中文電子化發展的新趨向

江德曜

中文發源久遠，詞藻豐富，字形之構成是多方面的，龐大繁雜，據中華大字典的蒐集多至四萬五千字，目前習用仍有八九千字之多。這些衆多的各別單元，在文學藝術的觀點上，也許是多多益善，但在資料處理方面，實在是一項冗重的負担。我們如何來簡化文字的處理工作，以求配合現代迅捷的資料機器，實是當前亟待解決的問題。

自從電子計算機發展以來，大批資料（包括數字與文字）送經機械處理，已成爲近代社會進展的必然趨勢。中文文字的機械化（卽電子化）更感迫切的需要。相對地，當今電子技術的迅速發展，也有助於中文電子化的推動。中文繁複字形的分析、組合、辨認、貯存、摘取、描繪等項步驟，以前試圖用半電機半機械的方式處理而無法順利完成的工作，目前藉電子之助，均能迎刃而解。現在我們把中文電子化的步驟，列如下表，隨後逐步說明：

㈠應用檢字法選取所需單字，採用鍵盤按鍵譯成電碼。

㈡電碼直接輸入計算機中或先鑿成紙帶、卡片、或錄成磁帶等然後輸入計算機中。

㈢在電子計算機中資料的貯存、分類、整理、計算、歸檔。

㈣所需資料的提取與顯出或印出。

一、文字的檢取與譯碼

文字必需譯成簡單而有系統的表達電碼，才能爲計算機所接受。西文的表達方式，由幾十個字母和標點符號拼湊而成，只要用一小小鍵盤，就能隨心所欲，打出任何文件，比手寫快速而整齊。而且由字母譯成電碼，機構簡單，動作快速。中文雖然也可用打字機打出，但字數衆多，檢尋不易，速率大打折扣。要想達到快速自如，人人會用的目的，勢非從文字本身的檢取方式上下功夫不可。

中文的檢取可從字音及字形兩方面著手。由字音檢字（如注音符號和五聲）主要困難爲同音字太多，而且一字數音或數聲的也累見不鮮。一般人的字彙有限，遇到不識的字或讀錯的字只好臨時檢查一番，中途阻礙必多。有人把字音和字形合併決定一字的符號，如美國Kansas大學C.Leban氏的SINCODE，及近日曾麗明氏所著「曾氏機器用中文字碼編譯法」。但我認爲符號太長，錯誤的機會多，用來做計算機的輸入，還嫌累贅。

我以爲中文檢字從字形着手，比較合理。許愼氏的說文，分字爲五百四十部首，是最早而有系統的文字分析法。後來梅膺祚氏作字彙，將部首減爲二百十四，依筆劃多寡排列，每部首中所包含的單字，也依筆劃多寡編排，沿用迄今。由於部首數目仍嫌太多，每部字數不均，數筆劃太費時，不適於機械化應用。近世王雲五氏的四角號碼，首先以筆劃形狀直接換成數

碼，每碼所含字數均勻，可惜未能達到一字一碼的程度，不能逕行作爲電碼。至於現行中文電報號碼，雖爲一字一碼，但記憶困難，不能普及應用。近時能直接送入機器的實用譯碼機，還是採用整片的字盤（如現行中文打字機）以常用、偶用、罕用及部首分類檢字，如高仲芹氏的中文譯碼機，日本輿國，富士的日文譯碼機都採用此式。這種字盤譯碼方式，依字的位置按鍵譯碼，能確認字形，不致差錯，譯碼機構也簡便經濟。但盤面字體小，字數太多，如非熟手，檢尋不易，其中單字的分組方式，也值得研究改良。此外林語堂氏的字首字尾檢字法，每字依首尾按鍵兩下，藉機械之助，即顯出一組字形，由其中選出所需之字，再依其位置按鍵，即能確定該字，予以打出字形，或譯成電碼。美國Ifek公司製作的Chicoder 中文譯碼機即是採用林氏的檢字法選字。此法的優點，爲縮小檢字的範圍，比整片字盤檢字容易快速多了，其缺點爲同首尾的字組有時字數太多不能一次同時出現。台北工專的教授陳舜齊氏，以字之首、次、末三字形檢尋一字，字組的範圍較林氏法更爲縮小，但仍不易達到一字一碼的目標，必需添加附碼，記憶上仍有困難。

