中文輸入法大勢評議

（兼及漢字組件化逆波蘭表達式）

澳門・梁崇烈

第三屆兩岸四地中文數字化合作論壇

專題演講

2005 年 12 月

sonnet@shuowen.net

摘 要

一種語言能否簡便地輸入電腦，又要效率高，又要全民皆懂，已成了使用該語言的國家能否提升國力的重要因素。大鍵盤並不能解決中文二維的先天特性，而語音和手寫輸入則效率低下，因此現時主流是基於鍵盤的三大派：拼形派、拼音派和形音派。如要落實到和英文輸入並駕齊驅，必將走向拼形派一條路上去。

英文輸入有兩大特點：簡易和快速。中文若歸納成十餘種筆劃分類，則其簡易可與英文相當，但效率則瞠乎其後，而且受到兩岸四地筆順規則不統一的困擾；若以部件為輸入基礎，則涉及大量部件的記憶，以及部件拆分規則的不易掌握。

舊式輸入法最受人詬病的是割裂漢字、違反筆順和錯配部件。中國人自小便花費大量時間和心力在學習漢字上，甚至影響了其他領域如科學技術的學習，這受到漢字先天特點的影響；但如果漢字教學能配合一種先進的中文輸入法，像砌積木般將漢字層層組合起來，那麼漢字教學、電腦輸入以至排序索引不難相輔相成。

要中文輸入普及到小學教育中，目前「萬碼奔騰」的現象必須作一定程度的「規範」，但在自由市場的社會中，人為的限制難免歸於失敗，這便使中文輸入的發展陷入進退兩難之中。完全的放任和嚴格的限制既是各走極端，由兩岸四地共同制訂一套「指引」，是中文數字化論壇與會同人應當努力的方向。

在現今電腦時代，從科學研究上太空，到電子商務和電子政務，到小學生上網聊天，電腦已到了無處不在的地步。一種語言能否簡便地輸入電腦，又要求輸入的效率高，又要求老少咸宜全民皆懂，已成了使用該語言的國家能否提升國力的重要因素。

壹．四大文明古國之一的四大發明之一

我國是四大文明古國之一，而中文(漢字)是現今世界僅存而廣泛流通的最古老文字。由於它二維的先天特性(意謂中文屬形系文字，其要表達的意思須在平面上的二維空間組合各種筆形而成)，不同於拉丁拼音文字例如英文的為一維線性文字，因此縱使中國書法藝術是如何的優美，中文表達的意象是如何的豐富，遇上由西方人發明的電腦便顯得有些格格不入。

其實所謂一維二維是相對的。每一個漢字的內部結構固然是二維，但如果將中文輸入的單位設定成一個個漢字，利用一個包含幾千字的大鍵盤，由工人逐個字找出來，那麼中文也可算是一維線性的。但問題是：第一、中文的字量有數萬之多，鍵盤無論多大也難以包容；第二、要做一個大鍵盤成本很高；而第三、最重要的是，在如此大的鍵盤中要找出想要的漢字，不要說「老少咸宜」，就是專業的工人也不容易做到，所以它的輸入效率極低。不過，自從北宋畢昇發明活字排版印刷術之後，一直到後來印刷廠的排字，以至中文打字機，其實我們已利用這種概念來輸入中文有近一千年的歷史。我們一直引以為傲的中國古代四大發明(指南針、火藥、造紙、印刷術)之一，現在竟然是我國現代化的其中一個難題！

話倒要說回來，我們不能因為古代的事物不能解決現今世界的需求，而怪罪到這種事物上去，更不應怪罪到它的發明人身上。要怪只能怪後代的人沒有日新又新，與時並進，將古人的智慧成果發揚光大，發展出新的科學技術。大鍵盤輸入，已經發揮過它的光輝，應該到了功成身退的榮休階段了。

貳．中文輸入的主流

電腦自從上世紀七十年代末，脫離了初期整個房間那麼大的型態，而踏入微型電腦、個人電腦的階段。事實上也不能說中文落後，在三數年間，中文輸入法亦步亦趨，也進入了微型電腦世界。其中最突出的，莫如朱邦復先生針對繁體字推出了倉頡輸入法，以及王永民先生針對簡體字推出了五筆輸入法。從此百花齊放，萬「碼」奔騰。再加上手寫輸入、語音輸入等，國人都在為中文電腦努力拼搏。中文之是否能用於電腦，在乎我們有沒有努力想方設法發揮電腦的功用和中文互相配合，而不是本末倒置地為了要使用電腦而妄想將漢字拉丁化去遷就它。

先說輸入方法的新產物——語音輸入，它是最自然的輸入方法，卻又是最不自然的。請大家細想，我們平常寫文章時是不是「我手寫我口」的？不是。第一、我們思考的時候絕少發出聲音，現在卻要將想到的東西讀出來，其中有多少是思考時尚未成熟的糟粕？第二、思想在未成熟的時候仍然是屬於自己一個人的，現在過早地朗讀，隱私不受保障。第

三、發出來的聲音會互相干擾，一般辦公室很難找到一個獨立的空間。因此語音輸入只是新產物，但不是新方向，更不是新趨勢。至於同音字太多，辨識率太低，隨著科技的發達加上人工智能的應用，語音輸入的效率將會得到改善，但問題絕不會完全解決。

至於手寫輸入，目前最大的問題也是辨識率太低，經常會認錯字或者無論怎樣小心寫都找不到要輸入的字，但和語音輸入一樣，會隨著科技的發展而改善。但另一個更長遠的問題是：手寫輸入效率太低，不能從手寫的速度中解放出來，如欲提高辨識率就必須慢慢寫，其速度有可能比用筆寫在紙上更不如。以使用電腦的目的在於改善人手操作速度低的立場來說，手寫輸入只能作為某些人的輔助，要用它來作輸入方法的主流是不可能的。

因此，在可見的未來，鍵盤將仍然會是中文輸入的主流，而本文亦會較著重這方面來討論。

叁・鍵盤輸入三大派

要評議中文輸入法大勢，不能閉門造車，到圖書館找資料也沒有用，必須真實去看看這個世界！從互聯網上的討論可見鍵盤輸入有三派：拼形派、拼音派、形音派。

拼音派說拼音最投合人類思維方式，會拼音的人完全不用學；形音派說兼顧字形和語音集兩家之所長……等等。不過，主流意見始終認為：論效率，仍以拼形為佳。

中國方言多，南方人多不會普通話；普通話聲韻組合理論上只有 417 個編碼空間，拼音碼難免同碼字多不勝數；尤其當使用詞組輸入，只採取每字首碼(主要是聲母)時，只有約二十二個(拉丁字母 26 個，減去漢語不用的 V，再減去不會用在拼音首碼的 I 和 U，合共 23 個，而 E 和 O 雖然會用在首碼，但是很少)，如何能夠組合成數十萬不同詞組而避免重碼過多？

至於形音派聲稱集兩家之長，但其實亦同時集兩家之短；我認為還應加上：將兩類不同概念融合時做成思想的混亂這個缺點，兩面不討好，可謂之「三短」。

肆・拼形派

我們要輸入的是漢字而非漢語(口語)，用拼形輸入漢字，筆對筆，形對形，是最直接的。漢語以至任何口頭語言本身就有「重碼」(主要指同音字，甚至同音詞也很多)，所以據此編出來的拼音碼就先天不可避免有重碼；漢字則不同，雖然近年統計顯示漢字多達九萬，甚至是一個無限的集合，但全部漢字本身都無「重碼」(意即沒有同形字)——如果有重碼就根本是同一個字，只是一字多義罷了。因此，理論上拼形輸入法也可以做到無重碼，問題只是編碼可能太長不是我們大多數人所接受而已。關於這點下面還會探討。

