中文輸入輸出系統之設計及效果之測定

周寧森*

一、前言

當今之世，學術研究日趨發達，各種文字、各種學科的出版品浩若烟海，自動化資訊系統之建立乃成爲當務之急。世界各國屬拉丁文字系統者，因字母數有限，自動化資訊系統已逐漸完備。中國文化歷史雖源遠久長，文字紀錄更爲舉世之冠，然中文尚不能全面以電子計算機處理，不但妨礙到與他國文化之交流；長此以往，學術研究及文化進展必遠落於他國之後。影響深遠，不容忽視。

資訊系統自動化之先決條件便是要能使機器辨認中文，也就是說，要能使中文輸入輸出電子計算機。但中文字構成複雜，字數繁多，全國各地更有不同的方言俚語；此先決條件，便成爲建立自動化中文資訊系統的一個極其嚴重的問題。

解決一個問題的方法很多，每一種方法都有其目標和隱於幕後不同的背景。甲解決某問題之方法，乃是根據他個人對此問題之知識及對週遭各因素之瞭解而定。乙對同一問題，因其瞭解及認識之差別，解決方法可能全然不同。二人所得之結果亦因之而異。墨子曾說：「是以一人則一義，二人則二義，十人則十義。其人茲衆，其所謂義者亦茲衆。」[1] 各人對同一問題觀點之差異，亦可能由於各人對此問題之解釋不同而起。

對中文輸入輸出電子計算機之問題，許多專家學者，見仁見智，曾提出了許多不同的解決方案。這些方案間的差別，皆源於提案者背景及其對此問題之瞭解而異。這些方案可以其目的之不同粗分如次：

一、部份解決方案。

二、全盤解決方案。

提部份解決方案者，因感到電子計算機應用之亟需；如科學研究、會計作業

* 著者爲美國 Rutgers University 遠東圖書館主任，現任國立台灣大學圖書館系客座副教授。

、商業應用等，皆能以電子計算機之快速精確達到節省人力物力之目的。據此方案而設計的輸入輸出系統，乃基於其目的範圍中常用的有限字彙。字數既少，問題較簡單，系統之設計便亦較爲容易。這種提案是基於下列理由及假想：

一、目前亟需應用電子計算機之事物，能儘早得到電子計算機之協助。

二、以同樣原則，稍加細節變動，便可建立各種不同目的輸入輸出系統。

誠然，如此做法，也許可以很經濟地設計某些既定目的的系統。是否能以同樣原則建立其他不同目的的系統，尚待時間及實用結果來證明。但若以設計一個有限字彙的系統之原則，來設計建立一個需用全部字彙者（如圖書資訊系統），便大成問題。換言之，提部份解決方案者，也許未曾深思熟慮及將這些不同目的之系統統籌運用時所可能發生的各種不同的困難。

提全盤解決方案者，爲求一勞永逸，乃以全部字彙爲範疇，來設計輸入輸出系統。對這些人來說，人類文化進步的過程中，各專科學術研究領域，必然趨向交綜錯雜，預爲籌謀，提出全盤解決方案，必然可以減少未來系統轉換應用時，及必須統籌運用各系統時之許多不必要的困擾。他們更慮及，若零星地建立各種不同目的之系統，必趨零亂；一旦必須統籌運用，牽一髮而動全身，改一錯而造十失。同時，在設計特殊系統時，很難期望設計者高瞻遠矚，顧及未來統一運用時所可能發生的一切問題。

筆者是圖書館員，自然關切圖書館的日常作業；諸如圖書採購、分類編目、資料檢索、圖書流通、以及資訊儲取等。若顧及這些圖書館的日常作業，便不能考慮採納部份解決方案。因文字紀錄浩如烟海，上溯遠古，兼及未來；死字也好，活字也罷，必須兼收並儲，必須要求一個具有高度彈性，而能包括已有全部字彙之輸入輸出系統。設計這樣一個系統，必須緊記後列基本原則：

