自動化處理中文信息研究

陶鴻慶著

全篇提要

文明社會利用文字，數字以及各種特定符號來作爲互相交換或係存信息之有形媒介，已有極長之歷史。自動化整理、分析、歸檔、搜查、編集以及核算此鉅量累積之資料，遂隨需要而變成必然之趨勢。近代電子計算機的出現及配合信息學、統計學之科學方法，使此孕育於百餘年前之構想迎刃而解，成爲事實。

本文主要在討論自動化處理中文信息時，吾人或將遭遇之難題，可採用之途徑，以及所應努力之方向。因恐國內對有關文獻不易覓得，凡有關知識、可使用機器等，均於文內酌予介紹或圖示，以使讀者僅由本文即可課得通盤瞭解而毋須猜臆想像。相關應用則或僅提出，或簡介。

陶氏中文電腦字母之組成、使用、應用及與其他中文索引順序法之比較，均有簡介討論。利用電腦對中英文自動機械化翻譯以及其他若干在中文上之應用，亦作廣泛之討論。

壹、前　言

世界文明社會不斷的演進發展，促使人與人之間以及人與社會國家之間的關係日益密切。構成這種密切關係的因素之一便是大量地與經常地互相交換及散播信息。從偏僻遙遠深山中的老農到繁華城市中的政府機構都是製造信息的息源，也都是信息的接收者。這樣經年累月的收送信息，使每一個公私機關、學校、團體甚至個人，對堆積如山的文件資料，都有不勝處理的感覺。欲在茫茫信息中尋找適當之資料，也變成千頭萬緒以至無從着手。大量的人力被消耗在整理、分類、核查及歸檔等工作上，但是並不能帶來理想的結果與效率。同時，在大量人口逐年增加之際，書報雜誌以及各式各樣的信息，也隨着社會的進步而愈形增加，使得這早已不堪處理的信息問題，走到束手無策的地步。

必然的，利用機器來處理大量信息，在迫切需要中產生了。在短短廿年內，由發明機器到逐步應用在處理信息上。因爲受需要的刺激，其進展之速是驚人有如神蹟。整個國家社會大小組織受其影響而改變了全盤信息收送系統。時至今日，在西方國家以及亞非若干先進國家，自動化處理信息應用在各種行業中，使人民生活在高度的科學化與效率化的社會裏。

在工商企業發展較慢以及公私機構向不重視效率的中國社會，近年來由於百業俱興，已很明顯的開始感到大量信息處理困難的壓力。利用機器輔助人工的需要不斷增加。顯然的，自動化處理信息在中國必將全面地興起，如同西方國家一樣。我中華民族人民也將生活在科學的、進步的及有效率的社會裏，享受文明與自由。但是這樣一個理想社會的建立，不是用空想與幻想，而是要用不知多少智慧血汗的累積去換取。不是各自行動，而要共同努力。

本文將討論如何自動化處理中文信息，各種應用，以及其對社會國家的影響與幫助。本文旨在報告作者研究中文信息自動化處理心得，以拋磚引玉。

貳、信息與中文信息

本文內所用「信息」二字，英文卽 INFORMATION 。這在十年前才出現之名詞，在今日則已建立其理論系統成爲一簇

新的應用科學。此學源出於控制論，但青出於藍，自成一學，且方興未艾。按控制論係研究在許多不同科學中所討論的各個不同系統的控制及聯繫問題，合數學、邏輯學、通訊理論、生物學、生理學、心理學、語言學、自動控制理論、統計學、機率學、物理學、化學、電機工程學等之大成。控制論研究發展的結果，產生了影響近代科學研究方法、社會組織形態最巨的電腦之發明。

「信息」的概念是十分廣泛的。古代的結繩記事、刻痕記時以及烽火傳息都是在表示一種特定的信息，因此也都屬於信息學研究範圍之內。

書報雜誌的寄送發售及讀者的閱讀，便是在傳播信息及接收信息。電視或收音機之播送及收看收聽，也是在交流信息。自然社會現象的變化，例如氣象溫度、人口生命，以及人爲事件的處理，如郵件、公共汽車行車時程、電話接通率等等，都帶給我們大量信息，如何有效地處理，便是信息學的內容之一。

人與一般動物用感覺器官來接收外界變化情形，便是在接收「信息」。而由大腦指示各部份肌肉作某一活動的指示，也是信息的一種。

因此我們瞭解所謂「信息」即是一種能使傳送與接收兩方面皆能相通之表示。或爲有形的文字、符號，或爲無形的聲光、神經。舉凡使用中文來播送或接受信息者，我們統稱之謂「中文信息」。

當我們談到自動化處理信息或言用機器處理資料時，在一般情形下，當即指利用電子計算機（或稱電腦，英文爲 COMPUTER。作者贊成譯爲電腦，不但簡短且更恰當。蓋在絕多數應用中，其用來記憶及處理資料多於計算也。）及其附帶設備或者自動資料處理機器（如政大公共行政及企業管理教育中心購置者是爲最簡單之設備，即打孔機、複核卡孔機、會計機、卡片次序整理機等）。此二者之主要區別在於前者可儲存資料及依指令進行計算及處理資料，後者則直來直去每次僅完成一件指令。是以前者之能量用途千萬倍於後者，價格自亦懸殊（註一）。

截至目前爲止，無論任何西方國家以及蘇俄、日本、僞中共等所製造之電腦及資料處理機器皆使用字母、數字及若干符號來表示信息。因此簡化其打孔機、卡片設計、內部機械表示等整個信息表示系統。但用這些機器處理中文資料時，如何用少數

的字母、數字及符號來表示上萬不同的中文方塊字呢？這難題當是影響國內遲遲沒有全面自動化處理信息的主要原因之一。在可以用外文表示信息的應用上，例如全國氣候預測、農作物生產控制、人口推計、全國鐵路、公路等交通中央控制系統之作業操作、都市公共汽車行車時間控制、海陸空軍作業控制等等以及在科學上之計算工作是沒有困難。但是如果所涉及之信息資料必須有中文存在的情形時，例如各種政府統計調查、戶籍、兵籍、學籍、大專聯招全自動處理、稅務行政、各公私企業政府機構各式檔案及會計財務、警局犯罪記錄、圖書館目錄編製……等等，西方國家或蘇俄、日本的高速電子資料處理機器，便須作必要的處置才能應用。

同樣的，自動化處理中文，在美國、蘇俄等各强國內也有迫切的需要。這是由於彼等要瞭解在有七億五千萬人口的中文世界中所發行的鉅量科學書刋、專門報告、政府公告文件以及論千論萬每日發行的報章雜誌。這項翻譯、分類、歸檔、搜集等工作是無法想像的繁重，實非有限的中文譯員及工作人員所能負担。在一九六二年卽供美國情報機構用之中文資料便達三千四百四十萬字，但是實際翻譯及處理者僅三百五十萬字（註二）。這項每年需要翻譯處理之中文將每年增加至少百分之廿五（註三）。機械化翻譯處理中文在美國及蘇俄於十年前卽大力進行研究，迄今仍積極進行中。

自動化處理中文的途徑顯然的有三條：第一條是建立及製造整套中文電腦系統，全盤中化。第二條是利用西方高速電腦爲主體，另設計中文打孔機及高速中文印刷機，配合運用。第三條是利用全套西方機器而用適當之中文索引順序來表示中文，進行處理作業。

在現代白話文中，通常使用的中文字約在四千字到八千字之間（註四）。如果每一個不同的中國字都單獨用一個符號來表示，將這八千個不同的符號全部設計在「中文電腦」內，在機械方面並不困難。因爲每一個符號僅需十三個信息單位（Bit），亦卽十三位二進位數字卽可表示。每一個信息單位爲一正負相反之二元單位，通常用〇及1來示。在機械上則可利用電流之强弱高低來表示。例如用三個信息單位串聯可用以表示八個符號，卽〇〇〇，〇〇一，〇一〇，〇一一，一〇〇，一〇一，一一〇，一一一（見表一）。但是在目前國內人才、研究製造經費兩缺，而其他國家在實際市場需要太小的情形下，「中文電腦」

之製造，在短期內將不可能實現。因此第一條途徑是應走而不能走。

單位使用信息位數	表示信息可表示不同數
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024
11	2048
12	4096
13	8192
14	16384
15	32768

表一　使用單位信息位數與所得表示不同信息數對照表

中文打孔機在一九六五年由美國空軍資助下委托波士頓之ITEK公司設計製造完成。費時數年，耗資數十萬之結果，並不理想，以至鮮少應用（此節詳見第三章）。中文高速印刷機聞國內中華資料處理中心設計完成，但如何使中文打孔機、西方電腦及中文高速印刷機三者合成爲一個處理系統，則尙需若干時日及大量人力物力的投注。可是由於將來使用量之限制，除非中美兩國政府合力支持，不計成本，則我們實不應坐待其成始採用自動化處理中文資料信息。由此，我們目前研究發展途徑，主要在尋求一種可以適合電腦接收處理及可以由現有之打孔機打出之良好的中文索引順序，從而利用西方國家現有的電腦設備，來自動化處理中文信息。

在從事中文索引順序研究時，必須從其息源着手。也卽是要先瞭解中文的特性與結構。中文字在總共約卅一種不同筆劃中（註五），利用排列及組合不同筆劃再用高低左右不同位置來表示上萬不同的方塊字，也南腔北調地形成不同的口語中文音韻。因此在應用自動化處理中文信息時，當應循此中文息源之特性，去發掘處理的原則及發展方向。這也是利用電子資料處理機器，去處理任何一種信息的基本法則。

要注意是：在上古時代的中文是一種象形文字，無所謂筆劃與順序。文字的本身由其象形來決定所表示之意義。文字與語言二者也無絕對的關聯。但是許久以來，中文文字早已失去象形的功能而演變爲一組用來表示中文話的有形符號。一個中文字的書寫已經無法用來直接象形地表示其字義而須人爲地設定。因此所謂部首之意義皆是世代相傳，音與字之配合也是傳統教育

