B) 18

漢字字碼與資料庫國際研討會，京都· 東京，1996 年 10 月

從缺字問題，談漢字交換碼的重新設計——第一部份

漢字的字形與編碼

謝清俊

中央研究院 資訊科學研究所 文獻處理實驗室

1996 年 10 月 4 日（修正版 85.12.20）

漢字的字形與編碼

壹、前言

現在用計算機處理漢字資料有許多問題待解決，其中最根本的問題是「字」不夠用，這就是所謂的缺字問題（missing character problem）。Wittern 和 App 曾說："In East Asia, the problem of missing character is ubiquitous, from individuals unable to type their own name to universities, companies, government agencies. In Japan, these missing characters are called " gaiji "..... It is clear from our work on electronic Chinese Buddhist texts that even Unicode (ISO 10646) will not significantly reduce this problem "【註一】。這段評語是很適當中肯的。我們從 1984 年起在中央研究院發展漢籍全文資料庫至今，對這段評語有非常類似的體會。【註二】

由於缺字衍生了許多管理上和技術上的嚴重問題，諸如【註三】：

1. 大幅增加了資料登錄的工作
2. 產生缺字管理和造字管理的困難
3. 字碼所允許的使用者造字空間不夠用
4. 缺字和異體字造成文件處理上的問題
5. 造成資訊共享的障礙

這些問題如今已達到非解決不可的地步。

造成缺字問題的主要原因，固然是現有的交換碼性能不夠，從另一個角度來看，計算機中漢字知識的貧乏則更是直接的因素。以往，大家認為擴大交換碼的字集可能是解決問題的一個方法，可是根據 Wittern 和 App 的意見和我們自己的經驗，這樣只能降低一些問題發生的機率，並不能徹底解決缺字問題。因此，要解決問題，還是應該從建立計機內處理漢字知識的能力著手；這包括將必要的文字學知識表達在計算機中加以利用，以及建構一個處理字形的機制。這就是本文的基本設想。

為了要將文字學必要的知識表達在計算機中，對漢字的一些基本性質，尤

【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

【註二】謝清俊，《中央研究院古籍全文資料庫發展概況》，AAAS Annual Conference, Hawaii, 1996 年 4 月

【註三】謝清俊，《電子古籍中的缺字問題》第一屆中國文字學會學術討論會，天津，1996 年 8 月

其是在基本定義和字形相關的知識方面，必須了解到相當程度。再者，這些表達在計算機中的知識必須與文字學和漢語語言學的知識能夠相容，否則所設計的系統將會失去擴充能力和必要的修飾彈性，也將會造成以後應用程式開發的困擾。上述的要求都是撰寫本文遵從的原則。

在這樣的工作條件下，由於資料甚多，本文分為兩個部分撰寫。在第一部份《漢字的字形與編碼》中，首先試圖將文字學中與編碼相關的知識整理出來作為解決問題的依據，對字形變化的分析也在文中一併介紹。第二部份《漢字交換碼的再造》，將提出一個能徹底解決缺字問題的方法，以及依據此方法所設計的交換碼再造（ re-engineering ）方案。相關的主要工程實務是一個「字形資料庫」的建立。此資料庫目前進行的狀況以及例子也將在文中報告。最後，以此計劃執行的檢討作為第二部分的結束。

貳、 綜論

漢字有字形、字音、字義三個要素，是眾所周知的。在此，討論的重點放在「字」與「字形」。這是因為目前計算機的應用，還沒能達到深入處理漢字音義的程度，而且在交換碼的問題上，主要涉及的是字和字形的問題。

自古至今，漢字大抵一字多形。遠古的甲骨文和金文一字數形，學者習以為常。公元 150 年左右《說文解字》收錄 9335 字中，即含有 1163 個重文。這些重文為古文、籀文及或文，都是六書造字的異書同義字。類此者，散見於周禮漢書之重文仍多，有待收錄整理。【註四】

