「第二次兩岸古籍整理研究學術研討會」，北京大學，北京，1998年5月11-13日

中央研究院古籍全文資料庫解決缺字問題的方法

莊德明、謝清俊、林晰

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。為了要保留這些字形，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：為了成千上萬新造的字，資料登錄的工作大幅增加；檢索文件時亦將面臨異體字檢索的難題；彼此分享資料時則更嚴重，可能的重碼將造成資料錯誤或文件讀不出來的狀態，以至於根本無法共享資料。這就是所謂的「缺字問題」。

目前，對於繼承漢文化的地區來說，缺字問題已是一個共同的夢魘，凡是遇到漢字的人名、地名、史料等等，都有相當嚴重的缺字問題；所以，缺字問題已是一個國際性大家都關心的問題。以中央研究院（台北）十四年來發展電子古籍的經驗來說，到目前為止集聚的缺字業已超過九千六百字，其嚴重性與求解迫切性可想而知。本文即針對此問題，將本院解決缺字的方法，包括理論和實務方面，作一介紹。

我們採取的方法是：運用字形資料庫來表達字形的結構和該字的屬性，在資料庫中，字形是以部件及字根的組合方式表達。目前已建立四萬餘字形的構字式。通常，遇到缺字時，該字的輸入碼、交換碼、字形等都沒有存在電腦中，是無法處理的，然而有了字形資料庫就可用它來提供缺字的輸入、描述、識別、查詢和作後續的處理工作。

理論上，字形資料庫所表達的就是字形的孳乳，每個字根都孳生出一棵字形的家族樹。在這些樹中，每個構字式的描述限制於只用三個分解方式中的一個來表達，此三個分解方式即：橫向、直向和包涵。在非末端節點上的字形，稱為「部件」，這是構成所有字形變化的基礎字形。本系統中的部件共有約2,370個字形：字根為625個；非字根的1,745個部件（其中1,419個是五大碼用字，326個為五大碼並未收錄）。以此基礎來對付前述本院9,600以上的缺字時，約有96%強的缺字即可解決，而殘留的問題字形，在增加部件後，即可全部解決。本文將對此缺字的分析，提出詳細的報告。

應用方面，我們正依據這個解決缺字問題的方法，來更新資料登錄系統和全文資料庫的檢索機制。這些構想亦將一併在文中說明。

莊德明　中央研究院資訊科學研究所研究助理
謝清俊　中央研究院資訊科學研究所研究員
林　晰　中央研究院計算中心分析師

中央研究院古籍全文資料庫解決缺字問題的方法

壹、前言

用計算機處理漢字資料時，文件中常有些字的字形是交換碼中沒有的，這情形稱為缺字問題。缺字問題在處理大量資料時（如國家級的資料庫和處理古籍時）特別嚴重。為了要呈現和處理這些欠缺的字形，常用的方法是在交換碼的使用者造字區內，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：為了新造的字，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，造字區的重碼現象將造成資料錯誤或文件讀不出來的狀態，若是在同一區域網路中，不容隨便更改造字區的內容的話，則會造成根本無法共享資料的窘境。

中央研究院利用計算機處理古籍已有十四年，其中以全文資料庫的發展最受矚目，目前上線的全文資料庫總字數已超過一億四仟萬字(詳本會議中黃寬重和劉增貴先生之論文)，其所用的技術則全由院內同仁自行開發，包括：全文資料庫的結構、文章的標誌系統、資料登錄之管理、缺字造字之管理等。然而，這些資料庫到目前為止集聚的缺字業已超過九千六百字，其嚴重性與求解迫切性可想而知。本文即針對此問題，將本院解決缺字的方法，包括理論和實務方面，作一介紹。

貳、表達缺字的方法

解決缺字問題的理論部份，業已發展完備【1, 2, 3, 4】。依此理論首要之務是解決缺字在電腦中的表達，即給予各缺字一個識別碼，並解決缺字的輸出入和後續處理等問題，以達成盡量減少文件中缺字的目標。此系統和目前所有的字碼均可相容，只要在各現有的造字檔中，加入適當的部件及運算符號約六百個，即可利用組字式來表達缺字，組字式分為構字式、部件序及缺字序號。以下，謹為此法作一簡介。

一、部件

當一個形體用來構成某字形的一部分時，我們稱該形體為某字的部件。如：「日」、「京」是「景」的部件；「景」、「頁」是「顥」的部件；「顥」是「灝」的部件；又如：「圍」的部件有「口」和「韋」。是故，部件是漢字的構形單位。部件是有層次的，如：「顥」可拆分成「景」與「頁」，「景」又可拆分成「日」、「京」。部件裡有的是字，有的不是字（是不是字是相對於某字集而說的）。最後不再拆分的最小形體叫作字根。

漢字最常用的構字方法為橫連(⿰)、直連(⿱)與包含(⿴)。如將上例寫作：『灝=氵⿰顥』，『顥＝景⿰頁』，『景=日⿱京』，『圍=口⿴韋』，則這些式子稱為「構字式」。在本系統中，一個構字式裡只能有一個拆分的符號，是故，構字式十分簡潔。再者，對缺字而言，構字式表達的是一個唯一的字形，不會與其他文字混淆；因此，以構

字式作為缺字的識別碼（identifier）十分理想。

茲以金剛經經文為例，『爾時世尊食時，著衣持金⿰本，入舍衛大城乞食，於其城中，次第乞已，還至本處。飯食訖，收衣金⿰本，洗足已，敷座而坐。』 文中的「金⿰本」即是一個用部件所表達的缺字；「金⿰本」不僅是此缺字的字碼，也指出了此缺字的字形結構。

我們拆分了四萬多字，並把這些構字式存放在字形資料庫裡。依此統計，橫連(⿰)的字約佔70%，直連(⿱)的佔21%強，包含(⿴)的佔9%弱。構字式在字形資料庫中可以表達字形的孳乳，每個字根都孳生出一棵字形的家族樹。