。中文之爲文字在於中文話配合中文字。這卽是我們可以利用中文索引順序來表示中文以爲自動化處理用之原則。由信息學觀點來看中文，中文方塊字只是一羣有特定意義的符號而已。

叁、現有中文索引順序可作爲中文字碼用者之評介討論

甲、人爲注引字碼方法

在現有的中文索引順序方法中，有三種方法經常用來作爲機器用之中文字碼。卽中文電報號碼、羅馬拼音以及王雲五四角號碼。

美國 Mc Craw-Hill Book Co. Inc. 出版公司在一九六三年曾出版專供機器用之中英文字典二巨册，卽用中文電報號碼來代表中文（字典全名 Modern Chinese-English Technical and General Dictionary）。另外用羅馬拼音來編印字典及教科書在國外甚爲普遍，許多大學的第一年中文，卽完全用羅馬拼音表示中文。

美國加州大學機器翻譯語文計劃（中文組）將這三種方法以及每一字的部首和偏旁筆劃數，加上兩部著名（包括Mathews' Chinese-English Dictionary）中英文字典中的編號，綜合地編印出版五巨册中文索引，每一字並有簡短英文解註以及各字對照索引（正楷與簡體、異字同義等）。這是目前在自動化處理中文信息範疇中最有貢獻的一部工具書。（註六）

茲詳細討論此三種常用的中文索引順序方法如下：

一、中文電報號碼

中文電報號碼是由四個數字組成，根據部首次序及總筆劃數依次排出。但是同一字時常有不同的寫法；例如「鄰」與「隣」、和無數的正體字與簡體字。此外同一字可以有不同的發音及不同的含義；例如「行業」與「行程」中的「行」字。在利用機器來處理中文信息時，如果採用中文電報號碼來作爲表示中文之媒介，則可以將之區別，以解決上述的困擾。

原則上，機器用中文電報號碼（作爲表達中文的媒介）即是一般慣常所用四位數中文電報號碼。凡一字有不同寫法但並無不同意思時，其出現於電碼簿或最常用者，用標準四位數電碼注引之；如 xxxx 。其餘則用 xxxx.1 、 xxxx.2 等類推注引。例如「開」爲 7030，「幵」爲7030.1。

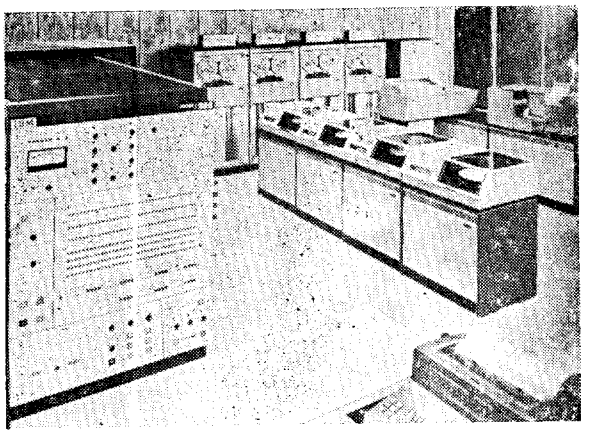
凡一字有不同發音時，則用 xxxx.OA, xxxx.OB 等類推注引。例如「省」字；用 4164.OA 表示「節省」的「省」，用 4164.OB 表示「反省」的「省」。

中文電報號碼之最大優點爲簡單且絕不重覆，但無法快速理解及隨讀隨注。因此可以作爲表示靜態中文信息用電碼；例如國家重要公文檔案、外交條例等等之保存。但無法採用在日常流動性大且經常需提出研究處理之動態中文信息中；例如全國戶籍、公教人員個人資料檔案等應用上。如用之於自動機械翻釋或語文整理上，則直接人工處理，猶倍速於先乞求人工注引長篇文章（可能錯誤率不會低），再打卡，再由電腦處理。

重要中文資料如譯爲電碼後存於磁帶或磁盤中，不僅節省大量儲存空間及最安全保存，同時在整理、歸檔及資料提存時，極爲快速準確。在政府或商業檔案處理上最是有效，且保存、攜帶或轉運最便捷。

按磁帶（Magnetic Tape）係電腦的外儲存器之一。其外表與一捲電影膠片相似，使用法則與錄音膠帶相同（見圖一上端，所示者爲四組 IBM 2402 磁帶機）。

磁帶最大的優點是可以獨立地儲存資料，而並不需要與電腦相連。這與錄音帶及錄音機二者之間的關係一樣。磁帶的儲存量隨製造技術的進步已有令人滿意的容量。例如圖一所示者爲IBM公司最新出品之磁帶設備，每英吋可儲存八百字或一千六百個數字（註七）。如數推計，則有五百萬字之資料僅需五百呎左右之磁帶即可完全儲存。目前許多所謂「自動化圖書館」即是由無數捲磁帶組成，在西方國家已不是新聞了。



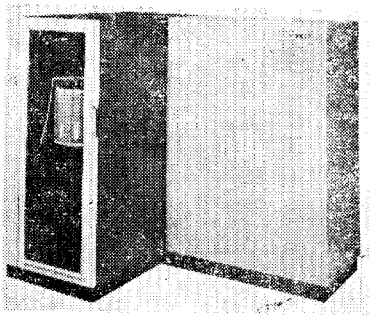
（圖一）

磁帶所儲存的資料之提出及存入的速度是驚人的，上述 2402 磁帶的資料提存速度每秒鐘最高可達九萬個字或十八萬個數字。

磁盤（DISK）則係電腦最重要的外儲存器。其外表狀如六張普通十四吋唱片之累疊。磁盤與電腦的關係大致如唱片與唱機的關係，但是磁盤可以直接提出或存入資料。

圖一所示者，爲IBM公司爲配合 IBM SYSTEM/360 電腦所最新出品的 2311 型磁盤機。其容量爲每一磁盤可儲存七百二十五萬個字或加倍的數字。若置備一具 IBM 2841 Storage Control，可以用來串連八個磁盤同時使用，則儲存量可增至五千八百萬個字或超過一億個數字。在有鉅量信息應用上，可以再串連若干個 IBM 2841 Storage Drive，則儲存量可以無限制地增大直至足夠爲止。其提出或存入資料的速度爲每秒鐘十四萬五千個字。

新式的電腦外儲存器不斷的出現在市場上。例如一九六六年中始推出市場的 IBM 公司出品 IBM2321 Data Drive 是最成功者。每一 Data Drive 可以儲存四億個字。最高可串連八部 Data Drive 同時使用，如此使儲存量增至三十二億個字。其提存速度爲每秒鐘五萬五千個字（見圖二）。



（圖二）

因此，電腦儲存資料量的顧慮，在處理一般中文信息資料的應用上是沒有必要的。

如果我們面對一項有鉅量信息存在之中文資料而欲採用自動化機器處理時，例如戶口普查、全國戶籍資料、兵籍資料、學生學籍登記以及稅務調查、普選登記審定投票等等，用中文電報號碼來表示，中文將顯然的無法滿足快速注引翻譯條件。

但是中文電報號碼在涉及遙控操作及電子信息交流時，却仍有其獨特的地位——簡明而正確。例如在全國自動化處理公務時，全國各地利用電動打碼機向上級機構彙總報告，由政府中央電腦系統接收處理。金門與國防部之間所有人事、配給、彈藥、軍事作業，事無巨細的經常交換資料（無礙地應使用密碼。密碼之編製是信息學一支流。利用高速電腦可以輕易地解析尋常的密碼，所以密碼的編製需極端的考慮）。尤其在映影附件（Display Unit）出現後，其與

電腦配合使用將遙控操作等應用，又發展到另一個天地。

由於中文電報號碼的特性，美國國家科學基金會（National Science Foundation 簡稱 NSF）在十一屆年會報告中明文指出並建議，在較妥的中文索引順序未提出前，希全美國研究機構統一使用中文電報號碼作爲機械化處理中文之媒介。

二、羅馬拼音

羅馬拼音是一種用拉丁字母來「拼寫」中文的方法。每一個中文由一至數個拉丁字母拼寫而成。其最大優點爲當譯員熟諳此法且對中文讀認發音極爲準確時，可以快速地將中文譯爲拼音字。同時其拼音字本身，顧名思義地表達着「口讀」中文。這與其他方法截然不同。

但是由於用「口讀中文」來代表中文字，因此無法辯認同音異字、異字同音、同字異音。同時尚有人爲因素存在；即同一個字，不同譯員可能有不同的譯法。

目前共有五種常用的拼音制度；即標準羅馬拼音、耶魯羅馬拼音（Yale Romanization）、俄文拼音（Cyrillicization 採用在Oshanin氏中俄字典內）、韋得羅馬拼音（Wade or Wade-Giles Romanization）及僞中共所採用的拉丁拼音（註六）。不同制度有不同的表示及發音。

因此，在自動化機械處理中文信息時，用羅馬拼音作爲表示中文之媒介有其事實上的困難；即誤差太大。但是在若干對中文原字之翻譯並不需要嚴格要求的應用上，則羅馬拼音仍不失爲一個可直接注讀的良好中文索引順序。

在若干中文資料處理上，可事先編定標準拼音，再由當事人自行譯定後固定使用，則較任何其他表示方法爲易捷，且易被人接受使用，例如姓名、街道地名等等。

所以羅馬拼音法可以局部的作爲中文之媒介，或輔助其他方法不足處。但是無法統一採用，尤其在機械化翻譯應用上，這是萬不可用的方法。

三、四角碼號索引

四角號碼是王雲五先生在一九三〇年首次提出後用之編印中文字典，以爲索引、編排中文用。

四角號碼的基本注引規則是：將所有可能的不同中文筆劃歸類併爲十組，分別用〇至九一個數字表示。每一個中文字，取其上左、上右、下左及下右之筆劃依次注引。每一個中文字用五位數字表示；前四位數字來自四角筆劃索引，如果同時有二個以上中文字皆有着相同的四角筆劃索引時，則用第五位數字來區別，否則留而不用。