一、輸入輸出代碼之獨一性。

二、自代碼還原至中文字之可行性。

三、錯誤察覺及訂正之機械性。

四、擴充字彙時不可重大影響原有操作程序之彈性。

五、學習、應用之簡易性。

六、經濟性。

二、實　例

中文字數繁多，但音節却少，無法以一音獨代一字。以四萬九千餘中文字[2]（不計新創字及未來可創字）之衆，國語音節僅有一千三百左右[3]，同音字之多可以想見。舉例來說，在韋氏漢英字典再版本中（Wade-Giles' Chinese-English Dictionary. 2nd. ed.），在「i」音下，有字一百零九；「shih」音下，

有字一百零五。而韋氏字典僅含中文字萬餘。由此可知中文字中同音字問題之嚴重。同時，多音字及各地方言使得這問題更形複雜。[4] 對一個輸入輸出系統來說，一字一碼是必然的先決條件。以音爲基礎，徒增困擾。分析字形及其構造，以形爲基礎來設計輸入輸出系統，似乎是較佳途徑。

舉最近設計的幾個輸入輸出系統來做實例，也許能將所牽涉的問題更具體而詳盡地表示出來

例一：

一九七三年，中央研究院的張系國先生及其同仁設計了一個輸入輸出系統，名之爲「中文音序檢字系統」。[5] 是以注音符號爲準。據稱此系統符合下列三原則：

一、輸入的方式，一定要簡單易學，適合一般使用。一般人不必接受長期訓練，也可以使用這個系統作業。

二、爲了避免輸入錯誤的資料，必須有適當的方法，能找出錯誤所在，由使用人加以改正。

三、這個系統必須能配合我國現有的各型計算機系統，構成完整而適用的中文輸入輸出系統。

此系統主要是爲通曉中文的國人設計的，先決條件是使用人對需用字之熟識。注音符號中廿一個子音，以廿一個英文字母（除Ⅰ、K、L、O、P外）爲代碼；十六母音，以十位數字（〇至九），加上A、M、Q、Z、，（逗點）及/（斜線）爲代碼；廿二個複合母音，以廿二個英文字母（除A、M、Q、Z外）爲代碼；以四個數字（一至四）爲四聲之代碼；ϕ爲輕聲之代碼。因基本音符只有卅七個，以常用英文打字機，略改字鍵便可輸入中文之工具，設計者並未說明重複使用某些代碼之原因。但重複代碼發生在不同的音組內，或表複合母音，或表子音；或表四聲，或表母音。各音組在輸入代碼中有其一定的位置，應該不會導致混淆。

爲了區分同音字，設計者先給同音字以筆劃多寡賦予序列號，但序列號是按序加值，不便記憶。此系統有一字檔，含字六千餘，以音爲序，儲於磁碟，再以「互作法」，另建字檔索引，索引各欄，指向字檔內各同音字組之起始位置。但此種互作法，必須人與計算機密切合作，必須「連線作業」（On-Line Operation），設計者認爲中文基本單位是「詞」而非「字」，若以詞檢法，詞檔之建立以「樹狀結構」（Tree Structure）爲原則，便可避免使用互作法，更可減少同音字問題至最小限度。是以另建二萬五千餘詞之詞檔及詞檔索引，儲於另一磁碟。此二檔可同時交互使用，亦可單獨作業，爲便利用者，更於使用手册後附「音序詞典」[6]，及「注音符號檢字表」。[7]

要使用此輸入系統，使用者必得知道各字之讀法及其正確的注音符號。因此不免常需查對手册或國音字典。但中文「詞」的定義很難下，更不易蒐羅俱全。從「使用手册」所附之「音序詞典」中所列各詞，我們可以很清晰地看出，設計者也爲「詞」之定義而困惑，爲蒐列而煩惱；例如：

