B) 43

中國文字的未來研討會

談中國文字在電腦中的表達

謝清俊

中研院 資訊所

中華民國八十年六月廿一日

談中國文字在電腦中的表達

自從民國六十年，當我們開始設法用電腦處理中文的資料以來，中文就和電腦結了不解之緣。雖然在今天，幾乎沒有電腦不能做些中文資料處理的工作，可是，似乎也沒有電腦的中文資料處理能力可以讓我們完全滿意的。造成這種現象的根本原因，是目前電腦擁有的中文「知識」不足，以至於用起來扯襟見肘、有許多限制。這使得一些「中文電腦」表現得呆頭呆腦，而且粗魯固執。

然而，電腦終究是處理文字的利器，對文化的影響尤爲巨大[1]。在討論中國文字的未來時，檢討工具(電腦)和文字之間的相互關係，應該是相當務實的做法。是故本文主要目的，是探討目前電腦中已擁有了那些中國文字的知識，並期望藉此能提供些線索供先進們構思：如何才能使未來的中國文字和它的工具之間能相得益彰。

由於「電腦處理中文的能力」的問題，本質上是中國文字的知識在電腦中如何表達的問題，所以，本文將先對這個問題略作說明；然而並不涉及如何將文字知識饋入電腦，或是如何在電腦中建構這些知識等較爲電腦專業的論述。這裡的說明只是希望能提供一個健全的概念基礎，以作爲文字學者和計算機學者之間能夠溝通和討論的環境。在此之後，本文將介紹並檢討目前電腦中對文字知識的表達和處理之能力，並將歸納一些現況以供讀者評議。

一、文字知識在電腦中的表達

文字相關的知識，本是文字學和語言學的範疇。這些知識是人發展出來供人用的。可是，當我們要電腦幫我做的事情涉及到這些文字知識的時候，我們就必須把這些知識用電腦可以接受的方式，表達在電腦之中。這樣，電腦才能據之來做我們要他做的事。所以，電腦工程師們常說「知識的表達」這句話，因爲他們工作上常常遇到這類的問題。

在文字學裡，最基本的問題是對文字的定義。我們且以此來討論。文字的定義是不容易表示的，而且根據不同的角度、情境、或是目的，便有不同的界說。然而，目前在電腦中文字又是如何定義的呢?

如果在一個電腦中只是爲了顯示或印出中文字樣，便只需要將各個文字的字樣儲存在電腦中(不管它用什麼方法，例如點陣或向量，這些都不重要)，此外還要給每一個字形一個識別的編號(可能是地址或指標或其他，亦無關宏旨)。這些看來只是些沒有什麼「知識」的數據(data，或稱資料)，對電腦而言，也算是「知識」，只是這些是極膚淺的知識罷了。有了這些，我們才可以寫程式將字樣依我們想要的形式顯示或印出來。當然，這些程式也是「知識」，是運用資料來完成任務的工作步驟，它是屬於程序知識(procedure knowledge)。

以這麼個系統做例子，它對中國文字所實踐的定義是什麼? 只不過是一個字形表罷了(字形加編號)！這樣的界定當然和文字學中的定義相去甚遠，可是目前電腦中對文字的界定比這個例子好不了多少，所以難怪顯得呆頭呆腦的。

從近代語言學的觀點而言，文字是一個抽象的概念；它有一些表徵，如它的形、音、編碼等等；有一些屬性，如它的語意、它隸屬的部首、它的筆劃、音標等等；有一些相關的知識，如它的孳乳，語法分類，語意從屬，相關的構詞等等。凡此種種絕大多數是抽象的，而電腦卻是工具，是相當具體的東西；因此，要把文字定義的知識完完全全讓電腦「知道」（表達在電腦中）是不可能的事。換言之，要給電腦一個完完整整的文字定義，是不可能的。

若是我們把文字知識作爲「體」，則電腦能做的只是其「用」。由此觀點，電腦對處理中文資料的問題，是一個緣用顯體的問題。換句話說，充其量我們只能將文字的表徵、屬性、以及相關的知識等等，也就是關於「用」的各方面，能具體表達在電腦中者，儘量去做，越多越好。這樣，庶幾乎在電腦中能匯集許多應用，以使得它的表現之能趨近於眞正文字定義的實踐。這就是目前我們能做的。