四角號碼索引有理想的聯想妙用，可以快速注引，也便於排成順序，是爲其優點。特別用之於字典索引上，有其歷史性的貢獻與價值。

但是四角號碼索引有着太多的重複使用一注碼的情形。用在自動化處理中文信息上，將使結果面目全非，不知所云。例如辦、辮、辨及辯字皆注引爲 00441。更有甚者；例如行、衍、衍、衕、街、衕、街、衙、衚、衝、衛、衙、衡及衢字全注引爲 21221。

由此，四角號碼索引應用在自動化處理中文信息上，有其不可克服的缺點——電碼重複使用過多，因而大大地限制其應用的場合。

乙、機器輔助注引字碼方法

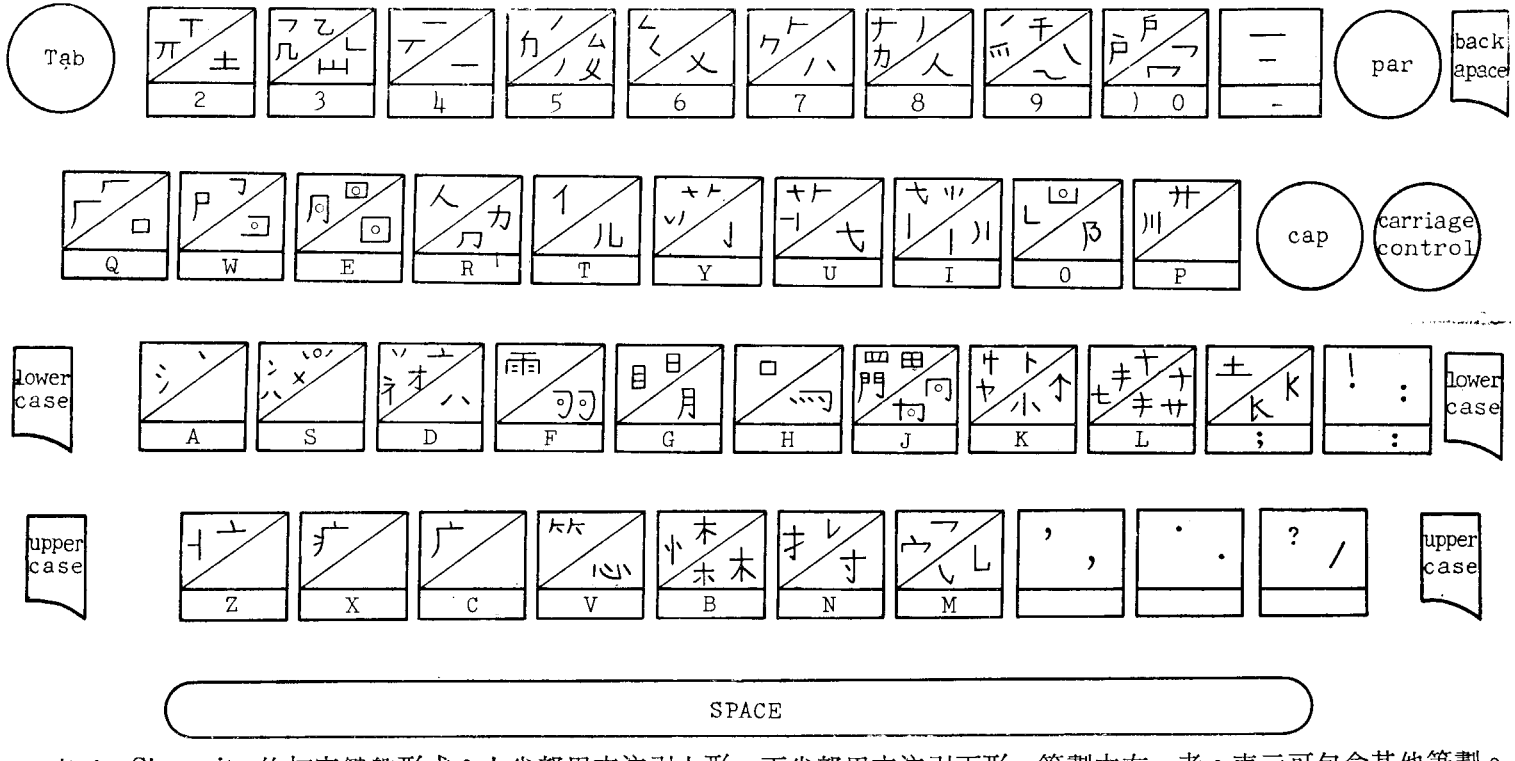
由於鉅量翻譯中文爲外文的需要，久久以來，美俄二國卽在進行利用電腦自動翻譯中文的研究。彼第一道難題卽在如何使中文進入電腦，正是我們也同樣面對待解的難題。因此，如何妥善設計一部中文電子字碼機來代替人爲的中文索引順序，向爲有關學者所努力者。

林語堂先生是我國最先着手研究，也是最有成就的發明家。自一九一六年起，彼經過五十年的思考，並傾家蕩產爲之，終於在一九四八年發明了中文打字機（註八）。這部中文打字機雖然沒有實際製造應用，但是由此發展出（相信用此中文打字機作爲理論根據及粗樣）下述二種中文電子字碼機的設計與製造。

在美國空軍資助下，一羣中美科學家與ＩＢＭ公司合作完成了第一部中文電子字碼機的設計。在一九六三年初，這部定名

爲「Sinowriter」的中文電子字碼機正式應市使用。

這部機器儲存中文數量依需要可以由六千五百個基本字擴充至一萬六千個不同的中文字。所有儲存在機器記憶系統內的中文字，依照它們的上下字形分類爲一千組左右。每一組內所包括的中文字都有着同樣的上下字形或設計時認作同類字形的中文字。例如「宏」、「安」、「蜜」、「宴」、「歹」、寥等字是同屬一組。因爲他們有着同樣的上形「宀」，與同類的下形（見表二鍵M、5）。



表二　Sinowriter的打字鍵盤形式：上半部用來注引上形，下半部用來注引下形。筆劃中有。者，表示可包含其他筆劃。

打字員每注引一個中文字時，所需要做的工作是：須要先認清這個字的上形及下形，然後在與一般英文打字機相似的鍵盤上依次按打上下（或左右，或左下）字形的鍵鈕。

機器在接收到這二個信號後即快速進行搜索，自約一千組字彙中選出所指令者，並將所有該組內的單字依次編號顯映在鍵盤上端的銀幕上。同一組內的單字可以儲存從一個字到最高十六個字。打字員在銀幕上選取所欲注引者之編號，然後依數按下數字鍵鈕。

這上形、下形、編號、三個鍵號組成一個中文字。每一鍵號由六個信息單位串連而成，亦即每一個中文字由十八個信息單位分三組串連而成（註九）。

一九六五年，美國波士頓ITEK公司將Sinowriter加以改良，使之更快更準，並改名爲「Chicoder」，即Chinese Encoder之意。Chicoder不但能將中文各字用打字機方法，打出電碼，也能打出中文原字。

上述二種已存在使用的中文電子字碼機，因爲不僅它們本身的租金售價極爲昂貴，同時對所聯用的電腦有諸多要求，包括機械方面及作業程序系統方面。所以使用者極有限，亦不宜在國內採用，故予討論其應用。但是在發展中文電腦過程中，它們佔着極重要的地位，特介紹如上。

丙、結　論

綜合上面對種種注引中文方法論敍，我們對一種切實可用及適合自動化處理用之中文索引順序字碼的要求或設計原則有如下三點：

第一：中文字與其索引字碼二者之間只有一對一的關係。亦即不能重複使用同一字碼，同時一個字也不能有用不同字碼注引的可能。

第二：須要可以直接注引字碼。亦即可以不借字碼書，也不須大量記憶要求，而能隨讀隨註，且要不受其發音、字體的限

制。

第三：使用人應該不受限制，亦即不論使用人對中文認識的程度，都能正確的注引字碼。

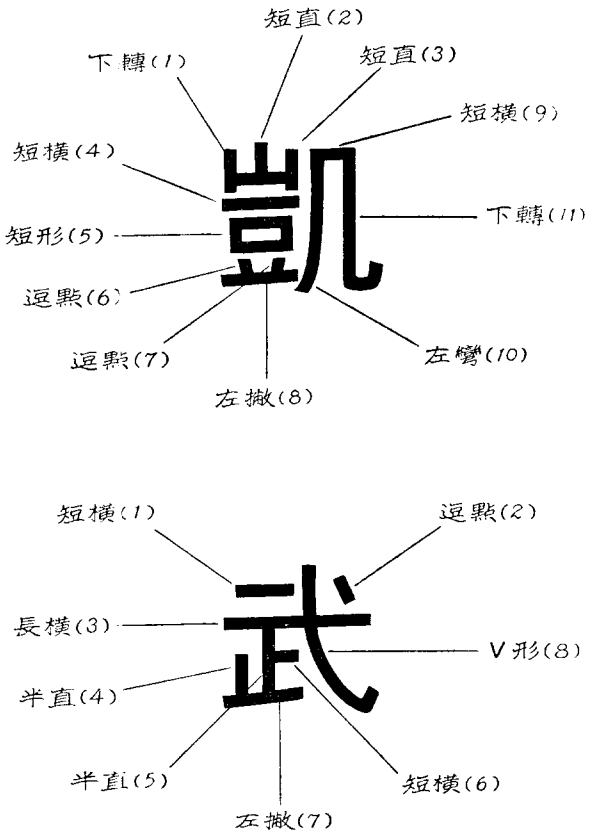
不待爭辯的，所有現有的中文索引順序方法都無法滿足這三點要求。除了文內所評介的三種人爲中文索引順序方法外，他如部首偏旁筆劃法、注音符號、字典編號等亦皆不能符合要求。

