第一屆中國文字學會學術討論會
天津

電子古籍中的缺字問題

謝清俊

中央研究院　資訊科學研究所 文獻處理實驗室

1996 年 8 月 25(修正版 85.12.20)

電子古籍中的缺字問題

謝清俊　中央研究院　資訊科學研究所　研究員

摘　　要

用計算機處理漢字資料時，常有些字的字形是交換碼中沒有的，這情形在古籍中特別嚴重。爲了要保留這些字形，常用的方法是在使用者造字區內，增加這個字形，可是這樣的做法不但要付出巨大代價，也沒能真正解決問題。例如：爲了新造的字，資料登錄的工作大幅增加；檢索文件時將面臨異體字檢索的難題；彼此分享資料時則更嚴重，可能重碼將造成資料錯誤或文件讀不出來的狀態，以致於根本無法共享資料。

造成這種情況的根本原因是目前計算機中預存的字形信息太少，如果我們能將字形的信息表達在計算機中，就可利用計算機來協助我們解這道難題。

本文用資料庫來表達字形的結構和該字的屬性，在字形資料庫中，字形是以其部件及字根的組合方式表達。目前已將兩岸的兩套字根建構在資料庫中，可相互爲用。

如果遇到計算機中沒有的字，此時該字的輸入碼、交換碼、字形等都沒有，通常計算機是無法處理，然而字形資料庫卻可提供以字根的方式來查詢。

爲了徹底解決字形的問題，我們建議將交換碼分爲三個層次，即字碼、字形碼及字體碼。字碼依字的語意分別字，字形碼依字的結構來分別字形，字體碼則選擇該字的字體型式。這樣的設計可依應用的性質提供適當的電子環境。

目前，此資料庫共收錄有約九千個字，以及這些字的部份異體字。我們也正嘗試利用此字形資料庫來解決電子佛典中的缺字問題，這些例子都將在本文中報導。

未來的主要工作是將字形資料庫中字根的結構再分解爲筆劃的結構。如果完成了，那麼在筆劃上構成的細微變化差異，也可以用制式的方式表達出來，電腦亦因之可以對這些變化加以區分並加利用。本文亦將報導這個制式表達的方法。

這個字形資料庫的系統是將放在電腦網路上的，所以凡可接到網路上的使用者，都可以共享它所提供的服務。此外，字形資料庫以後還可擴充到包涵字音和字義的部份。一個做法是和一部電子字典互相鉤連，並把文字學的資料一齊建立。如此，則可能發展成爲文字學資料庫，成爲漢學研究和學習的有力工具。

第一屆中國文字學會學術討論會，　天津　　1996 年 8 月 25-30 日

電子古籍中的缺字問題

壹、前言

漢字歷史淵遠流長，自隸書以後就有二千三百餘年，誠如北魏江式云:『世易風移文字改變』在所難免。《顏氏家訓．雜藝》云：『晉宋以來……不無俗字，非為大損。……大同之末，訛替滋生。蕭子雲改易字體，邵陵王頗行偽字……朝野翕然，以為楷式。畫虎不成，多所傷敗。……爾後墳籍，略不可看。北朝喪亂之際，書蹟鄙陋，加以專輒造字，猥拙甚於江南。乃以百念為憂，言反為變，不用為罷，追來為歸，更生為蘇，先人為老：如此非一遍滿經傳……』。可見自古以來漢字的「形」並未能定於一。雖經唐以後，官方刻石整頓約以範式，然而天下碌碌多士實難以盡入繩矩。是故爾後字書依然收錄各體字形，如《干祿字書》將字形分為通、俗、正三體。亦有收錄異體字為主者，如《龍龕手鑑》。雖然古籍的字形多變化,帶來許多麻煩，但也非全無好處，特殊字形能提供版本、校勘方面有用的信息。所以，對治古籍不能不妥善處理字形問題，即使用電腦也如此。

一、缺字的問題

現在用電腦處理古籍很辛苦，首先面對的就是缺字的問題。Wittern和App說得好；『在亞洲，缺字問題無所不在(ubiguitous)。無論用什麼國家標準交換碼，都免不了有這個問題，即使ISO 10646 也是一樣。……在可見的未來，這個問題似乎無解，只有套句老話「自求多福」了』【註一】。

為了應付缺字問題，一般治標的方法是：在交換碼的使用者造字區內，選一個碼位，並造上所缺的字形。這種做法固然可以在該電腦上顯示出該字形，但是付出的代價巨大，且沒能真正解決問題。茲將這種做法衍生的問題略述如次：

1．大幅增加了資料登錄的工作

當鍵入資料時，若逢缺字，目前的處理方式是以一個特殊符號，如：『●』，來表示所缺之字。這種做法，可暫時使鍵入工作不致中斷，得以持續進行；然而，必須增加第二道工作，來填改所有的缺字。這填改工作，卻必須等到稽查原文，把所缺之字造好後，重新校對原文，再一一改正打字稿。此補字工作之繁重，可想而知。

2．造字的管理不易

如果所缺的字不多，登錄工作煩複些也還罷了，可是，若古籍之量龐大，所缺之字動輒上千，再加上所造的字無法依交換碼的字序排列，那麼，這些新造的字是極不易查核比對的。因此在資料登錄時，查核及改正缺字是既費力又費時，一不小心便出錯，也常有重複造字的現象。再者，在工作中缺字表隨時在更新，要所有的相關人員都能及時同步更新此表，在管理與溝通上頗不容易，因更新的時差而造成重複的工作，也是常有的事。管理成千上萬的缺字，要付出許多代價。目前，電腦中還沒有一個理想的管理系統，來協助造字的管理，多靠鍵入人員以人工維護這數千字造字表，不僅浪費人力，也曠日費時，效率不彰。

3．造字的空間不足

通常,交換碼中允許造字的空間都在數十字至數百字之間,不會太大。以五大碼而言，造字空間是最大的了，也只有 5809 字。超越此數後，勢必造成碼位的重疊或衝突。

【註一】 Christian Wittern and Urs App.〈IRIZ Kanji Base : A New Strategy for Dealing with Missing Chinese Characters 〉世界電子佛典會議(EBTI)台北, 1996 年 4 月

4．異體字將造成文件檢索和處理上的困擾

缺字中有許多是與交換碼中字形不同的異體字，但是，這些異體字彼此對映的關係，並沒有告訴電腦知道。因此，電腦查「台灣」就找不出「臺灣」，並會判斷「拾元」不等於「十元」。這些現象嚴重影響到資訊處理的品質。再者，造字區字沒有交換碼用的字序，在排序上，引起應用程式的種種麻煩更不在話下。此外，現有的應用程式多半沒有考慮到加了一個使用者造字區而引發的處理問題，所以這現象又會造成程式共享上的障礙。

5．造成資訊共享的障礙

電子文件的重要特質之一，是可以幾乎不花什麼代價就可以抄錄分享。可是若有使用者造字所訂的單行碼，就破壞了交換碼的通用性，而使電子文件無法與大家共享了。即使把造字碼表提供給對方，也會因你造的字和我造的字用了同一位碼，而使彼此的文件無法共存。無法共存就失去共享的意義。事實上，缺字問題已是目前中文信息共享的最主要技術障礙。

由以上諸點看來，缺字造成的後果嚴重，而目前處置缺字的辦法實非解決問題之道。

二、解決缺字問題的原則：

造成缺字問題的主要原因，固然是現有的漢字交換碼性能不夠，從另一個角角度來看，電腦中缺乏漢字字形有關的信息，使得電腦無法順利依照我們所需要的方式處理文字，則是更直接的因素。目前的交換碼只收集字形【註二】，這點信息僅止於能做字形的顯示，對處理古籍來說，它的功能是不夠的，至少還需要各字形間的對映信息，如正俗、繁簡．古今等異體之間的對映，才能用電腦做好字形間的認別、替代、比對、檢索……等工作。此外，字的屬性，如筆劃、部首、聲韻……等也有需要用到的地方，收納在電腦中應是遲早的事。就理上說，要使電腦能較順暢地處理古籍，非把文字學的信息設法表達在電腦中作為基礎不可，至少應該把一些好的字書或辭典中的信息放入電腦才是。雖然，這些些息太多了，已非交換碼的結構可以容納，然而，現在的電腦缺乏一個機制配合著交換碼將這些信息納入電腦中，也是不爭的事實。

缺字問題雖然無處不在，人人受苦，可是要解決它卻不容易。以往，大家認為：擴大交換碼收集的字形可能是解決的方案之一。經多年各方的努力，包括中文資訊交換碼(CCCII)、GB 字集的擴充及大字庫的建立、JIS 的擴充、CNS 的擴充、以及 ISO10646 的擴充等等，根據我們處理古籍的經驗，這些擴充是有些紓解的作用，然而並不能完全解決問題，而所表現的效果並不怎麼好。這是有些原因的。首先，字形變化太多，難以收集完整。其次，是電腦中沒有字形的對映，缺乏文字學知識的表達，理由已如前述。再者，大字集使個人電腦負擔加重、成本增高；字碼變長了以後，現有的應用程式要全隨著更新，這幾乎是不可能的事。**雖然擴大交換碼的字集是該努力的工作，然而，僅靠它並不能徹底解決缺字問題。我們認爲，要解決缺字問題，還是應該建立電腦處理文字問題的能力，這包括將必要的文字學知識表達在電腦中並加以利用，以及建構一個處理字形的機制**。這就是本文的基本設想。

為避免重蹈目前解決缺字方法造成的窘境，本系統確立了以下的設計原則。

1．不可以為了解決缺字問題而犧牲了資訊共享的能力。

2．要能照顧到任何地區使用的漢字。

3．建立登錄、檢索、表達及共享「缺字」的能力，以便儘早利用電腦協助管理缺字。

4．建立電腦中關於字、字形、字體的制式表達，將必要的相關知識放入電腦，包括異體字之間對映在內。

5．建立漢字的屬性資料庫。

【註二】 中文資訊交換碼(CCCII)蒐集了異體字，並有系統的呈現在字碼結構中，是碼中唯一的例外。雖然美國圖書館界已用它十年以上，一直很順利，然而CCCII並非台灣的標準交換碼。