頁二：二三九中列有下列各詞：

臃（但未列「臃腫」等）

雍（但未列「雍容、雍容華貴、雍和、雍熙」等）

用器畫（但未列「用手畫、用毛筆畫」等）

運動神經（但未列「運動、運動精神」等）

頁二：九七中列有下列各詞：

好鳥（但未列「好馬、好狗」等）

好小哥（但未列「好小姐、好大哥」等）

好大臉（但未列「好大手、好大頭」等）

好極（但未列「好極了」）

好事不出門惡事傳千里（爲何不分爲二詞？四詞？）

由以上數例，可見「詞之長度可從一字到十字，倒底甚麼才是「詞」？

以此系統整體來看，可得下列推論：

一、若手邊字是熟識的，輸入速度則快；否，則慢。我們似乎應該要求學識淵博之人來做輸入工作。

二、每字注音代碼若長度一致，則檢錯改誤之程序簡易。但因中文字數繁多，讀音準確不易，常會導致含混不清之注音代碼。

三、若限用於某些專門學術領域或特別用途，此系統可能會經濟而實用，但擴大使用必生困難。

例二：

交通大學的謝清俊先生及其同仁，據字形及構造，於一九七二年設計了一個中文輸入輸出系統。[8] 他們將中文字以字根之形狀及位置，略分六類：

一、包含根：有⿴ ⿱ ⿰ ⿸ ⿵ ⿶ ⿷諸形。

二、左橫根：如扌、彳、系等。

三、右橫根：如攵、頁、欠等。

四、上直根：如忄、艹、宀等。

五、下直根：如氏、电、儿等。

六、中根：如口、日、田等。

經分析後，共得字根四九六個（內含常用整體字三〇五個）。編成二表；一

依使用頻率爲序，一依筆劃多寡爲序。設計者更提出「字根分析邊際效用原則」（Criterion of marginal utility for analysis of characters）將本可分解的最常用字廿五個及罕用字根廿三個列於「字根表」之下端，以利用者。

輸入時，用者可以四個「定位符號」（(）、⏃、⏄、△△）表示字根之位置，如：

休：亻△△木

類：米⏄犬⏃頁

盟：（日⏃月）⏄皿

葡：艹⏄勹⏃甫

更有四個「方便符號」（ꕥ、∞、⚮、8）以表示重複出現之子根，如：

品：口ꕥ

林：木∞

川：丨∞

炎：火8

據設計者宣稱，在現有四九九〇五字中[9]，此系統可包括四八七一三字。其餘一一九二字，不是已死之字，便是他字之變體，不會導致嚴重後果。

如果我們仔細觀察所列之「中文字根表」，我們可以發現一些令人不明之處。例如：「亠、乂、冂、一、口、丷、尸、弋、止、厂」等均列爲字根，而「文、方、立、亢、言、尺、武、足、才」等，亦被列爲字根。很顯然，這些並列字根有重複處。可能是設計者故意如此，以便增進輸入速度，便利工作。但我們不得不提出二個基本疑問：

一、這假想的利便，在實際工作中是否屬實？

二、這「字根表」究竟應該並列或重複到何種程度？

如果一個輸入輸出系統以基本筆劃爲基礎，[10]使用者只需記憶少數基本筆劃。選擇簡單，輸入工具之鍵盤亦較易設計。使用人更無需高深的中文知識。但這種系統之代碼通常較長，速度亦可能較慢。但設計時若稍用心思，較長的代碼卻可提供檢錯誤的機械性。這是謝先生等所設計的系統所無法做到的。

例三：

一九六五年至一九六九年間，筆者曾擬定了一個中文代碼輸入系統，[11]並設計了一個與其他系統作測效比較的統計實驗。[12]此代碼輸入系統很巧合地與謝先生等所設計的系統很相似。因獨力設計，個人智、力皆有限，缺點漏洞仍多，距「理想」仍然遙遠，但供「同好」者參考，也許尚能收「他山之石」之效。此系統之擬定源於下列二項假設：