所以，根據「字碼不重複使用」、「直接注引」及「人人可用」三個原則去設計一套較妥的中文索引順序是有迫切需要的。

肆、陶氏中文電腦字母論述

根據「字碼不重複使用」、「直接注引」及「人人可用」三大原則，利用中文字本身息源爲着眼點去設計一套較完善的中文索引順序以爲自動機器用，已經很成功地由作者在美國北卡羅來納大學信息研究所與電腦中心資助下設計完成。並且很圓滿地應用在用中英文機器翻譯之中文進入電腦技術上。

這套方法定名爲中文電腦字母（Chinese Computer Alphabet），於一九六五年冬開始設計，至一九六六年春定案，並開始一連串試驗以考驗其可用性，包括人爲注引及機器辨認（註十一）。同年五月首次成功地字譯一段中文報紙新聞，七月編成有一千二百十四個常用字之機器用中英文字典。至此，中



圖三　拆注中文電腦字母實例
括號內數字表示注引順序，順序的取決有機械化地由上至下，由左至右選定，與傳統的書法略有不同。

文電腦字母之設計遂告完成（註十）。

以下討論中文電腦字母之情形。

甲、原　理

利用中文字的息源——不同筆劃作上下高低的排列組合，將每一個中文拆開爲若干「筆形」，每一種不同的「筆形」用一個相似的拉丁字母或符號來表示。將這一羣「筆形」換寫爲「中文電腦字母」後，依照其「筆形」所在方位，依次串連，卽成爲機器用中文字碼。

每一種「筆形」或卽是一種傳統的「筆劃」，或是相等於數個傳統的「筆劃」的集合。這是由於所謂筆劃（註五）是由中文書寫爲着眼點而取決的，所以有「ㄱ」而無「ㄱ」，原因卽是在書寫時，「ㄱ」順手易寫，而「ㄱ」不易下筆。代代相傳，變成傳統的寫法，例如書寫「弓」字。因此「筆劃」只是爲書寫中文字時而定的，如果從辨認中文書上的各個中文字爲着眼點去看中文時，則我們所以能區別「士」與「土」二字是在於「筆形」，而不在「筆劃」。除了用羅馬拼音法外，「士」與「土」二字就字形看，只有長短筆劃的區別。

在考慮「人人可用」原則下，由字形變化爲着眼點去設計中文索引順序是惟一可行之途。因此才能適用於任何人——識與不識中文字者。

將不同「筆形」用相表的拉丁字母或符號來表示是爲符合「直接注引」原則。因此在注引時，只要依照規則拆開一個字爲數個筆形，然後直接換寫中文電腦字母便成。

乙、中文電腦字母的組成及注引規則

中文電腦字母是由廿二個不同的字母組成。其中有九個是拉丁字母、一個阿拉伯數字、八個標點符號及四個邏輯符號。卽

工、J、K、L、N、O、V、X、Z、3、*、)、(、=、一、丨、十、＼、／、—及「。所有的字母皆可用IBM 029打孔機打出（註十二）。

每一字母的名稱、字形、原名、相等的筆形以及舉例見表三。

每一個中文字依字形簡繁可以拆成一個至最多約廿個筆形不等。在全部不同的中文字中，約半數以上可以拆成左右或上下二半。由於中文字部首的存在，同一部首的中文字往往有着相同的一半字形。因此，二種注引形式被設計出來；即固定長度字碼及變動長度字碼。

在使用固定長度字碼時，不論一個字的字形簡繁，一律用八個中文電腦字母串連表示。而在使用變動長度字碼時，則依每一個字的字形簡繁，用一至十六個中文電腦字母來表示。凡原來字形超過十六個筆形組成時，一律捨去尾後的筆形。如何拆字及選用中文電腦字母見圖三。

簡單的注引規則如下（註十、十一）：

陶氏中文電腦字母				
字母	名稱	原名	相等筆形	用例
／	左撇	槓或除號	／一／／／／	千彩沟只化杉
N	右撇	拉丁字母N	\ \ \ \ \ \	木云立朴規原
)	左彎	右括號	)) ノ \ ) ノ ノ	斤片刷月用充
(	右彎	左括號	((((((	瓜久又是人支
<	左角	'小於'符號*	< < < < < <	災紅云法么鬼
>	右角	'大於'符號*	> > ㄱ ㄱ)) ㄱ	歹免又發定疋
—	短橫	減號	—	王士其正孝蓮
——	長橫	下線	——	王土其正孝車
工	短直	拉丁字母工	|	正臣則定功裝
|	長直	'或者'符號*	|	正臣中此工門
ㄱ	上轉	'不是'符號*	ㄱ ㄱ ㄱ ㄱ ㄱ ㄱ	中用也的功巴
ㄴ	下轉	拉丁字母L	ㄴ ㄴ ㄴ ㄴ ㄴ	中匕己也世亡
十	單叉	加號	十	花田土士哉王
井	雙叉	號目符號	井 井 井 井 井 井	井寒昔青佳甬
⁂	逗點	星號	، ، ، ، ، ،	馬定江鴻戊母
3	耳朵	數字3	3 3 3 3 3	了丞及李陶孑
J	直鈎	拉丁字母J	J ㄴ	才氏丁水排似
O	矩形	拉丁字母O	□ □ □ □ □	固雖四貝石的
K	K形	拉丁字母K	⊢ ⊢ ⊢ ⊢ ⊢	炊螢笠裡愁勞
V	V形	拉丁字母V	＼ ＼) ＼ ＼ ＼	代民狗豕氏幾
X	X形	拉丁字母X	×	凶殺學爾凶腦
Z	Z形	拉丁字母Z	Z Z Z Z Z Z	進之乙建迅亿

表三　中文電腦字母之名稱、原名、相等筆形及使用舉例對照表

固定長度字碼之注引規則

規則一：每一中文字，一律固定用八個中文電腦字母注引；多則捨去（見表四例5），少則補零（見表四例2）。

規則二：由左至右，由上至下拆開筆形後依次注引。

規則三：凡一字可拆成左右或上下兩半時，則先拆左或上半部（稱第一部份），並注引後再拆注右或下半部（稱第二部份）。彼此獨立拆注。兩部份使用中文電腦字母數如下：

一、先注引第一部份，依筆形數目最多用四個中文電腦字母注引。餘下數用以注引第二部份。如兩部份合共不足八個筆形，則用零補尾。超過八個筆形則依次注引，捨去尾形。見表四例3、5。

二、如第二部份由少於四個筆形組成，而第一部份由四個筆形以上組成；則若第二部份須由一個中文電腦字母注引，第一部份可用最多七個中文電腦字母注引，以此類推。見表四例3。

表四　固定長度的中文電腦字母注引舉例

○母字丁拉別區之以零即∅＊

	中文原字	筆形分析及次序 / 中文電腦字母字碼								
1	電	一	一	亅	一	匚	亅	一	L	
		一	＊	＞	一	L	亅	一	匚	
2	子	一	3							
		一	3	∅	∅	∅	∅	∅	∅	＊
3	計	丶	丶	一	一	□	＋			
		＊	一	一	一	○	＋	∅	∅	
4	算	冖	冖	□	一	一	——	）	一	
		K	K	○	一	一	——	）	工	
5	機	＋	／	，	＜	＜	＜	、	＜	
		＋	／	N	＜	＜	＜	＊	＜	
6	學	／	＼	一	＊	、	亅	一	3	
		／	工	一	×	＊	＞	——	3	

變動長度字碼之注引規則

規則一：每一中文字，依其筆形簡繁，用一至十六個中文電腦字母注引。捨棄超過第十六劃之筆形。

規則二：由左至右，由上至下拆開筆形後依次注引。

規則三：凡一字可拆成左右或上下兩半時，則先拆注左或上半部。全部注引後，再注引右或下半部。兩部份獨立拆注。見表五例解。

	中文原字	筆形分析及次序 中文電腦字母字碼
1	電	一 ノ 亅 一 一 一 丨 一 ⌊ ⌐ 一 L 一 ⅄ ⅄ 一 一 一 I 一 L ⌐ 一 L
2	子	一 了 一 3
3	計	丶 一 一 一 □ +
		⅄ 一 一 一 ○ +
4	算	六 六 □ 一 一 一) I
		K K ○ 一 一 ⌇) 工
5	機	+ ＼ ＼ ＜ ＜ ＜ 、 ＜ ， 、 一 ／ ＼ ⌐ ／
		+ ／ N ＜ ＜ ＜ ⅄ ＜ ⅄ ⅄ 一 ／ N V ／

6	學	／ 丨 一 × 一 一 × 亅 一 、 亅 一 3
		／ 工 一 × 一 一 × 亅 一 ＊ ＞ 一 3

表五　變動長度的中文電腦字母注引舉例

丙、結論

在經過長時間多方面的實驗後，陶氏中文電腦字母被證明爲一較所有現用中文索引順序爲適用的中文機械化表示字碼。

中文電腦字母符合了作爲中文電碼的三大基本要求：即（一）每一中文字有其自有的字碼，包括同一字的不同寫法與不同的印刷體。（二）可以直接有系統的注引，不須記憶字碼，隨看隨注。（三）無論何人只要能辨認中文字即可注引，不需要會讀寫中文（四位僅讀過一年中文的美國人曾作注引實驗，證明此結論，詳見註十）。

表六列出中文電腦字母與其他中文索引順序之比較。

中文電腦字母除可用作自動化處理中文時之機械化中文表示外，尚可用作中文字典索引。尤其用以作爲來華遊客及駐中國地區外國軍隊用之中文索引最是適合。

在情報機關用中文電腦字母來編製密碼，最能靈活運用。

比　較　項　目	電腦字母	電報號碼	羅馬拼音*	四角號碼
單一不重複使用字碼	99.6%	100%	無法做到	須加改良
直接注引程度	完　全	無法做到	部份可以	完　全
使用者瞭解中文程度	0—1年	3—5年	5—10年	1—2年
字碼組成記憶要求	不需記憶	必須記憶	一般人需要	不需記憶

注引方法學習時間	短時即可	不必學習	長時才行	短時即可
字碼長度（最長）	8 或16字	6 字	6 字	5 字
單一字母變化	22電腦字母	10數字	26拉丁字母	10數字
每一字碼最大信息數＊＊	36.7或73.4	13.3	28.2	16.6