6．以 ISO 8879 通用標準標誌語言(SGML, Standard General Markup Language) 來描述文件檔案及缺字，以達到文件共享的目的。
7．所建立之系統應考慮到以後往「文字學資料庫」擴充的可行性。
8．研擬「字碼替代」的技術。無論以後交換碼擴充到多麼長，希望能利用字碼自動替代技術做到維持目前 16 位元的漢字處理應用環境。

以上這些原則，也就是本系統的設計重點。

貳、文字知識在電腦中表達

如前述，解決缺字問題的關鍵在於如何把文字知識有系統地表達在電腦中。在以下第一節中將先敘述一些相關的背景，然後報告本系統對文字知識表達的做法。

一、背景

在本節中先討論二則本系統所引用以前的研究，並對目前電腦中的狀況作一說明。

(一)、中文電腦基本用字

1970 年前後，台灣地區開始重視利用電腦處理漢字文件的問題。然而，當時文字統計資料並不完整，也沒有一個適當的漢字字集可供電腦採用。1971 年在王安公司贊助下，交通大學計算與控制系委請林樹先生從事「中文電腦基本用字」的研究。

該研究從 1971 年 10 月初展開，以二千多人工作天，約經半年，於 1972 年 3 月底提出初步報告。現將該研究的一些特色及重要原則條列於次：

1．該研究綜覽過去漢字遞增的情形和 1856 至 1971 年間所有的字彙研究，輔以當代字書及重要工具與典籍的常用字數，作通盤的了解與整理。
2．以社會通用的資料為主要範疇，選擇十一種字集加權整理，匯集為《中文電腦基本用字》，詳如〔表一〕。此十一種字集亦多有匯集前人整理之字彙者，累計涵蓋之字集超過三十種。
3．整理異體字，並納入字集中。參考林語堂的〈整理漢字草案〉及國立編譯館製訂《常用字統一字形暫用表》（1968）的選字原則，訂定了異體字的處理原則如〔表二〕。依〔表二〕選出之形體當作「字形」，其他形體並不捨棄列為「參照字形」。是故「字形」中並不排除簡體字,但以當時曾收編在《注音漢字》及《國音標準彙編》中之簡體字為限。換言之，**此研究並不計較字體之正訛、本俗、繁簡、古今等,而著眼在電腦將要面對、將可能要處理的字形。這種態度是和以往字彙研究不盡相同的。**
4．此研究共蒐集 8532 字，另有參照字形 597 字。統計共 202 萬 2604 字次，其頻率分佈為：

常用字	最常用字	1857 字,	出現頻度爲	97.34%
	次常用字	2068 字,	出現頻度爲	2.27%
間用字		2182 字,	出現頻度爲	0.27%
罕用字		2425 字,	出現頻度爲	0.12%

此字集之熵值為 9.60，前 500 字使用累頻詳如〔表三〕。由於此後在台灣並沒有更好的文字統計資料出現，所以本系統仍以此字集作藍本。

表一《中文電腦基本用字集》匯集的十一種字彙

1. 莊澤宣,《基本字彙》,廣州中山大學教育學研究所,1930
2. 胡顏立,《小學初級分級暫用字彙》教育部,1935
3. 教育部,《注音漢字》商務印書館,1935 初版,1961 台一版
4. 蔡樂生,《常用字選》英文中國郵報社,1946
5. 台灣省國語推行委員會,《國音標彙編》,開明書局,1947 初版,1971 台二版
6. 王清波,《國民小學現行國語課本國字初現課次、重現次數之分析研究》, 高雄市政府,1963
7. 國立編譯館,《國民小學常用字彙研究》中華書局,1967
8. 台灣電信局,《電碼新編》,1967 增訂版
9. 星華打字儀器行,《中文打字機新版文字排列表》,台北,1969
10. 世界中文報業協會,《新聞常用字彙》,1970
11. 中南鑄字廠,《常用字表》,台北,1971

表二《中文電腦基本用字集》異體字的整理原則

1. 就已有字彙選取,不另創新字。
2. 一字數形,取其簡便者。而不計其本體抑俗體,古字抑今字。古字簡便者從古,如取「礼」不取「禮」,今之簡便者從今,如取「綉」不取「繡」。
3. 一字數形,取其結構適合電腦設計者。如取「略」不取「畧」,取「裡」不取「裏」。
4. 一字數形,取其通用者,如取「拿」不取「拏」。
5. 在世俗上已通行一體,而原字還有其他意義的,則兩者並存。如「尿」、「溺」。

按上列原則所選出之形體,當作「字形」, 其他各形體則列為「參照字形」。

表三《中文電腦基本用字集》前 500 字之累積使用頻率表

累積字數	累積頻率	累積字數	累積頻率	累積字數	累積頻率
5	9.24%	50	32.39%	372	+70%
10	14.20%	60	35.22%	472	+75%
15	17.79%	80	39.92%	500	76.27%
20	20.54%	100	43.74%	1000	89.38%
30	25.13%	141	+50%		
40	29.02%	232	+60%		

註：“＋”號表示「正好超越」, 如＋50%表示前 141 字的累積使用頻率剛剛超過 50%

(二)、交大字根系統

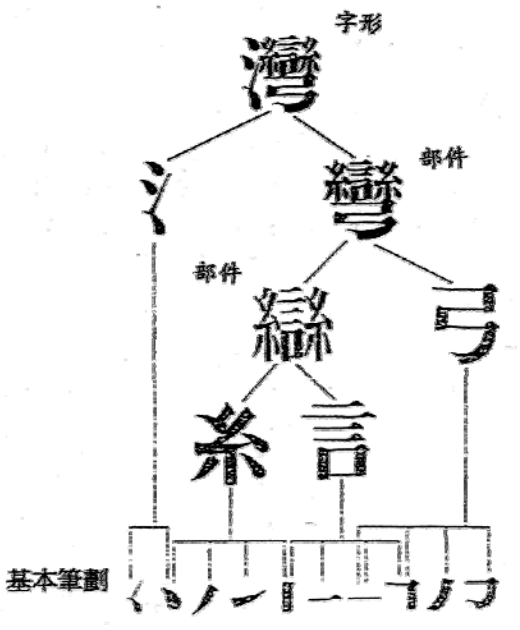
在台灣, 最早分析漢字字根的是交通大學, 這套字根就命名為《交大字根集》。1972 年, 倪耿在碩士論文《中國文字之結構模式及其分析》中, 測試了十幾種常用的構字方式, 並統計其使用之頻度, 發現只用橫向連接、直向連接和包含這三種組合方式, 就可以大幅化簡漢字的結構, 餘者甚少用可略而不計。據此, 漢字結構的字根制式表達方法確定了下來, 如〔表四〕所示。

表四　漢字字形結構的制式表達（以Bakcus Normal Form表示）

〈漢字集〉 ：：＝ 〈漢字〉|〈符號〉
〈符　號〉 ：：＝ 含標點符號、注音符號、英文字母、阿拉伯數目字及
　　其他專業符號等, 多少不拘。
〈漢　字〉 ：：＝ 〈字根〉|〈部件〉|〈漢字〉〈定位符號〉〈漢字〉
〈部　件〉 ：：＝ 〈字根〉|〈部件〉〈定位符號〉〈部件〉
〈定位符號〉 ：：＝ 橫連符號、直連符號、包含符號
〈字　根〉 ：：＝ 496 個, 詳見〔表五〕

〔表四〕中的構字系統是以BNF（Bakcus Normal Form）的制式語法（formal grammar）表達的。它是一種孳生系統（production system）, 亦即：可由最底下的具體符號, 如〈字根〉、〈定位符號〉這些符號集合(set)中的成員, 依表中的式子孳生為新的符號, 如漢字或部件。一般來說, 這種孳生系統能產生的, 常多於實用上所需要的, 所以必須加一些使用情境（語意）的約制, 來配合現實的環境。譬如：此系統產生的形, 究竟是部件？是字？或者什麼都不是, 是使用

者可以判斷選擇的。也正因如此，賦予了此系統不須更改便能應付未來造新字的彈性。在此系統中，〈字根〉是〈部件〉的子集合(subset)，代表最基本、不可再分解的一群部件。字根組成部件，部件組成漢字。漢字、部件、字根都可用來組合成構形更複雜的漢字。由於〈漢字〉、〈部件〉、〈字根〉這三個集合互有重疊之處，在使用上，〈部件〉常用來指不屬於漢字也不屬於字根的那一群成員。如〔圖一〕中的例子，灣和彎是漢字,絲屬於部件，弓、言、系是漢字也是字根，氵是字根。圖中字根與筆劃部份的關係容後說明。



圖一:灣字的構成

倪耿的工作是和林樹整理《中文電腦基本用字集》同時做的，字根的確定是雙方共同努力的結果。這套字根有一特色，它是經過反覆三次「最佳化」而得到的。所謂最佳化，是指在字根總數和平均每個字分解的字根數目（經使用頻率加權計算）兩者之間，求一近似最佳的結果。通常，字根越多，每個字分解後的字根數就越少；字根越少，則每個字的字根就越多。在使用方面，我們是既希望字根少，也希望每個字的字根少，所以在此不可兼得的情形下，只有求其最佳之組合。此最佳化是經過一些數學計算（稱為邊際效用原則）而決定的。其結論略如：凡一個字其使用頻率在萬分之 37.58 以上時，不應分解，在萬分之 18.79 至 37.58 之間者不可分解為兩個以上的字根，在 12.36 至 18.79 之間者不可分解為三個以上字根，在 9.39 至 12.36 之間者不應分解為四個以上的字根，餘者無論怎麼分解，沒有大礙。【註三】這個結論，也決定了分解漢字的底線。

依據《中文電腦基本用字表》中之 9129 個字形，逐字分解而得 496 字根，請見〔表五〕，其中含有常用熟字 305 個（其中 39 個是合於不可分解條件者），其總計使用頻率已超過全部的 50%。若依字根的使用頻率統計，最常用的前 25 個字根已佔全部使用頻度的 30%，前 50 個為 49%，前 100 個為 66.7%，前 200 個為 84.9%，前 300 個為 95%。【註四】

