中文字形資料庫的設計與應用

*謝清俊　**莊德明　**張翠玲　**許婉蓉

*中央研究院資訊所　研究員

**中央研究院資訊所　助理

摘　　要

本文介紹一個記載漢字字形的資料庫，它能記錄字形結構和筆劃上的變化，並與已經記載在資料庫內的字形做相似程度的比較。此外，它也提供簡單的統計，以量化數據說明一些字形或構字上的特徵。這個資料庫原本是為字體交換協會（AFII）所做。整個資料庫也可視為一個計算機使用的文字與字形的制式定義。

本文首先說明系統使用的文字與字形的制式模式，然後介紹此系統的設計與實作，最後，報告此系統發展的現況。本文採用的文字與字形模式，是根據民國六十一年發表的交大字根模式增益而成的。字形的變化可歸納為筆劃的變化、字根或部件的變化、和整個字的變化等三個等級。本系統是在IBM相容的個人電腦(PC)，中文視窗(WINDOWS 3.1)以上的作業系統下完成的。

國立中興大學中國文學系所

中　國　文　字　學　會　聯合主辦

中華民國八十四年四月二十九至三十日

中文字形資料庫的設計與應用

壹、緣起

本文介紹一個記載漢字字形的資料庫，它能記錄字形結構和筆劃上的變化，並與已經記載在資料庫內的字形做相似程度的比較。此外，它也提供簡單的統計，以量化數據說明一些字形或構字上的特徵。

這個資料庫原本是爲字體交換協會（AFII, Association for Font Information Interchange, 以下簡稱字協）所做。字協是爲國際標準組織（ISO）執行ISO 9541，即登記標準字體（standard font）的字樣（typeface）和設計（design）的財團法人。三年前，字協承印 ISO 10646碼本時，面臨由大陸、日本、韓國、和臺灣蒐集的數萬漢字，涉及許多頭痛的問題，其中之一就是這些字在構字和筆劃上頗不一致。

在一般人的印象裡，各地區的漢字應該有很多是相同的。果眞如此，計算機在處理時就可以省下許多儲存的空間、處理的時間和管理上的負荷，甚至編碼的位置也可共用而節省了。然而，這些漢字雖然都是仿宋體或類似仿宋體者，在構字和筆劃上卻不盡相同，仍有差異。更不幸的是，各地區都堅持這些差異處有理，不肯稍加修改以求一致。於是，判別這些差異，就變成了字體登記上極耗時耗力的問題。

其實，這個問題的根本所在，仍應歸咎於ISO 10646的設計。ISO10646的設計並沒有充份考慮到漢字的性質，亦即設計所用的文字學知識甚爲膚淺（甚至可說完全沒有！），以致於沒有辦法面對或解決這個問題。這種情形連美國的標準組織工作組也注意到了，幾經與ISO相關工作組協商都不得要領的情形下，發表了一份文件，說明文字(character)和字形(glyph)之間應該確立的關係【註一】。換言之,計算機中需要一個制式的定義(formal model)來界定什麼是文字、什麼是字形，以及此二者之間的關係，這樣才能好好的利用計算機來處理文字和字形。

【註一】請參照美國標準工作小組X3L2和X3V1在1993年9月聯合發表的文件<Character Glyph Operational Model>此文件經黃大一先生譯為中文，用於中央標準局與資策會合辦的資訊標準應用研討會(1994年5月, 台北)。又，此文件在ISO之文件編號為ISO/IEC JTC1/SC18/WG8 N1615並準備改寫成為ISO正式的技術報告(ISO Technical Report)。

關於文字與字形的制式定義，並不是沒有人研究，日本、台灣、大陸、美國等都有人做過，商品上也早已用過。只不過除了大陸以外，都沒有把這種制式模式提昇到國家標準的層次，因此，也就難以搬上世界標準的舞台。