表六　現有各種中文索引順序方法之比較
＊教育部定羅馬拼音法
＊＊見註十三

伍、電腦用中文索引順序之索引的編製

由以上討論結果，我們知道在廣泛的自動化處理中文信息應用上，沒有一種中文索引順序可以概全地採用。中文電腦字母有其精密準確的表示，可以應用在機器翻譯、以及種種非數學的信息處理上。但是在某些特別場合下，如中文姓名、地址的字碼表示，如果在不太妨害其信息眞實性時，採用羅馬拼音當勝於其他表示方法。

因此，我們在目前唯一可行之途是採用混合法。不僅中文進入電腦時應該隨機應變，在同一事件上採用不同的表示方法來表示不同的信息。在電腦處理完畢印出結果時，也應隨結果的使用目的隨機印出最合適的結果表示。比如，在研究某一中文著作中不同字彙，成語等出現的頻率時，某進入電腦之中文表示可採用中文電腦字母以爲準確及快速注引，而其處理後之結果則用羅馬拼音合同中文電腦字母一併印出以利閱讀。如果有音義不明時再查對中文電腦字母爲憑。

爲達到這目的，我們可以在電腦記憶系統中預存下各種中文索引順序之對照索引，如此便可以彼此相應使用。

圖四所示者，即是由美國北卡羅萊納大學信息研究所編製的電腦用中文索引順序對照及英文註解字典之樣張。每一個中文

字皆有如下信息：

1. 編號——作爲本字典索引，並無意義。
2. 中文字——留出空白，用手塡寫，以便參照。
3. 中文電腦字母字碼——圖三所示係採用固定長度形式。
4. 羅馬拼音——圖三所示採用教育部定標準羅馬拼音。
5. 英文註解——最多使用卅五個字母來解註。不同的解釋用逗點分開，相似的解釋用／分開。儘量用縮短字表示文意。在括號內的字，比如（SURNAME），表示這個中文字屬於這一類表示，如例卽表示此字是姓氏之一。
6. 部首偏旁數——在小數點前的數字是部首列序數（一般使用二二四部首者），其後爲除部首外總筆劃數。
7. 電報號碼——爲機器用電報號碼。
8. 四角號碼——爲王雲五四角號碼。
9. 字典編號——出現在 Mathews 中英字典上之編號。
10 備註——同字而不同寫法者，彼此相註。

CHINESE-ENGLISH SAMPLE DICTIONARY FOR MACHINE PROCESSING PAGE 1

NO.	CHINESE	CCA CODE	ROMANI	ENGLISH EXPLANATION	RADICAL	TEL-CODE	4-CRN	DIC#	CRX-REF
1		—————		—					
2		———OOOOO	SAN	THREE	001,02	0005	10101	5415	
3		—／(☆＞／*O	CHIAN	LEAD ALONG	093,05	3677.1	40502	NONE	
4		——)NI—)L	GUEI	COMPASS, RULE	147,04	6016	56010	NONE	
5		—／｜L／OOO	BEEI	NORTH	021,03	0554	11110	4974	

6	−／J−−) LO	FUU	TOUCH GENTLY WITH HAND, TO COMFORT	064,04	2329	58031	NONE	
7	−／J−I L＞ (	TOUR	THROW	064,04	2121	57047	6490	
8	−／J／−＞ LI	TUO	DRAG ALONG	064,05	2151	58012	6440	
9	−／J／+／L−	CHA	INSERT, STICK IN	064,09	2252	52077	0113	
10	−／J+＞ (OO	JIH	SKILL	064,04	2111	54047	0442	
11	−／J☆−N／−	LHA	PULL	064,05	2139	50018	3756	
12	−+／☆(−+−	BAN	CLASS/RANK,(SURNAME)	096,06	3803	11114	4889	
13	−+／I) ¬ LO	SHIANN	APPEAR,PRESENT/NOW	096,07	3807	16110	2684	A60150A
14	−+／L¬−−｜	LII	REASON/PRINCIPLE	096,07	3810	16114	3864	
15	−−) ｜OOOO	KAI	OPEN,START,OPERATE (VEHICLE) (VEH1CLE)	007,02	7030.1	10440	NONE	
16	−−JOOOOO	YU	AT/IN, (SURNAME)	007,01	0060	10400	7592	A2456
17	−−OOOOOO	ELL	TWO	007,00	0059	10100	1751	
18	−｜−OOOOO	GONG	WORK/LABOR	048,00	1562	10100	3697	
19	｜L¬／−II−	JEENG	IN GOOD ORDER, WHOLE,EXACTLY	066,12	2419	58101	0356	
20	／−−+／I L／	HWA	PLOUGHSHARE, SPADE	167,12	6985	84154	2220	
21	／−−I¬｜IJ	JYH	SYSTEM,MANUFACTURE,	018,06	0455	22020	0986	D5956

			REGULATE					
22	／－☆－＜¬☆O	MEEI	EACH/EVERY	080,03	3020	80507	4401	
23	／＋／N I ¬－）	IANG	SHOOTS/SPROUTS	115,05	4441	25930	7244	
24	／＋／N ⌊¬	O	JOONG	SEED,KIND, RACE (OF PEOPLE) , TO PLANT	115,04	4429	25906	1524 4467
25	／＋／NOOOO	HANN	WITH/AND,MIX TOGATHER,WARM,SOFT	030,05	0735	26900	2115	
26	／＋／NO／NO	JI	ACCUMULATE	115,05	4480,1	26980	NONE 4480	
27	／ I NOOOOO	GEH	GENERAL CLASS IFIER/A PIECE	002,02	0020	80200	3366	
28	／I－－OOOO	REN	HUMANE	009,02	0088	21210	3099	
29	／I－ L¬－） (	BIANN	CONVENIENT,THEN, EASE NATURE	009,07	0189	21246	5224	
30	／I／－I－－O	TZUOH	DO, MAKE	009,05	0155	28211	6780	
31	／I☆－☆／－O	WEY	AN. FOR PERSON, SEAT, POSIIION	009,05	0143	2021	8711	
32	／I☆／O－－O	SHEENG	FRUGAL, SAVE, PROVINCE	109,04	4164.OA	90602	5744	
33	／II／¬×／／	SHIOU	REPAIR, CULTIVATE	009,08	0208	28222	2794	

34	/JNOOOOO	SHEAU	SMALL, YOUNG	042,00	1420	90000	2605	
35	/O-/¬☆OO	DE	SUBOR.PART.-,S,-LY,-ER (NOM.) ,' IC.	106,03	4104	27620	6213	
36	.....							
37	<</-I-OO	HORNG	RED, BONUS, POPULAR	120,03	4767	21910	2383	
38	<☆I¬ L- L-	NENG	ABLE/CAN, ENERGY	130,06	5174	21211	4648	
39	(((((((((		(					
40	+->¬//OO	CHAANG	FIELD,PLACE	032,04	1034.OB	46127	NONE	
41	+/> LIOOO	DIH	GROUND	032,03	0966	44112	6198	
42	+/N-) LOO	JI	MACHINE, OPPORTIUNITY,SECRET	075,02	2623	47910	0411	2894
43	+/N/> (OO	GER	RULE,FRAME	075,06	2707	47964	3309	
44	+/N+/NOO	LIN	WOODS	075,04	2651	44990	4022	
45	+/N☆-#∨/	SHIEH	TOOLS	075,07	2750	43950	2538	
46	+/NOOOOO	MUH	WOOD, TREE	075,00	2606	40900	4593	
47	+++++++++		+					
48	+¬☆) ☆OOO	SHYE	COOPERATE, HARMONIZE	024,04	0588.1	44037	NONE	
49	+-OOOOOO	TOU	EARTH/DUST	032,00	0960	40100	6532	

50 +OOOOOOO SHYR TEN 024,00 0577 40000 5807

圖四 機器用各中文索引順序對照及英文註解字典樣本

注意，圖四所示爲經過安排後整齊依次印出者。但在電腦記憶系統中，只前述第三項至第十項資料緊密地儲存起來。各項目之間沒有空白，以減記憶長度。

由此，中文索引順序選用問題得到了滿意的解決方法。

陸、自動化處理中文信息應用泛敍討論

原則上，由於中文可以用中文索引順序來表示以之中文輸入電腦，凡是西方國家可以應用電腦處，經過適當的處理，也照樣可以應用在中文信息上。

今天在西方國家，利用電腦來處理各種信息的應用可能超出一千種。從天文、地理到人文、政治，從近代物理到古典文學。沒有人能下一個完整的定義，也沒有一本書能圓滿地解答什麼是電腦的應用。

因此作者與其空洞地提出一大堆應用名稱，不如專一討論其中一種。即在人文科學上的應用，同時較完整地討論在人文科學應用中最具有向人類智慧挑戰的應用—機械翻譯。

利用電腦的高速處理鉅量信息的特性以應用在人文科學上，在西方國家已有很高的成就。據作者的淺見，其中有若干也應在國內迅速地研究發展。茲擇其重要者討論如下：

甲、作者考證

歷代的名著往往發生原作者失傳或令人懷疑情形，例如紅樓夢作者。利用電腦來考證作者方法是先搜集有關作者的著作，選取與所懷疑之文相近性質者，使之輸入電腦。就每一作者的單字字彙使用，複字成語使用，虛字使用、句子長短、分段情形

，以及整段及整篇文章用字數，配合數理統計學及機率學決定其分配模型。然後將所懷疑之文輸入電腦，同法求得其分配模型。每一作家有其特有的寫作文字分配模型，因此不難決定誰是眞正的作者。

這種方法不僅可以考證古代著作的作者，也可以作爲今日科學測查翻印盜印他人著作時之方法。

作者考證研究在一九六六年始見於文字，現在由美國 Kent State University 的 Jacob Leed 教授領導研究中。

乙、圖書參考資料整理

大量的科學專門報告經年提出，使得世界最完善的美國國會圖書館也無從處理。其醫學部（Midlib）平均每日收到三千件新的發現、病例處理、藥品配方等等資料。息源來自全世界，使用着一百多種不同的文字。