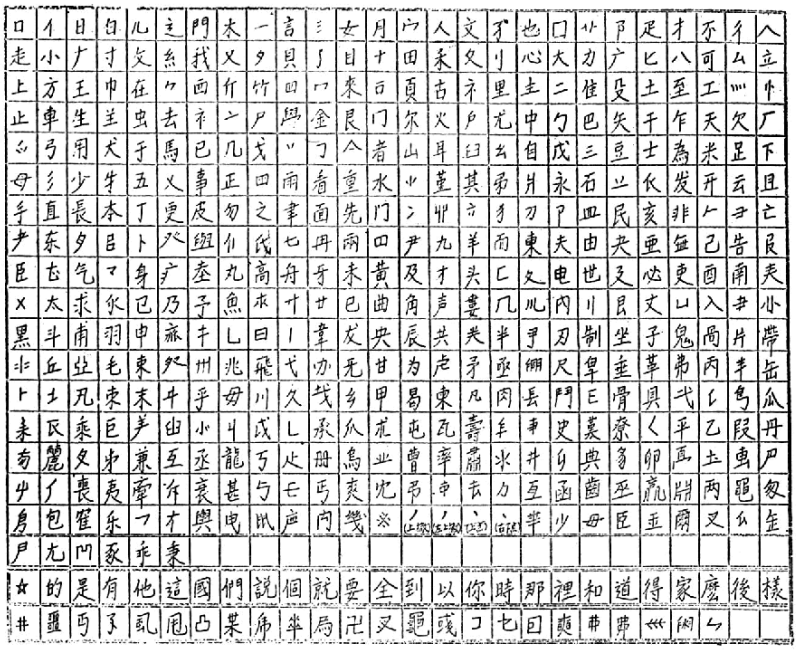
最佳化的結果是，依使用頻率加權後的平均每個字的字根數僅 1.9。至於這套字根構字的模式可以適用到什麼程度呢？將張其昀等編纂的《中文大辭典》中 49905 字逐字核試，計可組合 48713 字，僅 1129 字難以組合，這些字多為籀文，篆字，或一些古文、反文、圖騰之類者。今多有替代，或早已廢棄不用。經考慮後，並未將此納入交大字根系統中。如日後有此需要，再將此 1129 字再納入不遲。此構字的性質正是本系統所希望的，所以本系統亦以此為基礎並加以擴充。

字根系統是漢字字形結構中的一個基本部份，它是依楷書發展出來的，並沒有考慮到篆、隸、行、草等字體，也沒有顧及書法及各種印刷字型的變化。它能掌握的，只是從楷書外觀上來描述漢字的構成，即使如此，這構成的信息還是上述各字體、字型的共同部份，是可供各體參考利用的。至於各體的字形變化描述，詳如後文。

【註三】此邊際效用之計算，請參考：謝清俊、黃永文、林樹，《中文字根之分析》交大學刊，第六卷・第一期，1973 年 2 月

【註四】 關於 9129 個字形之分解及字根之使用頻度之資料，請參閱劉達人、杜敏文、謝清俊、張仲陶、蔡中川、林樹《漢字綜合索引字典》Asian Associates, Bedford, New York 1979

表五　交大中文字根表



說明：1. 本表依字根出現頻率之高低由左而右，由上而下順序排列。
2. ☆爲酌留常用字，"#"爲罕用字根。
3. 本表計收字根 448 個，酌留常用字25個，罕用字根23個，總計 496 個。

(三)、目前電腦的文字知識

時下電腦系統中已存入的漢字信息有：交換碼中的字樣、字碼和字序；字型庫中的各體字樣，以及它們和交換碼的對映；輸入系統中的輸入碼及一些漢字屬性，以及它們和交換碼的對映。這些信息及彼此的關聯有如〔圖二〕。

1	輸入碼(倉頡)	水廿中人	十子	人卜一口	竹山心	手一女	卜土廿手	輸入系統
2	中文碼(Big5)	BA7E	A672	AB48	AEA7	AAED	B946	
3	標準字樣(楷書)	漢	字	信	息	表	達	交換碼
4	隸書	漢	字	信	息	表	達	
5	仿宋體	漢	字	信	息	表	達	
6	明體	漢	字	信	息	表	達	字體庫
7	細圓體	漢	字	信	息	表	達	
8								

圖二、時下電腦系統中漢字相關信息的示意圖

交換碼中雖有字序的信息，但常常並不完整，如前述的造字表就割裂了原有的字序。此外，有些碼並無一致的字序，如 JIS 的前半照注音排序，後半卻照部首筆劃排，這並不表示字序的信息豐富，反而是各個排序的信息都只有一半，而使得電腦無法依某種規則對所有文字排序。這種字序的實用價值是很少的，幾乎有等於無。再說，幾乎所有的系統都沒有把字序的一些重要指標的信息整理出來。譬如：第十劃的字是從什麼碼排到什麼碼，某部首的字是從那到那兒等。沒有這些指標，使用者怎麼用字序信息呢？

有些輸入系統存有一些漢字的屬性，如注音輸入法中就有每個字的音標。可是，這些音標多半沒有收集全所有的破音，也沒有處理破音的方法。請問要輸入「長」字時，該如何注音？因此，這部份信息只是勉強來應付漢字輸入的，並沒有為各種應用設想。更有甚者，有些系統根本不讓使用者取用這部份信息。

由以上的情形看來，電腦中的漢字信息不僅少得可憐，更是先天失調。這就怪不得做應用時礙手礙腳了。這實在是中文電腦應用上必須排除的絆腳石。

二、文字的制式定義與表達

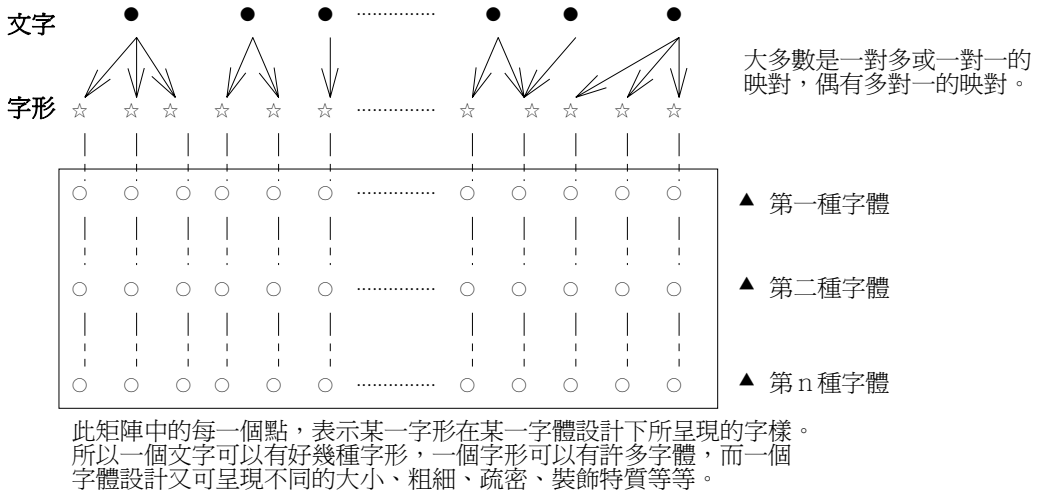
字、字形、字體這些名詞在我們語用中常常代表著不盡相同的意思，但佐以情境，我們並不覺得有溝通的障礙。可是，電腦遠沒有人靈活，也沒情境可參照，所以必需對這些名詞作制式的界定，電腦才能據以順利地處理文字信息。本節將說明一些關鍵詞在本系統中的工作定義。

字(character)是表達一種或一群概念的名相。它是抽象的，以語意區別。例如，對應的繁體和簡體是同一個字。在電腦中，字用一個識別碼(identifier)表示，此識別碼可以是交換碼，也可以是內部處理方便使用的「內碼」，或是輸入時的「輸入碼」。為方便計，以下的討論均以交換碼來代表字。目前，字所承載的語意還沒能表達在系統中，所以電腦並沒有方法可以直接處理語意信息。

如前所述，一個字可能有許多字形(glyph)。字形也抽象的，區別字形的關鍵在於它的組成結構，亦即構字，如前例，繁體和簡體屬於不同的字形。偶爾，也有些字會用同一字形的。所以，以數學關係來說，字之於字形大多數是一對一或一對多的關係，偶有例外。

字形只界定構字，並不關心該字好不好看。依同一規範製作的一群字屬於同一種字體(font)。字體也是抽象的，區別的關鍵在於它的設計規範。雖然字體有設計規範以表現其劃一的特色,但仍有藝術創作的空間，允許設計者表現自己的風格。所以，同一字體下各廠商設計的「字型（style）」會現出不同的表情、風貌。一種字型設計，通常有些參數來決定它呈現的大小、粗細、橫直粗細比列、疏密以及一些特殊裝飾的邊角等等。待這些參數選定了，才能借媒介呈現出這個字的面貌，此稱為字樣(typeface)。唯有字樣才是具體可見的。照理說，這些字體和字型在設計上產生的形狀變化（以下簡稱為字體變化）是不應該違反構字規律（即字形的定義）的，然而在實務上並沒有這麼嚴謹，也造成了些字形上的差異，詳細的分析如後文。

上述的關係可參見〔圖三〕。所謂文字的制式表達，即將〔圖三〕中的關係用電腦能了解的方式，表達在電腦中。字的表達已如前述。字體的信息存在字體庫(font library)中，這是大家熟習的，毋庸多言，參見〔圖二〕。目前電腦中無字形信息，或者說是字與字形不分，混淆著用，所以無法分別及處理異體字。字形資料庫就是要填補這個空缺，它擁有字與字形間關係的對映，以及字形的結構模式，如下文。