從時間軌跡觀之，以往漢字字形的演進主要經歷了甲、金、篆、隸、楷五種字體的嬗變，產生了不少字形結構的差異，然而嬗變之中其字理則同，字義則一。是故數千年來漢字體系始終一貫相承，生生不息：新生的字形賦予漢字當代的適應力與活力，而字形之兼容並蓄則延續了漢字的文化傳承。古今之字書無一不收錄異體字形者，實為漢字一字多形之直接佐證。

從文字學觀之，漢字一字多形固緣於造字五體六書之殊，在用字方面亦有還原通假之需而產生加聲加形者。此即所謂古今字或累加字。這些字固有助於字義的精緻分化，但也是致使漢字多形的根由。

【註四】 關於文中所談的文字學知識，散見於各文字學書籍，在此不特別引徵。然而，高緒价教授在第一屆中國文字學會學術討論會中發表的《意念造字》一文，(1996 年 8 月)，多有所涵蓋，且討論之重點切合計算機之需要，值得參考。

從歷史上觀之，歷代承平之時，多以行政力量規範漢字的字形以益當時語文的利用，如刻石示以正形、考舉約以範式。然而漢民族文化之絢爛發展早已深植於語文之性質中，文學、藝術、思想等莫不受惠於字、詞與字形之彈性與多樣變化。其直接影響者如詩詞歌賦、戲曲、文章、書法等，而日用器皿、建築、庭園、遊戲乃至於人們之價值觀、思考方式，亦難脫構字思維模式的影響。是故歷來規範字形者，僅止於申張文字的正統和導引文字未來的發展方向而已，並無必要將字形變異之自然發展盡納入繩矩。

字形變異之自然發展實源於生活上的需要。隸、楷、行、草和繁、簡之變，皆源於此；而自然發展的通則即眾所周知的約定俗成，這通則也就是漢字與漢語言世代綿延、生生不息的生命力根源。從物質文明發展的過程觀之，文字表達所用的物質和相關的技術革新也造就了不少新的文字形體。紙和印刷術的發明成就了印刷字的形體，近代美工與繪圖工具則創造了許多工藝字形體；這些例子均如實表現漢字字形多變的自然軌跡。

一時之字形規範，固有其正面之效益和作用，然而也不免產生新的形體而導致漢字字形之累增。近年來大陸推行的簡化字，二次大戰後日本推行的漢字改革，及新近台灣公佈的標準字形等，莫不如此，這些字形新標準的構形，也無一不是根據長久以來民間用字的約定俗成。若無約定俗成的歷史淵源，這些新字形將無由產生，也不可能為廣大民眾順利接受。就文化生命而言，因社會進步，生活環境的變遷而導致文字運用上形體改變的這個生命活力源頭，實應善加維護、珍惜和尊重，否則無異殘傷語文和文化的自然發展法則，並危及漢字適應生存的本能。

漢字交換碼的設計，都遵循各國標準字集，這似乎是理所當，然殊不知國家標準集的訂定，都有其政治目的和前述約制文字的企圖，並非完全依據使用者的需求或漢語文運用的自然法則而訂。處此情境，交換碼可滿足者就會有限度，不可能普及使用者所有的需求，凡涉及歷史、文化之資料，如地名、人名，或用於學術、古籍、藝術等領域，就常顯得過於局限而字不夠用。

計算機要面對的不僅是時下的國家標準字，從使用者的角度來看，計算機面對的是歷史上所有曾經出現過的字，無分正偽、繁簡、雅俗、古今……等區別。亦無所謂現在已不用的「死字」，這些「死字」在你指出它時，它就已復活了！只有用這麼大的胸襟氣度來討論、來設計計算機所需的字集，才不會在資訊科技(information technology)的運用上造成文字、語言和文化上的階級和岐視。

參、漢字的構字與字形變化

原始文字皆源於圖繪，漢字亦不例外。漢字之肇始多緣事物構形，以象形示意。之後，加以託形、變形為義，再以音相依附、音義相生，孳生了許多文字。

古時候，「文」、「字」各有不同之定義。鄭樵《象類書》云：「獨體為文，合體為字」。所謂「文」是指漢字中最初造字時每個不同的單體，包括六書中「象形」的日、月、山、川等，以及「指事」的一、二、三、上、下等。有些人稱「指事」為「抽象的象形」。象形和指事類的漢字是組成漢字最基本的完整個體。