一、中文字皆由一組或一組以上之「筆劃組合」（Component）結合而成。是以中文字便可以其「筆劃組合」結合型態而分類。

二、「筆劃組合，顧名思義，可知係由一個或一個以上「基本筆劃」（Basic Graphic Element）組合而成。一個「組合」若係由一個以上「基本筆劃」組合而成者，其中之「基本筆劃」必須彼此相連；否則，便形成不止一個「筆劃組合」。「基本筆劃」數目有限（暫定七十二），若能賦以適當規則，操作此系統應不是難事。

根據「筆劃組合」之結合型態，中文字約可分爲四種主要型式：

一、橫列型有⿲（代碼ＨＡ）、⿳或⿴等（代碼ＨＢ或ＨＢ4）等、⿱（代碼ＨＣ）、⿱（代碼ＨＤ）、⿱（代碼ＨＥ）等分型。

二、豎列型有⿰（代碼ＶＡ），⿰或⿰等（代碼ＶＢ或ＶＢ４等），⿰（代碼ＶＣ）、⿰（代碼ＶＤ）等分型。

三、包含型有⿵（代碼ＢＡ）、⿶（代碼ＢＢ）、⿷（代碼ＢＣ）、⿸（代碼ＢＤ）、⿹（代碼ＢＥ）、⿺（代碼ＢＦ）、⿻（代碼ＢＧ）、⿼（代碼ＢＨ）、⿽（代碼ＢＩ）、⿾（代碼ＢＪ）、⿿（代碼ＢＫ）、㇀（碼ＢＬ）、㇁（代碼ＢＭ）、㇂（代碼ＢＮ）、㇃（代碼ＢＯ）等分型。

（因此型字代碼通常較長，爲省時起見，取最常見之「外包」爲各分型之標準「外包」；譯碼時，「標準外包」無須另予代碼，有分型代碼便可。茲列各「標準外包」如次：

BA：㇄
BB：㇅
BC：㇆
BD：㇇
BE：㇈
BF：㇉
BG：㇊
BH：㇋
BI：㇌
BJ：㇍
BK：㇎
BL：㇏
BM：㇐
BN：㇑
BO：㇒

四、不可分型字其實與「筆劃組合」一般，所含「基本筆劃」互相牽連貫串

不易分解。此型字有四分型：

橫連型（ＩＨ）如：水，我等。

貫穿型（ＩＳ）如：册，車，申等。

半穿型（ＩＳＨ）如：甲，由，果等。

豎連型（ＩＶ）如：丁，土，王等。

筆者個人愚見，中文字之「基本筆劃」共有七十二，各以二數（0至9）爲代碼。「基本筆劃表」（Table of Basic Graphic Elements）以行列式法排列；表上端橫列「主碼」（Key Code）（0至9），表左手豎列「副碼」（Auxiliary Code）（0至9）。大致說來，筆形之代號與王氏四角號碼近似，以便記憶。作業時，主碼先譯。譯碼作業規則暫擬如次：

一、「字型代碼」（Pattern Code）後加引號（：）。

二、「筆劃組合」前加此組合之「構型代碼」（Structural Symbol）如ＨＡ、ＶＡ等。

三、譯碼時，由上至下，從外至內。數筆同高時，自左至右。

四、關係符號（Relationship Signs）：「筆劃組合」（Components）間以斜線（／）分之。再以橫線（—），小數點（·），句點（，），分號（；），分隔「筆劃組合」之「分部」（Sub-Components）及「分分部」（Sub-subcomponents）等。更以「重複號」（〃）代表重複筆劃，或「筆劃組合」。例如：「品」字應譯碼爲VD：60／〃／〃而非UD：60/60/60。

五、「包含型字」譯碼，先譯外包，後譯內涵。若外包爲二層或二層以上者，以最外或最主要之外包爲準。包含型之「筆劃組合」，譯碼時除無須加引號（：）於代碼後外，其餘作業與包含型「字」同。