圖三：字、字形、字體和字樣的關係

三、字形模式

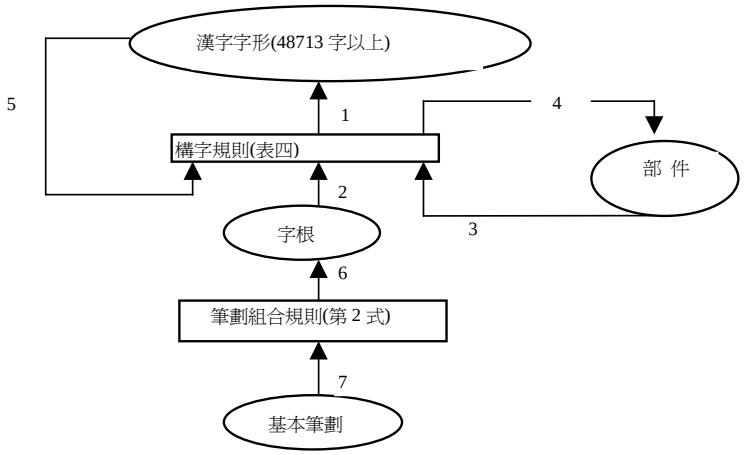
本系統之字形結構模式如〔圖四〕，〔圖四〕中, 12345 諸線連接者即〔表四〕中之構字規則。〔表四〕中之「漢字」即此圖中之「漢字字形」。67 連接者表示由筆劃組成字根的關係。一個字形的組成可用下面兩個公式表達。

假設 G 表示字形, R 代表字根, K 代表部件或 G 或 R, T 代表基本筆劃, 而 p 和 s 分別表示(T 和 K 在字形中的)位置(position)和大小(size), 則:

G＝ΣK （p, s）.....（1） 表示一個字分解至字根階段的字形構成。

R＝ΣT （p, s）.....（2） 表示一個字根和構成它的筆劃之間的關係。

其中(1)式中 G.K.R 的關係要符合〔表四〕的規則。至於基本筆劃, 其數目約為 30 至 100 之間, 依字樣美化程度及字體之設計而異。



圖四:本系統之構字模式

四、字形和字樣變化的表達與區分

依此字模式, 字形的變化雖多, 卻可歸納為筆劃的變化A、字根或部件的變化B和整個字的變化C等三個等級。其大要如下：【註五】

A、筆劃的變化函數

A_1：一筆劃位置改變, 筆劃數和構字的字根不變。 如:羊→羊 , 片→片

A_2：一筆劃尾部加勾, 筆劃數和構字的字根不變。 如:七→七, 不→不

A_3：一筆劃被另一種筆劃替代, 筆劃數與字根不變。 如：刃→刃, 言→言

A_4：增多一筆, 筆劃數增1。 如：者→者

A_5：減少一筆, 筆劃數減1。 如：德→德

A_6：一筆劃由另二筆劃取代, 筆劃數加1。 如：氏→氏

A_7：二筆劃由另一筆劃取代, 筆劃數減1。 如：此→比

A_8：一群筆劃由另一群筆劃取代。 如：四→四

B、字根或部件的變化函數

B_1：一字根R_1由另一字根R_2取代, 而R_1和R_2的差異只是筆劃上的變化
（如前述A_1至A_8之變化） 如：寺→寺, 吉→吉

B_2：一字根R_1由另一字根R_2取代, 而R_1和R_2之差異不屬筆劃上的變化。 如：耻→恥

B_3：一部件（一群字根）由另一部件取代。 如：却→卻

B_4：增多一字根 如：果→菓

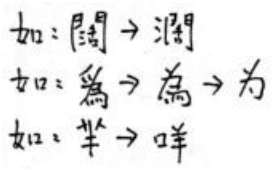
C、整個文字構字的改變函數

【註五】詳見謝清俊《On the Formalization of Glyph in Chinese Language》世界字體會(AFII)會議，東京，1990 年 2 月

C1：字根不變而組合改變者。

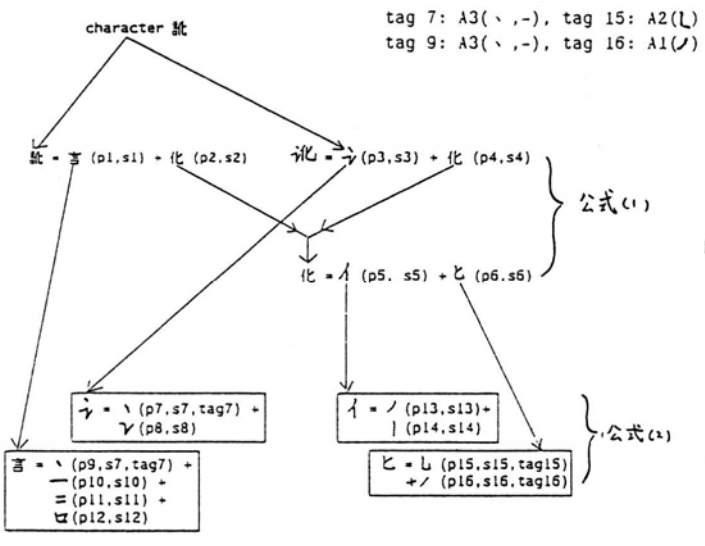
C_2：由簡化而改變者。

C_3：不規則變形者。

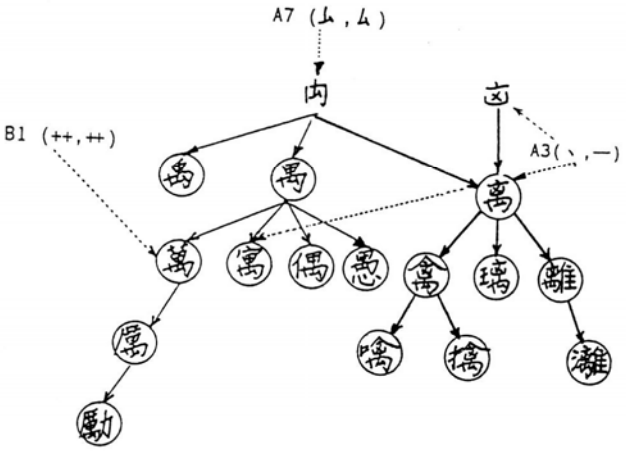


如果我們觀察一下各國的漢字交換碼,或比較一下各種設計的字體、字型將會發現很多字形的差異是屬於A_1至A_8及B_1類的微細差異（micro-difference）【註六】，一如《玉篇》中〈分毫字樣〉所列者。不同的是，〈分毫字樣〉所列的是不同的字，而A_1至A_8及B_1所造成的差異是同一個字，而這些差異，都是由於字體、字型設計的差別造成的。這些細微差異的「形」，若每個都造一個形、給一個碼，那會多得無法應付。所以，在本系統中它們都歸屬同一字形，只用A_1至A_8及B_1來標示差異，稱之為一個字形的「分毫字樣」（micro-difference variant）。至於B_2、B_3、B_4及C_1、C_2、C_3這些函數的變化，將產生不同的字形，即異體字。

如果一個字有許多字形，本系統允許字集選擇其一作為該字的代表，餘均稱為異體字。例如：大陸選用簡體字放在國家標準中，而台灣選用較傳統的字形，放在交換碼中，日本、韓國亦各有主張。這些使用上的彈性，都是本系統允許的，並無差別待遇。字和異體字之間的關係在電腦中用關聯資料庫的欄位表達。至於字形的孳乳則用樹狀資料結構表達而構成字形的家族。讓我們用兩個例子來說明上述的文字形知識表達方法。〔圖五〕中表示一個字「訛」，它有簡繁兩個字形(B_2類的差異)，字形再分解為部件、字根，字根再折成筆劃。在筆劃上的差異用標示（tag）注在該函數內。〔圖六〕是一個字根「冂」字形孳乳樹的一部份。圖中，有圈的是字形，沒圈的是部件或字根。虛線所指者，表示該形可能的筆劃差異，而凡在該節點以下的分枝中所有的形，亦均可能有差異。由此可知，筆劃變化的函數可以有系統地表達其變化，可以共用，不必每個形皆加注。〔圖五〕和〔圖六〕的形式都是從字形家族樹中推導出來的。



圖五：訛字相關的字形組成



圖六：字形孳乳樹之例

五、文字屬性的表達

【註六】micro-difference一詞爲Edwin Smura先生所首用。

所謂文字的屬性，是指字的一些性質歸屬。傳統字書及今日字辭典中所列者皆是，此外信息處理上用的，如各種輸入碼；統計上的，如使用頻次；語言學上，如詳細的詞性分類；諸如此類皆可納入字的屬性之列。

在本計畫中為缺字收錄的文字屬性如〔表六〕。這些信息是用關聯性資料庫製作的。這部份的電腦技術是相當成熟的，故從略。字集的屬性資料略同於〔表六〕所示，然其細節尚待斟酌。

表六　文字屬性欄位表　　（註:打"*"者，可以重複）

甲、缺字屬性表

1.缺字統一編號	* 5.筆劃數	* 9.注音
2.交換碼	6.首筆	*10.異體字交換碼
3.內碼(造字檔內)	7.次筆	*11.登錄日期及修改記錄
* 4.部首	8.末筆	*12.提供缺字之各單位欄位（含編號及內碼）

乙、字形結構屬性表

1.所屬字集編號	* 5.筆劃	9.部件二
2.交換碼	6.首筆	10.部件三
3.字形碼	7.分解方式	11.字頻次
* 4.部首	8.部件一	12.字根頻次(當用爲字根時)
		13.字根次(當用爲字根時)

參、系統設計與實務

前文談到問題的;背景和一些觀念、理論，在此報告解決缺字問題的實踐。我們將構字模式、校勘過的《交大字根系統》及《中文電腦基本用字》放在電腦中構成「字形資料庫」，希望以此為基礎發展。【註七】 談電腦的實踐，免不了涉及些較專門的工程技術，但本文重點不在此，故仍以字形相關的實務為此報告的重點。

