論中文資訊處理系統的發展

謝清俊

中央研究院資訊科學研究所暨計算中心
中華民國七十八年十一月十四日

摘 要

中文資訊處理的研究是一個獨特的研究領域，也是一個大型的科際組合學科。十八年來，此領域從無到有，由理論的研究，至工程的開發，而延續到一般的應用。如今，在國內，幾乎沒有一個計算機種無法處理中文資料，而工商業界和學術上的成就也在國際上舉足輕重。這個學科的研究領域和它的開發經過，完全是土生土長的，值得國人珍重和自豪。

本文將扼要的探討中文資訊處理的性質，包括：基本定義，研究所涉及的學科與範疇，以及基本問題之分類等。其次，介紹中文資訊處理的主要研究領域，其目前情況，並條列一些待研究的問題。在結語中，將强調中文資訊處理研究的特質，期爲關心人士之參考與政策發展上的建議。

目　錄

一、中文資訊處理的成長

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。十八年來，此領域由萌芽、成長、而開花結實；自理論的研究，延續到工程的開發和商業行銷，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，因爲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。

在這段期間，能供中文使用的計算機週邊設備甚少，圖形顯示器，點陣印刷機等均未問世。所以，引起一系列硬體的開發。其中以鍵盤、螢幕控制器、字形產生器、印字設備等爲最受重視的對象。在發展過程中，由於許多嚐試都須由最基本的電路設計做起，著實訓練了不少實務的設計人才，爲後期中文資訊處理商用產品的開發，奠定了良好的基礎。

從民國六十五年起，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爲「面」的發展，各種商業之應用，亦第次展開。譬如：各種軟體工具之開發，包括能處理中文資料的語言編譯器、檔案系統、資料庫、編輯程式、列表程式等等；又如各種應用程式，包括文書處理、人事、會計、採購、財產、庫料、以及一些特殊行業的應用軟體等，琳瑯滿目，式樣繁多。雖然這些系統並非至善，可是其所發揮之效果的確已提昇了我們社會上處理事務的能力。

當中文資訊處理的應用受到社會各界的肯定以後，廠商們開始設計和生產中文資訊處理的硬體，包括小型的計算機和週邊設備，而且有非常優異的成就。譬如：中文字型產生器就是全球獨步的產品，不僅物美而且價廉，爲日本、中共、香港、美國等地區所望塵莫及。當*IBM* 在民國七十二年推出*5550*型中文計算機時，其優異之品質引起國內的計算機製造業界相當大的震撼，繼而開始往高級產品的方向發展。目前國內能生產高品質個人用計算機的廠家已有許多家，有些產品的性能已較*IBM 5550*爲優，充分展示出國內計算機製造業的實力。此外，適合顯示中文資料的螢幕及其控制器，也是國內達世界級的產品 。

在學術研究方面，在過去十八年來也有明顯的轉向。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、中文語法模式之建立、和自然語言的應用等，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並瀕臨實用階段。

在學術活動方面，中央研究院數學研究所在民國六十二年底，召開了國內舉辦的第一屆國際計算機會議(ICS，73')。該次會議的主題是中文輸入和輸出的問題，會中的論文集被國內外學者引用頗多，至今仍是中文資訊處理的重要文獻。由於這次會議的成功，此後每二年循例召開一次，而每次之主題均與中文資訊處理相關。ICS 會議在國際上建立的聲望與地位，和國內中文資訊處理研究的水準實是息息相關。

受此會議之誘導，與 ICS 隔年舉辦的全國計算機會議(NCS)，亦沿襲中文資訊處理之主題。民國六十四年，參加ICS及NCS的國外華裔學者有鑑於中文資訊處理問題的獨特性，發動了成立中文電腦學會(CLCS)。此學會目前已有近千會員，二十餘國參加，並每隔一年半舉辦一次國際性的學術會議。在每次之會議中，我國學者發表之論文，無論在質與量方面，均甚受到重視。

最近二、三年，體認到在中文資訊處理的領域中，語言學的地位日益重要的學者越來越多，遂有發起成立我國計算語言學學會的倡議。此學會已在今年九月間獲內政部同意設立，目前正展開籌設階段的各項步驟，可望於今年底正式公開成立。這個學會的成立亦表示中文的計算語言學已在國內萌芽成長，對國內的語言學界和計算機科學界來說都將是個值得紀念的發展里程碑。中文的計算語言學也可以認爲是中文資訊處理推動了十多年來孕育而長出來的果實吧！

在國際上訂定資訊交換標準的活動方面，我國國科會，中美會和文建會發展的中文資訊交換碼(CCCII)，在民國七十年爲聯合國教科文組織下的國際標準局(ISO)正式納入標準技術文獻，七十二年起，美國研究圖書館組織 (RESEARCH LIBRARIES GROUP，簡稱RLG) ,採用CCCII作爲「東亞國家之文字碼」(RLIN EAST ASIAN CHARACTER CODE簡稱REACC 東亞字碼) 的結構模式與中日韓文字編碼。民國七十四年，國際標準局正式宣佈中文資訊交換碼在ISO 2022標準中的轉換控制碼。雖然我國非聯合國之會員國，國籍不被認同(通常國際標準須由國家標準提出)，然而中文資訊交換碼確已享受到成爲世界標準之實質。七十五年五月，全球最大且跨十四國之美國俄亥俄州線上圖書館資訊中心 (ON-LINE COMPUTER LIBRAPY CENTER, OCLC)採用CCCII/REACC 東亞字碼，作爲國際間中日韓文資訊傳輸交換之電腦網路用交換碼。七十五年十一月，美國國會圖書館正式公佈，以CCCII/REACC 爲基礎的EACC碼，並申請爲美國國家標準，作爲美國政府各單位間電腦網路用之「中日韓文交換碼」。七十七年底，EACC/CCCII/REACC於美國完成審議，正式成爲美國國家標準 (EAST-ASIAN CHARACTER CODE，簡稱 EACC-1989)。