根據鄭樵的蒐集，以形為主的文有三百三十個，稱為形母，以聲為主的文共有八百七十個，稱為聲母，合計一千二百文，進而孳生出許許多多漢字。孳生新字的法則包括了六書中的「會意」、「形聲」、「轉注」和「假借」四種方式。

鄭樵云：「六書也者，象形為本，形不可象則屬諸事，事不可指則屬諸意，意不可會則屬諸聲，聲則無不諧矣」，這段文字簡要地說明了六書造字的發展過程以及漢字構成的基本原則，迄今仍然適用。

鄭樵的形母和聲母已失傳，近代周何教授依據中文資訊交換碼（ CCCII ）第二集的22394字的字集重新整理的結果，得出漢字有869個聲母及265個形母，共計 1134 個。【註五】這些是從文字學角度，所得到組成漢字字形的基本元素，通稱之為字根。從漢字發展過程上看，甲骨文約有 150 個字根，小篆有 314 個【註六】。及至近代，有將漢字字形視為圖繪(graphics)而分析其構字者。這些工作僅依字形外觀分析，不涉及文字學，也得出一些為計算機組成漢字字形所需之基本元素，亦稱為字根（大陸稱為部件）。這方面發表的論述頗多，字根數目從約 100 至 1000 餘，所用的組成運算子(operator)亦由三個至 8、9 個不等。總言之，漢字構形由字根組成已是不爭之論，可是，漢字字形歷來卻有相當大幅度的變化。

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……』是一個明顯的證例。又如宋仁宗至和三年（公元 1056 年）郎簡《法寶壇經序》云：『六祖之說，余素敬之。患其為俗所增損，而文字繁雜，殆不可考。』這些都是敘述字形雜亂之記錄。

【註五】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
【註六】北京師範大學王寧教授，1995 年，未正式發表之資料。

潘重規教授在《敦煌壇經新書· 緒言》中【註七】 中，有這麼一段文字：

『荀子正名篇說：「名無固宜，約之以命。約定俗成謂之宜，異於約則謂之不宜。」荀子所謂名，即是文字。所謂約定，即民意所公認。所謂俗成，即大眾所通用。文字經約定俗成，足為標準，謂之正字。正字既已通行，復有人改變正體，斯為新造之字。新造字之，如得大眾認可，獲大眾使用，這亦是約定俗成。約定俗成的文字，便不容任何人把他抹殺。根據這一理念，加以觀察，許多敦煌寫本中我們認為是訛誤的文字，實在是當時約定俗成的文字。它們自成習慣，自有條理，它們是得到當時人的認同的。我們站在現代人的立場，覺得它違背了我們的習慣，我們認為它是錯誤。如果站在他們的立場，他們亦會覺得我們違背了他們的習慣，會認為我們是錯誤。』…… 『我把敦煌文字俗寫的習慣，歸納成字形無定、偏旁無定、繁簡無定、行草無定、通假無定、標點無定等等條例。字形無定，如人、入不分，雨、兩不分，瓜、爪不分等；偏旁無定，如木、扌不分，亻、彳不分等；繁簡無定，如佛作仏，舍作舍等；行草無定，如風作𠘨，門作冂等；通假無定，如是通事，須通雖等。標點符號亦和現代通行符號大不相同，如刪除符號作「卜」等。這種種現象，正和遼代沙門行均編纂的《龍龕手鑑》完全相符。原來《龍龕手鑑》是根據宋以前寫本編成的一部字書，所以人部的字和亻部不分，爪部的字和瓜部不分，木部的字和扌部不分。甚至同一個「雨」，可以用作雨，亦可以用作兩，並同收在雨部中。這種狀況，證明了敦煌寫本使用的文字，正是當時通行的文字，而不是近代人眼中心中的惡本訛字。』

從以上這些例證可知歷代的文獻中字形都多少都有當時的俗字，這是電子古籍中一個普遍的現象，也是造成「字」不夠用的原因之一。然而，依這些「字」的性質，它們大多是一些字的異體，所以計算機中並不是真正缺「字」，而是缺一些異體「字形」。因此，處理缺字問題時，不能不兼顧到異體字。這個現象，在我們製作廿五史及十三經等古籍時，均可證實。目前台灣製作電子佛經的各團體的缺字表中，約 70%的缺字都是異體字。