一、系統概述

本系統是在 IBM 相容的個人電腦中，台灣版中文視窗 3.1 作業作業系統下，用 Visual Basic 程式語言發展。系統中有兩個資料庫，《字形資料庫》及《文字屬性資料庫》，都是用 Foxpro 資料庫管理系統建立的。

這個系統有幾個用法。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。

目前系統已經輸入了三個字集，其一是校勘後的《中文電腦基本用字研究》，共收 8529 個字，另有參照字形 593 個，部件 629 個，字根 457 個。其二為前述字集再加上大陸的標準簡化字【註八】，字數不變，參照字形增加為 2284 個，部件 664 個，字根 492 個。第三個字集為《電子佛典補字集》，只收錄一些機構製作電子佛典時所缺的字，目前有 500 餘字，字數尚在增加中。

系統中的文字、字形、部件、字根、筆劃之呈現方式如下：

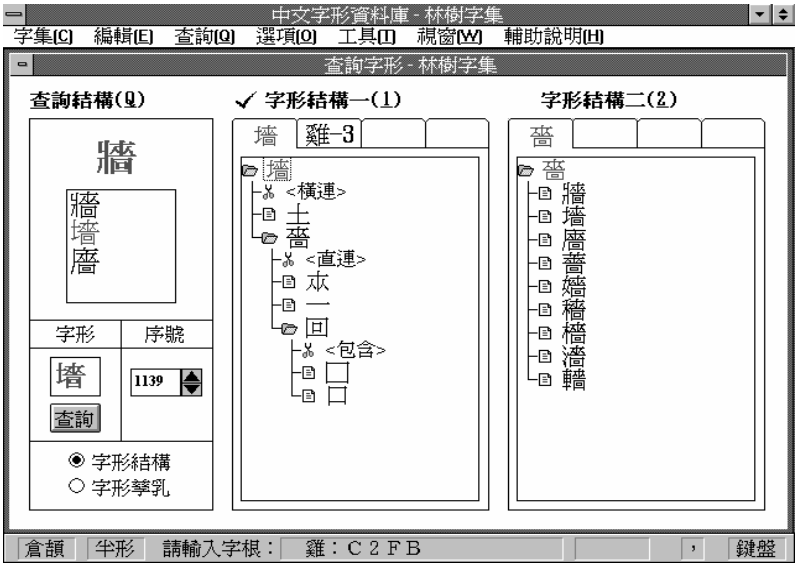
【註七】 謝清俊、莊德明、張翠玲、許婉蓉，〈中文字形資料庫的設計與運用〉， 中國文字學會年會，台中 1995

【註八】 依照《簡化字總表》(1986 年版)，共取簡化字 2235 個，及相對應的繁體字 2259 個。
蘇培成編著，《漢字簡繁體字對照字典》， 台灣珠海出版有限公司, 1994

1. 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。
2. 字形：原則上不造其形，以字根結構式呈現。
3. 部件：在系統造字區內造其形，並定其碼。
4. 字根：字根集合中有269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。
5. 筆劃：在系統造字區內造其形，定其碼。

依本系統的設計，可以由構字上觀察到字形，由標示中觀察到筆劃的變化；卻看不到實際的「字樣」。目前，系統只完成〔圖四〕中的字形結構部份，至於由筆劃組成字根的部份則尚未完成。

字形資料庫的檢索歸納為字形查詢與頻次查詢兩個介面。字形查詢時可查字形結構和字形孳乳，但請先注意字與字形的區分。例如查詢林樹字集中的「墻」字時，視窗畫面上會看到「牆」、「墻」及「廧」三個字形， 以「牆」為該字集的標準字。其次，並非所有的字形皆可看到實際的「字樣」，原則上，我們並不為這些字造形，而是以構字式表示。例如：查詢「雞」時，可看到「雞」、「雞-3」及「雞-4」三個字形；其中字形「雞-3」為「奚」橫連「鳥」，「雞-4」則為「又」橫連「鳥」。字形查詢的畫面如〔圖七〕，以下簡單列出查詢的方法及注意事項。



圖七、字形的查詢

1．字形結構區有兩個，每區可存放四個結構，結構區的切換可用滑鼠按一下結構區或是結構的標題。
2. 用滑鼠按一下「字形孳乳」或「字形結構」，可切換兩者間的查詢。
3. 結構的瀏覽方式，除了利用捲動軸外，尚可利用鍵盤操作；其中方向鍵↑或↓可選擇上一個或下一個字形；方向鍵←或→可捲動視窗以看到左邊或右邊的內容；<Home>或<End>鍵可選擇第一個或最後一個字形；<PageUp>或<PageDown>鍵可到上一個或下一個視窗。
4. 結構節點圖像 表示其下節點已經顯現，而圖像 表示其下節點仍然隱藏，至於圖像 則表示其下並無節點。用滑鼠按一下圖像 或 可切換彼此間的狀態。(或用<+>鍵以顯現, <->鍵以隱藏)
5. 結構節點上的字形，皆可用滑鼠拖曳到字區以執行查詢。
6. 用滑鼠連續按兩下**字形**區中的字形，即可作字形結構或孳乳查詢。
7. 用滑鼠連續按兩下字形結構中的節點字形，即可作字形孳乳查詢。
8. 用滑鼠連續按兩下字形**孳乳**中的節點字形，即可作字形**結構**查詢。

頻次查詢時，應先示明所選之字集、查詢之對象及排序方式。查詢的對象有字、字形、異體字、部件及字根等，排序則可依字碼、序號、字頻、字根頻、字根次等排列。至於字根頻及字根次的意義可以林樹字集中的字形「故」、「做」為例說明：

字形	字頻	字根頻	字根次
故	1685	6011	2
做	4326	4326	1

字形「故」、「做」的頻次分別為 1685 及 4326；在林樹字集的字形中，使用到「故」的只有「故」、「做」兩個字形，所以「故」的字根次為2，字根頻則為 1685 + 4326 = 6011；至於「做」的字根次為1 (「做」本身)，故其字根頻為 4326。字根次的計算單位為字形，部件及字根並不含在內 (除非其本身即為字形) 。頻次查詢的畫面如〔圖八〕。

圖八、部件頻次的查詢

頻次表的搜尋功能只針對目前的排序項目。例如欲搜尋字形，必須依內碼作排序；序號、字頻、字根頻、字根次的搜尋則須依序號、字頻、字根頻、字根次分別作排序。數字的搜尋並不需要完全一致，例如在下表的林樹字集中，搜尋字根次 100，此時會找到字根次 104 的資料。

字形	字頻	字根頻	字根次
犭	0	4023	97
鳥	0	2769	98
乂	0	20791	99
馬	1509	8060	104
彡	0	6470	104

頻次表的第二欄位為字形碼，可用來表達字和字形間的關係。例如林樹字集中的「雞-1」、「雞-3」(「雞」的第三個字形，字形碼 2 保留給相對應的大陸簡化字)及「雞-4」(「雞」的第四個字形)，其中「雞-1」即為「雞」，「雞-3」及「雞-4」必須由結構看出。它們在資料庫中的表達方式如下：

字形碼		分解	部件	部件
雞	1	⿰	奚	隹
雞	3	⿰	奚	鳥
雞	4	⿰	又	鳥

然而對於字形「牆」及「墻」，假定「牆」為標準字，它們在資料庫中的表達方式如下：

字形碼		分解	部件	部件
牆	1	⿰	爿	嗇
牆	2	=	墻	
墻	0	⿰	土	嗇

瀏覽頻次表或查詢結構時，只會列出「墻」，而不列出「牆-2」(小於2的字形附碼通常省略) 。字、字形、異體字、部件及字根於資料庫的表達中有下列的特徵：

1. 字　：字頻大於 0 而且字形附碼為 1
2. 字形：字頻大於 0
3. 異體字 ： 字頻大於 0 而且字形碼不等於 1
4. 部件：字頻為 0，其使用頻次顯示在字根頻欄位中，其結構可繼續分解

所以在資料庫中，可用來代表字者，必然是字形，也可能為字根。以下簡單列出查詢的方法及注意事項：

1. 選定類別及排序後，再用滑鼠按一下「重新整理」，即可更新頻次。
2. 輸入搜尋條件後，再用滑鼠按一下「查詢」，即可執行查詢。
3. 鍵入<Ctrl-PageUp>，可回到頻次表的第一列；<Ctrl-PageDown>，可到頻次表的最後一列；<Up>鍵，可到頻次表的上一列；< Down>鍵，可到頻次表的下一列。
4. 畫面右下角的數字表示字、字形、異體字、部件或字根的總數。

二、字形實務

本節討論的是與字形相關的一些重要問題，包括基本資料的整理、字形知識的表達、字形變化的標示，以及兩岸字形上的一些實務問題。茲分四節陳述如下。

(一)以往資料的整理

《中文電腦基本用字》和《交大字根系統》都是廿五年前資料,現在已經無法取得其電子版，所以只好重新整理。但這樣也好，順便可做一番校勘。這廿五年來，電腦的變化甚大，致使許多設計上的考量也變了許多，這些差異，也可在校勘時斟酌修正。