本文所敘述的資料庫，就是在上述的背景下，爲釐清文字及其字形的變化所設計的，整個資料庫也可視爲一個計算機用的文字與字形的制式定義。在下文中，我們將先介紹本系統使用的文字與字形的制式模式，然後介紹此系統的設計與實作。最後，報告此系統發展的現況作爲結束。

由於文字與我們的生活關係至深，文字累積的資料和知識理應因由國人共享，所以，我們發展的這個小小工具，願意免費提供給大家使用。有意者請洽作者。

貳、文字與字形的制式表達

本文採用的文字與字形模式，是根據民國六十一年發表的交大字根模式【註二】增益而成的。要言之，文字是表達一種（或一群）概念的名相，是抽象的。我們可以透過媒體，例如：聲音或字形，將它呈現出來以使人可以察覺。在本文中，暫不討論字音，只考慮字形部份。

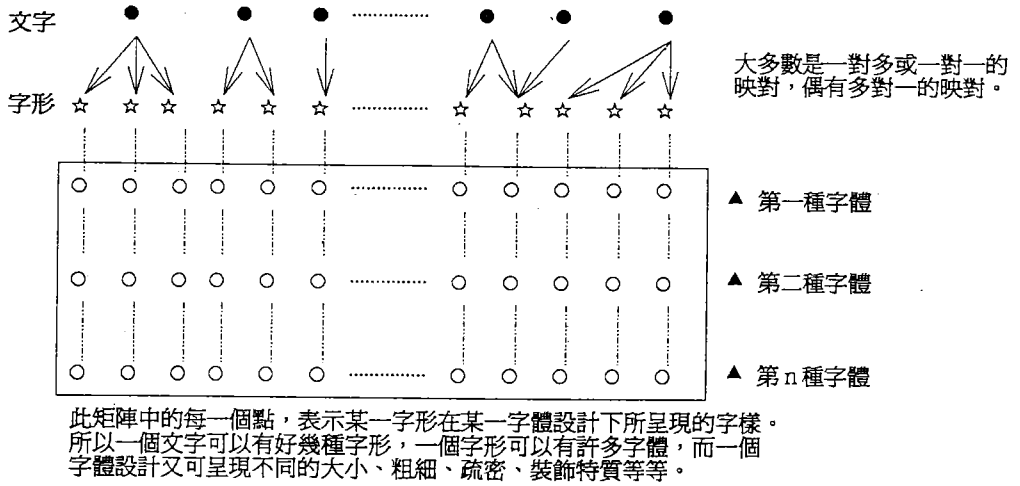
文字和字形的關係是：一個文字可以有許多不同的字形；偶爾，也有些文字會用相同的字形。所以，從數學關係來說，文字之於字形是多對多的映對，雖然大多數是一對多的關係。字形也是抽象的名相，區別字形的關鍵在於它的組成結構，亦即構字。所以一個字形可以有變化多彩多姿的風貌，然而其構字之規律卻不容更易。

【註二】 請參照：謝清俊等＜中文字根之分析＞，交大學刊第六卷第一期 pp. 112-121, 1972。此外，劉達人等編著《漢字綜合索引字典》,(Asian Associates出版，美國國會圖書館卡號：790525, 1979)也採用此模式。又，此模式曾改寫為 AF11文件<On the formalization of glyph in Chinese Language>1990, 2月。

印刷體的字形有一定的設計規範，遵從同一設計規範製作的一群字形屬於同一種字體。字體也是抽象的，區別它的關鍵在於設計的規範。一種字體設計，通常有些參數來決定文字呈現的大小、粗細、疏密、以及一些特殊裝飾的邊角等等。待這些參數選定之後，才能呈現出一個字形的面貌，稱爲字樣。這個面貌就不是抽象的而是具體可見的了。

上述的文字、字形、字體、和字樣的關係可如【圖一】所示。所謂計算機用的文字與字形模式，就是要把這些抽象的名相和相關的構字、設計規範，以及彼此間的關係和運作規律等，用計算機能了解的形式表達出來，以便計算機能夠運用這些知識，替我們做事。