成書約在 770 年的《干祿字書》，顏元孫在自序中曰：「所謂俗者，例皆淺近，惟籍帳、文案、券契、藥方，非涉雅言，用亦無爽……所謂通者，相承久遠，可以施表、奏、箋、尺牘、判狀，固免詆訶。所謂正者，有憑據，可以施著述、文章、對策、碑碣，將為允當。」可見古時已對異體字作了有系統的劃分與整理。近代的字典、辭典中，亦多有蒐集異體字，並將之分為古今、正俗、繁簡等類別。可見異體字的存在是古今共有的現象。它們並不因歷代規範「正字」的限制而消失。異體字的現象可以說是現代設計交換碼時忽略了的盲點。唯一有系統整理過異體字的交換碼是 CCCII，其異體字表例如〔表一〕。

【註七】 潘重規，《敦煌壇經新書》，佛陀教育基金會印，台北，1995 年 7 月

部首：鳥

序號：0196

頁次：446

B3		5154	5160	514C	5162	514F	5223	5169	5225	5221	5170
2	3	鶇	鶡	鶣	鶧	鶩	鶬	鶱	鶲	鶹	鶞
2	9	鸫	鹖	[illegible]	[illegible]	鹜	鸧	𫛭	鹟	鹠	䲠
2	F	[illegible]									
3	5	鶩									

部首：鳥

序號：0196

B3		5178	516B	516D	516E	5177	5173	5174	5231	5238	522D
2	3	鶻	鶼	鶿	鷁	鷃	鷇	[illegible]	鷖	鷕	鷙
2	9	鹘	鹣	鹚	鹢	鷃	鷇	鸮	鹥	鷕	鸷
2	F			鷀	艗	鷃				鷕	
3	5				榏						
3	B				鷁						
4	1				鶃						
4	7				鷊						

異體字非一時之產物，而是自古至今時時都在演變中的。例如：《說文解字》中「頁」居左者有三字，《康熙字典》中減至一字，至《字匯補》中，「頁」全在右側，而沒有居左的字了。

除了異體字外，在語言方面亦有異體詞。例如：梅雨和霉雨，筆畫和筆劃，蜡梅和獵梅，思維和思惟等。《現代漢語詞典》中共收有1000組以上這樣的組合。這些異體詞多半不會引發缺字問題，然而它造成了「依語義情境而定」的異體字。在試圖解決異體字問題時，不能不考慮到這些語言學上的現象。

其實，漢字是有異字同形現象的。六書中的假借：「本無其字，依聲託事」，就是異字同形之所以產生的方式。然而歷代字書都將這現象納入一個字的字義擴充中，並不把它們列為兩個字。所以，凡字典、辭典中，字義不同源者，皆屬異字同形之現象。如：「鎬」本為周武王建都之名，而今有「十字鎬」。又如：外來語之譯名卡（熱量單位），克（重量單位），亦皆屬此。而新近台灣流行之Cool作「酷」，Shaw作「秀」，亦假借也。

如果算上這種異字同形現象，「字」與「字形」的關係可以說是多對多的對映了(many to many mapping function)，許多字可以共用一形，而一個字也可有許多不同形的異體。然而，目前我們暫不處理字義，所以在「字」與「字形」關係上，仍是用一對多的數學模式處理它們。