自從民國七十年起，利用中文資訊交換碼爲基礎發展的中文圖書自動化系統，在國際上倍受矚目，也是美國、香港、日本等其他地區所羨慕和爭取合作的對象。經常被邀請在國際的重要圖書館學會議中發表論文，如國際圖書聯盟年會(IFLA)，亞洲圖書學會，亞太學術合作會議，以及美國資訊學會年會(ASIS)，美國亞洲學會 (AAS)等，又如專門爲交換中日韓文資訊技術所開的會議(1982 澳洲，1984 香港)以及AFFISO工作會議等。

二、中文資訊處理研究的性質

爲了對中文資訊處理有全盤性的了解，我們將給中文資訊處理的研究下一個較嚴謹的定義，然後討論其研究的範疇、問題之性質與分類。

(一)定義

用計算機處理中文資訊時，會遇到一些與計算機或中文資訊有關的問題。爲解決這類問題所做的研究工作統稱爲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爲中文資訊處理系統。在通俗的報導中，這類系統常以「中文電腦」名之。

「中文資訊處理」一詞的界定是廣義的。若要說得更明確些，可參考下列的定義。

定義一：中文資訊

> 凡是以中國語文表現其原始形態的訊息皆稱爲中文資訊。中文資訊依其表達的媒體和物理的現象不同分爲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。

由定義一，中文資訊在計算機裡的表達方式是一種人工形態的中文資訊，譬如：各種字碼，字形的矩陣或數學式子，以及數值化的語音訊號等等均是。語文表達的資訊可分爲兩個層次：其一是外在的物理現象，如上述例子中之字形、語音訊號等，這是較易在計算機中表達的部份，其二是內在的抽象結構。如：詞、句、文獻等的含義和音調所表達的情緒等是。這些抽象的結構是一些不易表達的訊息，因爲它們涉及到智識的表達以及語意的了解等問題。

定義二：處理

> 處理是泛指對資料的任何運作。例如：表達、轉換、計算、儲存、收發、傳送、檢查、排序、合併、分類、搜尋、查詢、識別、產生、分析、判斷、推理、了解、譯釋等等。

由以上的兩個定義可知中文資訊處理的研究工作範圍很廣。然而這些研究的問題之間有其共同性與相關性，且具有語言與文化上獨特的性質，因而構成了一個科際整合型態的新研究領域。

(二)範疇

從以上的字義，我們可以理解中文資訊處理的內涵，然而中文資處理的問題並不是孤立的，它經常涉及許多學門．以下，就讓我們談談中文資訊處理與這些學門的關係，以期對研究性質與範疇作全面的了解。

1.與語文學的關係

做中文資訊處理的研究固然必須計算機的知識，然而這方面的研究卻不是計算機專業人員獨自可以以的完整的。中國語文方面的知識也是不能缺少的。在早期的發展階段，就已經牽涉到很深入的文字學問題。譬如說：中文字一共有多少？使用的頻率分配爲何？中文字如何排序？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。

民國六十一年，林樹先生所著的<中文電腦用字的研究>提供了常用字集的分類、字的使用頻率等資料。此書一直是設計中文資訊系統者的重要參考資料。民國六十八年教育部公佈的4808常用字集，七十二年資策會提供的5404常用字集等等，這些工作與其成果都對中文資訊處理提供了極大的幫助。這些是文字整理工作對中文資訊處理的研究有正面貢獻的例子。此外，潘重規教授的< 龍龕寶鑑新編>，周何教授的<中國文字孳乳表稿>，周駿富教授的<中國文字通行字體表稿>，對異體字、文字結構與分析,以及文字之屬性歸屬等工作的品質有莫大的貢獻。由上面的敘述，我們明白，中文資訊處理的研究少不了語言文字學者的參與。

數千年來，語文是世界上任何一個民族延續文化的主要依據，也是人與人溝通的主要橋樑。自從計算機發明了以後，語文開始兼作人與計算機之間溝通的媒體。這現象明顯的指出語文與計算機有密不可分的關係。語文學與計算機科學的結合是件大事，在國外已行之有年。像計算語言學(Computational Linguistics)，計算機詞彙學 (Computer morphology)，自然語言處理(Natural language processing) 等等新學科皆在此結合下蓬勃發展。在國內除了計算機或資訊科學系偶有自然語言處理的課程外，上述之前二者尚未見萌芽。即使偶有開課者，所涉及的語文也是英文，和我們的社會環境脫節，不能完全配合我們的需要。爲了使計算機有處理中文語文的能力，或是要利用計算機來推動語文知識的應用，發展中文的計算語言、中文的計算詞彙學，和中文的自然語文處理等學科是必要的。

