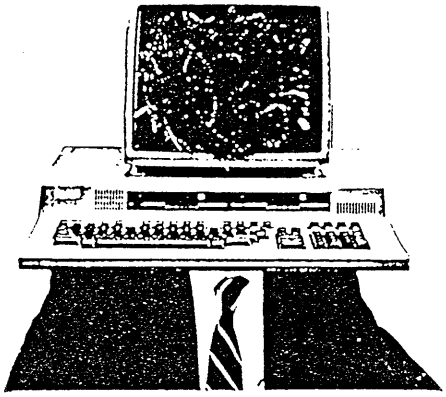


讓電腦說中國話

／謝清俊《中研院計算機中心主任》

語文與電腦

……當電腦有些語文能力時，它自然將擁有一些知識，以及具備某種程度的智慧能力……

談到國文和電腦，不免會意識到這「人文」和「科學」之間的鴻溝，他們之間似乎沒有什麼深切的關係可言。然而，由另一個角度來看，電腦是除了人之外，最善於使用語言的東西。人和人的溝通用語言，人和電腦的溝通也用語言。

當然，目前的電腦語言，像BASIC，FORTRAN，COBOL等，是和我們日常用的自然語言不同的。導致這種差別的主要原因是以往的電腦能力有限，更因自然語言過於複雜而無法在短期間發展出滿意的處理技術。所以，在電腦發展的歷程中，只好退而求其次；設計些語法和語意都甚為單純的人工語言，來因應人和電腦溝通的需求。由此觀之，電腦生來就有語文問題，他們倆的關係是至為密切的。

由於目前的電腦沒有能力使用人的語言和我們溝通，我們可就要屈就機器了：不花工夫去學電腦語言，就不會寫程式使用電腦。然而，科學家和語言學者們並沒有放棄使電腦能懂自然語文的希望，相關的研究工作，無論是語文學本身的、電腦本身的，或是對自然語文處理的，都一直在積極進行著。近幾年來，研究的成果甚為可觀，總希望有朝一日，電腦能有相當的語文程度，能用我們人用的語言和我們溝通。

一旦電腦有了相當程度的語文能力，它可說是破繭而出，脫胎換骨了，絕非今日之電腦可以比擬。試想語言、知識、智慧三者是相互交織密不可分的，當電腦有些語文能力時，它自然將擁有一些知識，以及具備某種程度的智慧能力，而這些特質正是目前電腦所欠缺的，也正是目前電腦對於人文的應用無法順利展其才華的癥結所在。

國文與國內資訊處理的成長

……遇到這些問題時，若無文字學者的協助，輕則所做的研究工作品質不好，重則對固有文字產生破壞，其後果難以逆料……

民國六十年，在國科會工程組兼

職的馬志欽教授策動了一項非常有意義的研究工作：利用計算機做中文資料處理的研究。在當時，這個倡議獲得了學術界空前熱烈的響應，幾乎所有大學或電子研究機構全投入這個行列。自此以後，建立了中文資訊處理這個獨特的研究領域，研究工作延續至今依然活躍。十五年來，此領域由萌芽、成長、而開花結實；自理論的研究，延續到工程的開發，而至一般的應用。如今，在本省幾乎沒有一個計算機機種無法處理中文資料。在國際上，我們的研究成果更受到重視和肯定。這個研究領域和它的開發過程，應該是國人珍重和自豪的，因為它完全是土生土長出來的。

最早期的研究集中在如何使計算機能夠接受和顯示出中文資料方面的問題。也就是要解決所謂的中文輸入和輸出的問題。在這個階段發展了許多中文輸入和輸出的方法。這些成果經多方測試、改良而逐漸商品化。在這個發展階段裡，就已經涉及許多文字學的問題。譬如說：中文字一共有多少？使用的頻率分配為何？中文字如何排序？如何檢索中文字？異體字如何處理？破音字如何處理？這些都是文字學方面基本的問題，而這些問題迄今一直沒有完全的解決。遇到這些問題時，若無文字學者的協助，輕則所做的研究工作品質不好，重則對固有文字產生破壞，其後果難以逆料。早期發展的系統中弊病頗多，譬如：錯字、字集蒐集不全、屬性誤植、無法排序、無法處理異體字和破音字，以及許多功能的限制等等，這些現象，到如今仍然可以在一些系統中發現。這些都是缺乏文字學者的參與所造成的後果。