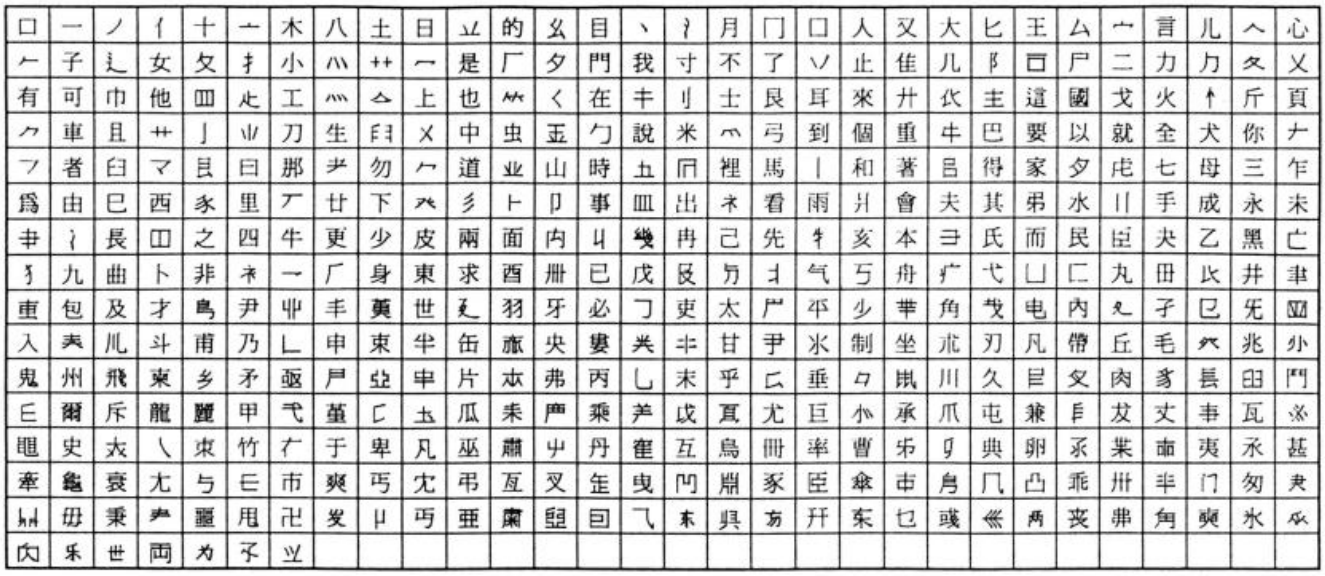
我們重新將《中文電腦基本用字》的 9129 字形逐一分解，檢查是否有分解得不很自然的地方。若有，則不惜加字根或部件。這是因為今日的個人電腦記憶體，已比當年大了一千倍以上的緣故。同時，也校勘錯字。由於當年匆匆付印，《中文電腦基本用字的研究》一書的字表中，有 40 字有誤。這些錯誤在編印《漢字綜合索引字典》時，曾予以更正，但未留下此四十字的勘誤表。這些也在此次整理中一並解決。【註九】。

整理後，發現有三個字（六個字形）重複並無法肯定，予以刪除。此外，有一異體字誤植，亦予以刪除。故實得 8529 字，共計 9122 字形。重新核對後的字根共 457 個，部件 629 個，分別如〔表七〕、〔表八〕。這兩表的字形，在五大碼中所無者，都造在五大碼中。校勘後的《中文電腦基本用字》之熵值為 9.60。加權後平均每字的字根數為 1.9，這些數據和往者差別不大。

【註九】詳見許婉蓉，〈林樹字集的更正及問題字〉，中研院資訊所文獻處理實驗室，1996

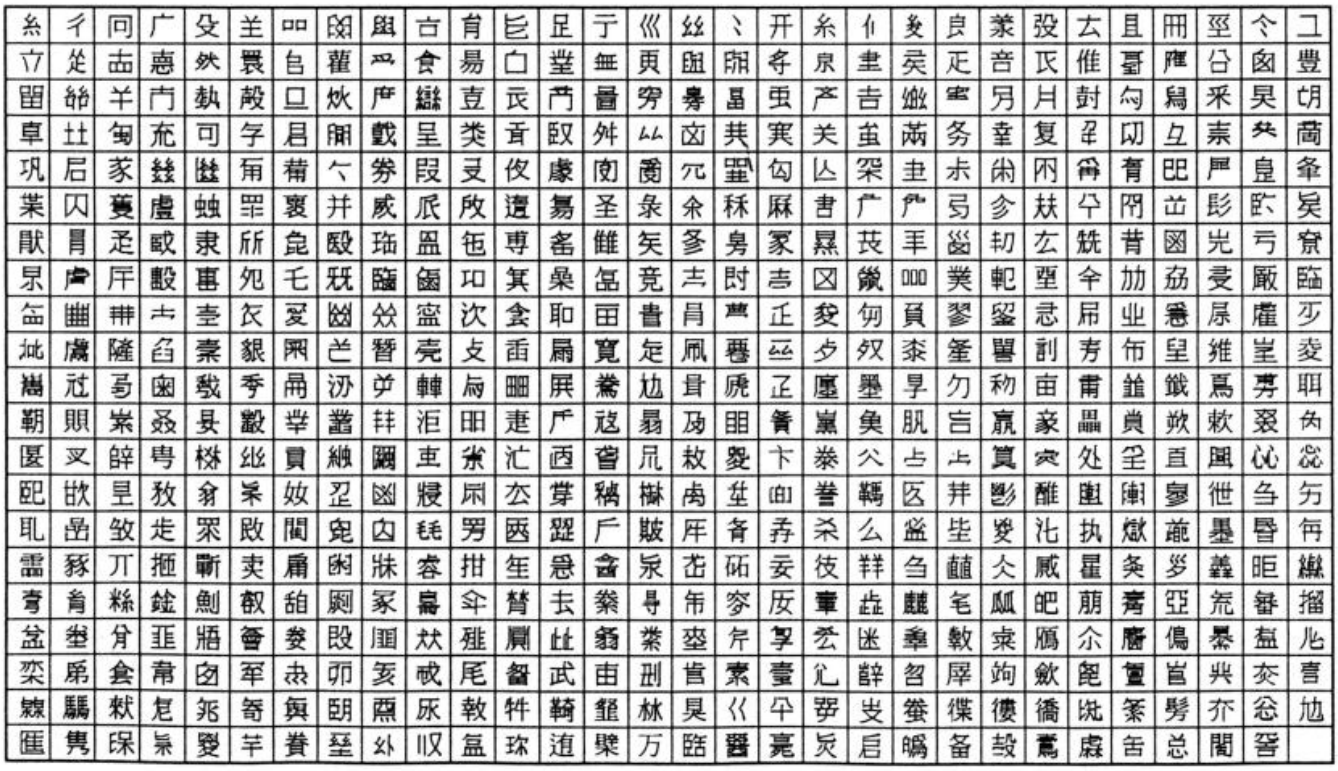
表七　校勘後《交大字根》表　（依字根頻次排序）

85年8月22日製



表八　校勘後《交大部件》表　（依字根頻次排序）

85年8月22日製



（二）字形的表達與呈現

讓我們先用例子，來說明一些名詞和符號，以及構字信息在電腦中的表達方式。〔圖六〕最右一枝中的一些字在電腦中的表達如下：

1.灘＝氵⿰又離
2.離＝离⿰又隹
3.璃＝王⿰又离
4.擒＝扌⿰又禽
5.噙＝口⿰又禽
6.离＝卤⿱厶内
7.禽＝人⿱今离
8.隹＝亻⿰又圭
9.圭＝𠂉⿱今土
10.卤＝𠂉⿱今卤
11.凶＝凵⿱㐅乂

1 至 11 這些式子稱為「構字式」，其中二嚞离分別表示橫連、直連、包含等定位符號。電腦可依構字式，逐次追蹤以消去式中的字或部件，而得到完全以字根表示該字構成的「字根構字式」。如(3) ：

灘＝氵⿰又（离⿰又隹）

＝氵⿰又（（（卤⿱今内）⿰又（亻⿰又圭））

＝氵⿰又（（（𠂉⿱今卤）内）⿰又（亻⿰又（𠂉⿱今土））

＝氵⿰又（（（𠂉⿱今（凵⿱㐅乂））内）⿰又（亻⿰又（𠂉⿱今土））……..(3)

(3)式中的「灘」是八個字根構成的，若省去定位符號， (3)式可寫為：

灘＝氵𠂉乂凵内亻𠂉土 ……..(4)

(4)式稱為灘的「字根序」。比照此法，1 至 11 式也可將定位符號省略，省略後的式子稱「部件序」，或「字根序」。

構字式中擁有字形結構完整的資料，部件序和字根序則否。雖然如此，字根序的判別能力還是很強的，在林樹字集中，只有八對字沒法以字根序判別，如（唄、員）。是故在判別字時，用字根序不失為一良法，它比構字式單純得多。可是，在實用上它還是嫌太長。部件序雖然所含的字形信息最少，但是它最簡潔：通常只有二個部件（偶有三個），而且省略的一個定位符號絕大部份可由前後部件的性質中判斷出來（用電腦）。所以，用它來表示所缺的字是很方便的。

我們可以把一個字集中所有字形的構字式放在字形資料庫裡，這樣便完整地將該字集的字形信息（或知識）放入了電腦。以《中文電腦基本用字》為例，共有 9756 個構字式，其中 8529 個屬字集的，593 個屬異體字， 629 個屬部件的。若將該字集中對映的簡化字也放進去，則有 2284 個屬異體字的構字式，部件加多了 35 個共計 664 個，而 8529 個屬字集的依舊不變，總計 11477 個構字式。

（三）字形變化的標示

用 SGML 來標示缺字的方法，是由 Wittern 及 App 首先提出的【註一】，稱為「漢字位標」（Kanji Placeholder，簡稱位標）。漢字位標規定以「＆」起頭，以「；」結束，而夾在其中的字串分為兩個部份，分別表示該字形在某一碼的某一位置。

利用此位標，就可以跨不同的碼來互補缺字。Wittern 及 APP 建構了一漢字庫（KanjiBase），包括 JIS, BIG5（即五大碼）, Unicode 及 CNS 等。其中 CNS 有 4 萬八千字，字數最多，Unicode 次之。若平時作業中遇有缺字，則首先在 CNS 中查尋，如果有該字字形，則用位標將 CNS 之碼標明在文件中；若無，再 Unicode。例如：「&U4AB5」表示一缺字；其中 U 表示 Unicode 字集，4AB5 表示該缺字字形的 Unicode 碼位。

這個方法可以減少缺字，是有用的；而且它採用了國際的標準作為換碼的控制，有助於大家採用流通。然而，依前言中的討論，這個方法並沒有完全解決缺字所帶來的諸多困擾。再者，由

於「＆」被用為控制碼，雙語文章中若要用「＆」時怎麼辦呢？善用 SGML 是解決缺字問題有力的方法，本系統將修正漢字位標的作法，提出更有效的方案。

漢字位標是以 SGML 的標示（tag）為載具，用其他字碼中的字形來表達缺字。仿此，我們可用構字信息，即構字式、字根序或部件序來表達缺字，稱為漢字形標（Kanji Glyphholder），簡稱形標。以部件序為例，說明如下。

