

中文資訊的處理（三）

科學月刊第十八卷第五期

中華民國 76 年 5 月

謝清俊

在末期中，我們將討論中文資訊處理的基本符號集的問題。它是中文資訊系統要處理的基本對象。在討論此問題時，我們將發現：中文資訊處理系統已是一個中英文的雙語系統；中文字集是一個開放的集合；中文字集有許多語言文字學上的問題，包括字數、字形、字序等，使得中文資訊處理系統所要處理的基本符號集變得十分複雜。

符號的世界

計算機裡的世界是符號的世界：任何不用的訊息，無論其原始形式是以文字、數字、公式、圖表、聲音所表達的，或者是自然界中任何的物理量，或是文明世界中的抽象概念等，當它們置於計算機中時，都是用一群基本符號（symbols）來表達及處理的。所以，計算機可說是一個處理符號的機器，而這群基本符號集也因之顯得十分重要。

在任何一種計算機系統中，這個基本符號集都是有限的，而且它包含兩種符號：一種是控制符號（control symbols），另一種是圖形符號（graphicsymbols）。控制符號是用來操作機器用的，通常不用它來表達被處理的資訊。一般而言，控制符號包括控制周邊設備和通訊設備的符號，以及表達資料格式（format）及資料間隔的符號（information separators）。圖形符號是表達資訊用的，它和各國的語言有關，包括字母、數字、標點符號，以及一些常用的數學及特殊符號等。

為了使計算機之間可以分享資訊，並且也為了使週邊設備及通訊設備可以適用於不同廠牌的機種，上述的基本符號集需要有個共同的標準。這個標準直到 1973 年，才由國際標準組織（ISO）設定：編號為 ISO 646，全名是：7 位元資訊處理標準交換碼。

在 ISO 646 中，用了 7 位元的空間，所以共有 $2^7=128$ 個位置可供安置符號；其中控制符號 34 個，圖形符號 94 個（圖表省略）。

ISO 646 是一個母法標準，各國可以依據其規定，設計自己的國家標準。為了達到資訊與設備共享的目的，控制符號是不可改變的。換言之，根據 ISO 646 設計自己的國家標準時，只可重新定義其 94 個圖形符號。這個標準對拉丁語系的國家是十分適用的，只要適當地修改圖形符號中的字母部分，就可言刻適用於該國，且可與其他各國的設備和資訊互容。

中文資訊處理的環境

對於使用漢字的國家而言，就覺得此一標準不能適用了，因為 94 個圖形符號的空間無論如何也容納不下成升上萬的漢字。於是 ISO 又制定了 ISO 2022 的標準來擴充。ISO 2022 允許將信元碼擴充至 8 位元，或是擴充為許多個位元碼一起來用。譬如：二個 7 位元碼的圖形符號就可以有 $94*94*94$ 個，這已超過了 83 邁個符號所需的位置，即使要容納全世界的字母和漢字也是足足有餘的。

在這兒必須注意的是 8 位元的擴充方法。雖然 ISO 2022 提供了這種擴充，可是 8 位元的碼和許多通訊設備無法相容，因此也就無法利用或通過這些通訊設備。譬如：非同步通訊介面就是一個例子。再者，由於 ISO 646 是母法，以 ISO 646 為基礎而孳生的 ISO 標準就有幾十個，其中包括磁碟及磁帶的存錄標準在內，當使用 8 位元碼時亦無法與這些生之標準相容。所以，8 位元碼的使用還是有許多限制的。

由於美國是計算機王國，而英語又是最通用的世界語，根據 ISO 646 所訂定的美國標準交換碼（American Standard Code for Information Interchange，簡稱 ASCII）就成為計算機界最用的基本符號集。幾乎所有的計算機都必須用它，甚至在我國設計的計算機也不例外。

以上所敘述的，就是目前中文資訊處理所面對的環境。為了與世界各國分享資訊與設備，我們必須遷就 ISO 646 的標準。如何在這樣的限制下有效地處理中文資訊，就成為非常值得研究的課題。

不平等的待遇

常有人說：「計算機的硬體是很公平的，它只處理 0 與 1 的訊號，所以無分英文、中文，都沒有歧視！」這句話只能說是部分正確的。說它正確的理由是：計算機的基本符號集是早操作系統來經營、管理和運用的，不是由硬體來執行，因而主要的罪過不在硬體。譬如說，當鍵盤和主機溝通時，操作系統會依據基本符號集檢視每一個收到的訊息，當發現有控制符號時，便依當時的情況立刻做適當的處置，遇到圖形符號時，則留待其他程式處理。由此觀之，似乎不要修改硬體，只需早操作系統改起，便可使計算機適用於中文資訊的處理了，其實並不盡然。

並不盡然的理由是：關於設備的控制部分是依據 ISO 646 的規定設計的，而這一部分卻是依賴硬體執行的。根據 ISO 646，所有的控制符號都是以 7 位元為單位，因而在硬體設計上即以此為本。這樣的安排使得 34/128 的空間用於機器控制，壓小了表示訊息部分所能使用的空間，使得它無法容下中國字。因此，不得不擴充，而擴充的結果是：操作系統無法直接有效地處理代表中國字的符號，必須依賴應用層次的軟體來做，這樣的安排對計算機資源的利用是很不好的。