無疑的，電腦又被付托重任了。今天世界上任何一位醫生，可以在二個星期內自Midlib得到任何細小問題的全部答案。而在往年使用人工處理時，可能在等候幾個月之後得到否定的答案。

中文書籍自古迄今，多不勝數。着手設計一套可以回答任何朝代、任何作者、任何事物參考資料的中文圖書自動化處理系統有其無法估計的價值。對發展漢學研究當有莫大幫助。至於科學上的專門報告，目前對國內似無利用電腦處理的必要。但目前有關單位或學校應着手編印科學文獻出版年鑑，分門別類簡評介紹，是不可再拖延的。若干年後則可利用電腦就年鑑所有資料作詳細的依人名、書名、範圍等分類工作。

丙、古文辨認

世界上有許多失傳及死去的語言，例如東方的甲骨文及西方的拉丁文。在許多情形下，對新發現的遺跡上刻着的文字，無人能夠辨識。無數語言學家經年鑑認結果，卻不能得到一致的答案。於是電腦又扮演着博學多才的專家的角色。在各種古代語言專家的合力「教育」下，電腦記憶着各種文字的有關資料。現在每當有疑問時，最高明的語言學家不能解答者，電腦卻能提

出它的見解。雖然它的智慧是人類稟賦的，但是它能作更縝密的考慮，因爲它記得所有的有關資料，而且一絲不亂，分毫不差。

用中文電腦字母來分析表示甲骨文或其他中國的古文、古體。並且「教育」電腦來學習這些古文，相信一定是一項很有趣味，但是非常有價值的研究。

丁、經典名作索引

當一本經典或傳世的著作廣泛地被人研讀時，索引的編製便有着極重要的地位。已經由電腦完成編印索引的例子很多。例如聖經索引，我們可以得到數十百處出現「彼得」二字的章節資料。當你查到「罪」字時，在這本索引裏列下了論百的章節資料，那些章節都是提到「罪」字的。

莎士比亞的名著索引裏，我們可以得到所有使用某一特別字眼的出現處，對研究文法，語氣、修辭等等的學者，當然有很大的幫助。

在中文裏，利用電腦將重要的　國父遺教等著作及基本文化教材中的四書五經編出詳細的，以字、短詞爲單位的索引，自然有重要的貢獻。電腦「學」會唐詩三百首後，還會自鳴得意的「印」出一些即席詩來呢！當然這是對中文學者對詩詞分析理解程度的一項最好的考驗挑戰。

在美國大學裏，無數以這些題目爲內容的博士論文已經發表了。國內中文研究所的學者們，這是新的嘗試。

戊、集韻研究

在中文裏，一個字可以有上十上百的發音，南腔北調。倒底有多少種？誰也弄不清。因爲實在字數太多，音調太雜。中文字有絕大部份恐怕一般人一輩子沒有看到過。除了常用字外，眞是還有太多太多的字。它們在語言上的地位如何？音

韻如何分析？

中國的詩詞賦都講究韻律，因意選字，都有一定的規則。那麼比如要寫一句描寫小橋流水時，究竟那些字可以用呢？再如我們想用廣東話來咏誦時，情形又如何呢？同時，某一篇古代的詩或詞，讓我們來推考作者在寫成後是用那一種方言來誦讀欣賞的呢？

民國以來，這些問題很少被深入地研究過，因爲被認爲是老學究的事。但是失傳的許多中醫治法及藥方，便是這樣論調的結果。用現在國人中文程度做準繩，即使二十年後的情形便值得憂慮。因此大規模徹底的整理中文漢字有其時代的意義。際此匪僞政權消滅中國古有文化之際，與其集會申討，大唱高調，不知默默地用科學方法有計劃的整理分析這被上百億人使用過及全世界四分之一的人在使用中的文字。

利用電腦來解決這些問題是值得研究討論的。

己、機器翻譯

利用電腦可以儲存鉅量資料及高速處理的性能，來應用在機械化翻譯文字上，已經有十年以上的歷史。這十多年來，機器翻譯一直是那樣的吸引著萬千位專家學者的注意，他們用各種方法去求得最簡捷、最完善的翻譯結果。來自世界各國的語言、電機、電腦作業設計專家經年的舉行會報，研討着共同合作的方案。無數的研究報告提出，無數的實驗結果發表。世界機器翻譯協會（註十四）按月按季發行通訊，報導全世界發展情形。

在蘇俄莫斯科附近，在十三年前便成立了「電腦機器翻譯語文中心」。上千的語言及電腦專家從事於①供電腦用之英俄字典編輯；②俄文文法編輯；③供電腦用德俄字典編輯；④供電腦用日俄字典編輯；⑤供電腦用中俄字典編輯；⑥中俄、英俄、德俄、日俄文互譯研究。

在日本，一個英日文互譯的特製電腦早在十年前開始使用着。如其他科學一樣，日本一直在電腦王國中有極高的地位。現

在是全世界第三使用電腦數最多的國家。超出所有美洲歐洲國家而僅次於美國及西德（註十五）。這是我們應感到羞慚的。

在以色列、捷克、墨西哥等許多地方都有着大規模的研究計劃。但翻遍所有文獻，尚未發現國內發表的研究報告，是爲憾事。

以下簡要的討論中文到英文機器翻譯過程。

整個進行過程可以大致分成四大部份，即

（一）中文進入電腦技術及編製須預先儲存的電腦用中英文字典。

（二）中文語法、語義分析，召集適當英文字彙。

（三）英文語法、語義分析，重行安排英文句型。

（四）翻譯結果之整理印出。

關於第一部份，在前文內已討論很多中文進入電腦的方法，而前敍 Sinowriter及Chicoder 即是特別爲此目的而設計者。在沒有這些中文電碼機時，採用中文電腦字母人工注引是最妥當的方法。注引時由二至三位打孔員依次傳遞注引，凡結果彼此相合時，即進入電腦。用中文電腦字母注引中文，平均每分鐘可注引十字。一萬字之中篇文獻，約需十七小時注引時間。

中英文電腦用字典是一件累人的工作。因爲中文與英文在文意上截然不同，每一個可能的中文組合都需要收在字典內，缺一不可。一個中文組合可能即是一個中文字，也可能是由二個中文字結合而成。同時，在不同的句子應用中，同一個中文組合，往往有着完全不同的解釋與說法。但是我們都要收在字典內，否則便有語義不清的結果。比如在字典內沒有收錄「多級彈道（Multistage ballistic）」這個新名詞，電腦在接收到一段中文句子而含有此四字時，在召集來的英文字彙中很可能是 many, grade, ballistic或者是many, grade, bullet, path 用這些英文字彙去譯成一個恰當的翻譯，其結果可想而知。

因此，在字典中的「一」字下，儲存的字彙需要有一、一定、一個、一件、一張、一條、……、一身、一進來、……、一飛衝天、一馬領先……等等無數字的組合。字典的周詳決定翻譯的結果。因此在實際應用時，可以按學科性質分別編製分科字

典、如數學字典、化學字典以縮小範圍。不同的文獻用不同的字典。

在中文進入電腦後便開始第二部份工作，即分析這一篇中文以句爲單位的語法及語義。

在電腦用中英文字典中，除了英文字彙外，每一個中文組合都應帶有語義的資料以爲分析中文時用。

美國加州大學機器翻譯計劃已經編妥六萬字彙的語義分析及二千餘條語法分析規則（取材一九六六年底文獻）。彼與美國德州大學語言學研究中心合作下發展自動分析系統，頗有成效。這是最基本也是最重要的工作。

語義分析即給予每一中文組合的每一註解一個特別的資料。比如單字或單語的語義表示爲

人　NALHOC即此字爲noun, animate, lowest level, human, any number, coman。

範圍NBLNMC即此語爲noun, abstract, lowest level, nanimate, mass, comon。

分化VILAO即此語爲verb, intransitive, a lexeme, animal agent。

一語或一字有二個以上註解時，則分別表示，例

處理
- VTLHO／NN　'trest' as in chemistry
- VTLHO／NH　'punish'
- VTLHO／NB　'manage' or make disposition of'

整句中文經過無數次的分析後，得到了適當的字語的組合，再尋找出每一個單字或單語所處的文法地位。然後再合乎文法的重新組合這句中文，以尋找出虛字的地位。

在機械化分析中文時，有二個字是非常特別的，即「的」和「了」字。它們本身沒有意義，專門用來與其他字搭用。因此在對整句中文進行分析時，「的」字往往用來作爲分析時的起點，而「了」字在一般情形下用來作爲一句的終點。「的」在中文裏的地位好比英文中的「THE」字，語言學上所謂功能字（FUNCTIVE）。