雖然上面所說的，給電腦能做的定下了個概念上的界限，然而，我們目前所做的卻仍然離這個界限甚爲遙遠。換言之，目前我們無需將這個界限放在心上。重要的是，當我們有了這個概念之後，我們可以以此檢查現有的電腦系統，較有條理地釐清目前電腦處理中文資料的能力，也希望藉此能明白文字知識與電腦間相互的關係。

以上之觀點在大陸的學者已有類似之體認[7]。值得參考。

二、中文字碼

在上文中，我們舉了個中文字表的例子，來說明文字在電腦中甚爲膚淺的表達。雖然是膚淺，但它卻是中文電腦中不可缺少且最重要的基本資料。中文字碼的內涵和字表差不多，它包括一個編碼的系統，給每一個中文字一個碼，此外它有一個字集，即是我們要用的一群字和它們的字樣。與上例中不同的是：我們希望能設計一個大家都認可的字碼，以供分享資訊之用。

由於涉及字集和字樣，編一個字碼便須考慮下列問題：

(1) 要多少字？是那些字？如何排列？

(2) 要什麼樣的字樣？

(3) 一個字有許多字樣時，怎麼處理？需不需要標清楚他們之間的關係？

如果深究下去，會涉及數不清的文字學裡的問題。即便是以上述的三個問題而言，二十幾年來，世界各地所有的編碼工作，都還沒有找到令大家都滿意的解答！令人驚訝吧，且讓我們看看表一目前各國已發表的字碼，並作一些說明。

表一中所列的是交換碼(interchange code)，這是專門給系統之間分享資料用的碼。這些碼，除了中文資訊交換碼(CCCII)以外，都是國家訂定的工業標準，換言之，它們是廠商生產「中文電腦」的依據。

表中的訊息可分爲二段來看，在1985年以前，只有三個碼：日本的JIS

表一：世界各國漢字交換碼編年表(1991,6.15製表)

序號	發表年月	發表國家	交換碼名稱	編碼字數與結構	補充說明
(1)	1978	日本	JIS X 6226	共6353字，分爲兩級：第一級2956字，第二級3388字	現已不用，爲⑤JIS X 0208取代
(2)	1980.3	中國，台灣地區	中文資訊交換碼，第一冊 CCCII,VOL.1	CCCII常用字集共4808字	爲正體字。
(3)	1981.2	中國，台灣地區	中文資訊交換碼，第二冊 CCCII,VOL.2	共33357字，其中常用字4808字(同CCCII,VOL.1)次常用字6025字簡用字11007字，異體字11517字	異體字中含中國大陸地區之簡體字(漢字簡化第一方案中之所有文字)
(4)	1981.4	中國，大陸地區	GB 2312	共6763字，分爲兩級；第一級3755字，第二級3008字	爲大陸用簡體字
(5)	1983	日　本	JIS X 0208	共6537字，亦同 JIS C 6226分爲兩級	取代JIS C 6226, 新增4字字形取日本戰後簡化之漢字
(6)	1986.2	中國，台灣地區	中文資訊交換碼，第三冊 CCCII,VOL.3	共53940字，是CCCII,VOL.2再擴編罕用字20583字	
(7)	1986.10	中國，台灣地區	CNS 11643	共13051字，分爲： 常用字集5401字， 次常用字集7650字	
(8)	1987	中國，大陸地區	GB 7589	共 7237 字	爲簡體字，是GB 2312之第二輔助集
(9)	1987	中國，大陸地區	GB 7590	共 7039 字	爲簡體字，是GB 2312之第四輔助集
(10)	1987	南　韓	KS C 5601	共 4888 字	
(11)	1988.6	中國，台灣地區	CNS 11643 增　補	共19199字，爲CNS 11643 增編6148罕用字	
(12)	1989	美國(及ISO)	EACC, ANSI/NISO Z39.64	共15686字，包含共用字集10934字，假名及符號1600個，其餘爲選自大陸、台灣、日本及南韓之漢字	採用CCCII碼之結構，故與CCCII相容(CCCII中有之字其碼相同)
(13)	1990	日　本	JIS X 0212	共 5801 字	爲JIS X 0208之輔助集又稱爲JIS第三級(level 3)字集
(14)	1990	中國，內陸地區	GB 12345	共6866字，爲GB231之對應繁體字(其中有103字是多重對應者)	GB 2312之第一輔助集