由另一角度來看，設若在訂定基本符號集時，即已考慮到中文，那麼此符號集的空間就可設計得大一些，毋需像目前用 7 位元的結構。例如用 16 位元的空間就可以有 65,536 個位置，足可供控制符號與中英文合併使用。這樣的話，處理中文資訊便不會像喻如此地委由和辛苦了。

然而，這樣更新事實上是極大易成功的，因為它改變了處理控制符號的過程，也影響到一些控制用的硬體規格。若要更新，目前在使用中的設備均將被波及，此伐價實在太大。

中文資訊處理的本符號集

看了以上的背景資料，再讓我們看看中文資訊處理系統需要那些符號。可以考慮的項目如下：

- 一、處理中文資訊所需要的控制符號
- 二、可處理的中國文字集
- 三、發音符號
- 四、標點符號
- 五、ASCII 集
- 六、其他日常常用符號

關於控制符號部分，相當於 ISO 646 控制符號中格式攔制與資訊分隔符號的擴充。這是由於中文文獻的格式、中文文字之處理方式以及中文編碼結構等，都與英文的不盡相同的緣故，所以必須增加一些控制用的符號以應處理時的需要。譬如，中文有直印與橫印的區別，中文輸入時修改的動作複雜，中文碼的控制亦較複雜，而橫印與直印所用的標點符號不盡相用……等等，這些因素都是增加控制符號的原因。

關於發音符號則包括注音符號、國語注音第二式等。在國外的系統中並不包括發音符號，然而在我們的應用中卻少不了它。譬如：教學用、輸入用等，所以也應收容。標點符號應略作改良，例如：書名號、私名號等與文字重疊印出的表現方法宜以其他方式表現為僅，這是便於機器處理的緣故。再者，為了應付橫直兩種印刷方式，標點符號亦應分為兩組使用。

關於納入 ASCII 集的理由已很明顯，國語注音第二式就非用到它不可，此外中英混雜的文獻比比皆是。將 ASCII 納入中文資訊處理的本符號集中的意義是很重要的，這樣做的結果是等於在發展雙語系統。而事實上，目前幾乎所有的中文資訊處理系統都是雙語系統。關於其他符號沒有什麼問題，只是選擇的工作而已。

中文字集的問題

在上列的項目中，最難決定的還是中文字集的問題。首先遇到的難題是：究竟有多少中文字？很不幸的沒有人知道有多少。目前最大的字書（即字典、辭典）是中華大辭典，共蒐集 49,950 個字。然而文建會的國字組在編輯中文資訊交換碼時，目前已發表了 53,940 個字，另外約還將增加 20,000 字將於年底發表。所以，已知的最多字數已經超過 74,000 字，可是仍可能有遺珠之憾。

在這麼許多字中，常用的字只占極少數。時下一般的字書約蒐集 8,000 至 13,000 字。國小和初中的所有課本總共用了 5,404 個字，新聞常用字（羊汝德先生輯）只有 3,000 字。根據林樹教授的分析，最常用的 4,000 字其使用頻率之百分比已略超過 99.6%，而擴充到最常用的 6,000 字時，使用頻率之百分比約為 99.88%。由此可知，中文字的使用頻率分布是極不均勻的。這種情形並不是缺點，反而可以利用它來設計一個效用高的系統。

正由於這種情形，有許多人建議只將一部分的中文字納入基本符號集內。目前我們的國家標準通用漢字交換碼，就只包含了約 13,000 字。可是這種做法

造成了文字間的歧視，不在符號集中的文字將永遠無法以計算機來處理，這種情形自非吾人所樂見。於是通用碼只好允許使用人可以自己增加新字。這種方式似乎解決了問題，可是卻留下兩個大難題：其一是這種做法違反了分享資訊的原則，自己加的字是難以和別人交換的。其二是這樣的中文字集變成了「開放」集合，而不是封閉有限的集合（ISO 646 即是如此）。在計算機的設計實務中，開放系統的結構比有限封閉系統的結構複雜得多，而目前國內要做到這樣的系統是十分不容易的事。

除了字數的問題以，是否需包容滿、蒙、藏等少數民族的文字也是爭議的問題。

再其次就牽涉到中國文字的特性，中國文字有許多異體字。例如：「十」元與「拾」元的「十」、「拾」通用，此外如「群、●」、「●、鄰」、「栖、棲」等在一般的應用中皆視為同一個字。可是「路不拾遺」不可作「路不十遺」，「梧棲」不可作「梧栖」、「唐三藏」不可作「唐參藏」。也就是說當用於人名、地名及作為專有名詞時，異體字就不可視為同一個字。簡體字亦有此特性。這種情形使得中文字集的蒐集更為複雜。

再次就是字形的標準問題。標準字集當然需要標準字形，否則就不成其為標準。可是，目前只有教育部公布的 48,172 個標準楷書字形，對於常用的仿宋體、粗體等字形則沒有標準，這些工作都有待努力。

最後，讓我們再談一固字序的問題。所謂字序是說字的排列順序。試想成升上萬的字放在一堆，若是沒有規則的話，是很難找出來的。所以，中文字排列的方法涉及到中國文字的檢索。然而中文字的檢索一直是一個研究中的問題，根據社學知教授的估計，自民國元年到五十一年止，已發表的檢字法已達百種以上，如果加上近二十年來為中文計算機系統所發展的方法當超過 150 種。可是，仍然沒有一種大家公認為滿意的方法。中文字的檢索問題，又涉及

文字的形、音、義三要素以為共定義的問題，關於這些問題將在下一期本專欄中討論。（作者保留著作權）