

缺字手冊・序

謝清俊 91.10.09

在 1994 的時候，中央研究院數位化的古籍已有相當的數量，因之而累積的缺字也已超過了五千個，形成了相當嚴重的問題。缺字是由於漢字字碼中的字形不足引起的。資料庫有了缺字，不僅影響到使用者對古籍的閱覽，更影響到對古籍的了解、檢索、列印、以及後續的種種處理，包括資料的交換和共享等在內。有鑑於此，本實驗室開始設法解決此一問題。

解決缺字問題的根本，在於解決現行漢字交換碼的根本缺失。漢字的字集是一個開放性質的；也就是說，漢字的字集依古今的變異、專業與應用環境的差異等而有字數、字形、以及字音、字義上的變化。音義且不談，僅就字數而言，即已不適合固定數量的限定；這與數量已定的西方語言的『字母集』，是不可以一概而論的。然而，現行漢字交換碼的結構，卻仿照西方語言的字母集的結構來設計，這不能不說是『削足適履』。

再者，語言是有生命的，任何活在當下的語言，都需要有創作新詞（word）的能力，否則必定無法配合社會變遷之需，而終將面臨被淘汰的命運。我們不時看見有英文新詞（字），就是個很好的例子。可是，在電腦中限定漢字的字數，就斬斷了漢語言電腦中創作新字（詞）的生命力，埋下了日後不能適用的因子。這個問題也不應等閒視之。

我們解決缺字問題的方法，是遵從漢字構形的原理，對漢字字形的結構做制式表達與詳細的分析。一個字的字形結構式，是該字極佳的識別符號與工具；因為字形若不一樣，則字形結構必不相同；反之，字形結構若相同，其形亦必同。事實上，漢字字形結構的拆分與分析，淵源流長。撇開文字學的研究不談，最早為電腦拆分漢字的研究，見於 1972 年交通大學關於『交大字根系統』一系列的論文和報告。由於交大字根系統的相關電子檔案已不能讀出，於是在本研究剛開始的時候，我們就決定沿用交大字根系統的方法重新再做一遍，並在過程中作若干的改進。這段工作最大的工程，就是建立了『漢字構形資料庫』。

為了建構漢字構形資料庫，我們澈底檢查了《漢語大字典》、《說文解字》和一些交換碼與各國所頒布的字集標準，分析了其中的每一個字。所以，目前建構的系統並不僅僅是做構字的表達與識別，還做了兩件大事：一是古今字形與古今構字的銜接和對映，二是異體字形的表達和處理。這都是電腦理多年來懸而未決的問題。

在古今文字銜接和對映的部份，我們在中央研究院歷史語言研究所古文字組的指導和協助下，已逐字分析了《說文解字》中所有的字形，建構了小篆字形的構形資料庫，並且已和『漢字構形資料庫』做了對映與聯繫。所以，從我們的系統，使用者可以從現代字形直接聯繫到小篆與楷阿化的小篆字形，也可由小篆的構形直接查核現代字形。換言之，使用者可不必具有任何文字學的識識，即可查閱到相當多古文字相關的知識和資料。我們相信，這對漢語的教學和應用是有幫助的。至於小篆以前的字形，主要是大篆（鍾鼎文）和甲骨文，則正在建構中。大篆的構形資料庫部份將於半年內完成，並將與歷史語言研究所古文字組所做的

『金文資料庫』銜接和對映。甲骨文部份則尚待努力。

在異體字形方面，我們的系統中已將《漢語大字典》中一萬兩千組以上的異體字表建構成了異體字資料庫。此外，系統也建構了異體字根與異體部件的對映關係，以使使用者能更人性化的使用本系統。我們相信，異體字資料庫能協助我們解決久懸未決的異體字問題，特別是在資料檢索方面，可以提高檢索的精確度和召回率。

由於本系統的設計理念與目前的交換碼完全不同，所以本系統可以與使用任何交換碼的系統相容。換言之，任何亞東文字的處理系統，無論是中日韓越、無論是簡繁，都可以附加本系統作為澈底解決缺字的機制。目前我們推出的是基於五大碼（Big-5）的系統，可將五大碼能處理的約一萬三千字形，立刻擴充到六萬以上，並擴及小篆和異體字等。更重要的是，這能力的擴大，並不需要佔用原交換碼的編碼空間，也不需要修改原來的應用軟體！

本系統的發展，歷時七年餘。此間得中央研究院資訊科學研究所的信賴與支持最多。如果資訊科學研究所對本實驗室的信賴不足，是不可能讓我們花七年多的時間，來做個這麼繁重墮長的工作。在此特別致謝。此外，為了要開發在網路與 WINDOW 下的應用系統，教育部（圖書館事業發展委員會）與國科會（數位博物館和數位典藏計畫）在最近兩年的支援，也有雪中送炭、臨門一腳的效果。這也是我們要特別感謝的。在系統工程方面，主要的工作人員是莊德明先生和許永成先生。他們兩位任勞任怨和不眠不休的毅力，是使得本計畫能堅持到最後一刻的關鍵，也是推動本計畫的主要動力。曾在本實驗室工作的所有人員也是應感謝的，因為在我們忙得不可開交的時候，總是他們挽起袖子來幫忙。此外，在此七年中，本計畫與新竹科學園區的易符公司經常相互切磋、交換心得，是一段令人懷念的愉快時光，值得一記。

最後，本實驗室謹以謙誠的心，將此系統免費地獻給所有的同胞，願您使用本系統獲得利益！畢竟，文字是我們文化的根，能在這個根上盡一份力量，已是我們最大的福報。

謝清俊

中華民國九十一年十月

於資訊科學研究所文獻處理實驗室