IBM公司在一九六三年所作的中英文機械化翻譯，因爲沒有將中英文作語義分析，所以結果並不理想。沒有語義，在翻

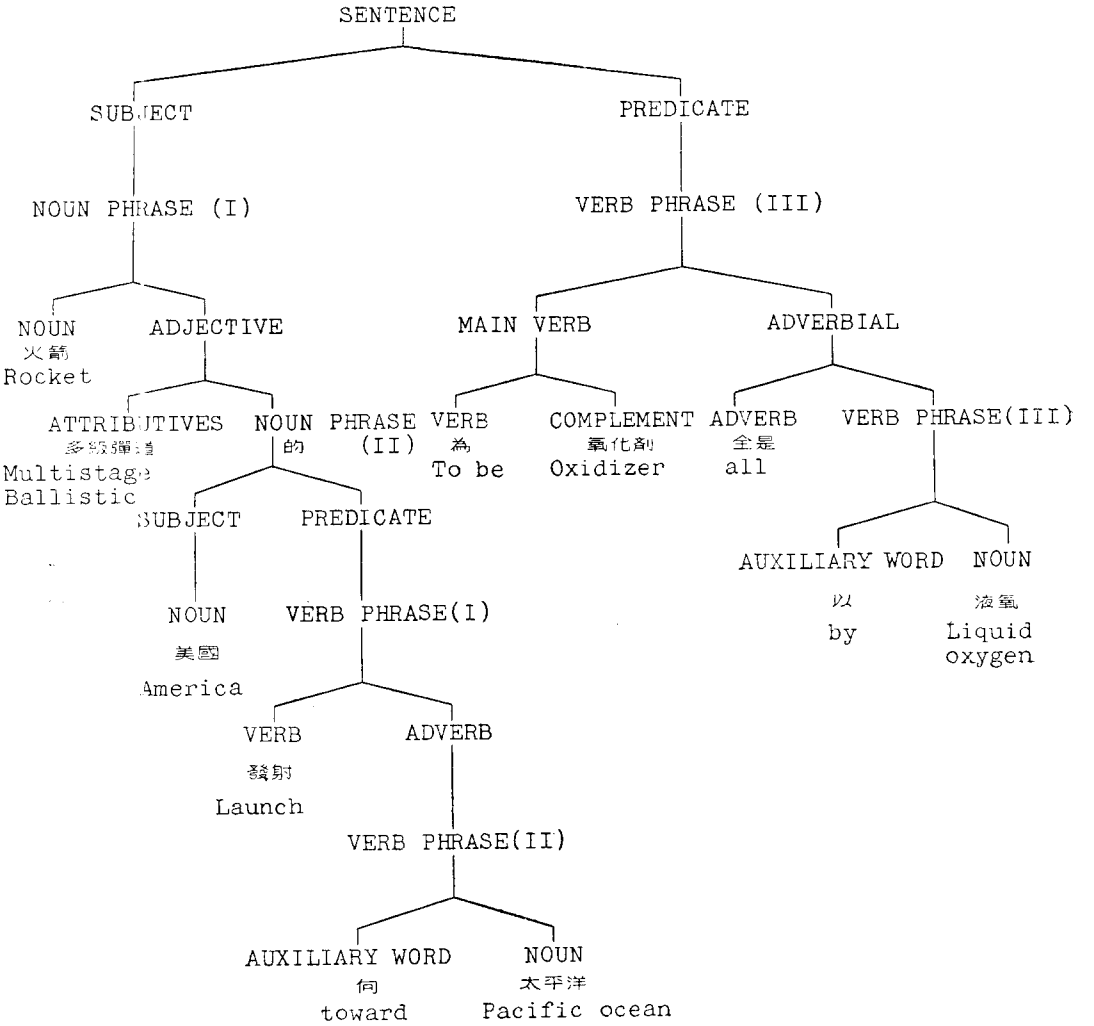


圖　五　　中文語法分析　一句中文"美國向太平洋發射的多級彈道火箭全是以液氧爲氧化劑"進入腦後"經過語法分析所得到的標準樹形句型。

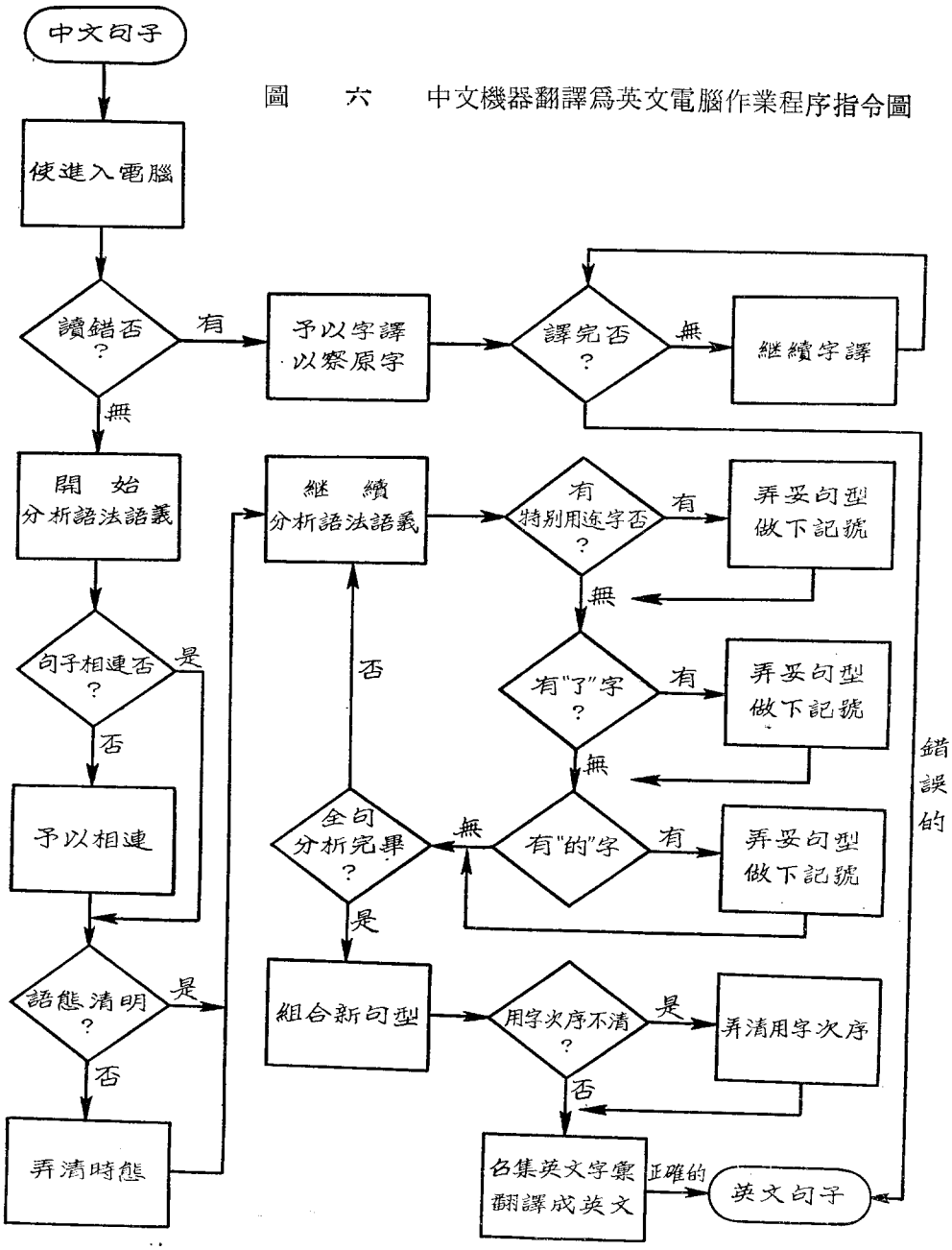
譯中文的命令句「不行！」時，便使電腦無所作爲了。因爲「行」字有許多不同的用法與意思。

在分析後的中文句子，每一個字都有着特定的文法地位及固定的語義，這項要求是絕對嚴格的。不能有一個字有不清的語義或無所適從的語法。無數語言學家終年埋首研究着這項問題，但是除非大規模的從事每一可能句型的分析研究，結果總是局部的。歸根清源，唯有在國內展開研究，才有成功的一天。

圖五所示是一句中文在用電腦分析文法後所有各字的文法地位的例子（取材自註九）

圖六所示是說明設計電腦

圖　六　中文機器翻譯爲英文電腦作業程序指令圖



分析中文文法、語法及語義的自動處理系統指令圖。每一步驟的進行都是需要長期的設計與測驗。圖七所示是字譯中文的全部作業過程。

中文沒有特別表示時態的用字。不像英文的動詞，可以分別時態。但是中文有不少習慣用法，可以大致用來決定時態。比如「我來了」，應該是現在進行式。「我來過了」，可以看作現在完成式。「我三天前來過了」，可以看作過去式。而「我就來了」，可以決定爲將來式。在無法估計的用字中，去作