拼形碼可以用足整個鍵盤(雖然一些偏遠的鍵，例如「１」和「＝」，應該排除不用作中文編碼，以符合人體工程學，增加輸入效率並避免疲勞)，而且可以平均分配，不像拼音的聲母偏向一些鍵而韻母又偏向另一些鍵，某些常用的讀音用得太多而某些又用得太少，所以拼形碼能夠做成的「編碼空間」比起拼音碼大很多倍，重碼很少，用在詞組輸入時這個特徵就更明顯了。

伍・數字鍵盤

編碼單位(鍵數)的多少，直接影響一個輸入法的好壞(前提當然是使用一般通用鍵盤而非大鍵盤，而且要符合人體工程學)。只用數字鍵的輸入法，就算同樣是拼形輸入法，且不論其他因素的設計好不好，只從鍵數說，由於組合空間小，如果不做成重碼多，便必定要令編碼加長。當然，由於鍵數少，學習曲線又可能會比較平緩，表面上似乎是有利有弊；但是好的數字鍵盤輸入法可以做到易學易用，但絕不能做到效率高超。相反的，利用主鍵盤(包括字母鍵、頂排的數字鍵和其他標點符號鍵)的輸入法可能學習時要記憶的鍵多些，但反過來也可以說每個鍵要負擔的字形比較少而較容易記，而且從鍵盤上每鍵印刻著的符號的聯想，可以幫助學習者記憶。

此外，使用字母鍵盤，十指如飛，當第一鍵未打完時，另一隻手指便開始為第二鍵做準備動作，因此縮短了兩鍵之間的時間差距，這又是能夠高速輸入的因素之一。因此，有些輸入法的宣傳說「只用一隻手便可輸入」，其實這句話的背面是「沒有利用兩隻手來輸入」。

總括來說，使用字母鍵盤的拼形輸入法有可能做到「易學易用，效率高超」(當然要輸入法本身設計得好才成)，而其他各種輸入方法則最多只能做到「易學易用」，而絕不能做到「效率高超」。

還有一點要注意的是：很多只用數字鍵盤的輸入法，其所以能夠做到易學易用，甚至可以號稱「三分鐘學識」，實在不是因為它的編碼本身很容易記，而只是因為它將字的筆形分成九類，讓用戶選擇了第一個筆形之後再列出這個筆形的九個子類，如是者一重一重地帶到要選的字——說它容易的確很容易，但付出的代價除了輸入速度很低之外，還會對視力做成很大傷害。這比拼音輸入更有所不如，因為拼音雖然有很多重碼，但用戶只須每一個字選一次，而不像這些輸入法要每碼都選。這點對保護小孩子的視力，尤其重要。

我不完全否定數字鍵盤的作用，尤其在現今手提電話、ＰＤＡ等逐漸流行的趨勢之下。我只是說，在絕大多數的情況之下，「效率高超」仍是輸入法的兩大目標的其中一個，而這個目標就只有可能用你的「雙手」在字母鍵盤上開創。雖然據科學研究，現代人的大拇指因為使用多了，比上一代人靈活，但總沒有人期望「拇指如飛」在手提電話上每分鐘輸入一百個字吧！要強調的是，這裏不是針對某些輸入法，而只是從整體上、理論上去分析，當然不能排除某某數字鍵盤輸入法比起某某字母鍵盤輸入法更快更好用。

陸・對重碼的看法

從我所蒐集的談論漢字輸入的文章中，讀到很多和我的想法不謀而合的理論(因為這是多年來在網上搜尋有關資料的結果，所以恕未能一一列出資料來源)：例如漢字雖然有數萬之多，實在無須為了重碼這個問題過於刻意遷就，而破壞編碼的理論體系或令編碼過長。我看的多數是大陸的理論，他們最照顧的是國標碼 GB 2312-80 的 6,763 個字，而國標擴展碼 GBK(GB 13000)裏的 20,902 個字以至 GB 18030 裏的 27,484 個字中的其他字大部分為繁體字，在中國大陸無須要求首碼出字。

照顧兩岸四地需要，繁簡兼備是我的目標，繁體簡體中文同樣重要已是既成事實，現代中國人必須「兩條腿走路」。但縱使如此，我們也不用強求將二萬多漢字做到完全無重

碼。「一視同仁」是「假公平」。好的編碼應該是照顧常用的字而不忽略罕用的字，但罕用字卻不一定要首碼輸出，在必要時從候選字中選擇是可以接受的。

好的輸入法比其他輸入法更進一步的，是不單止把字分成常用字、通用字、罕用字、備用字等幾級，而且編碼所用的字表，根本就完全按字頻排列。除了少數為了照顧同部類推做成一系列編碼之外，基本上排得越前就越優先，就越有條件享用較短的編碼。

柒・部件的多寡

部件輸入法有一個特點，就是部件的數量較多。但是有失必有得，如果中國人不想逗留在每分鐘十餘字的階段而想達致英文打字的輸入效率，就無可避免總是要花點精神去記部件。理想的輸入法的部件由鍵盤的字形一重一重地帶出，要盡量利用聯想。

由於中文輸入就如學寫字一樣，將會影響一生一世，因此花幾個鐘頭學一種高效的輸入法是完全值得的。

我在網上看到有人提出中文就是中文，何必要借助英文字母？其實，鍵盤就是鍵盤，ＡＢＣ也只是一些符號，無須因為大中國心理而認為利用這些符號就是崇洋。世界變得越來越小，國和國之間的溝通越來越頻繁。打字機畢竟是外國人發明的，但現在已流通全世界。我反對放棄現成的「ＱＷＥＲＴＹ」鍵盤而另外開發新的，也反對明明使用英文鍵盤卻忽略現成印在鍵上的符號，而要為每一個鍵賦與完全不同的意義，這將不利於聯想學習。但網上有人舉例說：「小」字橫看豎看都看不出有什麼地方像「W」，因而指出利用鍵盤上的字形純屬主觀想象。網上已經有人回答說：「小」字看來「像」「W」並不等如說「小」字看來「是」「W」。任何聯想都可以幫助記憶，縱使這種聯想對別人來說可能很荒謬，但只要你自己覺得有意思就成了。

《康熙字典》有 216 個部首，但那是部首，不是部件，不能涵蓋所有漢字的不同組成部分。《說文解字》有 540 個部首，那是兩千年前的事，不一定適合今日。所以我在網上搜尋人家的輸入法究竟用了幾個部件，以便研究部件的數量如何減到最少。

捌・先入為主

繁體字世界之有倉頡輸入法，猶如簡體字世界之有五筆輸入法。五筆名似筆劃輸入法，其實和倉頡一樣都是部件輸入法。它的出現稍後於倉頡，但仍屬輸入法的前驅，為王永民先生在一九八二年發明。它在國內給人一種「學中文電腦就是學五筆輸入法」的印象，和倉頡在臺港澳的情況差不多。有人說它也可以打得很快，但絕對可以肯定，它很難學(比較起大多數「後起之秀」來說)，然而它卻竟然流行了，原因只是：它出現得早！有人甚至(可能過激地)說這是中文電腦的「浩劫」，並指出，任何不合理的編碼在經過多番練習後都可以打得很快，但並不表示它好。對於這點，我也同意人的「惰性」(安於現狀不願意改變已熟悉的處事方式的特性)是很強的，「先入」自然就可以「為主」。儘管他們做成的結果可能很壞——學會了要跳出來已經很難，甚至幾乎不可能，但用不用由你，不能把責任怪罪到發明人身上。畢竟他們作為開路先鋒已經是一種很大的貢獻。

我還想指出兩點：第一、已經熟習舊輸入法的人固然難以跳出來，但何忍要下一代步自己的後塵？因此我最反對學校開班教舊式輸入法，這可以用「貽害子孫」四個字來形容。

第二、就算是已熟習舊輸入法的人，是不是真的就不可能跳出來呢？我自己是過來人，在用了舊式輸入法超過十年之後，自問已經到達「中等以上水平」，卻還要經常查「輸入法字典」。最後下定決心，還不是跳了出來？所以要視乎用和學的人的決心，以及新的輸入法是不是就真的很好，好到有足夠動機吸引原有一批用戶改變習慣？