圖一：文字、字形、字體和字樣的關係

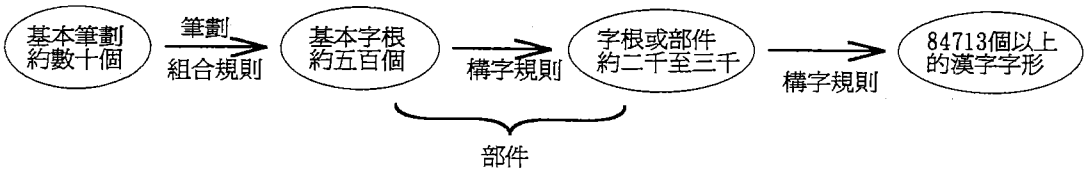


一、字形的表達

漢字由部件（component）構成。此構成是有規律的，可歸納爲簡單的制式規則。在交大字根系統中，將漢字字形以橫連(H)，直連(V)，和包涵(C)分解，得到496個基本的部件，稱爲字根。這群字根和組合的規則（亦即分解的規則）可以組成48713個漢字，除了一些類似圖騰的古字外，幾乎可組成所有的漢字，且具有創制新字的功能。所以我們採用這套字根系統作爲設計的基礎。

其次，字根由筆劃組成。印刷體設計所用的基本筆劃數目約三十至一百餘不等。其實，三十幾個基本筆劃已足資辨認，用到一百多個筆劃只是爲了做好細部的面貌以美化字樣而已。我們的系統重點在辨認字形，所以用的筆劃不多（事實上，可由使用者自己選擇筆劃之多寡）。字形、部件、字根、和筆劃的關係如【圖二】所示。

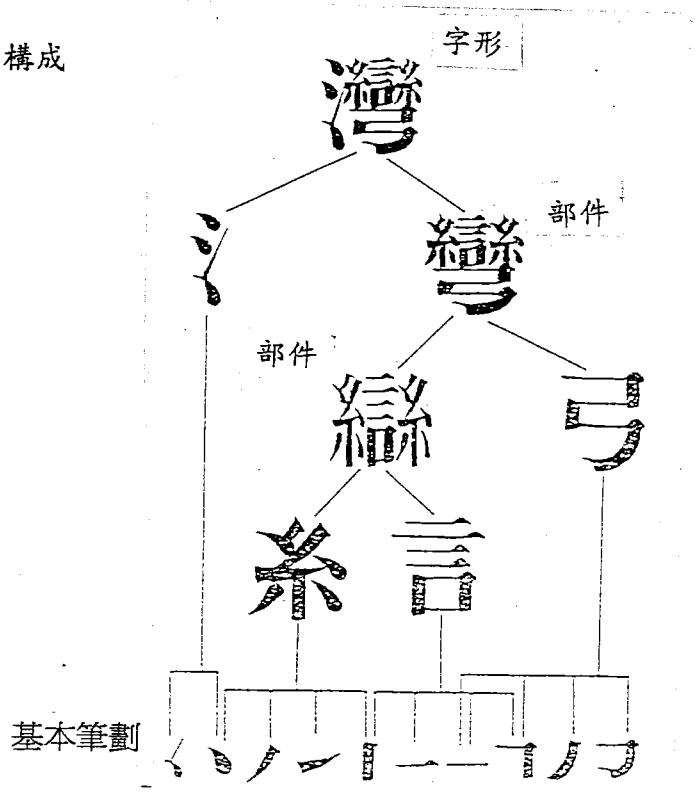
圖二：字形、部件、字根和筆劃的圖係



在實際運作時，這樣的系統會產生一些較複雜的表示式。例如：【圖三】的灣字。所以在設計時，我們儘量把構字規則簡化，以利執行。做法是：在組合（或分解）一個字形時，一次只用一個組合（分解）的運作，即選用橫連、直連、包涵其中之一。若還沒有分解成基本字根，則將生成的部件再分解，直到式中只有基本字根爲止。由於基本字根數目不多，筆劃組合規則就依其筆劃生成的位置逐一標明。