根據林樹字集分析而得的部件集（以下稱林樹部件集）共2,150個。從林樹字集觀之，是字的（即在林樹字集中原已有的字）共1,452個，不是字的則有698個【4】。這698個是要增入五大碼中的。然而，此698個字形使用的頻度並不一致，如下表所示：

表一：林樹部件集中非字部件的使用頻率統計

累計頻度	部件數(%)	累計部件數(%)	造字數(%)	累計造字數(%)
0~90%	628(29.21%)	628(29.21%)	182(26.07%)	182(26.07%)
90%~99%	662(30.79%)	1290(60.00%)	193(27.65%)	375(53.72%)
99%~99.9%	345(16.05%)	1635(76.05%)	89(12.76%)	464(66.48%)
99.9%~100%	515(23.95%)	2150(100.0%)	234(33.52%)	698(100.0%)

若希望99.9%的字都可由構字式組成（即留下0.1%的殘留缺字用其他的方法解決，以增進系統之效益），則吾人可省去207個非字部件（因234個非字部件中有27個是字根，而字根仍需完整保留，故234-27=207）。又，在464個非字部件中（即滿足99.9%條件所應加的字形數）；有22個字根重疊的字形，如下：

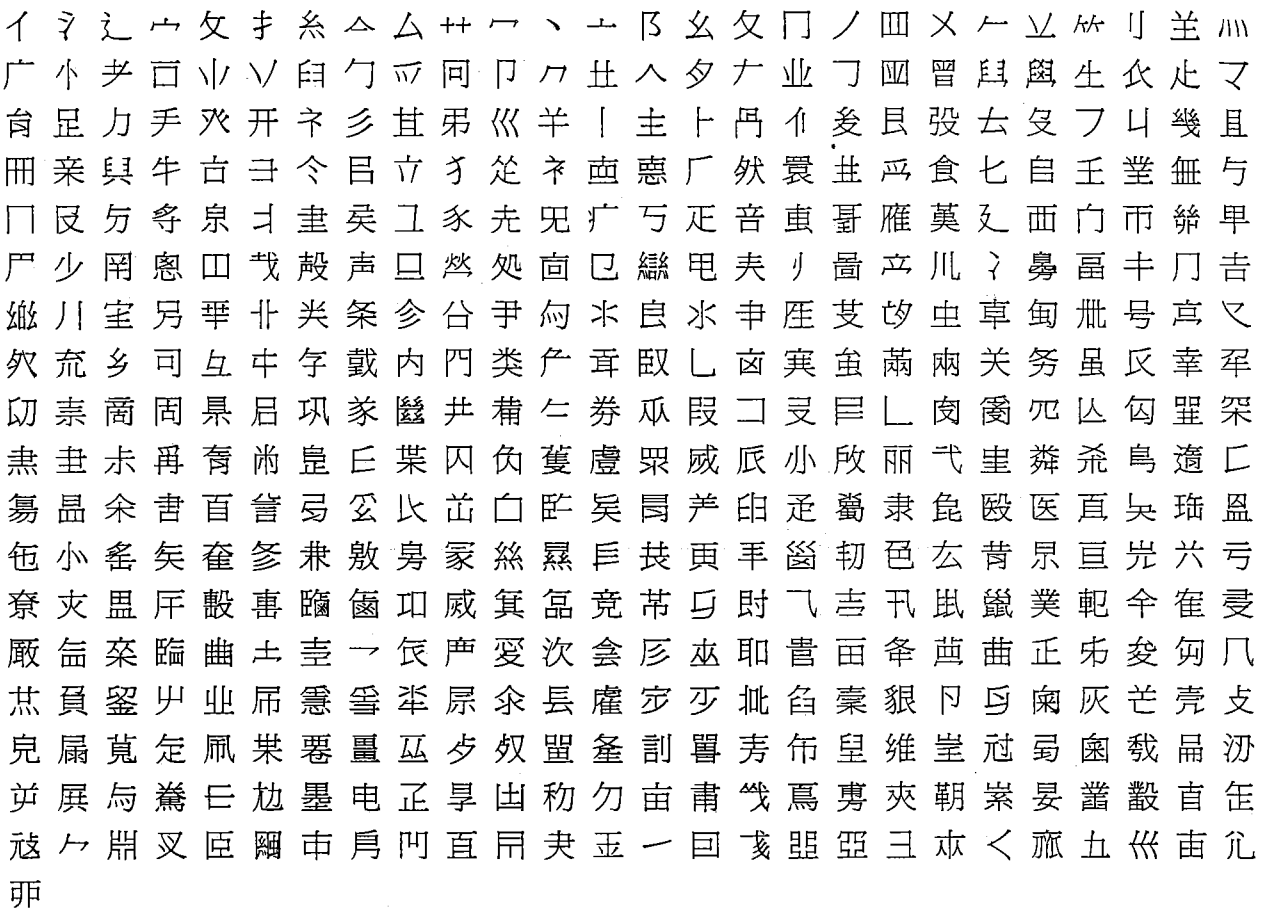
⿰口 ⿱乂 ⿱匕 ⿰人 ⿱丶 ⿰幺 ⿰火 ⿰土 ⿰厶 ⿱厶 ⿰巳 ⿰虫 ⿰夫

⿰斤 ⿰无 ⿰力 ⿰口 ⿰又 ⿰車 ⿰田 ⿱又 ⿰丰

這些以字根重疊而成之字形可用方便符號表示，不需另造新字，是故林樹部件集經上述最佳化後，需外加的總字數為464-22+27=469個，此469個非字部件如[表二]所示。

以上述的林樹部件集為基礎，將字集擴大至五大碼（Big-5）字集，並校以康熙字典214部首，及周何等導出的聲母869個，形母265個【5】，將林樹的部件集擴大至2,370個字形，稱為五大部件集。在五大部件集中，1,801個字形已為五大碼所收錄，非字者（在五大碼之外的字）有569個，五大部件集的全部字形則詳如附件一。至於簡筆字所需增加部件，共只有179個，目前因來不及造這些字形，並未加入 [表二] 林樹部件集中。