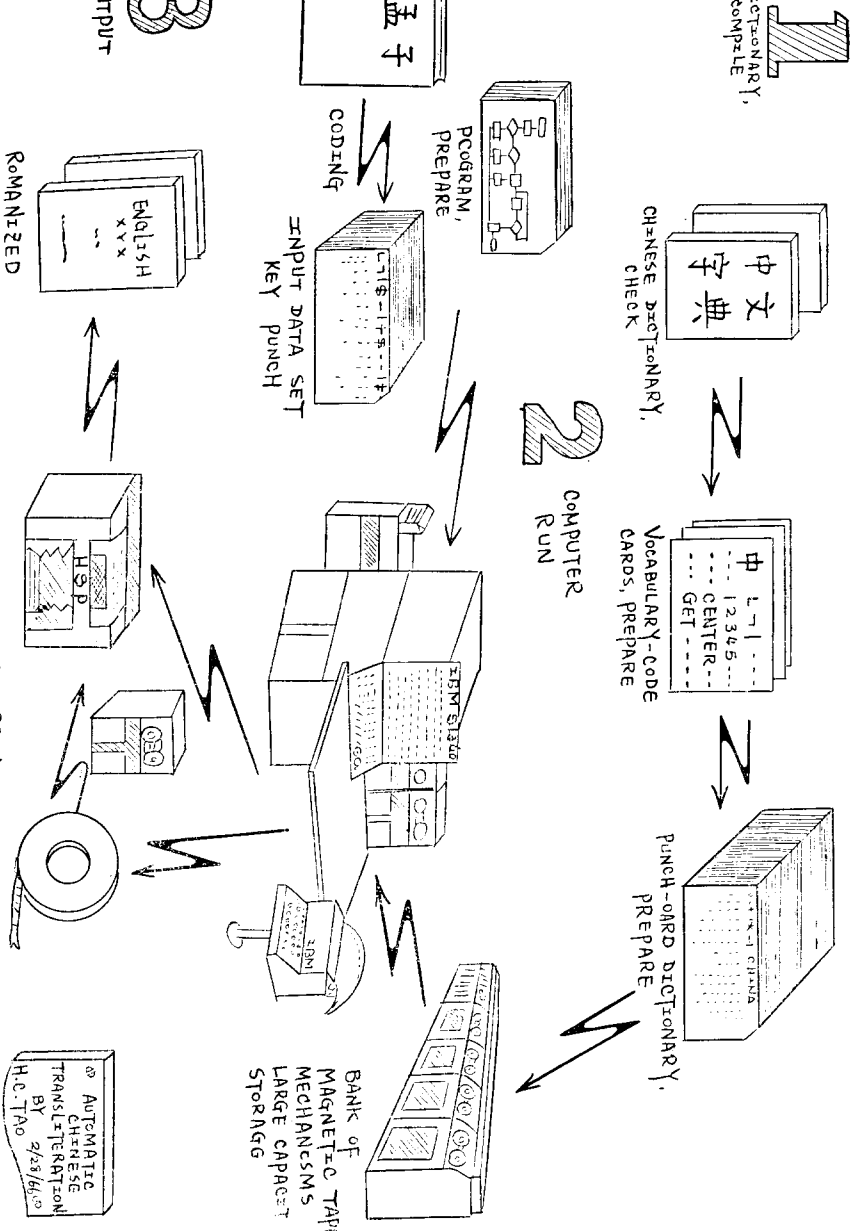


圖　七　字譯中文爲英文的全部作業過程

這樣時態的擇定，又是一件累人的分析工作。因爲中文中沒有絕對的時態文法，在較長的中文句中，照英文文法的規則分析，時

常會分析出不知所云的答案。唯有將所有可能的情形收集後，才可以尋求出一套通用的規則。

其他如動詞、動名詞及名詞的區別，也是需要重覆分析的。在作這分析時，有不少字是有特別用途的。即「之」、「爲」、「于」、「以」……等字之後的動詞是次要動詞，應用TO 冠之或視作動名詞、名詞使用。比如在「他致力於教育之推廣」及「他從事教育之改良」兩句中的情形。循此可以得到很多線索。

在完全分析一整句中文後，便須在茫茫大字典中選取最適當的英文字彙。這項「召集」的技術問題是十分複雜的。一般的方法是採用「查檢法（Look-up SYSTEM）」，即是由頭至尾逐一查檢。這是最簡單的方法。在用作測驗字典容量或字譯時，這是可行之法。如作者爲測驗中文電腦字母編成字典之效用，在一九六六年七月所作字譯聖詩時，即採用此法。

利用新發展的資料召集（Information Retrieval）技術，我們應該可以改良召集技術。使用「變換索引」（Key transformation）方法，可以只須查驗字典一小部份便可尋得恰當的英文字彙。召集技術的研究是信息學的一部份。雖在機器翻譯應用中佔有極重要的地位，但離本文主題過遠，不再討論。

第三部份的英文語法、語義的分析與第二部份的中文語法、語義的分析是有絕對相關的工作。有好的中文分析，才能導致好的英文分析。比如「……在清代後期……」，若能分解爲「在……後期」及「清代」，則可以召集「at the end of」及「Ching dynasty」英文字彙，自然可能有較通順的英文結果。否則，若分解爲「在」、「清代」、「後期」，所召集來的英文字彙不外「at」、「Ching dynasty」及「last period」之類。用設計最妥善的英文分析系統，也難作出理想的翻譯。

重新組合英文句時，我們的難題是：中英文語氣不同、文字表示次序不同、虛字的插入等等。作者未曾涉及這一步驟，不能提出討論。

最後一部份工作較爲簡單，主要在電腦作業程序的設計問題，茲從略。

中文譯爲英文大致情形，即如上述討論者。至英文譯爲中文的情形，並不即是還原，二者大有差別。主要前者以中文爲主體，後者以英文爲主體。前者的研究人員應熟諳中文，後者的研究人員應熟諳英文。

目前各處研究情形很值得研討。在國外的研究計劃，由於實際需要大都是從事中譯英方面，但是不能羅致足夠的中文人才。而將來在國內如果也從事於機器翻譯，無疑地主要興趣在英譯中方面。是否有足夠的英文人才來從事研究工作是不無疑問的。

最理想的情形是中英互譯合併同時研究。研究地點除電腦使用一項因素外，以國內爲宜。研究人員皆需熟諳中英二國文字，最佳人選配合當爲人數在三、五十人之間；半數須專家學者，半數可爲一般助理；再半數從事語言分析而瞭解電腦原理作業，半數爲電腦工作人員而瞭解此計劃之語法、語義分析系統，須從事長期的研究。

機器翻譯有光明的前途，有研究發展的價值。它對教育文化將有莫大的貢獻。國外新發行的科學書刊可以從此大量的譯爲中文供國內萬千學子研讀，從此國內對新知識的摘取如同在國外一樣，國內科學的水準將與國外一樣有立足點的平等。凡我有識之士，盼一同振臂提倡發展。

柒、後　言

本文沒有企圖說明中文信息自動化處理的全貌，僅僅討論一些原則及發展途徑。事實上，每一件談到的事物應用，都可以單獨地寫成一篇論文。作者儘量在保持本文爲學術性論文式體裁，但是不少機器及應用還是附加說明，這是想使讀者有更深刻與具體的瞭解。因爲有關文獻在國內恐不易得見。

感謝本校信息研究所F. P. Brooks, Jr. 教授給作者的幫助，他對作者設計中文電腦字母作了無數的建議及修正。沒有他高度有恆心地指導，作者將永遠在黑暗中摸索。

本校馬安尼（A. Marten）博士和她的學生們協助作者對各種中文索引順序方法作注引實驗，是作者感到難以爲報的。魏平、張德仁兩兄幫忙整理資料，友情感人。

更謝謝家父及姑父郭雁翎先生在國內幫忙剪報及搜集資料提供意見。沒有這麼多人的支持合作，永遠沒有本文的產生。

參考文獻

註一：可參閱文星書店出版范光陵著電腦與你及中央星期雜誌五五年四月十七日起所刊陶潛慶著電子計算機的原理發展應用。

註二：參閱Planning Research Corp. (1962). Survey of the Need for Language Translation, IBM Survey Dept. RC-634, March 12, 1962.

註三：參閱Casey, R. and G. Nagy (1966) "Recognition of Printed Chinese Characters", IEEE Transactions on Electronic Computer, vol. EC-15, pp. 91-101.

註四：參閱Yuen, R.C. (1948) Mandarin Priner, Harvard University Press. Chen, H. C. (1939) Modern Chinese Vocabulary, Commercial Press Ltd. China.

註五：綜合有關著作，不同的中文筆劃約共可分歸爲卅一種，即 ㇀ ㇇ ㇂ ㇁ ㇄ ㇅ ㇆ ㇈ ㇉ ㇊ ㇋ ㇌，可參見 Greel, H. G. (1943). Chinese writing, American Council on Education, Washington, D. C. Simon, W. (1959). How to Study and Write Chinese Characters, Humphries & Co. Ltd. London. Yee. C. (1954). Chinese Calligraphy, Harvard University Press

註六：Dougherty, C. Y. S. M. Lamb and S. E. Martin(1963). Chinese Character Indexes, University of California Press, Berkeley and Los Angeles.

註七：IBM SYSTEM/360 電腦是所謂第三代電腦。在一九六四年宣佈設計完成後，一九六五年中正式推進市場。但是在一九六七年二月份「電腦與自動化」月刊上調查資料顯示，全世界超過半數的用戶已換用此型電腦，且有急速增加的趨勢。我國有關單位在設計成立電腦中心時應注意及此。其計算單位稱爲 BYTE，由八個信息單位（BIT）組成，每一字母、符號或數字皆用一個 BYBE 來表示。但是每個BYTE也可以分別表示二個數字，即各用四個信息單位串連表示。在探討儲存量時，所謂一個「字」，是指一個字母、符號或數字。單獨討論「數字」時，是指前敍一個 BYTE 表示二個數字者。因此：例如在用中文電報號碼表示中文時，一個「中文字」將由若干個「字」組成。同樣地，一個「英文字」也是由許多「字」組成的。讀者宜注意之。

註八：詳見民國五十四年十一月廿九日中央日報國際航空版林語堂著：中文電子字碼機。

註九：簡要操作介紹，可參閱 King, G. W. and H. W. Chang (1963). "Machine Translation of Chinese", Scientific American. Vol. 208, 124-135.

註十：關於中文電腦字母之詳細設計經過，可參閱作者之碩士論文「on a Chinese Computer Alphabet for Automatic Machine Processing」1966 M. S. Thesis. Department of Information Science, University of North Carolina at Chapel Hill.

註十一：參閱作者著"How to Use the Chinese Computer Alphabet" work paper, Department of Information Science, University of North Carolina at Chapel Hill. April 1966.

註十二：如果用者有IBM 026打孔機而無IBM 029打孔機時，則可以用T、1、7、G及D依次代替在IBM0 26打孔機上所沒有的ㄅ、ㄧ、ㄑ、ㄏ、及ㄟ等五個字母。如果用者作如此替換，則須統一使用，不可混用。

註十三：每一字碼可表示利息數大小，表示它可以作不同表示可能性的大小。其計算方法及原理，詳見Brooks, F. P. Jr. (1963). Automatic Data Processing, John Wiley & Sons, Inc, New York.

註十四：Association for Machine Translation and Computational Linguistics

簡稱AMTCL。詳細情形可函下列地址：

Professor Harry H. Josselson

Socretary-Treasurer, AMTCL

Department of Slavic Linguistics

Wayne State University

Detroit, Michigan

U. S. A. 48202

註十五：根據美國Computers and Automation月刊十六卷一期發表的一九六六年十二月所作世界電腦使用量調查。茲檢列前十名及使用量如

下，以爲參考：①美國二八、五〇〇套，②西德二、七五〇套，③日本二、一〇〇套，④英國一、七〇〇套，⑤法國一、五五〇套，⑥意大利一、一五〇套，⑦加拿大一、〇〇〇套，⑧蘇俄一、〇〇〇套，⑨澳州四五〇套，⑩荷蘭四一〇套。全世界共有四四、四五五套。這裏所計算者是指一整套作業系統者，不包括小型單一式的電腦或插線式電腦。