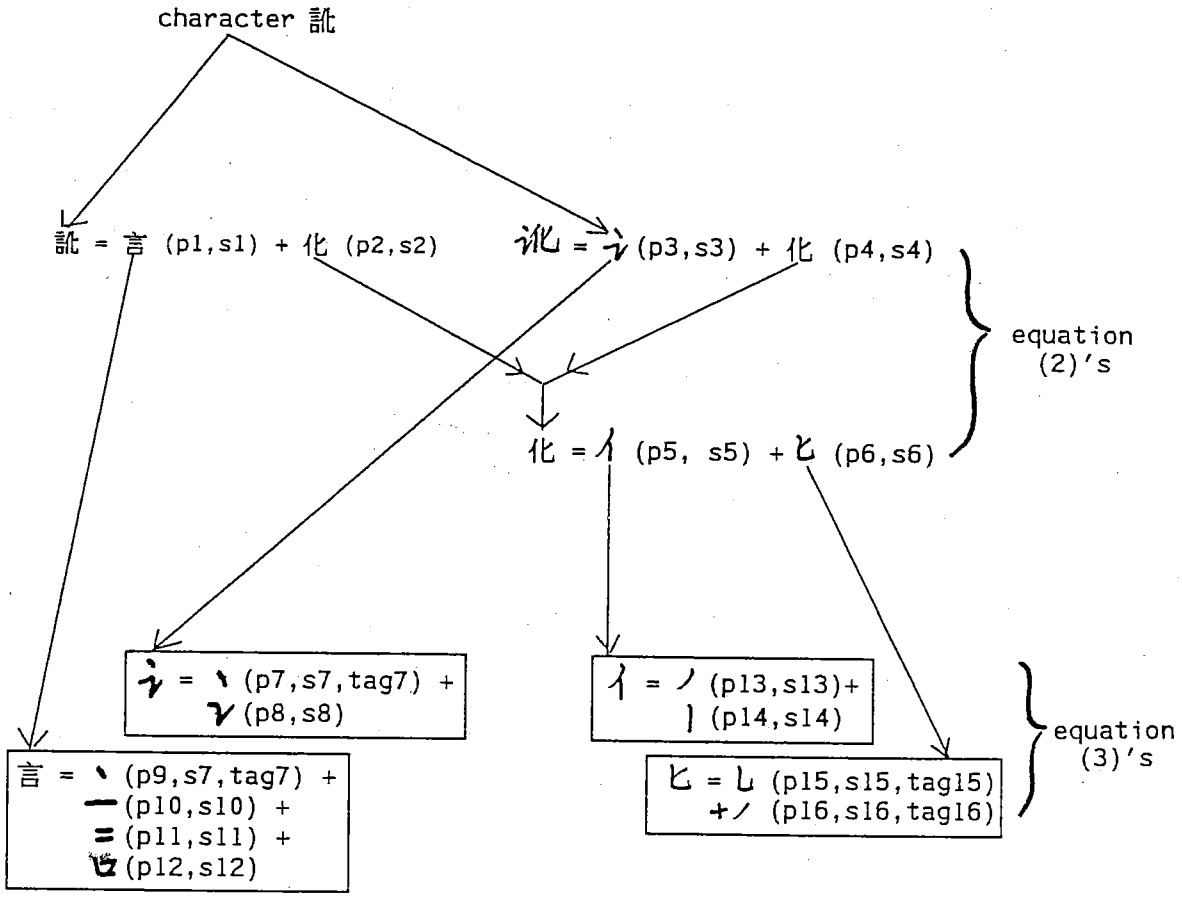


Figure 8: An example of generating different forms for the character 訛. Please notice that a tag field may be added to the parameters of each function to indicate micro-difference change functions of a glyph, where:

tag 7: A3(丶,-), tag 15: A2(L)
tag 9: A3(丶,-), tag 16: A1(丨)

In Figure 8, there are some things that need further explanation. First,the equation (2)s in Figure 8 is implemented by a binary tree. In other words, each glyph is divided exactly onto two compgnents, therefore finite steps of nesting may happen, such as where the two glyphs 訛 and 訛 share the component 化. A binary tree is not the only solution, but it is easy to implement and also saves memory space because the same component can be shared.

Secondly, a tag may be added to the parameter field of each function. A tag is a code for indicating variation functions mentioned in section 4. In this example functions at stroke level, A1, A2, and A3 are given. Functions at B1 level, or at the component level can also be implemented as tags in equation(2).