最早，有6353字；中文資訊交換碼其次，有33357字；大陸的GB有6763字。這些字碼是初期的產品，JIS和GB之結構和理念都類似，字數也差不多，然而中文資訊交換碼卻與之不同，不僅字多且把字劃分爲正體字與異體字，作有系統且相關的編碼[2,3,4]。用前文說過的觀念來說，它們之間對文字的定義不同，所蘊含的文字知識就不一樣，因此功能也就有別。換言之，JIS和GB只考慮了本節中前面所列的(1)及(2)兩個問題，而CCCII卻已經考慮到(1)(2)及(3)。以第(3)點而言，現在大家都明白他的重要性了，可是有許多碼卻因結構已定，而無法大幅改進，JIS及GB就是如此，他們只能把輔助集略爲修訂作爲補救而已(請參見JISx0208及GB各標準資料，不另列參考)。目前正在設計中的ISO 10646即對(3)作爲編碼的考慮[11]。

從1985年以後，編碼的活動比早期活躍許多，除表列正式發表的碼以外，正在編製中的還有：GB的第三和第五輔助集，這是對應GB75、GB89和GB7590的繁體字對應字集，字數比二者略多；ISO 10646和UNICODE，是試圖將中日韓所用的漢字納於一個標準碼之下的嘗試；而CCCII也正在編第四冊，將新增約22000字，而將其蒐集之字集擴充到約76000字，其中正體字四萬四千餘字，異體字約兩萬兩千字。

縱觀1985年以後的發展，有一個明顯的現象，那就是各地區都覺得字數不夠而努力地增加。日本由JISx0208擴充了JISx0212使字總數增加至12158字；大陸由GB2312擴充了GB7589,GB7590,GB12345，再加上第三和第五輔助集，其總字數將至四萬三千字以上；CNS11643由13051字增至19199字；CCCII將由53940添到七萬六千字。這些現象只說明一個事實：電腦的應用成長太快，字老是不夠用。然而，究竟要多少字才夠? 很明顯的是：除了像GB和CCCII這樣有蒐集盡所有字的雄心者外，恐怕使用時總會有些字不在其字集之中。

其次一個明顯的趨勢，是爲了避免中日韓文中相同字樣的重複編碼及促進資訊的分享，有將世界上各地區用的漢字合併編爲一碼的趨勢。中共對表一所列的各字集曾做了些重複字的比對[5,6]，日本亦然，而字樣交換協會(Association for Font Information Interchange，簡稱AFII，爲ISO下設之籌辦字形(glyph)，字體(font)，及字樣(type- face)標準的組織)亦廣集

各地區標準字樣而加以對比，並提供比對之結果供ISO 10646，UNICODE, EACC及其他團體參考。這些都是對整合中日韓所用的漢字所做的基本功夫。

有趣的是這些比對的答案並不一致(如中國大陸與日本所做的)[8]，可是字集之間相同的比例都很高，在50％至85％之間。CCCII在早期曾將GB2312及JIS C 6226的字編入CCCII中，結果只增加了約500個「不同」的字[9]。