六、不可分型「字」或「筆劃組合」，無須關係符號。但對下列事須特別小心：筆劃相連（或互貫）而不構成另一基本筆劃者，應按規則三將各筆劃」分別譯碼。否則，應視爲新而較複之筆劃。

此系統缺點仍多，尚待研究改進。但「關係符號」之引進，雖使代碼有時過長，因「構型代碼」確定了「分部」之數目，而「關係符號」之數永比其「分部」數少一，卻無形中也引進了「自動檢錯及改錯」之可能。此「累贅」（Redundancy）之引進，是好，是壞，尚未可知。若在實際作業時時發現譯碼過份冗長，而此「自動檢錯及改錯」系統並無大用時，不妨參照謝清俊先生等設計之系統改正之。

三、效果之衡量

爲求得一公允比較各系統之法，筆者於一九六八年冬，曾試行以統計實驗法比較了當時三個系統之優劣。此比較法尙有其可取之處，是以略介於次，以供參考。

因所擬納入此比較試驗中之系統有「三」，我們可選「三」個「具有同樣中文程度之人」[13]，作爲「被試驗者」（Subjects or Experimentees）。分整個試驗爲三「階段」（Periods），在各「階段」內，不同之「被試驗者」以不同的「順序」（Order）學習這三個系統之作業程序及規則。每習一系統，測驗一次。測驗題共分九組，每組有三類字（易、較難、難），每類有字若干。此九組測驗題（Question Package）亦以不同順序給與被試驗者作爲測驗。所有各種不同「順序」之制訂，均係根後列「模式三算術」(Modulo Three Arithmetic）公式（以012爲三個可能之層次）演算而得：

設：

a=Subject　　b=Period　　c=Sub-period

d=System　　e=Test administration order

f=test block　　g=Test sub-block

（fg together form Question Package 測驗題組）

則得公式：

$d=a+b+c \pmod 3$

$f=a+2b+2c+e \pmod 3$

$g=a+2b+e \pmod 3$

以上列公式可得下列三個「試驗程序表」：

程序表	階段	分期	系統	測驗題組及其測驗順序		
I	一	1	1	1	5	9
		2	2	7	2	6
		3	3	4	8	3
	二	1	2	9	1	5
		2	3	6	7	2
		3	1	3	4	8
	三	1	3	5	9	1
		2	1	2	6	7
		3	2	8	3	4

Ⅱ	一	1	2	5	9	1
		2	3	2	6	7
		3	1	8	3	4
	二	1	3	1	5	9
		2	1	7	2	6
		3	2	4	8	3
	三	1	1	9	1	5
		2	2	6	7	2
		3	3	3	4	8
Ⅲ	一	1	3	9	1	5
		2	1	6	7	2
		3	2	3	4	8
	二	1	1	5	9	1
		2	2	2	6	7
		3	3	8	3	4
	三	1	2	1	5	9
		2	3	7	2	6
		3	1	4	8	3

被試驗者可自此三程序表中任擇其一。

這種「試驗設計」(Experimental Design)，可得高度平衡，且易得對各系統較公允之評估。因在每一被試驗者身上可獲八十一個「觀察結果」(Observation Results)，被試驗者人數雖少，並不妨礙事後分析之正確性。又因已將測驗題分爲易、較難、難三類字，更可試驗出各系統對每類字之不同特性。

一個輸入輸出系統之優劣，通常應以操作速度及譯碼之準確性而定。故試驗時間必須正確控制，而速度及準確度評分應分別登錄。爲避免監試者個人之愛好或習性影響及被試驗者之成績，試驗時期內應不准發問。

試驗成績分別登錄後，可用「差異分析法」（Analysis of Variance）來鑑定此試驗中之主要因素（main factors）（a、b、c、d、e）以及各因素間之「交互作用」（Interactions）。或以所得結果之「平方和」（Sum of Squares）爲Y軸，「正常比例之對數」(Logarithm of the Normal Proportions)爲X軸，用座標法將各值標示，然後畫一直線穿過各值密集之中心。遠離此線之各值，便可能是具有影響性的「因素」或「交互作用」（Probable Influential Effects）。我們可將某些影響重大而非我們所關心的因素設法略去，便可更清晰地觀察我們所欲觀察的現象。