玖・國家規範

於是我首先要研究的是，如何減少部件數量？卻不料引出一個「規範」的問題來。原來正因為五筆難學難用，並且用熟之後再難以跳出來，中國國家語言文字工作委員會(國家語委)、國家技術監督局、教育部等機構，已經察覺到要達到中文電腦全民普及，輸入法問題已不能迴避，所以首先以政府的力量召集一些專家研製一個「最理想」的輸入法——認知碼。可是結果挾著「政府全力支持」的後盾，認知碼仍未能成功。究其原因，固然如前述的惰性問題，但更重要的是，認知碼是不是已經夠好？是不是在「有中國特色的社會主義」之下，任何事情都可由政府主導而忽略人民的自由選擇？

認知碼雖然不成功，但五筆的問題除了阻礙中文電腦全民普及之外，還侵入了非電腦的領域——教育問題。五筆最大的「敗筆」是：將漢字任意割裂，有人甚至形容為「肢解」。他們的文章雖然不能代表全部中國人的意見，但至少可以看出他們對有關問題的誠意，才會夾敘夾議，長篇大論。

拾・割裂漢字

是不是將漢字割裂這個問題，五筆和倉頡的情形非常類似。這引出了支持和反對兩派。支持的人說，輸入法就是為了輸入漢字，它可以使用能夠達到目的的一切手段，「不管黑貓白貓，會抓老鼠的就是好貓！」反對的人說，中文輸入法不單是為了輸入，而應該將輸入、排序索引和教育互相關連。

所謂割裂漢字，且讓我以倉頡為例子來解釋：「目」字的倉頡碼是「月山」，「月」沒問題，但為什麼「山」呢？卻原來倉頡輸入法不理會筆劃的編排，它要將兩條垂直線在最下面的兩條水平線之間割開，才能做成這個「山」(向上開口的兜)的形狀。網上有很多人稱這種分割方法為「剪刀式」。一些人認為並無不妥，並且認為它「乾淨利落」，甚為欣賞！但畢竟絕大多數人認為這種割裂漢字的做法不合中國人從小累積的中文知識。這不單在輸入法本身做成學習的困難，以及不用一段時間很快就會忘記，更有甚者，當用者輸入中文的時候，腦海中不停地以倉頡的思考方式來思考，和日常寫字的方式脫節，除了增加小孩子學習漢字的困難之外，還做成思想的混亂，未能達到「流水行雲」，順乎自然的效果。

拾壹・違反筆順

舊式輸入法除了割裂漢字之外，另一個不符合國人寫字習慣、容易做成困擾的問題是筆順。再以倉頡的「兩」字為例，它的編碼是「一中月人」，除了省略了其中一個「入」

之外，這個編碼的每一碼基本上都很容易找到筆形配對，問題是編碼的次序，為什麼不是「一月中人」？原來倉頡輸入法沒有針對中國人來設計，就算是外國人，只須拿起一把直尺，稍微傾斜，從左上角掃描至右下角，先遇到哪一個部件便先用它來編碼，因此最先掃到的是橫劃，其次便是直豎了。可是到頭來，這不但幫不了外國人輸入中文，而中國人自己反而歧路亡羊，不知何去何從！這教我們如何向孩子們交代？

漢字的寫法是中國人從小付出很大努力才學會的，輸入法應該善用這種知識而不是放棄它來另學一套。好的輸入法就如寫字一樣，只不過不用筆而用鍵盤而已。

由於海峽兩岸的交流日益頻繁，曾因各自發展而引起漢字寫法分歧的問題逐漸顯現，不利於漢字的輸入、索引和教學。關於這點，詳見拙文《筆順規則的兩難》(http://shuowen.net/BiShunGuiZe.htm)。

拾貳・錯配部件

說倉頡沒有將用者當中國人看，還有一個論據。

我記得小時學寫中文，老師在黑板一筆一劃地寫，要同學們一筆一劃地跟著唸，例如——就以「如」字做例吧，「一折，一撇，一橫，一豎，一曲，一橫」。根據這個概念的輸入法便是符合中國人學習方法的輸入法了，便是好的輸入法了？——錯了！問題在於，這種教漢字的方法已不合新的研究結果和新時代的要求。

漢字難，一直是外國人所認為的。其實連中國人自己也是這樣想：我們從小到大，花了多少時間學中文？如果把所花的時間用在學習科學技術方面中國不是更富強嗎？因此才有文字「改革」委員會不顧一切將漢字大量簡化，因此才曾經有人提出廢棄漢字，改用拼音文字。後一種想法已經被歷史證明是錯的了，在此不再贅述；就是前一種想法，是不是漢字的筆劃越少就越容易學呢？這是沒有科學根據的。我們學漢字，不應是一筆一劃地學的(那才是筆劃越少越容易學)，而應該是一個個部件地學：由小部件組成簡單的漢字，或者小部件本身就是一個漢字；再由簡單的漢字組成更複雜的漢字。這種樹狀結構更符合人類的思維方式，因此才更容易學習。舉一個例來說：當你要告訴人「想」字怎麼寫時，你不應說「一橫，一豎，一撇……」，而應該說「想」字是由「相」字下加一個「心」字組成(如果你能解釋古人錯以為心是用來思想的，所以採用「心」來作部首，而它的讀音是從「相」字來的，那就更好了)。那麼「相」字又怎樣寫呢？「相」字是由「木」加「目」組成。那麼「木」字呢？因為「木」是基礎部件，應該在學字的最初階段就已教曉，那才是一筆一劃地教(如果你能解釋它是一個象形字，由一棵樹的形狀逐漸演變而成，那又是更好了)。

這些不同的部件，是幼兒認識一個陌生的漢字時的「抓著點」——抓著點越多，越容易學。現在漢字簡化了，抓著點少了，字和字之間的不同特徵被削弱了，因此就更難學和更難認。「广」(廣)和「厂」(廠)看來差不多，兩者的分別純屬強制規定，完全沒有理據，其難於學習由此可知，是一個很明顯的例子。

我們輸入漢字，在思想時也是一個個部件地想而不是一筆筆地想的。試以「順」字為例，它是由「川」和「頁」組成的，在理想的輸入法中，「川」作為部件取一個編碼，而「頁」作為另一個部件取另一碼，才符合中國人的思想方式的。可是外國人不知道「川」是一個部件，他們只看到字首(直撇)和字身分離，而字身取「頭頭尾」，所以「順」字的倉頡碼是「中中中金」，但這種部件錯配是違反中國人的思想方式的。

有關漢字組件化的問題，請參閱附錄〈漢字組件化逆波蘭表達式〉。

拾叁・字形損耗率

研究輸入法的人還提出一個理論，叫做「字形損耗率」。它的意思是：將一個字拆開成幾個組件，而將每一個組件賦予一個編碼，如果所採取的組件極小，例如小至最極端的筆劃，那麼一個字將會包含很多個組件，亦即其編碼會很長。為了減少編碼，一般輸入法會訂立一些規則，例如只取「頭三尾二」，中間的筆劃不要，這就造成了極大的字形損耗率，亦說明為什麼一個字取五碼之多而仍然有相當多的重碼字，因為字形損耗率越大，就越不能表現每一個字的特徵。

上舉倉頡輸入法的「順」字是一個例子：倉頡採用的是小部件(字根)，所以是眾多輸入法中除了按筆劃的輸入法外字形損耗率較高的，例如「順」字中「頁」的上及中沒有採納作編碼。反觀理想的輸入法將「順」分成「川」和「頁」兩個組件，將整個字涵蓋，因此字形損耗率為零，是重碼率較小的原因。