從民國六十五、六年起，陸續有中文資訊處理的商品問世。之後，如雨後春筍，商品樣目之繁多已蔚為「面」的發展，各種商業之應用，亦次第展開。

計算機使用得越普遍，對語文處理技術的要求就越殷切，對處理能力和品質的需要就越高。要言之，語文處理的需求來自二方面：其一是藉以改善人與機器溝通的方式，其二是藉以提升對文獻和事務的處理能力。在改善人機溝通的需求方面，已如前述，最理想的情況是令計算機接受人類的語言，甚至看得懂我們的文字、文章，這樣的話，才能完全破除溝通上的障礙，使計算機能做到為人人所用的境界。只有這樣，才能充份發揮計算機對社會的潛力和功能。人工智慧加語言學知識正是使機器用「人」的方法和我們溝通的不二法門。

當輸入、輸出問題已解決至相當程度後，研究之方向指向提升品質和功能，而且明顯地擴大了科際合作的層面，尤其是加入了語文學、心理學等學科。語音處理的研究是較早發展的一項。目前對產生中文語音的研究已有良好的成績，且已由語音組合的研究邁入語音認別的範疇。在語文架構方面來說，從早期對字的處理，進步到對詞、句子的處理。在斷詞、剖句的文法處理方面，已有小成。至於語意之處理和自然語言的應用，目前也已逐漸展開。在字形的認別方面，用影像處理和圖形識別技術來教計算機看中文字，研究的基礎雖然較小，但也有良好的研究基礎。在這個基礎上想要更進一步發展，除了計算機研究者外，還需要語文學者的參與。

中文資訊的處理

……以目前研究的情形而言，範圍多局限於「句子」以下的簡單語文結構，許多基本問題尚待努力……

凡是以中國語文表現其原始形態的訊息皆稱為中文資訊。中文資訊依其表達的媒體和物理的現象不同分為自然形態和人工形態兩類。自然形態是指依語文的表徵以音（語音）與形（字之外觀）所表現者。人工形態是自然形態經過物理量或數值符號等的轉換以各種機器可處理的形式所表現者。用計算機處理中文資訊時，會遇到一些與計算機或中國語文有關的問題。為解決這類問題所做的研究工作統稱為中文資訊處理的研究。具有某些中文資訊處理能力的計算機系統我們稱之為中文資訊處理系統。在通俗的報導中，這類系統常以「中文電腦」名之。

中國語文有其特色。做中文資訊處理時依語文之特性不同而演繹出各種的處理方法。所以中文資訊處理可以據此分類。依文獻結構的元素來分，處理的對象可分為：字、詞彙、片語、句子、段落、文章等。由文法結構的層次來說，則可分為語法和語意兩大部份。目前的商用系統，還停留在只能處理「字」的