近時的研究，除基本字形外再加上字形間的定位符號，如橫排、直排、包含、交叉、重複、對稱等。如日本東京大學的O.Fujimura氏及R.Kagaya氏，採用21個字根，每字根以二至三個連接點，以不同的定位符號來表示其間相連的關係。國內專家學者，近來研究者也頗不乏人，如民航局的冀家琳氏用三十字根及十種定位符號連貫排列，據謂可達到一字一碼的目標。成功大學的王惠然教授陳述該校的研究工作，由一萬六千字中，採選幾百個字根，希望精選到二百個左右，用作按鍵選字。交通大學的謝清俊教授指導研究生倪耿採用四種定位符號用樹枝結構法來代表各種字形模式，並利用計算機統計每字符號的平均長度，以常用六百字爲範圍，如字根數爲一九六，每字的平均長度約爲三個符號，如字根數減至一三四，那麼長度將增加至三、七個符號。這是純由統計的立場來選擇適宜的字根及字根數，但爲適合人們的動作習慣，其他種種因素也應該考慮入內。

在六十二年暑期中由中國電機工程學會聘請旅美學人返國在台舉行的工程研討會中，鄭國賓氏曾發表他最近研究完成的中文電子化計劃，採用四百個字根，依陳立夫氏的五筆檢字法原則（改用六種筆劃）分字根爲十二類。利用現有英文鍵盤按鍵選取一類，並利用示波器顯出該類字根，由其中找出所需字根，再依其位置（與鍵子排列相當）按鍵決定該字根。循此反復行之，可把一字所組成的字根及其相關位置全部打出。此法手續雖較繁複，但每次所選字根可經計算機自動予以組合，最後完成一字，中間毋須經翻譯工作，是其長處。

由於字形之複雜，字體之不統一（有宋體、楷書、及古、俗、簡、異之分），筆劃順序之歧異，要求人人易學方便而不致差錯還需要繼續研究和不斷的試驗。本人蒙中正科學技術研究講座基金委員會及交通部電信研究所之資助並承台灣大學電機系同仁之合作及應用國家科學會補助之儀器，從事此項中文電子化計劃已有三年之久，現已能按鍵打成卡片，逕行輸入計算機中摘取字形。爲求易於記憶起見，只用三十二個鍵子，每次僅按一鍵即能選出一組字形，由簡便的顯字機顯出。如此將

檢字範圍縮小至三十二分之一再依細節分類檢字，方便得多。選字手續簡單，錯誤機會少，即使有誤，改正亦便。本法採合字盤檢字法和字根檢字法之長，既可避免記憶衆多字根及其冗長組合之規則，又減少整整文字檢尋的困難。且每次檢字有字組逕行顯出（不經過計算機，不消耗其工作時間），有自行校驗的功用。