上述的構字概念可以用簡單的數學表達，是故可以用計算機處理。在【註二】所引的各參考文件中，有詳細的敘述，請各位參考。

圖三：灣字的構成



二、字形變化的分析

根據上述的字形模式，字形的變化雖多，卻可以歸納爲筆劃的變化、字根或部件的變化、和整個字的變化等三個等級。這種劃分法，筆者曾在中華民國文字學會報告過（民國78年12月）。現謹將其大要節錄如下：

（一）、筆劃的變化

A1：有一筆劃位置改變，但文字的筆劃數和構成的字根沒有改變。
例如：丸→丸，丩→丩，才→才，匕→七等。

A2：有一筆劃尾部加勾，筆劃數和構字之字根不變。
例如：不→不，月→月，也→也，七→七等。

A3：有一筆劃被另一種筆劃替代，筆劃數與字根不變。
例如：丶被丨，一，丶，㇀之替代
一被丿、丶替代，丨被亅替代等。

A4：增多一筆，筆劃數增1。例如：者→者，少→少等。

A5：減少一筆，筆劃數減1。例如：广→厂，次→次，衆→衆等。

A6：某一筆劃由另二筆劃取代，筆劃數加1。
例如：工→工，厶→厶，⺊→⺊等。

A7：某二筆劃由另一筆劃取代，筆劃數減1。
例如：止→止，了→了等。

A8：某一群筆劃由另一群筆劃取代。
例如：雨→雨，艮→艮，兌→兑，育→育等。

（二）、字根或部件的變化

B1：一字根R1由另一字根R2取代，而R1和R2的差異只是筆劃上的變化（如前述A1至A8之變化）而引起的字根變化。
例如：吉→吉，寺→寺，產→産，壬→壬，周→周，內→内等。

B2：一字根R1由另一字根R2取代，而R1和R2之差異不屬筆劃上的變化。
例如：多→多，覓→覔，恥→耻，秘→祕，雋→雋，耽→躭，
磚→塼等。

B3：一部件（一群字根）由另一部件取代。
例如：寿→寿，由→由，市→市，足→辶，灸→矣，鳥→隹，
⺗→业，曲→林等。

（三）、整個文字構字的改變

C1：字根不變而組合改變者。例如：啟→啓，廉→廉，滙→匯等。

C2：由簡化而改變者。例如：爲→爲→为。轉→転→転。

C3：不規則變形者。例如：㕣→咩，裡→裏等。

以上這種分法，是完全從字形著手劃分的。其實，尚可加注從文字學角度的觀察，例如：聲符的取代多屬字根或部件的變化（亦即屬B級函數）。這些工作，留待以後補充。

三、舉例

現在，讓我們用一個較完整的例子，以說用前述的表達方式。

假設G代表字形(Glyph)，R代表部件(Component)而R_0代表字根，T代表筆劃，p和s分別代表T和R在字形中的位置(position)和大小(size)，則一個字形的組成，可用下面二個再生式(recursive equation)表示之：