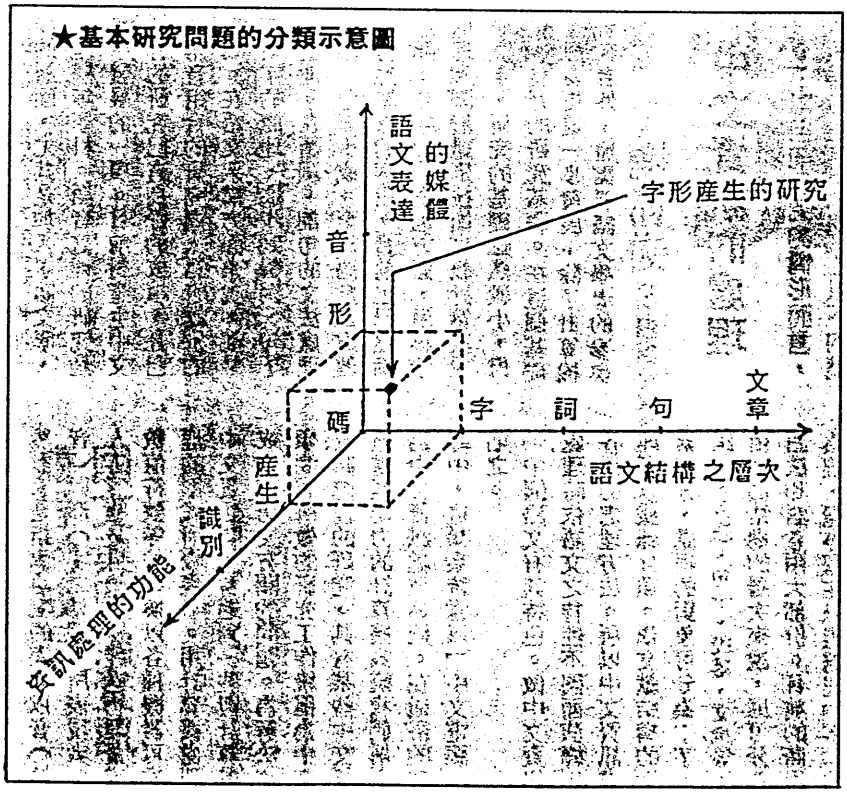
階段。較詞彙更複雜的形式，則無理論的模式可用，只能依應用問題的性質寫些依附資料（data dependent）的程式作特殊的解決之道。這樣的做法使得一個程式只能解決一個問題，而無法與同樣性質的問題共享，其投入之成本自然高漲，且對複雜的問題則無法做有系統的解法。至於以語法與語意二者，還只是研究中的對象。在研究室裡，上述各層面的問題雖均已有小小的涉獵，可惜計畫之規模較小，且欠缺長期的恆定性，以至於目前的成就不大，離實用仍有距離。

另一個角度的分法是由語文表達所用的媒體來區分。可分為三類：形、音、與碼。「形」是指人類以視覺功能處理的文字外觀，它包括各種印刷、顯示、或書寫的形態。以「音」表達者就是以人類聽覺可以處理的語文形態，稱為語音。語音的表達欠缺參考的標準是研究工作目前無法克服的問題。所以，語音處理之研究成果多局限於一小小的封閉範圍之內。「碼」是指以計算機可以閱讀的方式所作的表達。它包括數位化的字形點矩陣、數位化的語音訊號、字的認別碼、交換碼、檢索碼等等。

第三個角度的做法是以資訊處理的基本功能來分。計算機本質上是自動機（Automata），所以由自動機的基本功能上來說，研究的問題可分為產生和識別二大類。若由應用的角度來分類，則可更細分其功能。

上述的三個分類角度是各自獨立的，可以組成一個三度空間。在此空間中之一個點則可代表一種中文資訊處理的基本研究問題。例如做字形產生的研究是位於「字」、「形」和「產生」的交會點；又如語音識別為「字、詞、句子」、「音」、和「識別」等交會處之總稱。此結構表現如下圖。下圖中的三個座標分別是：語文結構、語文表達的媒體和資訊處理之基本功能。以此分類，可以對中文資訊處理面臨的種種基本問題作一綱領式的了解。

以目前研究的情形而言，許多基本問題尚待努力，而研究範圍多局限於「句子」以下的簡單語文結構。若是我們把功能概略納入產生和識別兩類，把語文結構約略分為目前常見的處理對象——字、詞、句，則下圖可化簡為分別以形、音，及碼之三個平面，而每個平面上則有「產生、認別」與「字、詞、句」等組成之六類問題。這個約略的分類可以含蓋了目前所有中文資訊處理研究的基本問題。