一般說來，「被試驗者」（Subject）及「試驗執行順序」(Test Administration Order）對試驗結果多具重大影響，且非我們所欲觀察之對象。爲便於觀察不同系統之成效，我們可以後列公式將之略去：

設：　O =Observation
　　M =Grand Mean
　　Es=Subject Effect
　　Et=Test Administration Order Effect

而　$Es=\overline{S}-M$

　　$Et=\overline{T}-M$

則　Residual=O—M—Es—Et

如此，我們便可將三個試驗階段中各系統之表現成績（速度及準確性）分別計算出來。再用座標法，便可清晰地顯示出各系統在不同試驗階段中進步（或退步）的情況。

這個統計試驗法雖是用以比較「三」個不同系統之法，但以同樣原則，略將公式變化，便可應用於比較任何數目之系統。

四、結　語

有人曾建議將中文拉丁化，以有限字母設計代碼，中文輸入輸出問題便迎刃而解。名文字學家約翰·第·弗蘭西斯（John De Francis）曾說：「中文若能以音符代替，任何中國人便能寫出他能說的話，便能看懂他能聽懂的話。對於『甚麼可以拉丁化？』這個問題的答案是『幾乎沒有甚麼不可以的。』[14]對此種看法，我們有下列疑問：

一、如何解決同音字問題？

二、如何解決方言問題？

三、若以「詞」爲基本單位，如何決定甚麼是「詞」？又如何建立一個「完整」的「詞檔」？

四、如何省去檢索程序？（若遇見不識之字便如何？）

中國幅員廣袤，方言種類繁多，加以數千年來積累之文字紀錄，即使中文拉丁化可行，也絕非一朝一夕之功。我們知道，若要將中文拉丁化，必先將「口語」標準化；國人民性保守，推行了數十年之國語，尚未能完全普及；倘若强制執行，必遭强烈反抗。是以，中文拉丁化之實現，尚有待時日。在此之前，全盤中文之輸入輸出，似乎只有以中文字形態及構造以譯碼代之一途。但以中文字形爲本以設計輸入輸出系統，我們也有下列疑問：

一、如何選擇「字根」或「基本筆劃」，以使學習容易、作業單純、代碼簡短？

二、如何建立「自動檢錯改錯」程序，而不過份增長代碼？

三、如何保證由「代碼」歸還原「字」之可行性，而不使譯碼程序過份複雜？

在設計一個「全盤」的中文輸入輸出系統時，不論以「音」爲本，或是以「形」爲源，都應將上列七個疑問常掛心頭。

〔附　註〕

1.墨子尚同上第十一。

2.謝清俊等，中文字根之分析，原載：中央研究院數學研究所中文輸入輸出系統參考資料第一集，p.108.

3.John De Francis, *Nationalism and Language Reform in China* (Princeton, New Jersey: Princeton University Press, 1950) p.158.

4.Nelson L. S. Chou, "A New Alphameric Code for Chinese Ideographs, its evaluation and applications." (Ph. D. Dissertation, Graduate Library School, University of Chicago, 1972) p.2—3.

5.張系國等，「中音序檢字系統」，原載：中央研究院數學研究所中文輸入輸出系統參考資料第二集。

6.見前，2：1—2：240.

7.見前，3：1—3：10.

8.謝清俊等，p.108—117.

9.張其昀編，中文大辭典。

10.Chou, p.33.

11.見前， p. 20—39.

12.見前， p. 47—72.

13.因中文程度之高低很難測定，不妨選三個同等學歷不識中文之外國人。因彼等中文程度皆爲「零」也。

14.De Erancis, p. 190—191.