拾肆・部件拆分

主流意見認為割裂漢字是舊式輸入法的嚴重缺點，不但令輸入法本身難學難用，而且在輸入時要不斷思考如何將部件甚至筆劃割裂，極大地阻礙了寫文章的正常思路，以致效率低下。更有甚者，因為先入為主，投入了舊式輸入法就再難跳出來，結果一生一世受其牽累，再加上因為部分家長和校長、老師的無知，令本來猶如一張白紙的小學生接受舊式輸入法，於是將問題一直延續至下一代。割裂漢字不但損害到輸入法本身，而且由於這些輸入法和漢字結構背道而馳，違反了中文的造字和寫字規則，令人無所適從。影響所及，不但中小學生的中文水平下降，甚至越來越多成年人出現執筆忘字的現象。根據報章登載，海峽兩岸越來越多人能夠在電腦上輸入中文，但要他們用筆書寫反而大感茫然。究其原因，一是拼音輸入法的流行，尤其是中國北方，人們只知其音不知其形，二是舊式拼形輸入法將漢字割裂，筆劃次序顛倒，令人再難記起字的正確寫法。

為了解決這個問題，1997 年 12 月 1 日中國國家語言文字工作委員會發佈了《信息處理用 GB13000.1 字符集漢字部件規範》(GF3001-1997，簡稱「漢字部件規範」)(可在 http://www.moe.edu.cn 搜尋「汉字部件规范」找到)。GB13000.1 字符集亦即是國標擴展碼 GBK 字集，涵蓋了簡體字和繁體字兩種，因此臺港澳雖然主要用繁體字，仍須相當重視這個規範的推出。

該規範的起草單位是北京語言文化大學、北京信息工程學院和上海交通大學，起草人都是有關方面的學者。他們將 20,902 個漢字逐個分析，將其中的「部件」歸納成 560 個。

由於將漢字拆分成部件就猶如拼音文字用字母標音一樣，對於中文的教學、索引的編排、電腦的輸入以至其他科學研究都有莫大的裨益。有了「部件拆分」，便不應再有人妄提放棄漢字而改用拼音文字了。因此，臺港亦掀起了對部件拆分的研究熱潮，例如臺灣中央研究院的資訊所和語言所便提出《漢字部件及組字規則》(http://www.ndap.org.tw/2_techreport/index.php?pid=215)，將 Big-5 的 13051 個繁體字以及一些「外字」，包括簡化字，分拆歸納得出 783 個部件。香港浸會大學也不甘後

人，特別針對識字教學，發表了《小學中文科常用字研究》(http://alphads10-2.hkbu.edu.hk/~lcprichi/)，提出 980 個部件！

拾伍・因減得加

我本來想參考人家的研究成果，看看如何減少輸入法的部件數量，但是卻適得其反，部件越來越多！

其中的原因，是研究者很多不是電腦從業人員或專門研究輸入法的，他們的目標也不全為了電腦輸入。無疑，漢字教學、排序檢索和電腦輸入是部件拆分的三大目的，但它們三者是不是一定可以同步呢？這是值得商榷的。

漢字本身就已是相當複雜的，經歷數千年的演變，有加繁的，有簡省的，有為了實用目的，有為了美觀目的，有要求書寫方便，有要求辨識容易，有隨著人民的智慧自然演變，也有某些人逞其私智而人工「改革」，於是形成了今天這個豐富多姿但卻不易學習的體系。作為教學的目的，能夠將千頭萬緒歸納出 980 個部件亦已不易。不過，話說回來，《小學中文科常用字研究》的對象是《小學中文科常用字表》，共 3000 個字，而歸納出來的部件竟然將近字數的三分之一，顯然未能發揮「以簡馭繁」的作用，它在教學方面的助益如何亦很值得研究者深思。

影響部件數量多寡的另一個因素是：我們要求的是平面拆分還是樹狀拆分？是否只計算基礎部件(又稱末級部件)還是要將合成部件(又稱複合部件)也計算在內？每一組只計算主形部件還是將附形部件也計算在內？這沒有絕對的標準。部件和非部件的界限是漸進式的，是模糊的，而不是非此即彼。一些筆劃的組合是否應該稱作部件有時可依據「有理據拆分」，但縱使是《漢字部件規範》的專家也須要容忍「無理據拆分」，因此，比《規範》中的部件小的筆劃群能不能升格成基礎部件，比《規範》中的部件大的部件群能不能認為是合成部件，既然沒有絕對標準，只能就中國人寫字的習慣去推想，但這種推想總少不免有很重的主觀成分。

拾陸・規範好還是自由發揮好

於是，《漢字部件規範》的推出，有譏有譽，有彈有讚。反對的人認為：在市場經濟之下，應該百花齊放，各展所長，由人民去汰弱留強，最後剩下來的輸入法便是好的，而《規範》卻箝制了人的思想，不能發揮最大的創意。有誰敢說研究《規範》的這幾位專家就可以代表國人的全部意見？如果他們的想法有偏差豈不是全國人甚至將來的子子孫孫都會受害？贊成的人認為：舊的輸入法正是在不受規範的情況下產生，雖然說時代不斷演變，不應用新的研究成果去指摘舊事物的不足，但客觀上卻的確有很多人陷入舊輸入法的窠臼中「不能自拔」。而且人的喜好各有不同，要用自然淘汰的方法選取最好的輸入法，那麼這種輸入法永遠都不會出現，萬碼奔騰就將永遠萬碼奔騰，這將不利於漢字教學、排序檢索和電腦輸入——沒有一種「標準」的輸入法，學校又怎能開班教授，而如果小學生不能儘早接近電腦，又談什麼電腦可以提升國力？要有一種統一的輸入法，大前提就是要有漢字部件規範。

這些紛爭的確不容易解決。有一位對輸入法很有興趣，就輸入法問題「著作等身」的中醫師知道有專家小組正在研究部件規範，他說本著「無私奉獻」的精神，興高采烈地寫

了很多信和將自己的著作寄給小組的袞袞諸公，但《規範》推出時發覺自己的意見沒有被採納，大感失望，在網上發表了很多文章指出這些專家的出身沒有電腦輸入法的背景……。我不懷疑這位中醫師的誠意，但他對《規範》的批評，卻未必中肯。說「自相矛盾」其實他自己也很多，例如既反對部件規範，另一方面他又贊成輸入法統一，殊不知沒有較溫和的部件規範，何來更極端的輸入法統一？

拾柒・規範標準和參考指引

除了上述的幾篇關於部件規範的文章之外，還有香港特別行政區政府資訊科技署與法定語文事務署合編並在二〇〇二年二月發佈的《香港電腦漢字楷體字形參考指引》和《香港電腦漢字宋體字形參考指引》(http://www.info.gov.hk/digital21/chi/structure/glyph.html)，參考了上述的文章和規範，歸納出 644 個部件，但專用於簡體字的部件沒有列入而由另表處理。

我很反對香港又訂一套漢字標準，例如「周」字和「告」字中間的直豎在大陸和臺灣都沒有穿出橫劃的，但香港的「標準」卻要穿出。這不但在字形上造成差異，書寫的筆順也有不同，引起學習和各種電腦化工作很大的混亂，就算在字源學方面有什麼理據，也是不應該的。但這可能是一些「歷史遺留下來的問題」，香港既不跟隨大陸使用簡體字，又不依附一個在政治上沒有關聯的臺灣，便造成這種不合理的現象，實在是無可奈何的。不過，隨著海峽兩岸的關係趨向融和，就如「兩岸四地中文數字化論壇」的展開，國人應該就「書同文」進行討論和研究。

《參考指引》依循 ISO/IEC 10646-1:2000 的國際標準，和商業上流通的 Unicode 標準一致，並且考慮《香港增補字符集——2001》的特殊需求，因此亦可算符合兩岸以及香港區的實際應用。只是《參考指引》根據的是香港教育學院編訂的《常用字字形表(二零零零年修訂本)》，其代表性頗有疑問，但是正如上述是一種政治現實，也就不能求全責備了。