相關的研究

國文與電腦的關係，既然日益密切，那麼，以下這些研究工作是必要而迫切的：

前文已經談到許多國文和電腦的

關係，在此，我們就目前重要的研究工作作重點說明：

●語文基本資料的整理

目前，等待整理的語文基本資料很多，工作量異常龐大，擇其要者如下：

1.一般性質者：

・字與詞之蒐集、整理，各種屬性之確認與整理，及其使用頻率之統計

・常用的檢字法之標準化與其標準鍵盤之安排

2.關於字形者：

・各種字體之標準字形與寫法

・字形點陣之各種標準（依大小、字體、美觀程度等之分類）

・字形結構的模式和字形之標準表示法（字形之定義與描述）

・各種媒體呈現字形時之版面規格

・異體字之認定與標準

・各種計算機儲存媒體存錄各種字形資料之規格與標準

3.關於語音者：

・字與詞之標準語音錄音

・字與詞之標準數位化語音檔案（以上皆須依性別、年齡、地區等採樣環境之不同，以及錄製時所採之發音方法之變化分別製作）

・破音與又讀之認定與標準

・各種語音資料（包括參數）在各種計算機儲存媒體中存錄之規格與標準

・各種音碼之標準表示法及其間之轉換

・方言之相關資料檔案

●語音處理

語音處理的研究工作已有十年以上的歷史，可是重要的進展是近五年的事情。早期的研究集中在語音的組合（Voice Synthesis），也可說是以數位化的資料仿造語音的工作。

這方面的研究已由單字音的產生進步至詞和句的連續音的組合。目前在單音間主振頻率以及前後音四聲變化的銜接上已有很好的成績。其組合之語音已經減少了許多機器的「鄉音」。

最近的研究情況顯示，對語音識別的興趣已大為提高。語音識別較語音組合困難許多，其涉及的變化因素亦複雜許多。但是若能成功，則應用的價值亦將大增。目前，語音的識別已從單音的識別進入連續音的識別研究。然而由於語音樣本的來源變化的程度很大，是故研究工作必須依性別，說話者是否固定、說話者的年齡條件、詞彙有多少等等外在因素對研究的範圍加以限制。連續語音的識別一定需要語音學以及語法方面的知識來提高其識別的正確程度。因此語音識別的研究已成為訊號處理（Signal processing）與語文學的結合型態，是典型的科際組合問題。

目前語音識別的研究是朝著以「語音至文獻的轉換」（Speech to text conversion）為目的的方向走。這是發展以人說話的方式與計算機溝通的必要研究。

●字形的識別

關於字形識別是近五年來的熱門研究，目前，以日本對漢字的認別技術而言，已經出乎意外的成功（在四千漢字內，識別率可達99%以上）。現在的問題是：所需的計算能力太大，以至於產生經濟上的困難。除非等值的計算機其計算速度增加千倍以上，否則，考慮經濟上的可行性，已發展的技術仍是無法達到實用的境界。國內此方面研究的力量甚為薄弱，全國只有清華大學一個研究室而已。雖然該研究室成績甚佳，然而杯水車薪，進展總是緩慢。

字形的識別率可利用字或詞的結構資訊加以提高，也可經此協助而減少計算的負荷。因此，字形識別的研究也從單純的圖案識別（Pattern Recognition）學科中轉變為與文字學相輔相成的領域。

此方面的研究，能認別已印製好的文件內容，並高速地輸入計算機中。對建立各種資料庫及檔案庫有其大的功效，尤其在光碟（Optical Disc, or CD-ROM）技術迅速突破的今日，此研究項目尤顯重要。

●中文的自然語文處理

目前這方面的研究尚無明顯的績效，研究的方向有：

・中文語法的研究。尋求適合計算機用的中文語法模式，作為文法上分析的基礎。

・中文語意方面的研究。尋求適合計算機用的語意網路，建立詞彙間語意的關係，作為語意分析之基礎。

・發展有語法分析能力的程式

・發展有語意分析能力的程式，能了解語意的程式

・構詞模式的研究

・機讀式字辭典的設計

・推廣利用語文知識的應用程式：譬如：能找別字、錯字的程式、能發現語意混淆的程式，能改正文法錯誤的程式等等。

・翻譯系統的研究

・其他有關人工智慧的研究。例如：會了解文章內容和結構的程式、會造句的程式、會做詩的程式、會寫小說的程式等等。

●中文的全文處理

全文是指文獻中全部的原文。它和計算機中傳統的格式化的記錄是相對的。一般來說，文法是研究句子以下的結構為主，而全文則以處理句子以上的大架構為主。在排版系統中，就需用到全文處理的規格與技術。譬如，章、節、段落、圖表、公式、標註等之版面安排就是一個典型的例子。