2.與計算機科學及人工智慧學科的關係

計算機使用得越普徧，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。

早期的計算機能力有限，更因語文的問題過於複雜，無法以自然語文作爲人機溝通的媒體。所以，只好退而求其次，設計些人工語言來因應溝通的要求。初期的機器語言、組合語言，以及稍後的高階語言，都出自這種環境的限制。也正

因如此，人便要屈就機器：不花功夫學這些人工語言，就不能寫程式去使用計算機。同樣的理由，能使用計算機的人口因而受到極大的限制，阻礙了計算機功能的發揮。最理想的情況是令計算機接受人類的語言，這樣的話，才能完全破除上述的障礙，使計算機能做到爲人人所用的境界。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是使機器用「人」的方法和我們溝通的不二法門。

早期計算機的使用，除了計算以外，大多是處理一些整理好的表格化資料。例如處理各種單據、查詢一些整理過的資訊等等。目前，這樣的使用方式已經無法滿足接踵而來的要求。人們開始要求處理完整的原始文獻，如信件、公文、法律條文和判例、新聞報導等等。甚至要求多媒體型態的處理，包括圖表、影像、聲音 …… 等的資料。對處理的基本功能也起了變化，不再僅囿於計算、邏輯判斷等，而是擴大到一些較智慧型的能力，譬如：推理、計畫、識別、學習，甚至到解題、聯想、創新、理解等等。其實，語言、知識、智慧三者是相互交織，密不可分的。當計算機有些自然語文處理的能力時，它自然將擁有一些知識，和具備某種程度的智慧能力。這正是人們所追求的心目中理想的機器。

由以上的說明，我們已可知道中文資訊處理和人工智慧之間的牽連。衆所周知的日本第五代電腦計劃想發展的是人工智慧型計算機，而自然語文處理是它最重要的基本功能之一，由此亦可窺知二者間的關係。其實，自人類有史以來，所有累積的智識和資料，絶大部份都是用語文表達的。因此，計算機和語言結合後，自自然然地涉及知識的表達，取得、組織、以及應用。對計算機來說，以機器來做這些事就是人工智慧研究的主題。因此，中文資訊處理和人工智慧之間亦關係密切。

再者，在研究中國語文的語意時，已經涉及智慧及意念之表達。爲了要了解及認定什麼是知識？它的範圍和界定是什麼？知識如何產生與取得？認知的過程如何？知識如何分類？如何運用(如推理、歸納、解題、聯連、創造、規劃等等)？……等等問題時，就已經涉及了知識論(哲學)、認知科學、心理學、分類學(圖書館學)，甚至腦神經醫學等等學科，而國外時下發展的趨勢也正是往這個科際大結合的方向邁進。

以上所舉的一大堆基本問題，雖然在學術界還在探索之中，迄今仍無定論，可是工程界已經將一些有限的獨立環境劃分出來，並且已經開發出一些用人工智慧產品，譬如：擬似自然語文的程式語言、各種專家系統、以規則爲基礎的程式系統(Rule-based systems)，符號解題法等。由於這些系統的基本功能，像是推理、聯想、理解、創新、學習、規劃等與語文的表達相關，而在我國我們也希望能用中文來做這些事，所以將無法避免地會涉及中文資訊的處理問題。

3.與其他學科的關係

前文已經提到，改進人機溝通的方式是做中文資訊處理研究的主要原動力之一。在考慮人機溝通因素時，不可避免地要用到一些專門學科的知識。例如：心

一。在考慮人機溝通因素時，不可避免地要用到一些專門學科的知識。例如：心理學、人體工學 (Ergonomics) 、傳播學等。爲了使機器表達資訊的方式能爲工作人員接受和喜愛，並且要維持一個良好的工作環境，不要讓工作人員受到傷害，這些專門學科的介入是必須的。試想在中文輸入方法發展的過程中，曾有多少稀奇古怪的的輸入「發明」被使用的人「拒絕」過？這就是一個明鑒。有許多輸入輸出的標準，是經由人體工學的設計，以及心理學對接受程度和使用效率的測試而訂定出來的。譬如：在不同媒體中字體點陣的大小標準、字間距、行距和畫面的標準、鍵盤上符號的安排標準、以及輸入方法的評估等等均是。

由以上的討論，我們已經了解：中文資訊處理的研究涉及許多專門學科，是相當複雜的一門科際整合形態的新研究領域。在這樣的情形下，我們還需要一些知識來駕馭這些錯綜複雜的科際關係，才能做好中文資訊處理在科學上的研究和工程方面的發展。這些知識就是模控學(Cybernetics) 和系統科學，以及管理科學，此外在工程發展方面則須重視系統工程(System Engineering)的學養。模控學和系統科學都是觀察和歸納複雜的系統現象，並以之推導出對系統的了解 (知識) 的學問。雖然它們之間有對自然系統和人工系統的分野和差異，然而對中文資訊處理的研究而言，二者的觀念和素養均爲必須者：因爲中文資訊處理的研究不止是涉及機器系統(人工的)，更涉及到「人」以及在「人的社會組織形態下」，如何使這機器系統能夠良好的運作以充份發揮其效能。是故它涉及與人有關的系統(自然的)和管理科學的知識了。此外，由工程的立場來看，要開發這麼複雜的產品，系統工程的知識當然重要。