因為《參考指引》後出，可以根據前人研究的成果而後出轉精，所以還是很有參考價值的。須要注意的是，它叫做「參考指引」而不叫做「規範」，我認為較符合實際情況。孔子到七十歲時才能做到「從心所欲，不踰矩」，可見規範的有和無之間實在是很難拿捏的。

大陸推出了規範，以後凡不和《規範》配合的將不易得到官方的認可，這對各種輸入法來說也不能說不是一個顧慮。但如果遇著規範中不合理或不能配合高效輸入法的要求的部分，又應該如何？這是研究發展輸入法時會遇到的兩難局面。

拾捌・尊重理據

國家語委的《漢字部件規範》〈１．１〉開宗明義指出：『本規範是根據漢字的構形規律、現行漢字的發展現實和漢字的歷史承襲性，採用「從形出發、尊重理據、立足現代、參考歷史」的原則制定的。』所謂「從形出發」，正如上面的分析，漢字始終是形系文字，在不排除《漢語拼音方案》對漢語的科學整理所起的貢獻之餘，從漢字的字形分析去整理出一定數量的部件，已是大家的共識。「尊重理據」是什麼意思呢？在〈３．１０〉的解釋是：『根據字源或參考字源，從漢字的部件組合中分析出的造字意圖稱「結構理據」。

例如：「旦」的理據是像太陽(日)從地平線(一)升起；「架」的理據是從「木」，「加」聲。』亦即「理據」指的是漢字的造字原理，包括(但並不止於)「六書」的理論。雖然時至今日，「日」字並不是圓的，已不能象太陽之形，而很多形聲字也不能完全表達每一個字的讀音，但總的歸類仍是有跡可尋的。而且正因為形聲字只表達「聲類」而非「聲值」，才可以在幾千年來統合關山分隔的各族人民使用相同的文字，漢字之功正不可沒，而亦正是我們中國人對於「字字有根，字字有解」一直所引以為傲的。但是上面提到的中醫師認為部件的拆分可以就字論字，就形拆分，要拆的就拆，不可拆的就不拆，完全可以不考慮「理據」，也無須「參考歷史」。

那麼，我們要明白，什麼叫做「漢字部件」？《規範》〈３．６〉：『由筆畫組成的具有組配漢字功能的構字單位。簡稱「部件」。例如：木、心、口、也、氵、亻、刂、礻。』假如我們只用字形去分析，那麼「木」是一個很整合的結構，例如「相」字中「木」和「目」就分別得很清楚，固然沒有什麼爭議，但是「心」又為什麼要算是一個部件呢？如果根據該中醫師的看法為什麼不把它當成四個部件？如果是這樣部件還有什麼意義？有些人就是忽略了漢字有三大要素：形、音、義，很多人只知用前兩者去分析漢字，而完全忘記了「義」的重要性。不單像「日」有它的象形意義，很多形聲字的聲符也表達一定的意義。就以「架」為例，吳潤儀《漢字詳解字典》解釋為：『形聲、會意字。從木加聲，指「用木料搭起的棚子」，故從木，又因「加」有「增加」之義，「架」是許多木料配搭而成，故架從加聲並會意。』當然，並不是所有形聲字的聲符都可以這樣解釋，我們絕不能望文生義，穿鑿附會，但多數形聲字的形符，亦即該字的義類，是絕大多數中國人的常識，而非獨文字學家方懂。就算本來不懂，難道就不可以當成一種教學時的聯想或提示？理想的輸入法正應利用這種知識幫助國人很快速地分析一個字的結構，亦因此才可以解釋為什麼小學生也可以接受「心」是一個部件而不是四個部件。因此，結論是，以「尊重理據」、「參考歷史」作為部件拆分的重要依據完全是有科學根據的。反之，如果只從形去拆分，以為任何人，甚至外國人都可以做到，表面上是顯淺易明，實際上將漢字拆分得支離破碎，失去了部件拆分的意義。

拾玖・拆分原則

在《規範》〈５〉列出「漢字基礎部件表使用規則」，看似是給各界人士使用《規範》時的「規則」，其實更應說是起草《規範》各先生們決定每一個漢字的各部分要拆還是不拆的「原則」。由於「原則」是死的，人是活的，漢字隨著人類的進化亦衍生出不同的形態也是活的，因此這些原則訂了下來雖然比沒有訂好，但很多原則是互相牽連而互相抵觸的，如何取捨只可以由人去決定，正如香港的《參考指引》是根據《規範》而來，而且因後出轉精，本來應該更成熟，但仍不免要指出：『以上原則無法兼顧時，酌情處理。』「酌情」也就是加入了人的主觀決定，因此我認為根據幾個人的主觀決定而作出的《規範》，其「規範性」是值得懷疑的。如果政府強行根據這個《規範》而排斥所有不順從的設計意念，是對學術自由的扼殺。因此，《規範》只宜看成一種「指引」，而不是操生殺大權決定某一輸入法的存亡的「標準」。

那麼，且讓我們來看看拆分的原則是什麼？《規範》〈５．１〉：『相離、相接可拆；交重不拆(可拆成筆劃)。極少數不影響結構和筆數的筆劃搭掛，按相接處理。』例如「明」可拆分成「日」和「月」。「日月為明」是小孩子也懂的，就算像「盟」拆分成三部分，也應該沒有什麼異議。但是，是不是相離就一定可以拆分呢？又不盡然，像剛才所舉的「心」，還有像「言」等，雖然相離成幾個部分，卻只當一個部件，為什麼呢？

貳拾・康熙字典

研究輸入法的人，多數都不大看重《康熙字典》，原因是《康熙字典》是清康熙年間由張玉書等人編修，距今已三百多年，當年根本不知電腦為何物，自然不能滿足今日新時代的需要。再加上它本身無可避免有不少疏漏之處，例如與「光」有關的「輝」字和「耀」字為什麼分別歸入「車」部和「羽」部，而不另闢「光」部？

但是，《康熙字典》不是張玉書一個人編的，也不僅僅是他和他的同僚關起門編的，而可以說頗能代表中國古代，包括東漢許慎以至明朝楊慎等人，對漢字部首編排的成果。而這種成果，影響所及，就算今時今日臺灣、香港出版的詞典也仍多採用這種編排方式。究其原因，《康熙字典》是中國歷史上第一次大規模由國家進行的漢字整理，動用龐大人力，而收字量 47,035，基本上包羅至清朝為止的所有漢字，其影響力自然是巨大的。《康熙字典》的部首，基本上就是漢字的「義類」，是我國學子從小就須要認識的，雖然其中有不合理的地方，例如上舉「輝」字和「耀」字的部首就不能反映這些字的義類，但大原則總是對的。這詮釋了漢字是「形音義」相結合的道理，三者缺一不可，因此我很不贊成稱中文為「象形文字」，同樣不贊成稱中文為「表意文字」(ideograph)——任何文字都是表意的，脫離意義就失去文字的目的。

這樣就解釋了為什麼我們將「心」當成一個部件而不是四個，因為筆劃並不是一種隨機結合，而是有「理據」的。一種理想的輸入法不應讓任何人都可以使用，並不是外國人可以按照一些簡單的規則就可以組合出漢字，而應該充分利用中國人自小培養的對漢字的認識。有人說中國人花費了太多時間學習漢字，而西方人卻可將這些時間用在其他的學習上，因此而感到悲哀。姑勿論這種想法對不對，我只想提出一點，無論怎樣我們都已經花了那麼多時間學中文，為什麼平白無端地浪費這種知識而當使用中文電腦時又要另起爐灶重新學習另一體系？

貳拾壹・以部件為輸入單位

總結以上所述，部件是拼形輸入法的基本單位。雖然我們也可以利用大鍵盤將整個字輸入，但上文已經提過這必受時代的淘汰。在另一個極端，我們也可以將字的筆劃逐筆輸入，但因為它的字形損耗率太高，效率一定很低，所以只能作為一些僅偶然要使用中文電腦的人的輸入方法，例如在網上進行電子交易時輸入姓名、地址等少量文字。如果要作為一種現代化的處理工具，中文電腦必須以使用部件的輸入方法為主流，才可以做到像歐美國家般幾乎全民都可以有效率地將文字輸入電腦。要打一個漢字就只不過是將它的部件像字母一樣逐個打出來罷了。(不過，要對漢字作科學研究就不能只管部件而不理部件間的關係。詳見下面附錄。)