三、字、字形、字體與字樣

爲什麼字集之間的比對會有不同的結果呢？這是因爲同異之分的條件尚未建立起共識。大陸曾對字的認同規劃作了下列的建議[10]：

A.同一漢字字形近似的可按下列細則進行認同：

A1. 字形結構不變，只是筆劃小有差异的字，認同

A1.1 筆形變化，如：文，文，文。

A1.2 筆劃類型不同，如：户，户，戶。

A1.3 筆劃曲率不同，如：示，示。

A1.4 長度有異，如：天，天。

A2.斷筆與連筆，如：免，免。

A3.筆劃之增減，如：者，者。

B.偏旁有細微差別的字(不含簡化偏旁)，可以認同如：糸，糸。

C.下列的差異不予認同：

C1. 因簡化造成的差異，如：单單，对対，团団。

C2. 下列情況，應涵蓋現有各標準之規定。

C2.1 同一字，因結構方式不同形成的變體，如：峰峯。

C2.2 同一漢字，在現某個標準中已列入兩個或兩種以上字形的，不再認同。如：曾曾 劍劍劒刄劒刃劒，苺莓。

在JIS X 0208中，亦對認同之標準加以說明(參閱JIS X 0208說明書第3.4節)，其內容與上列者大同小異。然而這些規則都有毛病，例如：沒有一條規則沒

有例外。復次，這些工作只是純就文字之外觀加以歸納，欠缺理論上的依據。如此一來，勢必有欠週延。

要解決這些問題，就必須在電腦中對文字的定義或是其所蘊含之知識，作較深入的規範，不能只在編碼上想辦法。首先，要弄清楚什麼是一個字(character)。在語文學上說，它只是一個抽象的概念。因此，概念相同者，或經語文學上認同的字，在電腦裡也應該是一個字。上述的C2.2就顯然違反了這個原則。若根據C2.2，把羣和群認爲是兩個字，那麼就弄擰了文字學和語文學，使它跟著科技工具走了。這是絕對不可以的，我以爲科技要配合文化的需要走，而不是委曲我們的文化傳承，讓它隨科技而逐流。學科技的人，往往會因人文素養不夠犯下這種錯誤。

其次要了解一個字常常有許多形(形、音、義的形)，而不是一字一形。早期電腦的應用不廣，而性能又不高，所以其中對文字的定義都是一字一形。當電腦有必要處理一字多形的問題時，這種設計註定會報銷掉。所以，在電腦中必須給形下個定義，如此才能將字與形之間之關係(知識)清楚地在電腦中表達。

其三是除字形的變化外，還有字體之變化。例如，群和羣都有楷書、仿宋體、宋體、明體(日本)、隸書、行書、草書甚至鍾鼎甲骨等字體的變化。最後，還有字樣的變化，同樣是仿宋體，聚珍閣的就是與眾不同。此外在同一字體中字也有大小、粗細等等的變化。而這些都是字樣變化之列。

字(character)，形(glyph)，字體(font)都是抽象的概念[7]，只有到字樣時，當各種參數的值，如寬窄、高矮、粗細等等都定了下來，才會產生能讓我們看到的實體字貌。所以，在未來電腦中，一定要對上述之字、形、字體、字樣等作適當之界定與表達，才能處理與這些相關的問題。

筆者曾協助AFII解決了界定字和形的問題，亦即是解決了字的認同問題[12,13]。其方法是由文字之孳乳和構字之法則爲基礎，仔細將筆劃之差異分類，以字之抽象概念爲基準，並考慮各國文化上之差異，而歸納出判斷不同的字和不同的形的規則。此規則已獲各國一致之贊同，而AFII正循此規則整理並整合各地區之漢字。

四、正體字與簡體字

在以往，正體字和簡體字之間的爭論是大家都知道的事，當時爭論的情形似乎是存亡絕續之爭；也就是要在二者之間擇一而行。然而在電腦裡，並沒有這麼尖銳的衝突：因爲此二者都是存在的實體，電腦必須要能處理二者，才不致有歧視，或應用上的偏差和缺憾。

CCCII在設計之初就蒐集了簡體字[2]。當時雖然有些人對此惡意攻擊(其攻擊實是另有目的)，然而並沒有造成問題。對正體字而言，增加了簡體字只是增加了處理上的成本，而在技術上是沒有什麼困難的。美國1982年採用了CCCII作爲圖書界的標準之後，不只可在一電腦中同時處理正、簡兩種字形，連日韓用的漢字均可交互地輸入、輸出、或是做彼此間的轉換[14]。例如，一個不懂中文的日本人可以用日文輸入法找簡體字的書目資料，並可將之用正體字印出來。這個例子不僅說明了正簡字體在處理技術上不成問題，也說明在一個字碼中，若能包涵字與形的關係是有好處的(參閱本文第二節中文字碼)。

照前所述增加了簡體字，似乎只是增加了「多一個字形」的麻煩，其實不然。簡體字改變了文字的屬性，譬如：部首、筆劃、構字成份等等；亦改變了些文字相關的知識，譬如：在構詞上它使有些詞變成有兩個以上的意義；又如使某些字的語法分類和語意構成變成更爲複雜，如：干字的語法和語意成份便成爲原來正體字干、乾和幹的聯集，而干字相關的詞彙亦爲原來三個字的總和。