表二：林樹部件集中之469個非成字部件 (依使用頻度排序)



二、部件序

讓我們先來看兩個例子，第一個例子是「牖」字，此字無法用構字式表示，因為構字式限定只能用一個運算符號。「牖」字的左邊為「片」，右邊的「戶⧋甫」並不在五大碼中，字形組合的方式是『牖=片⧊(戶⧋甫)』，其中同時用到兩個運算符號，所以不符合組字規則一。(其實部件「戶⧋甫」在我們系統中原是一個部件，然而因為使用度太低而被刪除)。第二個例子是「奭」字，由「大」、「百」、「百」三個部件組成，但是卻無法採用橫連、直連或包含的方式來分解。

當缺乏適當的部件，如：「牖」字； 或缺乏適當的運算符號時，如：「奭」字，則可完全摒除運算符號，用部件序來表示。此時，直接按照書寫的順序將部件打出，(次序不同也沒啥關係，因在計算機中均將化成字根式來識別。) 並在前後加上起始「囧」及終結「⊙」標示 (Tagging)。例如：『牖=囧片戶甫⊙』，其中「囧」及「⊙」也是自創的符號。要注意的是『牖=囧片戶甫⊙』中，部件「片」、「戶」、「甫」出現的順序按照書寫的順序較好。如此，在表示缺字時，可以方便往後校對。

本系統的設計是以使用者方便為首要考量，這也是前文中刪去0.1%使用度207個冷僻部件的原因之一，因為即使系統中有這些部件，輸入的人都不一定會熟稔其形和輸入

碼。為了同樣的原因，亦允許用部件序來表示缺字，下一節中，將再提出一些好用的方法來減底使用時之困難。

三、方便符號與從缺符號的運用

並不是所有的缺字，都可以利用上述的規則打出來，這些缺字中有可能出現一個部件都沒有提供或是不知道如何輸入的情形，所以再造一個符號「❓」，表示從缺，例如：『竊＝⿰穴釆❓⊙』。目前，一個組字式中，最多只允許出現一次從缺符號。若是使用倉頡輸入法的人，或許會覺得「❓」與倉頡中的「難」有些類似。

為了方便輸入，另外還造了一些方便符號。方便符號需放在部件的前頭，如：「∞」表示兩個相同的部件橫連（例如：『兢=∞克』）；「 ⁸ 」表兩個相同的部件直連（例如：『叁= ⁸ 克』）；「⁘」表三個相同的部件橫連，（例如：「⁘去」，中文大辭典編號3171)；「 ⁞ 」表三個相同的部件直連，（例如：「 ⁞ 戶」，中文大辭典編號12029)；「⁂」表三個相同的部件三角頂立（例如：『轟=⁂車』）；「⁜」表四個相同的部件橫連；「 ⁝ 」表四個相同的部件直連；和「⁸⁸」表四個相同的部件分佔四方（例如：『燚=⁸⁸火』）等。

從缺符號和方便符號是可以用於構字式，如此可擴大構字式的適用範圍；例如，原先只能用部件序表達的字即能改為構字式，如：『瞿=∞目⿱仜』、『俎= ⁸ 人⿰且』、『桑=⁂又⿱木、『啜=口⿰⁸⁸又』等。方便符號也可用於部件序中，讓組字更為方便，例如：『歠=⿰⁸⁸又酉欠⊙』。方便符號也讓一些字的組字式既不需橫連、直連及包含符號，也不需起始及終結標記，例如：「⁸⁸魚」(中文大辭典編號47603)。

四、缺字序號

會不會有些字，連一個部件都拆不出來？有的，在中文大辭典49,905個字中，就有約二千字很像圖形，幾乎無法拆分，這時候就得利用缺字序號來識別他。缺字序號的型式如同部件序，需要用<起始>和<終結>的標示，但是標示中的部件改為編號，如：『⿰5⊙』表示這是第五個無法利用前規則表達的缺字。

五、其他的標示

我們正在設計一些其他的標示來表示異體字和異寫字【6】，包括多一筆、少一筆等的變異在內。原則上，不讓異體字佔碼位。本系統中有管理異體字的資料庫，異體字的字形則利用字體庫（font library）來顯示。

參、漢籍全文資料庫缺字的整理和分析

為了解決本院的缺字問題，本計畫從計算中心取得了本院目前所有『登記有案』的缺字資料，並加以整理、歸納、分析，以期了解本院缺字的真實狀況，和作為設計與測

試『解決缺字之系統』(以下簡稱為本系統)之數據和參考。本報告即將此整理、歸納、分析之結果，作一簡要之說明【7】。

一、中央研究院的缺字

計算中心經管之缺字共分兩部份:納入五大碼使用者造字區者4,553字,未補者5,174字，共計9,727字（至1998年3月前之統計資料），這是歷年來本院所有缺字之累積。目前的五大碼造字區可容納5,809個字，已造的4,553字全在此區內，另外1,256個空碼分佈如下：

1. FAB5-FEFE有702個
2. 9DF6-A0FE有480個
3.另外74個零星散佈

為了不使本系統干擾到既有的造字區，也讓舊系統能順利的轉移到本系統，只利用計算中心目前撥給我們使用的702個空碼（FAB5-FEFE），加入非字的部件。

二、造字區內4,553字之分析

1．扣掉五大碼重複字13個、字體變異者6個及符號9個外，實際上的總缺字數為4,525字。

2．若以字數來看，4,525個缺字中，可用構字式表達者有3,903個(佔86.25%)，用部件序者515個(佔11.38%)，另外尚有107個(佔2.37%)需要加入字根方可拆分。

3．根據已上線的一億三千八百餘萬字的資料庫字數統計，4,525個缺字總字頻次為 517,891，可用構字式表達的3,903個缺字頻次為411,698(佔79.50%)，515個可用部件序表達者的缺字頻次為89,638(佔17.31%)，其他的107個缺字頻次為16,555(3.19%)。