因此，訂立一套完善的漢字部件規範或指引，從而避免「群雄割據」的局面，已是很急切的工作。很可惜的是，到目前還未有這樣的一套「完善」的規範，其中還有很多基本問題引起很大的爭議。例如「交重不拆」是不是真的不可以拆？網上有論者說得很好，他舉出一個例子：假如有一個小孩問「中」字的寫法，很自然的，我們會說：「哦，『中』字就是『口』字中間有一直豎穿過。」但是主張「交重不拆」的人就無疑只能說：「『中』字嗎？就是『中』字這樣寫！」這不是說了等於沒說嗎？

理想的中文輸入法應該是以部件為輸入單位，而漢字部件拆分是一個須要全國廣泛探討的問題，而不應只是幾個專家說了算。在未有達成共識之前，國家語委的《部件規範》

可以作為「指引」，但不應作為「標準」。根據這些指引來發展的輸入法，應是未來中文電腦的主流方向。

由於部件的數量有 560 個，甚至多至 980 個，失去了「以簡馭繁」的意義，無論在教學上和輸入法方面都不能起到應有的作用，所以有識之士應該摒棄「交重不拆」的原則。我認為拆還是不拆，不是絕對而是相對的，而且有層次的先後次序，例如要將「忡」字和「忠」字拆開兩份，自然分別是左右和上下拆開，但顯然兩碼的輸入法不能涵蓋全部漢字數萬個之多，所以如果要進一步編碼就要將「中」分拆，就不能教條主義地說「交重不拆」，反而拆開之後更符合人的學習認知過程。又如許慎《說文解字》將「東」字解釋為「日在木中」，固然已有文字學家指出其字源上的錯誤(「東」字的本源是一個兩端束起的囊)，但我們仍然不得不承認，它的本義已經消失，而用「日在木中」反而符合「立足現代」的原則，而且和「束」、「柬」等字比並而觀，更是條理井然，利於教學，因此理想的輸入法必須利用這些中國人已經普遍接受的基礎知識。

部件的總數以多少才算是正確的呢？這沒有絕對的答案。上文已說過有些筆劃的組合能不能算是部件可以有不同的詮釋，而且交重要拆還是不拆也是相對的——盡可能不拆，如果有相離相接的部分應該優先分拆，但是如果要進一步拆分取碼的話就非拆不可了。我認為作為輸入法的編碼單位不可以多至數百個，而應該大約只有兩百個左右，以利學習和記憶，但是由於中文不是先有編碼才據而製作漢字，而相反是先有漢字才歸納出編碼。以漢字有數萬之譜，結構千變萬化，是頗難僅以兩百個部件統攝的，這就造成了一種矛盾。經我反覆研究的結果，我認為問題在於一般人將部件的劃分和編碼的設定等同，而事實上卻可以是兩回事。在部件的劃分初步確定之後，接下來的工作是研究各部件在字形上的共通點，從而歸納出一些共同的編碼。亦即是我們要求的是見部件即能聯想起編碼，但是無須見編碼即能確定所代表的是何部件，皆因部件和編碼之間是一種多對一的關係。由於理想的輸入法的每個字應由一碼至三碼組成，再多的已不能符合高效輸入的要求，而利用近年各界學者對中文字頻的研究，可以先照顧常用的字，所以三碼已足以涵蓋數千個常用漢字，而我們實在不須為罕用字是否能夠首碼輸出而煩惱。

貳拾貳・充分利用現有鍵盤

另一個引起矛盾的地方是使用哪些鍵來作為編碼。上文已經提到單用數字鍵沒有可能產生高效的輸入法。而在使用主鍵盤的輸入法中，有些人主張只使用英文的二十六個字母，例如倉頡輸入法，好處無疑是可以做到和英文狀態互相切換，而保持數字及符號鍵可以隨時使用，亦避免出現像「a;[3」等所謂「怪碼」。實則這些人並不明白ＱＷＥＲＴＹ在一般鍵盤上的排列法並不是天經地義的，而只是一種「偶然現象」，甚至是發明人為了拖慢人打字的速度以符合當年的機械能力而設計的，在二十世紀初葉已有人明白其不足而另設計出更合理的 Dvorak 鍵盤，無奈已是積重難返，更合理而效率更高的設計並未能得到普遍的接受，這就和舊式中文輸入法做成的壞影響如出一轍。現在我們既然要推行理想的中文輸入法，當然無須步英文鍵盤的覆轍，而應該根據各鍵的實際位置來設計。鍵盤上所印出的英文字母和符號，只是作為按鍵記憶的一種提示而已。因此，我們的左手小指既然可以用來按「Ａ」鍵，沒理由我們的右手小指不可以用來按「；」鍵的。至於中英文之間的切換，一來因為中英文使用相同的按鍵，要先按「熱鍵」來切換實屬難以避免，而既然由中文跳至英文須要切換，那又何吝乎按數字亦先按一個熱鍵？其次，反過來說，中英文之間的切換是不是真的一定要按熱鍵呢？其實也不是，隨著電腦智能的提升，已有足夠條件「中英合璧」，輸入時可以中英文照打，電腦系統可以判斷何種編碼的組合是中文還

是英文，尤其當中文的編碼最多只用三碼，而英文平均每個 word 要五碼就更能分得一清二楚了。

為了增加「編碼空間」，加長編碼既然並不理想，那麼就唯有打破英文二十六個字母的限制，而盡量使用雙手能及的所有按鍵——除了過於偏遠的鍵，亦即將手指伸直而仍然達不到的鍵，如「｀」、「＝」等。

貳拾叁・繁簡兼備

由於本文只探討中文輸入法的大方向，一種理想的中文輸入法的細節要求就不在此一一列出了，不過最後仍須提出兩點：其一是姑勿論漢字簡化是否恰當，但已是海峽兩岸四地的既成事實， 因此爭論何者效率更高、何者更能傳承中國文化等已不是普通人應要探討的問題，就留給語言文字學家來研究吧！我們須要做的，是「兩條腿走路」。亦即理想的輸入法須要「繁簡兼備」。不單要繁體字、簡體字都要按各自的特徵來編碼，而且為照顧海峽兩岸的隔閡，還應加入「互見」的功能，即是打繁體字時容許從而查找簡體字，反之亦然，以利於人民的互相溝通。

要達到這點，除了輸入法須要考慮繁簡字的不同之外，內碼的使用標準更是關鍵因素。目前國際標準 ISO 10646 與統一碼(Unicode)基本上一致，已能照顧繁簡字的要求，而香港和澳門為了照顧本身的特殊用字，亦已進行將其增補字符納入 ISO 10646 的工作。大陸為了兼容以前推出的簡體字 GB 2312-80，雖然至今沒有完全採納 Unicode，但至少做到一一對應，以現今電腦的處理能力完全可以實時進行內碼的轉換，因此問題不大。唯獨是台灣方面社會上主流仍是採用大五碼(Big-5)，只能處理繁體字，與兩岸四地大融和的格局不協，理應積極進行大五碼的擴展工作(或將已經進行的工作積極推廣)。

貳拾肆・詞組輸入

理想的中文輸入法另一項要求是「字詞雙修」，即是為了提高中文輸入的效率，已不能止於逐字的敲入，因為逐字的輸入法充其量也只能達到每分鐘一百二三十個字左右，雖然這樣可能已令慣於舊式輸入法的人瞠目結舌，但是和英文的輸入速度仍然有一段距離。由於人的思想是以「詞」為單位的，因此輸入法亦應以此為依歸，這樣才可能達到每分鐘二百個字以上的速度。雖然有些人可能覺得如果能夠每分鐘打三十個字已經很滿足了，但這是因為慣於過去中文輸入的低效而不敢作出「奢望」。實則這不是個人的問題，而是全民能否利用中文電腦提升國力的問題，而電腦處理中文的效率超越英文，完全是有可能的。