$$G = \Sigma R\ (p, s) \quad \ldots\ldots\ldots\ (1)$$

$$R_0 = \Sigma T\ (p, s) \quad \ldots\ldots\ldots\ (2)$$

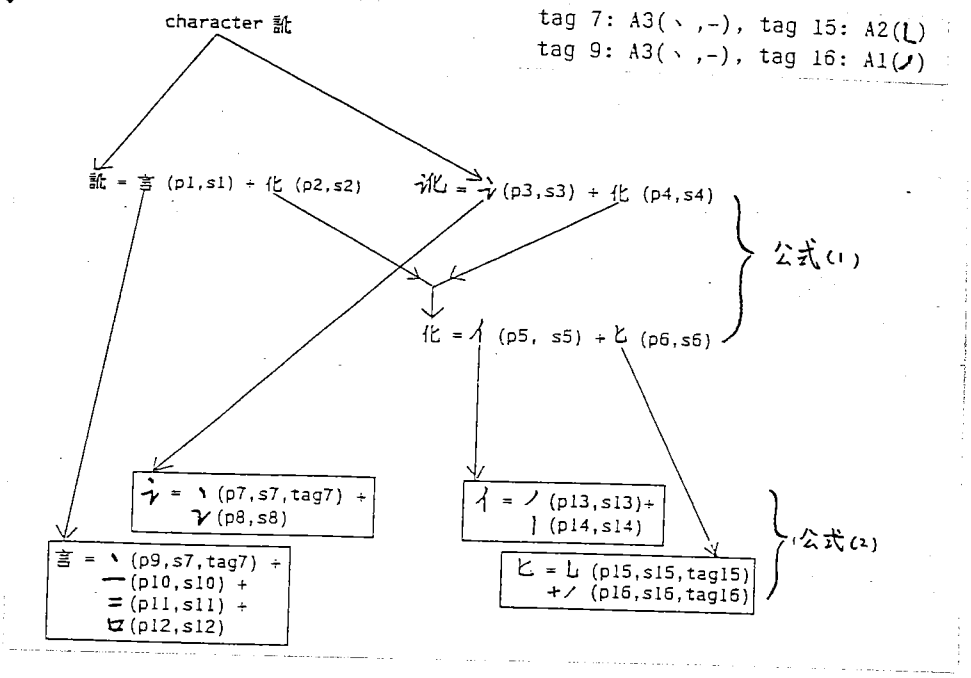
（1）式的意思是：G由許多（以Σ表示）R組成，而每個R在字形中都有它特定的位置和大小（即p和s）。（1）式可重複使用，直到所有的R都化簡到R_0爲止。而R_0，根據（2）式，可再分解爲一群筆劃T，每個筆劃亦有其位置和大小的關係。

了解這些符號後，【圖四】和【圖五】即分別從兩個角度，來說明這種模式的應用。【圖四】表示的是一個文字「訛」、它的兩個字形、一些部件和字根，和基本筆劃之間的組成關係；以及字形變化所加的標識(tags)。

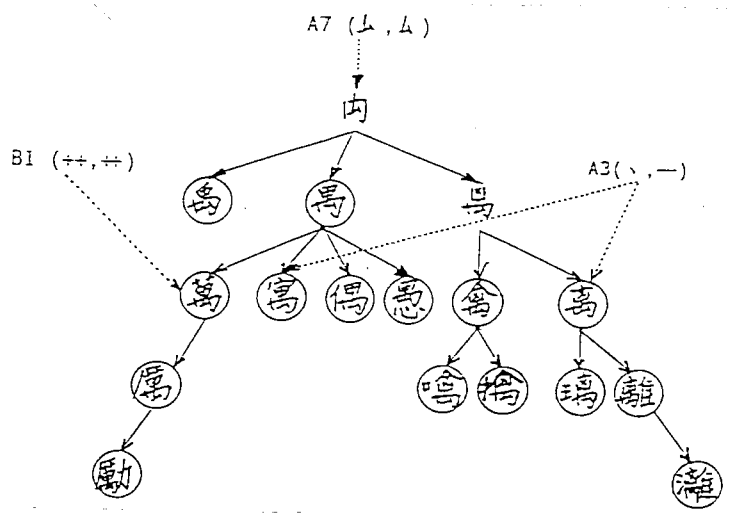
【圖五】是一個字根「内」的字形孳乳樹。在圈內的是字形，沒圈的是字根或部件。圖中字形變化的標識，以虛線分別表示。