Because of the sharing property of this model, a set of trees for sharing the same variations can be organized and implemented. An example is given in Figure 9.

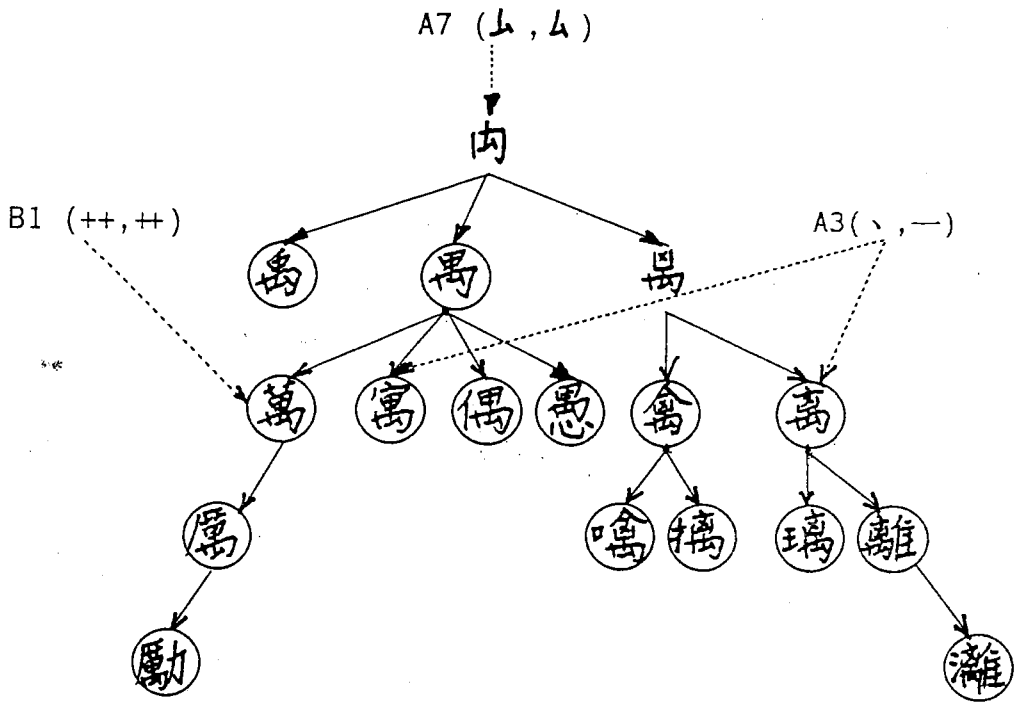


Figure 9: An example of component sharing and variation function sharing on a GLYPH DERIVATION TREE.

In figure 9, glyphs (encircled) are located at the leaves of the tree or at the nodes next to a glyph. Therefore, this tree is the reverse of the tree in figure 8. Those nodes that are not circled are components. It is easy to see that all these 15 glyphs/characters share the same A7 (厶,厶) function of the component 内.

Also, some glyphs with a leading dot are subject to the A3 (丶,-)function, and all glyphs that begin with the component are subject to the B1 (忄,艹) variations. These relations can be represented in the computer with the GLYPH DERIVATION TREE (字形孳乳樹) illustrated in Figure 9, also.

To conclude this section, the model proposed has the following memebers:

1. a set of characters
2. a set of glyphs
3. a set of components
4. a set of strokes

and also, they entail the relations listed below

5. a mapping function from character to glyph
6. a set of glyph definition functions $G = \sum R(p,s)$
7. a set of component definition function $R = \sum T(p,s)$ in terms of stroke set T

and, while implementing the system model, we may consider the following implementation:

8. the trees in Figure 8, or
9. the GLYPH DERIVATION TREE in Figure-9.

Although there are no procedures explicitly listed in this paper for processing the related information about the visual forms of character, glyph, etc., we have discussed the processing procedure through the possible data structures and internal representation schemes. It is shown that this model is feasible for implementation with existing computational technology.

7. Remarks

The authors had presented the ideas and examples of this paper to the Society of Chinese Character Studies (中國文字學會, at Taiwan) in December, 1989. Their comments are appreciated. Therefore, the work of this paper can represent

certain common ground for glyph definition from linguistic point of view.

Also, the model presented here can be useful for design character/glpyh generators, for machine recognition of characters, and for data compression studies of character images.

Any comments are welcome and will be appreciated.