詞組輸入除了提升輸入速度之外，還有另一重更重要的意義。「易學易用，效率高超」，可說是中文輸入法的兩大目標，多少人殫精竭慮都在想如何發展一種中文輸入法又快又易用，但這好比一條軸上的兩端，向其中一方靠攏就必然離另一端更遠，如取中庸之道，則只是一種折衷，到頭來不是太難學，但效率又不能很高。

一個字的碼數，和這種輸入法的難用程度是成幾何級數增長的，亦即如果一種輸入法每個字打三碼，其難度並不是每個字打兩碼的輸入法的一倍半。由於增加了對何者是字首何者是字身的「分析」，和將字身進一步的「分拆」，所需的判斷思考將會倍數增加。如果一種輸入法要求每個字打五碼，那麼其難度更加不是打兩碼的兩倍半，而輸入一個漢字將會是一件極艱辛的工作，所以難怪學倉頡輸入法不成的人要退而轉學簡易/速成。

無奈只打頭尾的倉頡既不簡易又不能速成。原因如上所述，只打頭尾捨棄中間的編碼大大地增加了字形損耗率，於是造成太多重碼，而倉頡基本上只利用 24 個英文字母，打兩碼的編碼空間是 576，面對數以千計的漢字，就算只計常用字亦有二三千個，可知要選字實屬無可避免；更有甚者，可能要跳頁多次才能命中目標，其效率之低可以想見。

打太多碼令學習難度幾何級數地增加，只打兩碼又令重碼字太多，那麼，怎樣打破這種矛盾，保持每字最多兩碼而又要能夠做成數以萬計的組合呢？我們必須開放思想，打破以字為單位的桎梏。

事實上我們思想本來就不是以字為單位的，而應該是以詞為單位。一個漢字不是和其他漢字作隨機的組合，而是有著一定的思想上的關聯，例如我們只會說「美麗」而不會說「美硬」，縱使這件東西是又美麗又堅硬。因此，我們可以根據某些簡單的規則為「詞」編碼，我們可以使用三碼、四碼、五碼甚至六碼，造成天文數字的編碼空間，而每一個字所要分擔的編碼卻不會超出兩個。更進一步，詞和詞之間的關係也不是隨機的，例如「陶冶」和「性情」是兩個詞，但是它們卻經常走在一起，我們不會說「陶冶興趣」或「陶冶學問」，雖然從意義上說並無不可；又再進一步，我們經常會將「陶冶性情」和「音樂」掛鈎，於是人們常將「音樂可以陶冶性情」掛在嘴邊，如果我們只打幾碼就能將整句輸出，不是既簡單又快速嗎？

有人或會將「詞組輸入」和「相關字詞功能」混為一談，其實後者要利用一個字的編碼帶出數十個詞組，要在第一頁命中，無異緣木求魚。

利用詞組輸入，本來已經有不少人研究，尤其在大陸利用漢語拼音方案，在詞組中每個字，或者指定的頭和尾的幾個字，只打拼音的第一個字母。不過，事實驗證，這種輸入法的效果並不很好，撇開南方人很多不懂漢語拼音不說，最主要的問題是重碼的詞組太多。究其原因，漢語拼音理論上採用 26 個拉丁字母，但「v」在標準漢語不用，「i」和「u」不會用在首碼，「e」和「o」雖然可以用在首碼，但是很少，如此以少於 22 個編碼，編碼空間自然不足以應付數以十萬計的詞組了。更何況中文的「詞」佔絕大多數是雙音節詞，只取聲母或首碼是不可能效率高超的。

貳拾伍・教學責任

上文提到中文編碼是輸入、排序索引和教學的三結合。最後再談一談教學的問題。現時港澳普遍人的中文程度不太令人滿意，除了因為華洋雜處，受到東西方文化的衝擊之外，而兩岸四地越來越多人執筆忘字，錯別字連篇，已越來越嚴重，例如最近澳門某賭場開幕，而它的橫額卻寫上「……娛樂場……」，招搖市街，這不啻是中國文字以至中國文化的污染。究其原因，論者認為部分是現時無論繁體字還是簡體字，主流的中文輸入法不是拼音，就是將漢字割裂得支離破碎的輸入法，前者令人只記起讀音而記不起寫法，後者則令人只看到細部而看不到全部。

輸入法就算不能提高國人的中文水平，至少也不應降低。詞組輸入法的好處是所收錄的詞組是經過整理的，如果製作上面那幅橫額的人用的是理想的輸入法，便不會造成這個錯誤，因為詞組庫中根本沒有「娛樂場」這個詞/詞組。

※　　※　　※

為了更科學更強大的中華民族，請不要再使用舊式的中文輸入法了！讓我們一起為開創理想的中文輸入法而努力吧！

附：漢字組件化逆波蘭表達式

在研究理想的中文輸入法的過程中，我體會到不單輸入法的目的是為了便於電腦處理，而反過來在開發輸入法時，電腦的輔助也是絕不可少的。

簡體字的 GB 2312 收字 6,763 個，而繁體字的 Big-5 收字 13,053 個(一說 13,051)，照顧繁簡體的國標擴展碼 GBK(GB 13000)收字 20,702 個，和 ISO 10646/Unicode 對應的 GB 18030 收字 27,484 個，康熙字典收字 47,035 個，而 ISO 10646 本身亦在擴展之中，連著擴展區 A 和 B 字數達七萬多個。據說現時統計得出的漢字超過九萬個，如果將佛經壁文、韓國、日本的經藏等字，以及各變種異體字加起來，漢字可能上百萬！……「海量」已不足以狀其多！

正因為漢字可能是一個無限的集合，有人主張「無字庫漢字」，即不事先製作個別漢字的形體，在使用的時候才即時用向量或其他方法拼湊出來。以現代電腦處理能力來說，這不是不可能的事，但這樣臨時拼出來的字形能否符合審美的要求，則是未能肯定。無論如何，對於字庫所無的字，這不失為一個可行辦法。問題是如何將字庫所無的字形容出來？

撇除極罕用的字不說，就是我們常用的字，如要科學化地對其結構作出統計分析，漢字的平面二維特性也是一個障礙。舉例來說，什麼部件(或組件的「手柄」)用什麼鍵來編碼，固然說應該利用鍵上的標記去聯想，例如「人」形似「A」等，但這只是「聯想」，必須承認這是沒有絕對的規則的。又如一種輸入法起初用「H」去表示「工」，但後來發覺「H」已是一個極度擁擠的編碼，將會產生大量重碼，於是改用另一個鍵「I」。輸入法編碼的分配無可避免是帶著一定程度的主觀因素，但理想的輸入法仍應盡可能尋求較客觀的依據。除了用嘗試錯誤然後從頭再來的方法之外，如果藉著電腦的分析，預先知道或至少能若干程度預測「工」作為部件在所有漢字中的分配情況，以及「H」和「I」作為編碼的使用情況，必定可令輸入法的制訂更容易而有效，其認受性亦必將提高。

要將漢字組件化，首先要做的是制訂漢字的部件。正如前述，這除了有賴一些學有專精的專家學者努力之外，我們必須明白一國一族的語言文字是歷代全體人民逐漸演變的智慧結晶，任何少數人以一己之私或為了某些政治目的而強加於人民的改革如果不被人民唾棄，亦可能會引起文化上極大的衝擊，結果得不償失。

交重要不要拆？例如「中」字可否拆成「口」加直豎？沒有交重為什麼有些能拆有些不能拆？例如「因」可拆，「田」為什麼就要當成末級部件而不能再拆成「口」加「十」？我認為這些仍須進行廣泛的諮詢，現有的研究成果可以作為討論的起點。