假設⿰表示標示的起點（open delimiter），⿸表示結束（close delimiter），則佛經中的阿閦佛可用：阿⿰門人人人⿸佛的字根序表示；如果字集中有「⿱亠从」字，也可用阿⿰門⿱亠从⿸佛的部件序表示。這樣的表達方式，可以免去查其他字碼的工作，而部件式所含的信息足夠作為認別該字缺字之用。

同樣的方法也可用來標示異體字。所用的方法是在構字式（或字根序、部件序）之後用一「・」作區隔，再接上該字的字形碼。所用的標示與上相同。如，「艹⿰樂・3⿸」表示「艹葯」，若葯是藥的第三個異體字。

形標並不排除位標，Wittern 和 App 的漢字位標也可以放在⿰和⿸之間的。所以，形標包含位標的功能，比位標更方便，是位標的延申。

(四)兩岸字形的共享

為了兩岸字形能夠互通共享，我們嚐試將簡化字及大陸的字根與「部件」納入系統中。在簡化字方面，是採用蘇培成根據 1986 年《簡化字總表》所編的《漢字繁簡對照字典》中收錄簡化字。目前已填入與《中文電腦基本用字》中能相互對應的 2017 個，餘下未收的 242 字正在造字樣，尚未填入【註十】。至於字根方面，則採用武漢大學《現代漢語定量分析》書中收錄傅永和先生〈漢字結構及其構成成分的統計及分析〉一文中所列的字根，以及 ISO/IEC JTC1/SC2/WG2/JRG N319 號文件〈ISO 10646 Additional Components 〉的字根。這些字根並未一次放入系統中，目的是要測試二者的相容程度。結果顯示兩岸的字根是十分相容；加了 2235 個字形後只增加了 35 個字根和 36 個部件，其中多屬簡化的偏旁及部件。

在字形結構方面,遇到了一個問題，那就是簡化後的偏旁或部件有使用情境的限制，如「讠」不單獨使用，亦不出現在字形的下方。又如，蘭簡化為兰，攔簡化為拦，讕卻簡化為谰，不是⿺。闌是簡化為阑的，是不是字典錯了，攔應簡化為搁？因為蘭不是闌呀。由於文獻不夠，我們很難查證，也無法獲得簡化字時所定的規則加以判斷。這種依情境而作不同簡化的情況，使得在孳乳結構中產生了些例外。也引發了界定「界體字根」的課題，換言之，沒有辦法簡單地用一字根替代另一字根而得到繁簡的對映，只有逐字登記其構字。我們相信簡化必有依據與規則，若可獲得這些規則或將之納入電腦，以便更有系統地表達簡化字形。由於有這種情形，也使得有些可以「類推」的簡化部件，在類推時產生猶豫，如婁簡化為娄，嶁、褸、廔等三字在字典中找不到，是不是可類推呢？或是這些字已刪除了？

其次，據聞大陸整理過異體字表，在簡化過程中合併了些異體字，如鉋、⿰、砲、炮合併成炮。這情形也因為我們缺乏了解而產生問題：究竟⿰、砲還在不在 GB 字集中？它們是消失了？或是仍算作異體字？如果能獲得異體字表及相關信息，我們就不會如此無知而不知該怎樣處理這些字。

【註十】許婉蓉，《林樹擴充字集中的簡化字》，文獻處理實驗室，中研院資訊所, 1996

此外，在筆劃方面也有些差異，如：「辶」與「辶」，「呂」與「吕」等。這些差異用分毫字樣的函數盡可表達，應該不會造成問題。只是《漢字繁簡對照字典》中找不到吕字，雖然在鋁、閭兩字中可以印證，但總是不放心。也許是我們用的《漢字繁簡對照字典》不太好罷，在資料收集上我們應更努力。

總之，繁簡字形合併在字形資料庫中，目前雖有些問題，但大體上是可行的、成功的。前述的困難加大了字形結構上的複雜程度，並不是不能實踐。

肆、應用

在本章中，將舉三個與缺字有關的應用依次說明。首先報告用本系統設法解決電子佛典缺字的計畫。其次，介紹利用字形資料庫改良資料登錄系統的規劃。最後，觸及根本問題，希望從本文的理論與經驗中，重新考量交換碼的設計。畢竟造成缺字問題的元兇，就是目前的交換碼。

一、電子佛典缺字的解決方案

佛教典籍多屬古書，其中常有當時通用之字形，與現代的「標準」字形並不一樣。在台灣，有許多機構正在製作電子佛典，一則為了流通，二來為了保存典籍。可是，缺字問題卻造成流通和保存上致命的障礙。為此，台灣大學文學院佛學研究中心邀集各機構，擬共編一套《電子佛典補充字集》來解決問題。幾經商議，同意由本實驗室負責技術工作。這個工作規劃實是一個典型解決缺字的方案。

參加機構所製作的佛教典籍，計有：《大般若經》、《禪藏》、《佛光大辭典》、《丁福保佛學辭典》、《律學辭典》、《大正藏第九冊》，以及近代的套書《妙雲集》等。各機構先匯集缺字，並查明其屬性後，送交本實驗室合併處理。

莊德明先生曾就送來的缺字予以分析，發現甚多是由於字體不同而產生的分毫字樣。這些字目前暫不處理，餘者分類如〔表九〕。表中，除第二類屬結構變化外，餘皆屬字根與部件的變化。這些字大多是異體字，可登錄在字形資料庫中，目前已存入 500 多字。【註十一】

這些字存入 Big-5 碼中，亦有規定，詳如〔表十〕。由於造字空間只有 5809 個，在使用上至少應分成公用造字區及專屬造字區。若造字使用頻次較高者，應置於公用造字區；頻次較低者，則依個別的需要而置入專屬造字區。然而在造字初期，若無頻次的統計資料，可先收錄佛經中非咒文的用字。佛典用字除了漢字以外，還有特殊符號，包括外文字母、流水號及佛典常用的符號。流水號的編號有「１、２、３……」、「一、二、三……」、「壹、貳、參……」、「甲、乙、丙……」、「子、丑、寅……」等，外框樣式則有「○、（）、●……」等，以五大碼有限的造字空間，自不可能將這些流水號通通收錄，目前只為流水號預留 100 個位子。

常用的外文字母包括梵文、巴利文、藏文及日文。造字區第四段 C6A1-C8FE 在倚天中文系統下另有定義，其中就包括了日文字母，為兼顧倚天及中文視窗使用的一致性，這一段我們不用。另外依據臺大佛學研究中心「梵巴藏羅馬字轉寫與中文軟體的相容問題」討論會議結論，有鑒於五大碼造字空間有限，目前只收錄 32 個梵文、巴利文及藏文的轉寫字母。規劃下來，真正能給補充字集用的有四千個字位，應該可以用一段時間。如果這四千個位置都滿了，則可用字形位標來借用其他碼的造字空間存放。

【註十一】 詳見莊德明，〈佛典共用造字集的規劃〉1995，及〈佛典共用造字集的整理〉，1996，文獻處理實驗室，中研院資訊所

表九：異體缺字的分類

類別	說明	字形
1	字根	「厂厽尉」→「廚」、「宀厽龍」→「寵」
2	位置	「山厽夆」→「峰」、「阝厽犇」→「鄴」
3	筆勢	「秊厽牛」→「犁」、「去厽刂」→「劫」、「乃厽木」→「朵」
4	古今	「屯厽阝」→「村」、「豸厽苗」→「貓」、「石厽导」→「礙」
5	累增	「艹厽果」→「果」、「小厽吝」→「吝」、「蕊→「蕊」
6	錯字	「亻厽且」→「但」

表十：Big5 造字空間的劃分

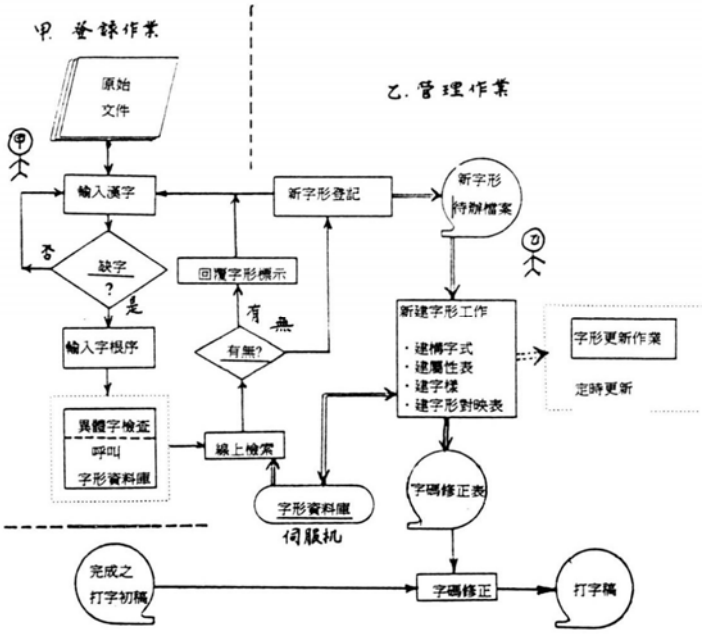
使用區	內容	使用序號	字數	保留序號	字數	合計
公用區	漢字			1-4000	4000	4000
	流水號	4001-4100	100			100
	轉寫字母	4101-4132	32			32
	其他符號			4133-4200	68	68
	字根與部件	4201-4908	708	4909-5400	492	1200
專屬區	倚天			5401-5809	409	409
合計			840		4969	5809