三、未補者5,174字之分析

1．扣掉五大碼重複字7個、自己重複字10個、空白字形14個及符號199個外，實際上的總缺字數為4,944字。

2．若以字數來看，4,944個缺字中，可用構字式表達者有3,760個(佔76.05%)，用部件序者864個(佔17.48%)，另外尚有320個(佔6.47%)需要加入字根方可拆分。

3．未補的4, 944字和199個符號的頻次是由十三經（八百六十萬字），諸子（舊的，含十九種古籍，共五百八十六萬字），古籍十八種（八百零五萬字），台灣方誌（七百五十四萬字）等（共計三千零五萬字）統計而得。

若以字頻來看，4,944個缺字總字頻次為16,598，可用構字式表達的3,760個缺字頻次為13,375(佔80.58%)，864個可用部件序表達者的缺字頻次為2,409(佔14.51%)，其他的320個缺字頻次為814(4.91%)。

四、本系統可處理之缺字

根據以上之資料，本院之缺字總數雖已有9,727字，但扣除因重複而產生之錯誤，異寫字，古字（如：甲骨、金文、小篆等）和符號外，實得9,469個缺字，其出現之總頻次為 517,891+16,598=534,489字次

以目前138,500,000餘字次之資料庫，對五大碼而言，缺字之機率約佔萬分之3.85。用目前的方法經4,525個造字補充後，則仍有16,598字次無法解決，亦即有約萬分之0.12弱的機會仍有缺字，若以本系統的方法來處理，則可直接解決者有517,120字次，約與舊系統相當，只餘萬分之0.125的機會仍有缺字。 然而，若廢除既有的造字檔，把目前本系統所無法直接處理者427個缺字（107+320）補上，則所有之缺字得以圓滿解決。

伍、缺字解決方案與缺字管理

漢籍全文資料庫現行的製作流程分為五個步驟，分別是資料登錄、校對、缺字管理、標誌與資料庫建立。各步驟大體上依序而行，唯獨缺字管理的工作另又散見於資料登錄及校對。以下略述現行的缺字管理作業，再詳述引進缺字解決方案後，將採行的新作業規範。

一、現行缺字管理

現行的資料登錄，每遇缺字，就代以固定的缺字符號「●」。於最後一次校對時，凡遇缺字，就填寫缺字表，按出現順序列出缺字的出處和字形。

缺字管理的終端處理比較複雜。首先要去除缺字表中的重複記載，篩檢出新字，同時累計每種新字出現的次數。早期新字篩檢完全仰賴人工，非常費時。現在雖因輔助程式，效率大有提昇，仍不如新法，因為後者通常不需要篩檢新字。

新字篩檢完畢，開始製作新字（造字）。受限於五大碼的造字空間只有 5,809 字，已知的近一萬個新字無法都造，只能取其中字頻高者優先造字，剩下的只好置之不理。

造字終了，展開補字。依據缺字表所載出處，於各資料檔補上新字。補字過程中的錯誤因素包括漏補、補錯字，以及編輯器指令操作失誤，導致正常資料受損等。因造字空間不足而從缺的字無法補回，自然無從檢索。

二、導入缺字解決方案的缺字管理

在【2】中提到一種資料登錄系統，結合字形資料庫、網路、具備組字式處置能力

的編輯器及中文系統，期能即時供給缺字字形，自根本消滅缺字問題。其中涉及的資訊技術繁多，一時之間不能盡解，缺字的困擾卻迫在眉睫。是故自目前的系統轉變為理想中之系統時，勢必經過一些中間過程，如此方不致於產生人員、技術、設備等等變更太大的困擾，以下之構想即循此而設計。

新的缺字管理比舊的省了許多事，造字檔只包含組字式的部件、運算符號與組字式不能表達的新字。在登錄階段，標準字集之外的字不是用組字式表示，就是直接用造字檔中的字。除非遇上組字式表達不了，造字檔又尚未供應的罕見字，否則沒有代以缺字符號的機會。既無缺字符號，輾轉抄錄缺字然後回補的現象消失了，伴隨的人工錯誤源跟著消失。效率提昇、錯誤反而減少。遇到前述之罕用字，可由專人查詢字形資料庫，找到缺字序號，記錄於資料檔中。字形資料庫若缺此字，則立刻新增，並將此字納入造字檔中，再將序號改為內碼。

到了校對階段，逕自檢查缺字的組字式是否正確，或者更進一步，藉著字形資料庫及相關程式之助，產生組字式的字體檔，再把資料檔的組字式映對至字體，呈現出來。此法所以可行，關鍵在字形資料庫能聯接到四、五萬字的大型字庫來提供字形，如：Kanji Base等【8】。本方案以字體檔來安置能用組字式表達的缺字之字形。藉助於組字式—字體對照表，解決了字體檔裡的缺字內碼和標準字集重疊的問題，提供了幾乎無限的空間，而無礙於缺字字體的正確顯示。

本系統中缺字處理相關的程式包括：組字式蒐集程式、字形資料庫查詢程式、字體檔管理程式與組字式—字體轉換程式，其用途與運算步驟如下：

1. 組字式蒐集程式掃瞄各資料檔，尋找各種組字式，並報告初次出現之處。
2. 蒐集到的組字式，轉由字形資料庫查詢程式核對。根據構字原理，再加上方便符號的介入，一個字常見有兩種以上的組字式，卻總能推導至共同的字根序。用它來查詢字形資料庫，得到此字的訊息。如果此字不屬於標準字集，記其序號、字根序、正規組字式及字體位址；否則，記其內碼。假使字根序不存在，留待校對者判斷究竟是發現新字，或是組字式有誤。即刻將所有新字的訊息納入字形資料庫，再次執行字形資料庫查詢程式。
3. 字體檔管理程式接續處理前一步驟濾出的訊息。如果字體檔尚未產生，取得缺字字體，組成字體檔，並產生組字式—字體對照表。對照表的欄位包含序號、正規組字式、字根序及字體訊息(字體檔名/內碼)，內碼係自動設定或人為指定。倘若字體檔與對照表業已存在，維護程式查出對照表所無的缺字，據以擴增字體檔與對照表。遇上大量缺字時，會有更多的字體檔，對照表仍維持一份。
4. 組字式—字體轉換程式透過對照表把組字式應對至缺字字體或標準字內碼，配合字體檔，即利於校對。校對工作包含訂正組字式謬誤，也許還有把內碼改成組字式的狀況，所以末了再以程式查驗這些組字式。