在計算機裡，【圖四】和【圖五】都是從一個結構表推導出來的。【圖四】是由字形推導到筆劃，而【圖五】則相反，由字根推導到所有相關的文字。

圖四：訛字相關的字形組成



圖五：字形孳乳樹之例



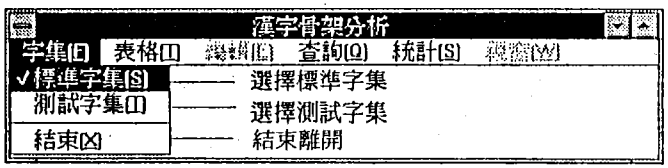
參、實施

本系統是在IBM相容的個人電腦(PC)，中文視窗3.1(WINDOWS3.1)以上的作業系統下，用Visual Basic程式語言發展出來的。系統的功能分爲字集的選擇，系統內各種表格的檢視、修正，結構或統計的查詢，以及視窗的管理等項目。系統的一些視窗畫面如【圖六】所示。

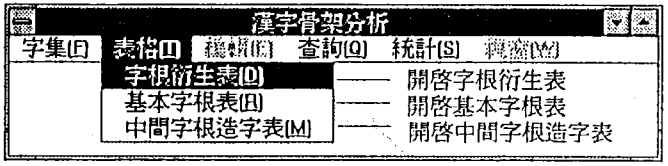
圖六、系統主要視窗

一、選單

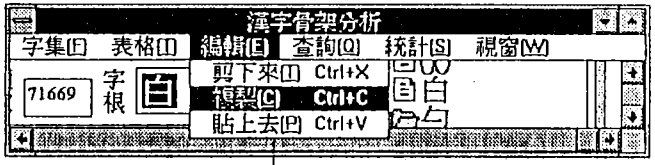
字集選單



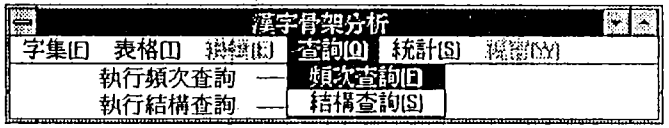
表格選單



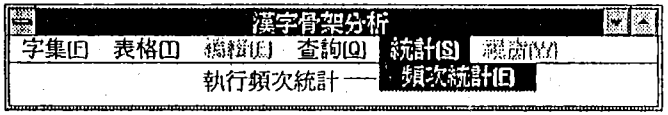
編輯選單



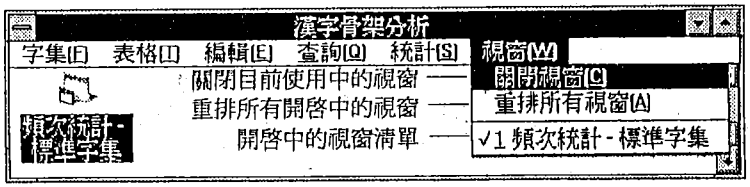
查詢選單



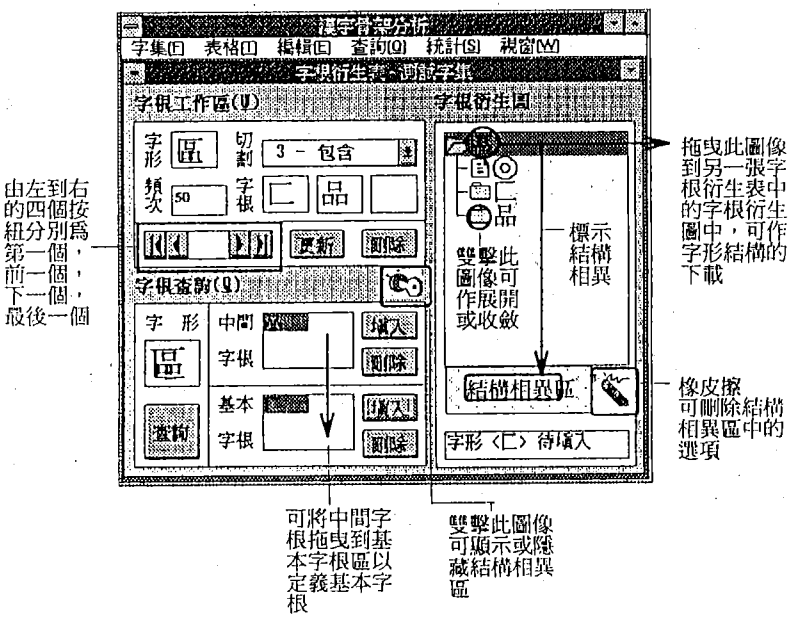
統計選單



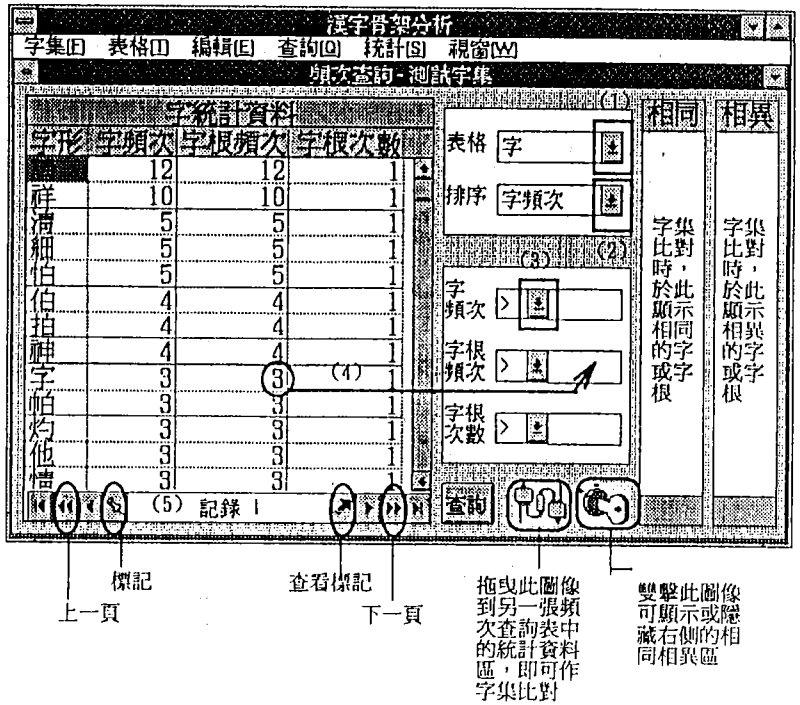
視窗選單



二、字根衍生表

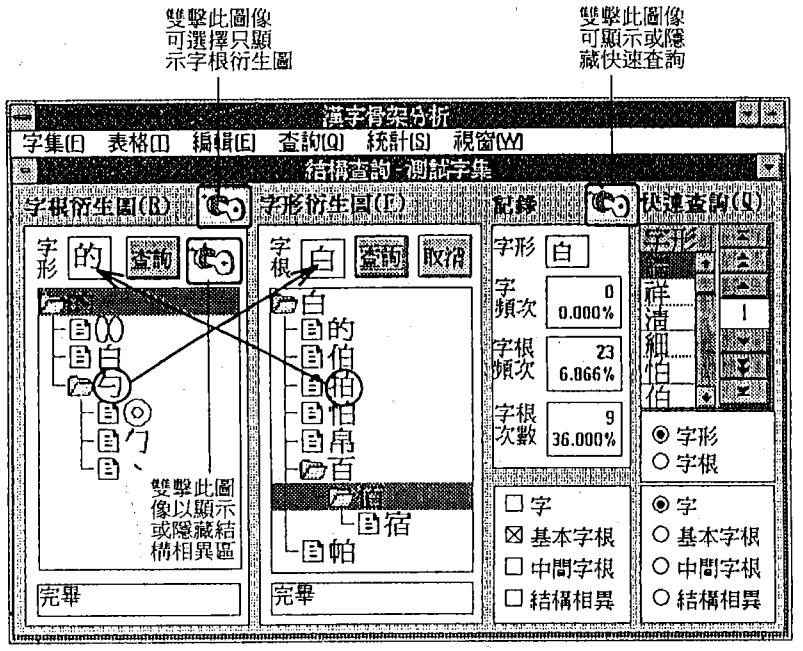


三、頻次查詢