肆、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。

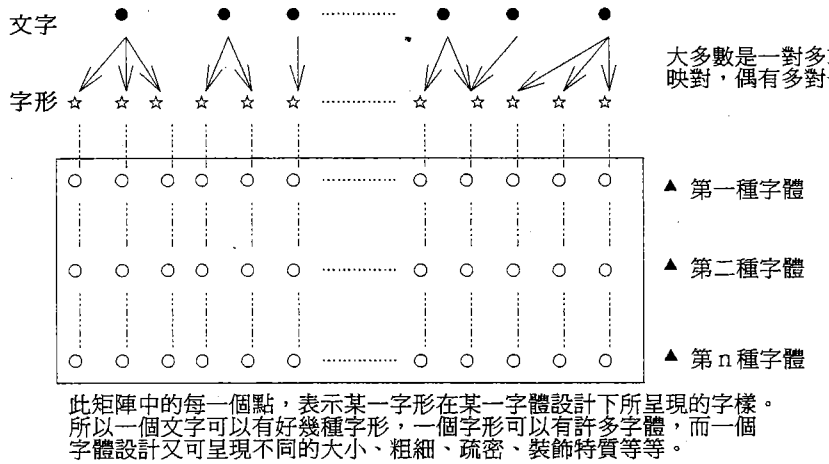
字(character)是表達一種或一群概念的名相。它是抽象的，以語意區別。例如，對應的繁體和簡體是同一個字。在電腦中，字用一個識別碼(identifier)表示，此識別碼可以是交換碼，也可以是內部處理方便使用的「內碼」，或是輸入時的「輸入碼」。為方便計，以下的討論均以交換碼來代表字。目前，字所承載的語意還沒能表達在系統中，所以電腦並沒有方法可以直接處理語意信息。

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〔圖一〕。所謂文字的制式表達，即將〔圖一〕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。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。我們設計的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。



圖一：字、字形、字體和字樣的關係

伍、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：【註八】

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。

A_2：一筆劃尾部加勾，筆劃數和構字的字根不變。

A_3：一筆劃被另一種筆劃替代，筆劃數與字根不變。

A_4：增多一筆，筆劃數增1。

【註八】 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

A_5：減少一筆，筆劃數減1。

A_6：一筆劃由另二筆劃取代，筆劃數加1。

A_7：二筆劃由另一筆劃取代，筆劃數減1。

A_8：一群筆劃由另一群筆劃取代。

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）

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。

B_3：一部件（一群字根）由另一部件取代。

B_4：增多一字根

B_5：減少一字根

C、整個文字構字的改變函數

C_1：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）【註九】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。最近，Unicode及ISO10646在統一字形上均做了許多努力，他們把中日韓台等地的交換碼中的字形「相似」者于以合併，【註十】所使用的原則頗與上述者相容。是故在字形的統一上，各國應是已有相當一致的看法。可是在漢字一字數形的關係上，以乎仍未見標準組織有積極的作為【註十一】。本文花費甚多篇幅說明

【註九】 micro-difference 一詞為 Edwin Smura 先生所首用。

【註十】 請參照 ISO 10646-1, Annex S, 〈 Procedure for the Unification and arrangement of CJK Ideographs 〉。另在 Kenneth W. Whistler 〈 Unicode Approaches to the Extension of Han Character 〉, Chinese Unicode Workshop, U.C. Berkely, 1996 年 9 月 7 日

【註十一】 在 ISO/IEC/TR 15285 的 Annex E 〈 Examples of character to glyph mapping 〉(1996 年 8 月), 中略述及此問題, 然而對漢字而言是不夠的。

了字與字形的關係以及漢字字形的變化，應可知字與字形的對映是改進交換碼必須面對的問題。

如果一個字有許多字形，字集可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、韓國亦可各有主張。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可在電腦中用關聯資料庫的欄位表達。

陸、結語

交換碼是一種標準，所謂標準者，實是指工業標準，處理工業標準的態度是現實的，是只顧當下的，所以凡過時之技術規格，均于淘汰棄置不用而讓它走進歷史資料之範疇。如果用這種態度來對付文字相關的標準，並不妥當，因為文字是和歷史文化緊密結合的，我們是生活在歷史文化所建造的社會中的，除非我們要摒棄所有的歷史文化而生活，否則，應用計算機處理文字訊息時，怎能棄先人所用的文字而不顧？目前的交換碼標準，徹頭徹尾的是工業標準的做法，實缺少歷史文化體認，本文強調文字學中漢字一字多形的現象並述及國家標準字集的只顧現在不顧歷史的現實，都是造成目前交換碼不敷使用的主要緣因。要改善交換碼的缺失，必需要有這樣的體認。