可嘆的是，在過去的十八年中，國內沒有體會到(也是沒有能力)以這麼大的規模和魄力來從事中文資訊處理的科學研究和工程開發的工作。在短視的利益導向心態下，產業界不僅沒有做好應該做的最最基本的工作，許多廠商反而捨棄了對高品質的追求，漠視許多學術界的努力成果，以劣品廉價的方式，爭奪即回之利。當然，我們今日已可看出，眞正能屹立不搖的是些品質較高的產品。可是在發展過程中，已是創傷累累，以社會整體利益來衡量，是得不償失的。

綜合以上所述，中文資訊處理的研究不僅是計算機科學和語文學的深入結合，還經常涉及下列三類的專門學科，是典型的科際整合的形態。

(1) 心理學、人體工學(Ergonomics)、傳播學(科技、語文傳播)

(2) 智識論(哲學)、認知科學、圖書館學(分類學、資訊系統)、腦神經醫學

(3) 模控學 (Cybernetics)、系統科學 (System Science)、管理科學、系統工程。

(三)基本研究問題的分類

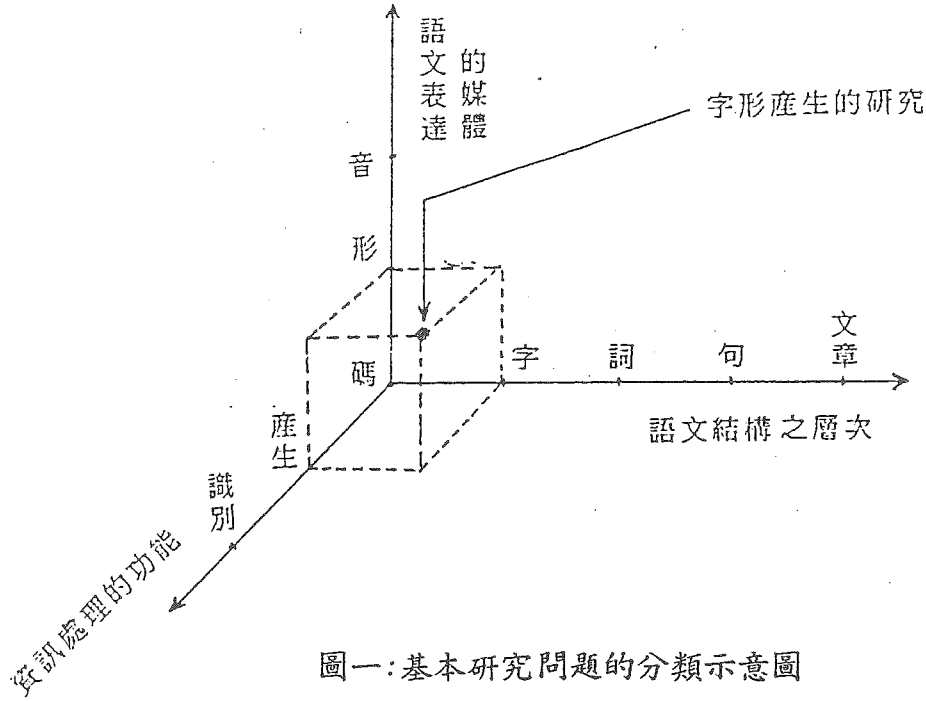
中國語文有其特色。做中文資訊處理時依語文之特性不同而演繹出各種的處理方法。所以中文資訊處理可以據此分類。依文獻結構的元素來分，處理的對象可分爲：字、詞彙、片語、句子、段落、文章等。由文法結構的層次來說，則可

分爲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，以致於成就不大，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。可分爲三類：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。稱爲語音。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。譬如，男女發音有別，老少亦不同，由單字的發音組成詞彙的發音必須經過複雜之修正，而更甚者爲地區性音色的變化。所以，語音處理之研究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，研究的問題可分爲產生和識別二大類。若由應用的角度來細分其功能，則定義二中之各項運作都可視爲資訊處理的基本功能。

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；又如語音識別爲「字、詞、句子」、「音」、和「識別」等交會處之總稱。此結構表現如圖一。