這些程式都不複雜，短期內能夠完成。一旦字形資料庫具備所有組字式的字體，校

對者的作業流程簡化如下， 其中第1、2、4 各項非常快速，它們在校對階段佔用的工時比例，少得可以忽略。

1. 用單一指令執行有關程式，產生或更新字體檔，並獲致經過字體轉換的資料檔
2. 若字體檔變動，重行安裝
3. 校對
4. 用單一指令查驗校妥的組字式

柒、缺字解決方案與全文資料庫工具

缺字解決方案牽動漢籍全文資料庫的內部結構和程式工具。首先面臨的問題是究竟組字式應該原封不動的擺進資料庫，還是加以變形？慮及組字式並非唯一，又長度不定、格式稍顯複雜，對緇銖必計的檢索效率略有不利，決定將它轉換成定長的編碼，名為組字碼，格式如下：

<組字碼標><字形資料庫字號>

構字碼標佔一位元組(byte)，它的值有別於中文內碼任一位元組的值，也有別於 ASCII 內碼，以資區分。以五大碼環境論，可定其值為 255。字形資料庫字號佔三個位元組，考慮照CCCII 的做法，每個位元組只用 ASCII 有形(printable)字碼的值，計 94 個，而三個位元組能供應的值多達 830，584 個(94*94*94)，應足敷已知、未知的全部中文字之需。既然字形資料庫是集中管理的，每個字的字號自屬固定，所以構字碼雖是內部所用，仍有交換共享的潛力。

漢籍全文資料庫的程式工具，以資料庫建製子系統及資料庫檢索子系統為核心。這些工具需具備組字式和構字碼的處理能力。此外，如實載錄異體字形的難題，已因組字式而徹底達成。倒是存真之後，同樣的詞夾含不同異體字的情形更普遍，增加檢索的複雜度。譬如：雖是檢索「煙葉」，應將「菸葉」和「火厽因葉」一道帶出。可惜異體字的判準並非一成不變的，在周延的擘畫出現以前，暫將大體上完全通同的字合為一組，叫作通同字組，並用它來解決部分問題。

資料庫建製子系統讀取標誌過的資料檔，經對照表把組字式轉化為組字碼或內碼，建成全文資料庫，並產生索引來加速檢索。組字碼對索引建構是有影響的。這個索引採用完全逐字反轉架構，記有每個中文、外文字在資料庫內的位址。常用中文內碼的結構固定、總數有限，有別於外文字的長度變動不一、總數不明，所以分開處理。構字碼的長度雖固定、而總數未明，宜比照外文字處理。至於異體字方面，宜把通同字組的位址叢聚在一起，以便循序快速讀取整組字的位址。

資料庫檢索子系統接受使用者的檢索條件、實施檢索並輸出符合條件的資料，組字式對輸入、檢索、輸出三者都有衝擊。在輸入方面，遇到使用者下達的指令含有組字式，便轉為組字碼或內碼。不過，組字式的部件、運算符號有那些？如何輸入？對多數使用人而言當屬陌生，需要一種組字式的輸入輔助介面，詳列部件與運算符號，供使用者揀

選而組成組字式。再者，組字式無效，必須採缺字序號時，也應有相關的輸入機制。

檢索子系統有兩種檢索機制：經索引檢索或逐字搜尋。前者怎樣因組字式調整，已概略提及。後者若要同步搜尋異體字，可將一詞展開為各種異體字組合，如把「煙葉」展成「煙葉」、「菸葉」、「火⿰因葉」三詞，再做多詞搜尋。

有的同義字幾乎在甚麼場合都互通，像「饑」與「飢」；有的只在特定場合中通同，像「煙」與「菸」出了「煙草」的範圍就不通了。通同字組純就字的範疇考量，難免不周，然而勝於沒有。大體上，凡是檢索兩個字以上的詞彙，就能發揮很好的效果，鮮少副作用。例如：搜尋「煙囪」，檢索機制連著「火⿰因囪」、「菸囪」都找並無妨，「菸囪」本不存在，找不到就是了，而該找的「火⿰因囪」並未遺漏。現階段宜讓使用者調整通同字組，滿足個別需求。

在輸出部分，若是純供閱讀，把構字號映照至字體即得。倒是使用者常做的裁文(cut and paste)，若發生在缺字字體上，必須留意。譬如把裁得的缺字字體貼至檢索條件，檢索程式要能利用組字式—字體對照表，使它還原為組字碼。更有甚者，使用人取得含有缺字字體的檔案，想對它進行檢索或統計等，會遇到怎樣的問題？似乎，提醒使用者區別自己的用途，在單純閱讀以外的情況，直接給他含組字碼的資料檔、對照表與相關的基本程式，是現階段較好的方法。

捌、結語

本文報導了本院解決缺字問題的一個方案，依目前的分析，此方案是可行的，可以徹底解決缺字問題。本院原有系統的更新工作正在進行中，其中字形資料庫的使用者介面業已完成第一版雛型，如有機會當可展示或由使用者自行操作試用。

至於後繼的工作，在研究方面：當可以本系統為基礎，將文字學中之資料和程序陸續加入，以使本系統得發展為漢學研究時的文字學輔助工具。在應用系統方面：則以開發中文文件共享機制為優先，使得有缺字的文件亦能共享和作後續的處理，這些都是急待努力的。