(1) 表格部分的選項爲字、基本字根、中間字根、結構結異。
(2) 排序部分的選項爲字頻次、字根頻次、字根次數。
(3) 字頻次、字根頻次、字根次數部分的選項爲等於、不等於、大於、大於或等於、小於、小於或等於。
(4) 可拖曳資料區的任何數字到數字輸入區。
(5) 記錄區其他符號的意義同字根衍生表，分別爲第一個、上一個、下一個、最後一個。

四、結構查詢



(1) 如同在字根衍生表中，可從字根衍生圖中拖曳字形的結構到另一個字集的字根衍生表中的字根衍生圖區，以下載該字的結構。
(2) 使用快速查詢時，先選擇資料的來源爲字、基本字根、中間字根、或是結構相異的字或字根，然後設定是字形結構查詢或是字根反向搜尋，即可從資料區雙擊該字或按下<Enter>鍵自動進行搜尋。

這些系統中的文字、字形、部件、字根、筆劃之呈現方式如下：

文字：以系統既有的字形呈現，若無則在系統造字區內造其形，定其碼。
字形：原則上不造其形，以字根結構式呈現。
部件：在系統造字區內造其形，並定其碼。
字根：字根集合中有300餘是文字，故可用系統字形，不是文字的字根則在造字區內造其形，定其碼。
筆劃：在系統造字區內造其形，定其碼。

至於因筆劃導致的字形、部件、字根的變化，亦不造其形，僅以A1~A8和B1函數表示在其變化的標識中。

所以，在本系統中，可以由構字上觀察到字形，由標識中觀察到筆劃的變化；卻看不到實際的「字樣」。

這個系統可以有幾個用法。當系統內尚無資料，或是要分析一套尚未載入系統的字形時，可以逐字將字形的結構資料，包括文字、字形、部件或字根、基本筆劃等。以人機互動的方式，建立電腦內部的結構。此時，若有文字或字形的字頻統計，亦可一併載入留待後用。