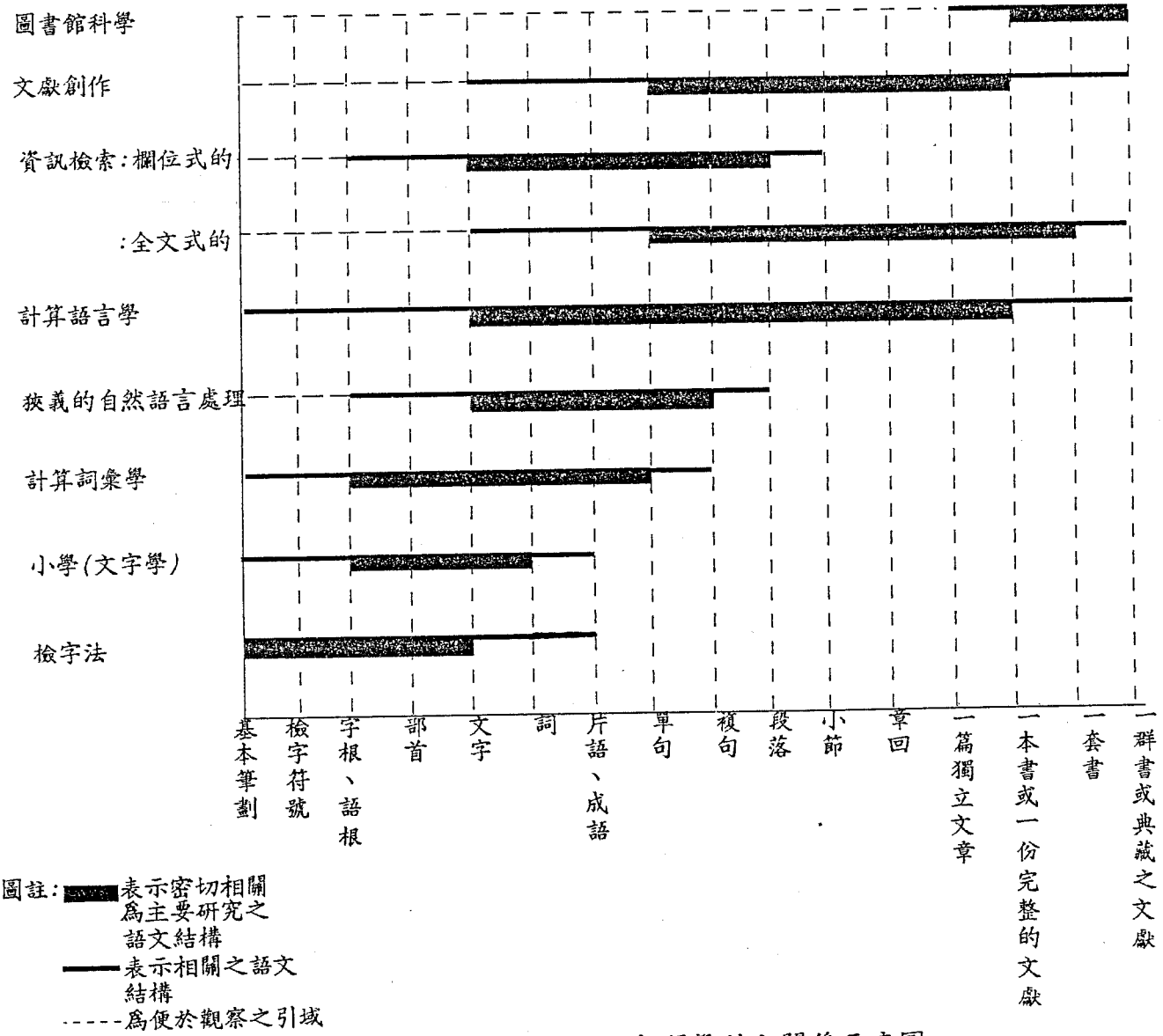


圖一:基本研究問題的分類示意圖

圖一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，而研究範圍則多侷限於「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，把語文結構約略分爲目前常見的處理對象——字、詞、句，則圖一可化簡爲分別以形、音、及碼之三個平面，而每個平面上則有「產生和認別」與「字、詞、句」等組成之六類問題。這個約略的分類可以涵蓋了目前所有中文資訊處理研究的基本問題。

若是觀察各學科和語文結構的關係，則如圖二。在圖二中的橫座標，詳列了語文結構上涉及的「語文元素」，由左至右是由簡單而複雜，而在語文上表達之訊息，亦由淺而深而廣。由此圖，我們不難明白縱列的各學科之間都涉及到語文學。從中文資訊處理的角度來看。處理的對象就是橫軸所代表的各種資料，當中文資料處理的應用越來越廣時，涵蓋表中之各項研究是指日可待的事。



圖二：語文結構與各相關學科之關係示意圖

三、目前主要的研究領域之檢討

由上面的敘述，我們已經把中文資訊處理的基本研究問題作了一種分類的敘述。現在，我們將列出在中文資訊處理的研究領域中，較活躍的幾個研究方向，並條列些正待解決的一些問題。當條列這些問題時，雖已盡量列舉，然而限於所見所學，難免有疏漏之處。因此，雖然這些條列有參考的價值，但並不是涵蓋完整之清單。

除了在理論上做基本研究之外，很少研究工作能像上一章裡所列的基本問題那樣單純。以中文鍵盤輸入方法的研究爲例，看起來這麼單純的問題，除了計算機相關的問題以外，還涉及到許多文字的基本問題，人體工學的問題，以及心理學的問題等等。所以本章中所談的研究題目幾乎都是科際整合的型態。

(一)語文基本資料的整理問題

和語文有關的研究目前都會面臨缺乏語文基本資料的困難。語文資料不僅量多，而且因時間、空間及許多社會和文化的因素而變化。在國外，利用計算機來協助解決這個難題是近卅年的事。各種機讀式的辭典、主題詞表、詞彙檔、索引典、和語文屬性資料庫等等是國外語文研究上不可缺少的工具。這些工具不僅可以協助語文有關的研究，凡是以計算機做推廣語文知識的應用時，也是不可缺少的。