二、電碼的輸入方式

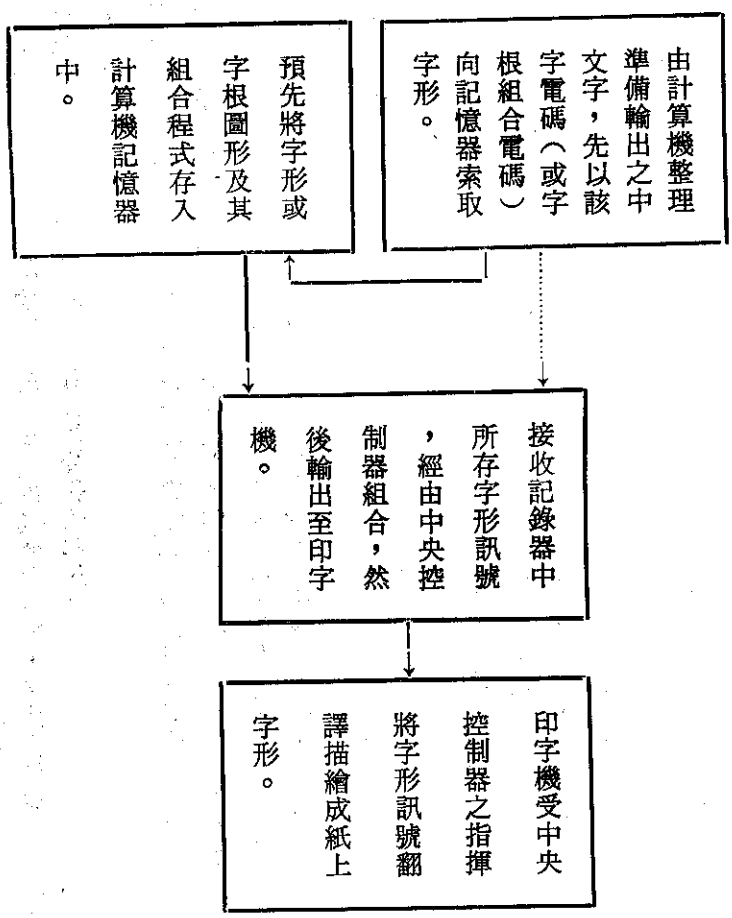
上述由鍵盤打鍵經譯碼電路譯出的電碼（爲斷續或正反的電流訊號）可以直接輸入計算機中，或先經打卡機打成卡片，或經鑿孔機鑿成紙帶，或經磁帶錄碼機錄成磁帶。前者稱爲上機輸入方式（On-line），鍵盤直接和計算機相連接，每打一鍵，訊號輸入一次。由于人工打鍵速率遲緩，本法佔據計算機太多，除非採用分時裝置（Time—Sharing System），本法殊不經濟。後者打鍵時與計算機脫離，稱爲離機式（Off-line），電碼先錄在紙帶，磁帶或卡片上，然後累積以高速一起送入計算機中，可節省計算機時間。

三、資料的貯存、處理和輸出

盈千累萬的資料，包括數字和文字，如人文、地理、事物項目等，都能經由電碼的形式貯存在計算機中，經計算機加以分類、整理、統計、歸檔，最後以圖表的方式顯示結果。以前計算機中的資料祇能以英文表達，究竟是外國的東西，不能深入我們的社會。現在我們既可把中文資料輸入，也能以中文的形態把所需結果顯現出來。如銀行存款記錄、學生成績報告、營業統計、帳册、以及戶稅、水電費等通知單等，其中人名、商號、地址必須以本國文字填寫，如能由計算機逕行印出，節省人力，將不可以道里計。

四、顯字機或印字機

以前計算機文件的輸出，是用打字機或字模型逐行印字機印出。若改用中文，字模太多，必然結構笨重，動作遲緩，無法與現行高速計算機配合應用。現時有一種靜電感應描繪兼印字機，採用點陣圖(Dot matrix)的方式能印出任何圖形（包括漢字），圖形的分配，除送紙外純以電子方式進行，故速率可高達每分鐘英文字一萬二千行，如以每行八十個字母，每四個字母相當於一中文字的面積，，那麼每分鐘可印出二十四萬個漢字，即是每秒鐘四千字。此外有鋼絲衝擊型（Wire Printer）、烙印型（Thermo Printer）噴墨型等也能在紙上印字。另外也可用陰極示波器顯示，並隨時能把字形複印下來（Hard Copy），這些裝置都靠電子電路控制，不受機械呆滯的影響，所以雖字數衆多，字形繁雜，仍能迅速印出，處理自如。不過每字字形，或字根的組合，要預先在計算機記憶器（Memory）中存貯起來，每一字形訊號之前編一號碼（就是代表該字的電碼）。印字時由計算機中央控制器依碼摘取字形，經由印字機將字形印出。凡此種種動作，純由電子控制，所以說來雖長，實際動作祇是剎那間事。最後把計算機的印字步驟，用流程圖表達如次：


由計算機整理準備輸出之中文字，先以該字電碼（或字根組合電碼）向記憶器索取字形。
預先將字形或字根圖形及其組合程式存入計算機記憶器中。
接收記錄器中所存字形訊號，經由中央控制器組合，然後輸出至印字機。
印字機受中央控制器之指揮將字形訊號翻譯描繪成紙上字形。