全文處理的研究主要涵有下列項目：

・全文元素的認別

・全文結構的分析

・全文資料庫的結構與檢索技術

・全文的表達（（text representation）問題

全文處理的技術與線上資料庫系統以及圖書系統息息相關。近四年來，國外的全文線上檢索大為風行，國內亦多有此應用之需求，立法院與行政院法規會所作的全文法律資料庫就是很好的例子。全文處理的技術亦可用於文學與歷史的領域。中研院發展的史籍自動化系統就是一個例子。它與格式化的欄位系統可以相輔相成。如何截長補短，將傳統的欄位記錄資料結構與全文資料結構合為一體，以求其最佳之組合，仍是目前研究的熱門問題。

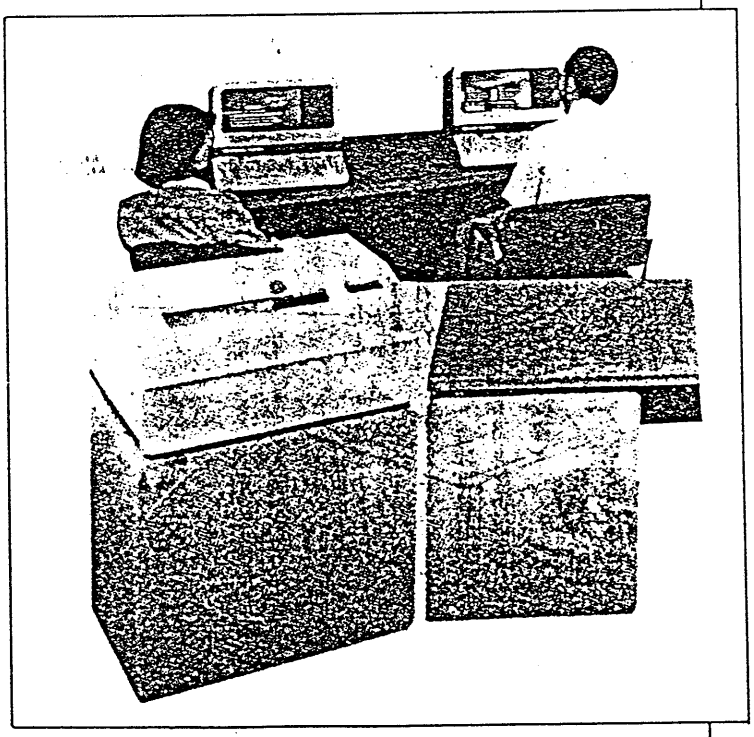
結語

……提高中文資訊處理的能力就是提高了我們的國力

中文資訊處理的研究涉及了語文，而語文本具有民族的文化色彩。因此，中文資訊處理的研究自然帶著濃厚的國家意識和文化特質。在這方面的研究只有靠我們自己的努力，無法像自然科學一樣可以全盤借重國外的知識而自國外引進。

中文資訊處理的應用和我們的社會唇齒相依，禍福相共。提高中文資訊處理的能力就是提高了我們的國力。如果電腦處理中文資訊的能力不足，沒有外國人會同情我們，只會認為我們程度不夠。今日的電腦對自然語文處理的能力仍然有限，可是在五至十年內，局面將完全改觀。屆時，電腦將有相當好的自然語文能力，會造成計算機科學與文化結合的場面。此影響將較以往電腦給我們帶來的任何一個衝擊為大、為深、為遠。試想，有一天電腦能和我們用說話的方式溝通時，您願意說日本話呢？還是英語？

在國外，文、史、社會科學等的研究也常利用電腦，可是國內才剛起步。中文資訊處理的研究在這個應用上扮演決定性的角色，理由已很明顯，不須多說。文、史、社會科學的水準對國家社會的影響如何，也不須在此贅述。重要的是須認清這個因果關係，未雨綢繆。



■有一天，當我們能用語言和電腦直接溝通時，你願意用那種語言？