在中文資訊處理的研究裡，基本語文資料經常扮演下列的角色：

(1)作爲研究的對象，或設計上的依據

深入的了解這些基本資料是研究上必須的態度，也是確切解決問題的要件。

(2)作爲測試時之準繩

依不同環境或性質所整理的基本資料是測試系統性能、以及評估系統優劣必備的客觀準繩。

(3)作爲制定工業標準之依據

制定工業標準的好處毋庸多言，然而有關語文之工業標準的訂定，確非徹底了解語文基本資料的特性之後，無法做適當的選擇。

(4)作爲應用上的基本資料

當我們利用語言學的知識爲我們做些事時，參考語文基本資料有時是必須的，例如Spell Star程式中就存有一部機讀式的字典，可利用它來檢查拼錯的字。這就是一個典型的例子。

雖然基本語文資料是這麼重要，一定不能缺少，然而，我們在這方面所做的努力實在不夠。早期的研究工作品質較差，其中基本資料的欠缺是主要因素之一。

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
- 各種句形與典型範例之蒐集，和中文語法在計算機中之表達和定型 (formulation)

(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
- 各種音碼標準表示法及其間之轉換
- 方言之相關資料檔案

在日本，與漢字處理有關的工業標準已經超過十種，各種公衆認同的測試資料集更是比標準更多。音與形兩方面的測試資料集，是做語音識別和字形識別不可缺少的試金石：沒有這些資料集的測試結果作參考，將無法客觀地比較各種方法的優劣，日本也就不會有今日這麼令人羨慕的成就。在國內，一個交換碼的標準已經爭吵了十年，此外沒有任何標準及規範之設立，更沒有一個爲大家認可的測試資料集。若是中文資訊處理的研究對我們眞正很重要，那麼，鼓勵中國語文基本資料的整理工作是獎勵研究發展的必要措施。

(二)和計算機科學相關的問題

在目前所有的計算機裡，其系統軟體的設計並未包含中文字碼(包括符號)的資料型式(data type)。換言之，系統軟體是無法直接接受代表中文字的這種資料形態。原本在系統層次必須要做的中文資訊處理工作，都降到既有的系統之下，以類似應用程式的階層來處理。這樣的做法，猶如在一大廈之內再加蓋一違章建築，在此違建之下才能處理中文資訊。雖然這種做法，可以解決部份的問題，可是對計算機系統資源運用之浪費，自不在話下，何況有些問題是無法以此變通的方式可以解決的。有識之士，早已認出此情形之嚴重性，甚至在十餘年前，即

有碩士論文從事中文操作系統(operating systems)或中文系統程式的設計，也有些研究工作在撰寫能接受中文資料型式的語言編譯器。可惜這些成果並未能順利地轉移至工業界，是故目前機器中雖能處理中文資訊，卻絕大多數是「違建户」。所幸目前已經在這方面投下了些資源在開發，然而其主要性應再予肯定，而開發之步調亦應加快才是。

其實，所謂中文資訊處理系統，其性質是屬於多語言系統(multi-linqual system)，因爲除了處理中文資訊以外，無可規避的要處理英語，或是更多語文的資料。所以，由此角度觀之，原來機器中的系統程式(以單一英語形態設計者)必須做系統根本上之修改才能適合處理中文。也就是說，必須自行計設多語言的系統程式才能徹底解決計算機處理中文資訊的根本問題。所以，中文資訊處理之於計算機，涉及軟體系統整體的基本設計，舉凡操作系統、檔案系統、資料庫、語言編譯器、編輯程式、偵錯程式，以及許多系統公用軟體等，皆應該考慮中文資訊的特性而重新研究和發展。

再者，在硬體方面雖然有各種中文卡，字形產出卡之發展，然而卻沒有從基本的計算機結構上徹底的檢討一下：中文資訊處理系統究竟有那些使用的很頻繁的運作是值得以ASIC來開發設計的。我們不必直接去修改微處理機中基本的結構和指令群，至少我們可以設計一個像協同處理機(co-processor)來加速中文資訊處理的瓶頸，或是設計一些具特殊功能的週邊卡來加强中文資訊處理的能力。在這方面，可以做的很多，像利用平行處理的結構、多處理技術、Associative Memory和Associative Search技術，神經網路技術，資料庫機技術，各種人工智慧技術等等，都是可以考慮的對象。在這方面的工作，目前做得幾乎等於零，難道國內眞的這麼缺乏計算機結構設計的人才嗎？

(三)人機介面的問題

人機介面的問題本質上是輸入與輸出方面的問題。茲將輸入與輸出的問題分述如下：

1.鍵盤的輸入

鍵盤是淘汰不掉的。這是由於英文資料仍需用它，而且它仍有自然(習慣了)及簡捷等優點的緣故。可是，以鍵盤作中文資料的輸入卻始終沒找到一個大家都喜愛的方法。這方面的研究或將永遠存在，直到一個公認的最佳方法出現爲止。由於目前沒有一標準的中文輸入方法，所以也沒有標準鍵盤。這種情形當然會造成製造、銷售、維護、學習等成本的增加和使用時之不方便。因此，訂定標準的輸入方法和標準鍵盤依舊是目前的課題。爲達到這個目的，改良現有的系統，值得研究的問題有：