一套字形載入後，即可對這套文字作各種查詢，包括每一個字形的構成、字根孳乳表、文字孳乳表、以及字形、字根、部件等的統計資料。

其次，若有某字形的構字待比較，則可叫出計算機內既有的字形由人比對，或將該字的結構輸入由計算機比對。

若是有兩套字形均已載入計算機中，則計算機可以詳列此兩套字形的差異。可載入計算機中字形的套數不限。蒐集越多，則越具參考價值。

系統操作的情況實一言難盡，請參考現場操作的展示和說明。至於此系統的應用，則依使用人活用上述的用法而定。

肆、結語

這套系統的開發，並未接受任何單位的資助，是故進展緩慢，目前已具雛型，略可應用。此外沒有接受補助也有好處，那就是不涉及產權問題，是故可以讓大家共享。

此系統中已輸入一套字形，乃是根據民國六十一年林樹教授的《中文電腦用字的分析》一書，所收錄8532個字彙及其使用頻度的資料。現場展示亦此一套字形而已。

文字資料是資訊處理最基本的資料。由於文字資料的整理一直未能規劃完備，在處理方面不知已浪費多少人力、物力和時間，甚或製造了不少文字、語文上的混亂。我們甚盼有一天，每個電腦中都配備了充份的文字資料和知識，不只能幫助提升文件處理的水準，亦將有益於民眾語言水準的提升。

文字資料是一切文獻的表達基礎，目前所納入系統中僅字形這一部份。其實，其他相關的例如：文字學、聲韻學等資料，也可以和本系統整合在一起運用。可是，這樣的發展實不是我們這個小計劃可以完成的，必須以文字學爲領導，輔以資訊技術才能竟全功。

謹以此小小系統拋磚引玉，期盼有心人一同來推動。