由於上述的問題目前欠缺具體的研究，其影響究竟有多麼大實在難以定論。然而，可以肯定的是：由於簡體字而產生的多義詞彙問題，已迫使將原先一些不需參照上下文就可以解決的語法問題，變成爲非參照上下文來解決不可。這個現象將使得在中文句法的分析、語意的了解、語法的構成等等問題上都變得非考慮使用參照上下文的模式(context sensititive model)不可。這個代價不可謂之不大。

五、結語

本文從知識表達的觀點來看文字在計算機中的生態。由文中之討論可知，目前文字在計算機中的表達實在是先天不足——有些文字學語言學方面的問題並未能解決而影響到計算機處理中文資料的能力；後天失調——目前「中文電腦」的設計不是很好，文字學者和語言學者的參與不夠，計算機工程人員又過份自我澎脹，造成文字知識在計算機裡頗爲貧乏。這現象嚴重地影響到計算機處理中文資料的能力。爲未來計，這些缺點是終究要克服的。

簡體字帶來的麻煩不只是添一個字形而已，它對句法、語意，以及文法上都有影響。這方面的研究是有必要進行的。以正體字做的中文自然語言處理的研究[15]似乎可以作爲解決簡體字做中文自然語言處理所帶來的一些問題的基礎。這個關係亦有待研究澄清。

在本文中，我們對於中文相關的屬性資料，和相關的文字學和語言學方面的知識如何在電腦中表達之事，未能詳爲探討。事實上，這方面的研究工作已做了許多[6,15]，希望以後能有機會撰文將這部份補齊。

在本文中，我們亦未能就各種應用的角度來討論中國文字的問題。其實，由許多應用的文字統計上應該可以供給我們許多計量的資料，使我們更正確的了解中國文字在應用上的各種現象和問題。

中國文字在電腦中應如何表達？文字在電腦中的定義應該如何實踐？這些目前都沒有理想的解答，現有的系統更是毛病叢生。這些問題似乎應該作全盤檢討並重新做起才是。

參考資料

1. 謝清俊，＜論中文資訊處理系統的發展＞，中文資訊處理研討會，資策會主辦，台北，民國78年11月14日。
2. 國字整理小組，《中文資訊交換碼》第一、二、三冊，台北，國字整理小組，1980至1985。
3. 謝清俊等，＜The design and application of the CCCII＞，中文圖書資料自動化國際研討會，台北，1981 2月14-19日。
4. 楊鍵樵等，＜The design of the CCCII and its application in library automation＞, IFLA（國際圖書聯盟協會）1981年會，東德柏林，1981,8月17日。
5. ISO-IEC JTCI/SC2/WG2，在北京臨時會議及在安曼第十五次會議，中共官方提供之N513號資料，1989 10月16日。
6. 通用中文代碼國際聯合會 (Association for common Chinese code, International)，《中文信息處理技術的現況與進展》，北京1991，3月。
7. ISO-IEC JTCI / SC2 / WG2，漢城臨時會議，中共官方提供之N561號資料，1990, 2月16日。
8. 中國電子計算機及訊息處理標準技術委員會 (CCIPS), ＜Proposal for creating unified CJK Ham Character Collection (HCC)＞提供給ISO/IEC JTCI/SC2/WG2委員們之建議書，北京，1989,7月22日。
9. 楊鍵樵等，<The CCCII and its feasibility for international adoption>，中文書目自動化會議，澳洲坎培拉市，1982,8月29日。
10. ISO-IEC JTCI / SC2 / WG2，中共提出之正式建議＜漢字認同規則＞，1989, 9月，修訂版。
11. 張軸材，＜關於新字符集中簡體字／繁體字的編碼＞，個人意見供ACCC /TC-C會議參考，ACCC/TC-C/90012，北京1990。
12. 謝清俊等，＜On the formatization of glyph in Chinese language＞，AFII文獻編號AFII/90-10。1989年11月。亦曾於1990我國文字學年會上報告。
13. 同[6],pp.73-75。
14. 謝清俊，黃克東，＜國字整理小組十年＞，國字整理小組出版，台北，1989, 12月。
15. 陳克健，中文詞知識庫小組，＜中文詞知識庫計劃與中文電子詞典＞中央研究院資訊所，台北，1991。