- 理想的鍵盤輸入系統模式和各種現有鍵盤及輸入方法之確實評估
- 檢字法的再研究，並應擴大到詞及片語的層面，以及充份利用交談式操作之功能

・各種鍵盤鍵位安排之最佳化

・改進現有的編輯及校驗資料之公用程式

・資料登錄系統的設計

・適合人體工學之鍵盤、終端機、及其家具

・螢光幕對婦女工作者引起的生理及心理可疑症狀之了解

2.語音處理的研究

語音處理的研究工作已有約十五年的歷史，可是重要的進展是近七、八年的事情。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爲提高。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須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爲訊號處理(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)爲目的的方向走。這是發展以人說話的方式與計算機溝通的必要研究。

3.字形的識別

關於字形識別是近七、八年來的熱門研究，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，以致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爲薄弱，較活躍的不過三、四處。這方面的研究應予以重視和支援。

字形的識別率可利用字的結構資訊和前後文字在詞、片語或句子結構上加以提高，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)學科中轉變爲與文字學相輔相成的科際整合形態。

此方面的研究並涉及文獻版面識別的問題，若能認別已印製好的的文件內容，並高速地輸入計算機中。對建立各種資料庫及檔案庫有莫大的功效，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。

4.字形的產生和輸出的問題

前文已談過，中文字形產生器是我們獨特的產品之一，成績很好，可是在字形、字體、以及編排版面的功能上，仍有許多值得研究的問題，值得去嚐試。

- 在字形產生器中，增加字形大小變化、字體變化、方向變化等能力。
- 將圖形處理(computer graphics)之能力與字形產生器相結合以便產生商用的藝術體字與平面圖案的繪製能力。
- 智慧型中文排版系統的設計。
- 智慧型中文報表語言及報表印製系統之設計。
- 提高目前各級計算機中文印刷機之速率，開發或應用新型印刷設備，研究高速、高品質大型印刷系統。
- 螢幕之改良和螢幕與印刷版面之映對技術。
- 計算機輸出設備的多樣化，如傳眞、底片，或與各型通訊設備之連接等研究。
- 在上述各種硬體發展上開發超大型積體電路(VLSI)的應用。

(四)與語文學關係密切的研究

前文已談到許多計算機與語文的關係，而且已經說得相當仔細。所以在此節中，我們只就與中文的自然語文處理和全文處理兩方面的研究，作條列式重點的說明。

1.中文的自然語文處理

- 中文語法的研究。尋求適合計算機用的中文語法模式，作爲文法上分析的基礎。
- 中文語意方面的研究。尋求適合計算機用的語意表達方式，建立詞彙間語意的關係，作爲語意分析之基礎。
- 發展有語法分析能力的程式，並作爲上述語法、法意模式之驗證工具。
- 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，並進一步發展利用計算機編輯字典或工具書之技術。

・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。

2.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爲主，而全文則以處理句子以上的大架構爲主。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含有項目：

・文獻元素的認別

・文獻結構的分析和在計算機中的表達方式

・全文資料庫發展系統的設計與全文檢索技術的開發

・多媒體全文的表達(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。近七年來，國外的全文線上檢索大爲風行。全文處理的技術亦可用於商用領域。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爲一體，以求其最佳之組合，仍是目前研究的熱門問題。

(五)應用的系統

在國內的計算機應用系統，大部份都需要有中文資訊處理的能力。有些系統中，對中文資訊處理能力的要求並不高，多半做些檔案、報表、單據、通知等之類的工作爲主。像這樣的系統，它需要的是一些處理用的工具—中文資訊處理的套裝軟體。這些軟體的功能、效率，以及價格直接影響到工作的品質和投入的成本。這種軟體的發展，以目前市場上的人力而言，是可以自行發展的。所以，我們不再往這方向探討。

在本節中，我們要討論的是些較大型且較複雜的系統，在發展的技術上需要較多的人力與較高的技術者，而且是在前文中沒有談到的。這類系統，其在國內之市場也許不很大，，可是其開發出之技術，具有火車頭的作用，可以帶動我們資訊處理技術的提升。

1.中文圖書系統

圖書系統是相當複雜的一個系統。在國外，一個成熟的圖書系統都需十年以上的發展時間。中文圖書及典藏資料的性質與西洋者差異頗大。直接利用已有的西文圖書系統改裝成中文圖書系統的則由於語言不同而困難頗多，這樣的做法經常會產生削足適履的情形，無法完全適應國內的環境。所以，即使發展中文圖書系統是一項艱鉅的工作，仍須我們自己來做。

中文圖書系統的發展和中文圖書館學關係密切。例如：中文圖書的分類法、編目規則、主題詞表、權威詞檔(authourity file)，索引典(thesaurous)、機讀目錄格式、等等，都要利用圖書館學方面的知識和成果。當利用計算機設計自動化的圖書系統時，則需資料庫、資料儲存與檢索(information storage and retrieval)，人機介面設計、計算機網路、分時多工作業系統等方面的知識配合。當然，以上所論及之各計算機學科，無不涉及中文的問題，使整個系統相當複雜。

圖書系統是超越國界的，它需要和國外的圖書系統交換和分享資訊。因此，它的設計又須要遵照國際上的標準與規範，因而更加多了設計上的限制。

圖書系統對國內文化及學術界的影響頗大，其投資效益的分析是無法以商用的方法來計算的。它像是文化建設與學術研究領域中的高速公路，是一件必要的公共設施，也須要以政府的力量來推動。