本系統歡迎各界試用，有興趣者可與本研究室聯絡，網際網路地址為 http://www.sinica.edu.tw./~cdp

【參考資料】

【1】莊德明，〈字種與組字式〉文獻處理實驗室技術報告，1997.1222

【2】謝清俊，〈電子古籍中的缺字問題〉，第一屆中國文字學會學術研討會，天津，1996.08

【3】謝清俊，〈漢字的字形與編碼〉，漢字字碼與資料庫國際研討會，京都及東京，1996.10

【4】謝清俊，〈A Descriptive Method for Re-engineering Hanzi Information Interchange Codes〉漢字字碼與資料庫國際研討會，京都及東京，1996.10

【5】周何、邱德修、沈秋雄、莊錦津、周聰俊等，《中文字根孳乳表稿》，國字整理小組，

1982
【6】王寧主編，《漢字漢語基礎》，北京市高等教育自學考試委員會組編，科學出版社出版，1996.07
【7】莊德明、許永成，〈中央研究院 造字分析〉文獻處理實驗室技術報告，1997.12
【8】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

壹、附件一：五大部件集字形表（1998年5月4日製表）

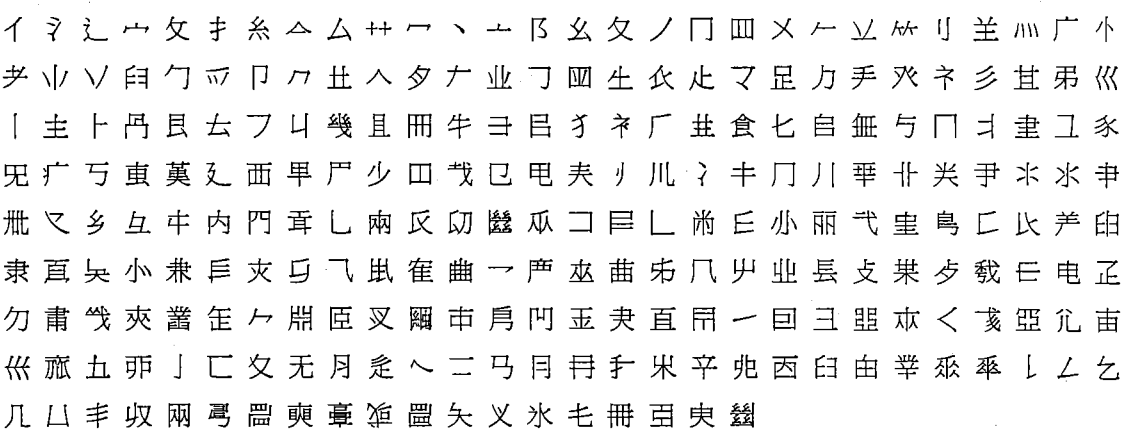
五大部件集的漢字字形係拆分自林樹及五大（Big5）字集，並加上說文和康熙部首及中文字根孳乳表稿的形母及聲母，部件總數為2370，分為下列四類，並依使用頻度高低排列於後：