二、資料登錄系統之重新設計

資料登錄時常遇缺字，其後果與代價已如前述。為改善這些缺失，我們規劃了一個利用字形資料庫、網路以及視窗多功能等特性的新資料登錄系統，其流程示意如〔圖十〕。

圖中有兩種作業，即資料登錄和登錄管理。配合的作業員，甲的工作是打字，乙是缺字管理。一個乙可以同時和許多甲配合工作，數量多少視原始文件缺字情形而定。兩種作業用相同的視窗操作系統，並用區域網路連接，是故資料及控制的信息皆可線上互相傳遞。當缺字發生時，打字員只要鍵入字根序或部件序的形標，其餘的事便全由電腦或乙來完成了。

乙在螢幕上隨時可見到剛發現的缺字形標。如果許多甲都發現相同的缺字，電腦在「新字形登記」時會自動將重複排除。當乙收到缺字的形標，可以立刻翻閱文獻，將各項新建字形工作完成。這個程序要花些時間，但比起現在的人工作業已經快了許多，我們估計，就算半個小時應付一個新字，已經和以往費時數日的情形，不可相提並論。以後將字書存入電腦後，可以在線上直接查缺字屬性，這些工作還會加快。新增的字形，可定時更新字形資料庫，並修正打字初稿。這個系統目前仍在設計階段，尚未完成。



圖九：資料登錄系統流程圖

三、改良交換碼的建議

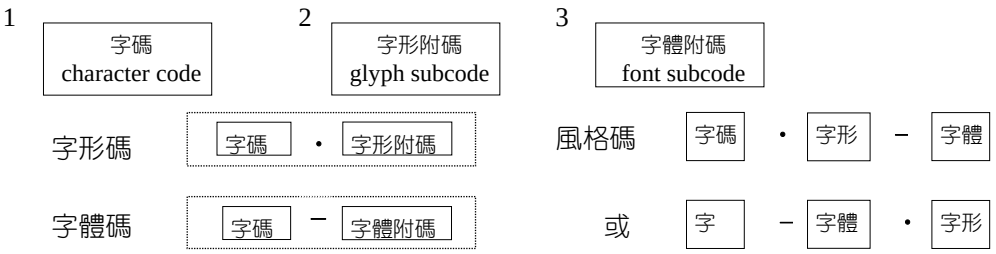
目前交換碼的根本病源，是錯把字形當作字。在此名實不符的情形下，名叫字碼（character code）而實際上卻以字形的差異來判別字。例如，五大碼中饑（C4C8）和飢（B047）分別佔兩個碼位。那麼它們是不是同一個字呢？依碼中的定義，不是；但在語用上實在是同一個字的兩個形罷了。這種情形在各交換碼中比比皆是，Unicode 和 ISO10646 也不例外。從文字學和語言學的角度看來，這種文字知識的表達根本是錯誤的，不健康的。從科學的角度來看，定義都錯了，往後推演可以勿論。像這種不健康的設計，導致目前應用上捉襟見肘，毛病百出，並不是意外的事。要改善交換碼，自應從此處下手。

此外，應知漢字集合從集合論的觀點來看是一個開放集合（open set），隨時可能增加字。它和封閉式的字母集合（close set）不同，字母是不會再增的。這些基本性質的不同，將導致其不同的性質和處理方式。可是，目前的做法卻是套用 ISO 646 處理字母的方法，來編漢字的碼。這真是削足適履。以上的說明和批評是希望大家了解，要改良現有的交換碼必須要爭脫既有的思考巢臼，否則不會有徹底解決的辦法。以下就是我們根據上述的原則，對改善現有交換碼所提出的方案。

（一）三段式的編碼

首先，我們建議對字、字形、字體分三段編碼，如〔圖十〕：

基本元素：



圖十：字碼、字形碼、字體碼、風格碼的關係

若字形附碼與字體附碼都用時，則前後不拘，此時若只選字體而無字型參數，則稱風格碼（style code），若含有字形參數，則稱為字樣碼（type face code）。字碼的使用與現在使用交換碼相用。字形碼、字體碼，風格碼或字樣碼等，則可以用前述之漢字形標載之與字碼一起用。字樣碼可能較長，因為其中可能包括許多字型的參數，這無關本文主旨，詳情略而不談。字形附碼只是一個正整數的編號。在本系統中它可以用來區別異體字，也可以用來存放某字集中所用的字形。例如，我們把《中文電腦基本用字》表中的繁體放在字形附碼為 1 的位置，把大陸標準簡體字放在 2 的位置上。這樣使用字形附碼的方式可以存放更多的文字信息。

(二)各碼表達的忠實度

這種編碼方式的使用可由其傳輸失真的程度上分為下列四種組合：

1．只用字碼時，失真最大。此時不在乎所用的字形或字體，如台灣和臺灣都是一樣，只求明白語義就好。許多應用至此已可滿足。
2．用字形碼時，表示還要求正確的字形；用字體碼時，重在選定字體（不在乎字形）。這情形比上列的情形失真少。
3．用風格碼時，失真又比前者少。不止字形要對，字體也不能錯。
4．用字樣碼時，幾無失真，字的大小、粗細、……使用者與接收者完全一樣。

以上四種情形可斟情使用以應付不同的需求。這是分成三段碼能提供的彈性。

（三）構字式的運用

每個字形的構字式都不相同，所以它可用作字形的識別碼。這情形在討論漢字位標及形標時已述及。此節說明此觀念在設計交換碼的可能做法。

用構字式作字形碼最大的好處，是它是一個封閉系統：只要有一個封閉的字根集合，配以〔表四〕的規則，就行了。如此便可省去成千上萬的碼位。《交大字根系統》用 496 個字根就能產生 48713 字形的碼，並且還有不需要修改系統就能對付新字或缺字的彈性，就是最好的例子。其次，構字式是一種知識表達，它不僅較數字碼易讀易懂，所孕藏的構字知識更有利於應用程式的處理。然而，它的缺點是構字式長短不一，也嫌它太長，以致用起來較不經濟。

部件序或根序較構字式簡潔許多，用它們來替代構字式是很自然的想法。以《中文電腦基本用字》的 9122 個字形，加上 629 個部件，457 個字根，亦可產生 48713 個字形。此時 9122 個形只用一字碼，而其餘約 4 萬字，每個構字序卻可簡潔如 15 頁的那 11 個式子。換言之，可用漢字形標來表示剩下的四萬字形，每個字形之長度僅二或三個字碼。由於 9122 個字形的累積使用頻率已高於 99.9%，餘下千分之一以下的機會用較長的碼，對系統之效率影響極為有限。所以善用構字式實是改善目前交換碼的一個好方法。

伍、結語

本研究進行兩年餘，始終秉持第二頁所列的原則，試圖解決缺字問題。至目前，雖然後繼工作仍多，大體看來這是可行的路。要言之，我們以增加電腦中漢字字形知識為主要手段，用 ISO 的標誌標準，利用網路的溝通能力，多工作業系統的方便等，在不犧牲信息共享之下，建立了登錄、檢索、表達缺字的能力。這能力是不分國界地區的，所以有利於各地區漢字的整合，以達到相互參照的目標。

在技術方面，本文所提出的字形模式是封閉性的孳生系統，比條舉式的開放系統所設計的交換碼無論在性能上、效率上均高一籌。我們改進了《交大字根系統》，把部件集找出來，使得構字式可短到兩個部件，這也是一貢獻。此外，我們給電腦字、字形、字體、字型、字樣等清晰的定義和表達，這是一項創舉。我們並不執著這些定義及表達是完美無缺的，反而，我們希望大家多批評、建議，以使這些基本更精確、更切合大家的需要，因為這些基本定義和表達方式是日後大家共享資料，共享程序的基礎。

我們延伸了漢字位標的觀念，創造出漢字形標。在觀念上，我們以構字的知識作為字形的識別碼（identifier），和傳統的數字碼相輔相成、相得益彰。在系統方面，在《電子佛典補充字集》和改善目前資料登錄系統的構想和嚐試，都得到不少人士的支持。在這情形下，雖然整個工作尚未完成，我們覺得有需要發表出來，以廣納各界的批評來改進可能的缺點。

在未來工作方面，完成字根與基本筆劃間的構字關係是一要務，此中包括用SGML標示來表達分毫字樣的一些細節。再者，沒能將傳統文字學的信息納入，是本系統的弱點，也是以後要努力的地方。周何等 1982 年的《中文字根孳乳表稿》中，從文字學的角度，由《中文資訊交換碼》的二萬二千餘字中，歸納出聲母 869，形母 265 個，並詳列字形與聲母，形母的樹狀孳乳關係，是值得參考的文件【註十二】高鴻縉的《中國字例》【註十三】杜學知的《孳乳叢考初稿》等，就字源及語意角度推演漢字孳乳形成的家族，也是很值得參考的。

【註十二】周何、沈秋雄、周聰俊、沈德修、莊錦津，《中文字根孳乳表稿》中央圖書館，台北,1982
【註十三】高鴻縉，《中國字例》，三民書局，台北, 1960 初版, 1981 六版

這些孳乳系統的數學結構與交大字根系統極為類似，可說是同質異形。因此，用相同的數學模式，將文字學的字形孳乳建在電腦中，並與字根系統相互參照對映，並不是難事。如果能做到這一步，就可使古今對照，承先啟後相輔相成了。這是吾等深深盼望的。

本系統收納的字形知識和文字屬性資料不能算多，只是最基礎的部份而已。但是，若能善用這些資料與知識，相信可以開發出許多漢字信息處理的程式，甚至可以發展出一些漢學研究和學習的工具。這也是我們希望能一試的。目前，本系統尚未正式成為網路上的伺服者（server），是故諸多網路上的介面亦待開發。對於這許許多多未來的工作，我們的態度是開放的。也就是說，歡迎有興趣的同仁加入我們的行列，一起分工合作。畢竟，這個系統涉及漢字信息處理的根本處，應由大家開發、大家共享。

誌　　謝

感謝本實驗室的同事莊德明、張翠玲和許婉蓉，沒有他們數年的努力，這篇文章不可能和大家見面的。本系統所有的程式開發是莊先生一人研發，勞苦功高。文字資料的整理許小姐出力頗多，張小姐除整理資料、稽核文獻外，還包辦了協調、檔案管理，以及本文繕打排印等工作。而本人，似乎只動了動口而已。