2.線上資料庫(on-line data bases)

線上資料庫是快速提供資訊的最重要設備。它可以作爲一個機關的線上檔案，也可以作爲公衆資訊的分享設施。在今日的社會中，資料庫的運用越來越多，幾乎有：要把資料整理好，就要有資料庫的趨勢。

資料庫有小至在個人電腦中供一人使用者，亦有大至供全國公衆使用之資料庫。用資料庫存取中文資料，也有因語文而產生的問題。如何製作運作和維護中文的線上資料庫，目前有很多問題是很值得研究的。設計中文資料庫的事，在三第(二)節中已說明，在此不討論。然而，在這方面的實務上，如何交換各單位之心得，使大家能分享技術和經驗，也是目前極值得做的事。換言之，應該結合MIS、資訊系統管理和Information System Science之專家學者，共同來開發各種線上資料庫，例如户政、警政、土地、房屋、法律、新聞、等等相關的大型資料庫，在技術上、管理上皆有共同可借鏡之處。

3.辦公室自動化

各項辦公室之功能，鮮有不涉及中文者。各項功能均有其特性，當使用中文資料時，都成爲一獨特的問題。譬如：如今的中文文書處理軟體，其功能應可加强就是一個典型的例子。又如發展電子化的公文檔案與網路系統，則是一項有意義的挑戰。

4.排版與印刷

在國外，計算機與排版印刷界的結合，已發展出了全電子化的排版印刷系統。它們被利用在報界和雜誌書刊的印刷界，甚至有許多海報和簡報之圖表，亦以計算機輔助設計。排版與印刷在國內的市場應該是相當大的，這種情形，對於國內發展中文用的排版與印刷的系統是很有利的。

發展這一類的系統有一件非常有用的附加產品，那就是：這些經電子排版或印刷的資訊都已變成了機器可閱讀的形式。它已經爲建立線上資料庫，做好了費時、費力、費錢的資料登錄工作。國外有許多大型資料庫，其原始資料的來源，都是由電子出版界直接拿來應用。所以有些資料庫能在報刊出刊後數分鐘之內，即可提供線上檢索服務。這個工作可直接促成線上全文資料庫和電子圖書館的發展。

四、建議與結語

中文資訊處理的研究涉及的層面很廣，和計算機科學有關的問題幾乎都牽涉到了，一如增加了一維(one dimension) 的變化。在本文中，選擇了它重要的部份，並依其在學術研究相關的性質，分門別類地檢討。要言之，目前各種研究發展的水準都還有很大的空間可供改良和創新，圖一和圖二所表達的方向和本文中所討論之各點可以前後呼應，以供有志於此發展之人士參考。

在中文資訊處理的研究領域裡，計算機科學、語文學與認知科學是三個最主要學科。在計算機科學中，人工智慧扮演了很重要的角色，它是結合三個學科之間的橋樑，它使三者相互交織爲經爲緯，構成了中文資訊處理的主要棟樑。中文資訊處理是一門典型的新興學科。由日本發展第五代計算機的聲勢可以提供一個很好的佐證：這門學科將是未來計算機發展的主流之一。

然而，這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。如果計算機處理中文資訊的能力不足，沒有外國人會同情我們，只會認爲我們程度不夠。今日的計算機對中文自然語文處理的能力仍然有限，可是希望能在五至十年內，將此局面改觀。屆時，計算機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。此影響將較以往計算機給我們帶來的任何一個衝擊爲大、爲深、爲遠。試想，有一天計算機能和我們用說話的方式溝通時，您願意說日本話呢？是英語？還是國語？

在國外，文、史、社會科學等的研究也常利用計算機，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，不需多說。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，運籌帷幄。

中文資訊處理的研究可以協助我們開發我國獨特的資訊產品。例如中文的計算機輔助教學設備、中文的圖書系統、中文的博物館系統、中文的文物民俗資料庫、中文的文史資料庫、中文的人文和社會學科研究用工作站，中文和外語間的輔助翻譯設備、中文(或雙語系)的地圖繪製設備，以及中文的文、史、社會學科智識庫等等。這些研究不僅可以提高我們的國際學術地位，更可使我們的文化在科技發展的衝擊下承先啓後，日新又新。

以往在中文資訊處理方面的研究，多半是依教授學者們自己的興趣而發展，缺乏長期策略性的規劃和領導。因此，在發展科際合作方面，在研究的延續性和穩定性方面，在研究的規模方面，以及在相關計畫間的溝通方面，都不理想。這

都是我們應該注意和改進的。然而由本文的說明，中文資訊處理的確是一個與我們切身相關的新興領域，此中有無數的機會等待我們去研究、發展，和開創。

〔誌謝與後記〕：本文承陳克健博士、黃居仁博士和曾士熊先生對圖一之架構作深入之討論，特在此誌謝。本文之校對、修稿和照顧打字工作承顧秋芬小姐細心的協助，實是呈現本文的大功臣，應該特別致謝。打字是由童素珍、林清美小姐幫忙的，排版及繪圖是由李惠君小姐幫忙的，在此一併致謝。本文因參考資料眾多而未列，有興趣者請與作者連絡。