	字根	非字根	合計
已在五大字集中者	382	1419	1801
不在五大字集中者	243	326	569
合計	625	1845	2370

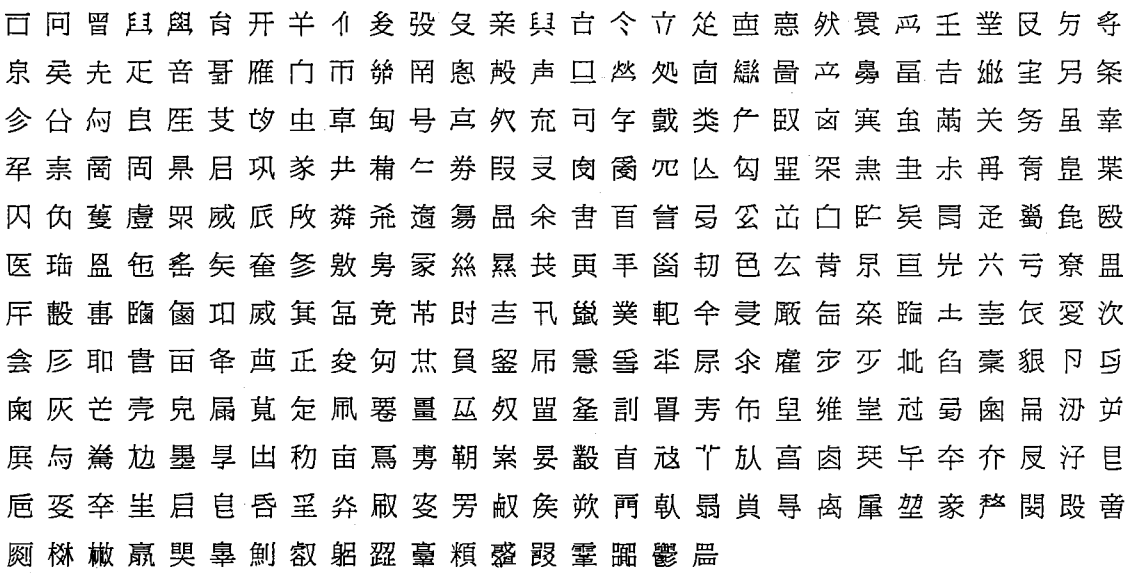
一、已在五大字集中之字根共382個

口 一 的 日 土 木 十 八 月 人 言 心 子 大 又 女 是 我 王 不 寸 了 隹 貝 夕 力 彳 目
有 可 他 方 小 工 立 巾 匕 止 田 白 儿 上 也 二 士 在 門 口 禾 來 耳 這 國 火 戈 艮
們 殳 斤 豆 頁 車 見 生 刀 金 說 中 尸 戶 到 個 弓 重 比 矢 要 以 就 走 全 廾 虫 犬
音 米 用 兀 你 父 巴 厂 自 戊 那 青 勿 欠 道 且 山 時 口 裡 馬 臼 和 著 得 曰 家 虍
麻 母 三 乍 爲 穴 里 巳 下 五 行 事 屮 雨 衣 月 几 示 年 舌 夫 皿 石 永 未 糸 長 之
至 四 辛 水 更 少 皮 手 面 羊 冉 丁 己 亥 本 老 氏 而 民 果 臣 乙 由 黑 亡 文 西 九
曲 卜 非 兩 七 身 東 求 廿 酉 夬 谷 歹 已 气 舟 弋 匚 丸 毋 高 聿 包 及 黃 才 禺 骨
尹 囪 世 羽 牙 必 豕 吏 干 太 卯 平 魚 角 內 无 辰 食 內 鳥 入 丰 予 斗 甫 乃 牛 申
風 束 半 韋 缶 齊 央 婁 甘 釆 制 坐 疋 朮 丘 毛 足 兆 鬼 州 飛 柬 矛 亟 倉 亞 刃 革
片 弗 千 舛 支 末 乎 鹿 垂 色 川 久 夜 肉 多 鬥 爾 斥 龍 甲 瓜 耒 赤 鬲 髟 乘 鼠 麥
戉 尤 巨 畢 承 爪 首 屯 犮 丈 香 玉 瓦 黽 史 鹵 井 妻 尺 束 竹 存 血 玄 于 卑 鼓 巫
肅 丹 齒 夾 烏 冊 率 鼻 丑 串 典 函 酉 阜 喪 夷 永 甚 牽 龜 衰 乂 尢 市 爽 丐 宄 弔
韭 黍 亙 禹 邑 脊 走 曳 龠 熏 豕 傘 幽 凹 叉 凸 乖 臾 丗 鼎 帶 匆 爻 鬯 毋 兜 噩 秉
卍 丫 甩 丌 丐 孑 乜 豳 奭 弗 孓 网 艸 襾 囧 丗 王 羊

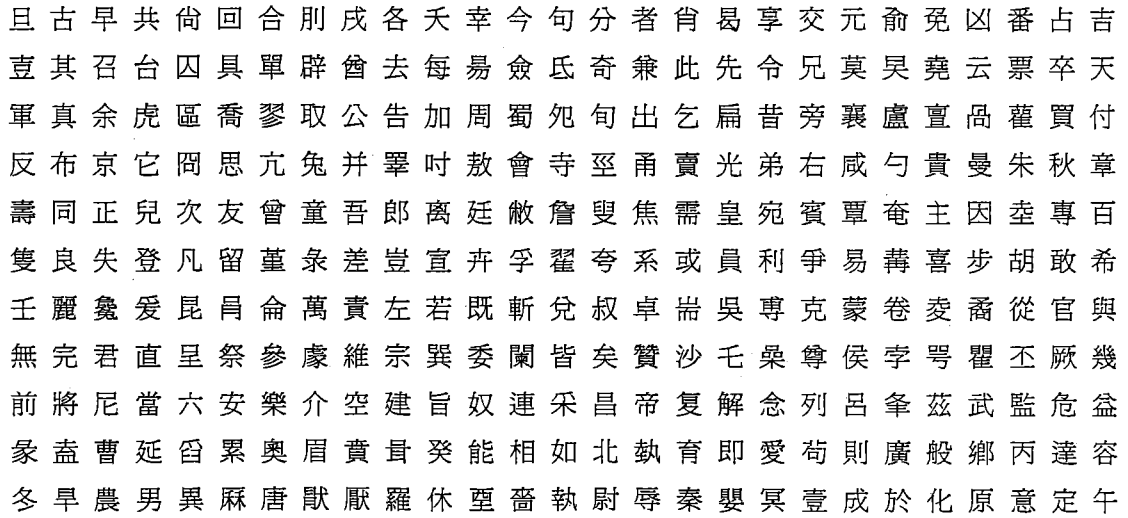
二、不在五大字集中之字根共243個



三、不在五大字集中之非字根(非Big5)共326個



四、已在五大字集中的非字根(Big5)共1419個



南 帚 代 袁 死 華 敬 呆 英 司 盡 居 奉 苗 奚 疑 雲 敕 匈 歷 胃 尋 暴 匈 遂 追 賴
貞 柔 卯 畏 雍 契 敦 后 鐵 朔 盾 匡 蔑 宁 匿 雋 發 義 催 氣 間 條 難 害 星 帶 昏
勞 保 景 屋 任 式 習 歲 奐 舍 憂 善 查 族 折 堂 秀 固 困 勻 豐 寧 盜 荅 乏 妥 屈
庶 匍 朁 閒 刑 虩 稟 庸 隋 留 皋 咎 牟 知 孝 貫 路 豐 戢 向 表 封 府 產 妾 帛 強
某 泉 助 亦 印 斯 岡 考 戕 屬 規 亭 引 段 展 爵 灰 朵 逢 戾 畜 孟 湯 佰 奏 幼 雷
戎 虛 犀 敏 扇 燕 鬳 翁 厲 貢 罔 廉 惠 羔 葛 丞 戚 晉 宅 臭 彭 頻 桑 豪 巢 郭 析
栗 樊 厄 耆 彗 晃 肴 胥 邕 矍 麃 弇 過 便 名 最 並 位 美 師 部 象 竟 度 壯 局 阿
斿 造 服 射 節 封 除 息 志 夏 春 退 臧 孫 耶 波 冒 背 寅 黨 散 奠 松 威 朝 卦 嚴
毒 素 否 靈 宿 彥 魯 匈 堅 榮 伐 藝 藝 覽 審 磨 允 獻 忝 旋 慮 隆 伏 殿 焉 廬 亨
春 桀 洹 渠 禽 卸 崩 狄 窄 崔 蚤 粵 攸 禿 靡 邪 匝 睿 臬 頡 翕 虐 纍 戢 麼 明 然
動 常 只 市 親 書 哥 受 烏 臺 罷 張 充 數 兵 答 流 忍 畫 縣 客 須 弄 養 切 復 肯
戲 感 科 宓 淮 穌 預 質 匪 店 疔 荒 另 頃 康 勇 升 敞 鄭 施 季 賞 孰 刺 耤 互 孔
陰 聖 遣 禁 劉 疾 逐 戒 沓 卵 聚 擇 嘗 恩 署 迷 莧 狂 貌 僕 遷 廚 溥 困 徵 羲 獄
隱 鮮 虐 悉 宏 邦 黎 欣 霸 屏 雁 霍 寇 盈 劾 潘 芒 顛 氾 謁 虜 敕 蕭 夐 薛 粲 咨
貳 卞 彌 滕 蚩 兮 眇 莽 匿 庚 斂 厘 叟 虞 蒿 遽 羌 徙 嬴 弁 嵩 筮 寔 圂 雝 垔 尨
芮 雩 耍 學 都 故 羔 問 法 對 外 聽 新 再 活 快 海 業 許 邊 何 別 通 寫 收 江 界
哭 號 拉 錢 叩 清 昭 卻 運 舉 熱 約 遠 讀 務 送 量 滿 苦 款 研 備 賈 諸 離 察 洛
歸 忽 裹 眾 彎 推 島 資 團 茶 衛 隊 乾 蘇 待 項 越 藏 亮 巷 致 領 絲 野 覺 雇 溫
拜 絕 守 躬 冗 律 微 靜 席 負 杲 豬 普 絜 欲 肥 楚 筑 幹 頓 拍 均 姑 旅 幕 緣 載
掌 尾 含 侵 勤 移 伊 馮 惟 刷 戴 錄 奔 宜 憲 愈 索 汎 閣 捕 孤 董 盆 繁 炭 患 悶
涂 洪 倍 殼 桂 牌 寬 奈 浦 衰 匍 轉 黍 熊 貪 純 柰 賜 敘 毀 猶 匹 哀 效 嵩 歇 葬
艾 庫 肩 魏 欽 疏 空 芳 匱 勃 晏 漂 肋 軌 賤 宰 泊 宇 芬 澡 淑 珀 浸 舜 杉 贛 削
衡 毅 梨 閏 糞 匐 霜 札 御 逮 扈 斑 忌 侃 羞 籥 泙 贏 殷 唯 悠 矩 沃 沾 剌 脾 衍
纂 赦 釜 沛 耦 淫 隼 吝 閔 冀 佘 翰 閑 粥 竄 弘 緇 淩 脣 奢 攀 頹 齋 洼 苛 勘 蕊
閻 弧 坎 藺 啟 恣 閵 閭 稜 仄 鞏 瑩 窒 瓠 奎 孱 昱 彪 滂 脩 陟 縈 戍 痂 剡 儷 倩
雉 戛 昀 欒 宕 岑 豚 遝 殄 曇 耿 睢 湛 梹 梟 吁 皓 畀 熒 翦 荀 防 燮 翳 瞢 沮 猗
弭 貫 胙 諼 卣 舋 齧 輅 鬩 顥 叵 訇 斌 斐 鴦 鷸 焄 圣 觜 隡 汇 掌 品 醀 恥 夔 畬
冢 拑 絫 貍 涌 么 狊 婿 檗 犀 斂 薦 粖 尒 鴈 豔 銎 竘 于 岔 玳 壺 卡 匠 劣 夙 印
宋 牢 吹 阜 役 尿 肘 刪 杳 杰 毒 岩 恆 牧 炙 習 乳 岳 佳 宦 穿 祝 突 厚 青 省 苜
胤 看 班 桌 拿 閃 冤 軟 覓 甜 舁 婦 寒 粟 焚 詈 集 飧 塵 煞 煩 塵 望 奩 罰 慶 徹
褭 駮 暹 興 毅 聯 簋 蠆 鬻 奧 雙 彝 斷 繭 爨 刁 仁 勾 仕 功 打 丟 好 字 扣 朴 作
吟 序 彤 忘 抄 技 扶 杏 汞 沈 沐 沒 沂 皂 私 味 岸 底 弦 往 拔 抽 抱 放 昊 昇 河
沼 泓 治 版 花 芷 俊 俎 勉 勁 哉 室 怨 恬 括 拾 政 昨 柱 柯 津 洞 洗 派 紅 茅 虹
述 修 倫 屑 庭 徒 恭 料 朗 柴 泰 浪 涅 爹 畔 盎 租 紡 紛 缺 脅 脈 茸 草 荃 虔 軒
逆 勒 商 圈 基 寂 密 屠 彬 措 救 斛 晨 梁 梵 臬 殺 淵 笠 終 翌 聊 茶 設 陳 雀 雪
頂 博 圍 報 惡 惰 戟 智 畧 替 棠 欺 渝 稍 策 筆 筍 結 絮 費 貿 貸 辜 進 鈞 陽 雅
亂 傷 剽 塞 慈 愁 搆 楊 溪 瑟 罪 羨 肆 葉 賊 遐 零 夢 搴 斡 旖 滯 漸 熙 甄 睽 算
箸 綿 翠 舞 蓋 輕 銀 頗 魁 寮 廢 慧 慰 摩 敷 樞 潭 盤 穀 窮 箴 蔡 蓬 賢 適 墨 禦

積 融 遺 隨 營 牆 糜 翼 薄 鍾 鞠 檗 叢 壘 熹 職 藍 藉 贅 雜 壞 懷 瀝 繫 類 孽 寶
闌 夔 蘭 蠡 囊 顯 靈 釁 鬱 邛 忒 忡 旰 狃 昝 朐 洴 耆 苕 恝 涅 烝 罡 庳 眯 稊 豥
歆 歒 滎 窨 綦 銛 隤 圜 縠 褰 廚 蹇 黈 蘄 虋 屳 吊 汙 呴 昦 拏 党 娄 峹 焊 埜 窐
巠 湡 笵 亜 滈 蔿 歁 蕈 瘣 蔍 藍 蓸 犪 蟺 瀸