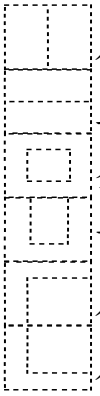
要得出普遍接受的結論殊非易事，尤其這還涉及兩岸對繁簡體字不同的觀點和要求。在未達共識之前，我認為對部件的規範宜寬不宜緊。我主要的觀點是部件數量不可太多，否則會失去部件「以簡馭繁」的作用，因此上述的「中」和「田」我是傾向拆的。

既然「部件」一詞肩負著如此沉重的使命，在未有結論前我姑且用一個較寬鬆的名詞去討論——「組件」。

要將二維的漢字變成一維的表達式，並不是將組件臚列出來就算了事。舉例來說，「匯」和「滙」中組件的排列次序不同，所以問題還不大，但是「咒」和「呪」的組件都是「口口几」(「几」在兩字中略有差異)，單是列出組件的次序便不足以表現兩者的分別。其他如台灣的「感」中的「心」一般嵌在裏面，而大陸將它寫在「咸」字的下面。「埠」的「土」字有寫在左面，也有寫在左上角的，是另一個例子。如果說這些寫法的不同只是異體字的

問題，對漢字的認知尚不至於構成障礙，那麼「條」字和「悠」都是從「攸」的形聲字，如果結構用錯了恐怕便不容易認得出來了。再進一步，如果說雖然辛苦一點，但總還是可以分辨得出來的，那麼「另」如果將「口」寫在左邊，「暉」如果將「日」寫上邊，那就根本不是想要表達的那一回事了。

所以說，除了制訂一套廣泛接受的部件之外，接著要有一套規則表達各組件間的位置關係，這可以用八個字來總括：「左右上下，包圍支架」：


左右(清新)

上下(安基)

全包圍(圍回)

下包圍(凶函)

左上包圍(廣民)

左下包圍(連建)


左中右(術倒鴻)


上中下(實裏寧當真)


上包圍(鳳問)


左包圍(區匹)


右上包圍(勾式包)

支架(乘乖坐巫幽兩)

左中右是左右結構的變種，上中下是上下結構的變種，甚至有分成四個組件的。包圍可以分成全包圍和半包圍兩大類，而半包圍又可以分成三面包圍和兩面包圍兩子類。前者又可再細分為上包圍、下包圍和左包圍三小類，但是沒有右包圍的漢字；後者又可再細分為左上包圍、右上包圍和左下包圍三小類，但是沒有右下包圍的漢字。

大陸在兩岸四地中最強調漢字結構對認知漢字所起的作用，所以很多字詞典都有標明組件間的位置關係。大陸甚至向 ISO 10646 提出將表達位置關係的符號加入到字符集中(如上所示)，但是很可惜，有關方面似乎遺漏了「支架」這種位置關係。例如「乘」和「乖」兩個字，大陸當成是獨體字(亦即只由一個組件組成)，殊不知「獨體字」這個概念是不可輕易使用的，否則部件便不能起到對漢字分析的作用了。「坐」字大陸當成是上下結構，但卻如何表達兩個人分坐兩旁的意思呢？「巫」字大陸當成特殊結構，但是「特殊」(或「例外」)是解釋事物的現象時最無可奈何的下策！為了彌補這項不足，我提出「支架」這種結構——一個位於中間的支架承托著左右兩旁相同或相對的兩個組件(或者一個組件重疊在中間的支架上)，書寫時從上到下如果先出現中間的組件的一部分則先寫該部分，然後分寫兩旁的組件，最後將中間剩餘部分寫完。試以「乘」字為例，「禾」是中間的組件，「北」的左右兩部分分別位於兩旁。書寫時先寫「千」，再寫「北」的左右兩部分，最後將「禾」的末二筆寫完。

正如人類發明了加號可以表達「1+2=3」、減號可以表達「3-2=1」一樣，如果我們為每一種位置結構設定一種符號，那麼也可以解決前述「叻」和「另」要分辨清楚的問題：

叻：口  力　　另：口  力

不過，並不是所有電腦都能顯示這些形象化的位置結構符號，而且用鍵盤輸入亦有相當困難，所以我提出以下符號：

左右：＜　　上下：＾　　包圍：@　　支架：$

要列出表達式，並不簡單地只是從左到右或從上到下。正如我一直強調理想的中文輸入法應該是和手寫對應的，漢字組件化表達式亦然。就算是大陸亦不排除如「樂」字等的「先中間後兩旁」，支架式結構的字就更是如此。因此，除了左右結構用「＜」來表示之外，還應加上寫完中間再寫左面的「右左」結構，並且以「＞」來表示。

由於包圍的各種子類已可由字形表示出屬於何種，因此位置結構符號無須採用不同符號表示不同的包圍結構，例如「聿@又」就一定是左下包圍結構。這點無論是由人去觀察還是由電腦的人工智能去分析都是一樣。

正如數學上有四則運算混合加減乘除一樣，漢字的結構也不盡是簡單的左右或上下，而可以是多種結構的綜合。在數學上我們可以使用括號來表達要優先處理的運算，但是如果運算太複雜的話，用盡小括號、中括號和大括號也未必夠表達，算式將會非常混亂。

有鑑於此，於是有所謂「逆波蘭表達式」。在詞典中查找「Reverse Polish Notation」，可見它的譯名可以是「反向波蘭表示法」，也可以是「逆波蘭記數法」、「逆波蘭表示式」、「逆波蘭式」之類。有「逆」當然有「不逆」，如果只查「Polish Notation」，可見它的解釋是「波蘭表示法」、「波蘭記法」，又稱「prefix notation」和「Lukasiewicz notation」。在 http://www.calculator.org/Lukasiewicz.html 中有 Lukasiewicz 的小傳。

逆波蘭表達式(RPN)根據波蘭表達式反過來而成，用過 HP 計算機的人對此應該不會陌生，而在電腦堆棧(stack)的處理中 RPN 也是一個很重要的概念。它的其中一個好處是可以完全免除括號，而運算的先後次序可以一目了然，例如 3x(2+1)可以表達成「3 2 1 + x」。

應用在漢字組件化方面，先以簡單的「叻」和「另」為例，前者是「口力＜」，後者是「口力∧」，這和「口＜力」及「口∧力」並不顯得有任何優勝之處。但是再以「樂」字為例，如要維持手寫的順序，用一般的表示法是「((白＞幺)＜幺)∧木」，可見使用了很多括號，相當混亂，而且(白＞幺)並未能表達左邊的「幺」向「白」靠攏的形象，反而好像「白」在左而向在右的「幺」靠攏。如果用逆波蘭表達式便會是這樣：

樂：　白幺＞幺＜木∧

省掉了四個括號，而各組件的歸屬和先後一清二楚，意謂：先寫「白」，再將「幺」向右靠攏它，然後是另一個「幺」向左靠攏，三個組件結合之後，再加一個「木」向上靠攏，這樣「樂」字便出來了。其他如：

乘：　禾(北之左)匕＜$2

坐：　人人＜土$1　　(必要時可加上1、2等表示從上到下哪一個空間嵌入)

再以「鬳」字為例(鬳，音闡，笑貌)：

鬳：　口口＜田十∧∧尸(共之上)(喪之下)∧@＜

意即「口」和「口」左右結合，加上「田」和「十」上下結合之後再上下結合。這裏「田十∧」沒有加上括號，但仍然很清楚表達了它們要優先作上下結合，然後第二個「∧」才表示和上面的兩個「口」結合。另一邊是(共之上)和(喪之下)上下結合，然後和「尸」作包圍結構，最後將「單」和「展」左右結合起來。

漢字組件化逆波蘭表達式對一般用戶關係並不很大，因為輸入法本來就不應該將全部組件打出來，更不會將它們的結合位置打出來，不過這種理論性的表達式對漢字的研究應有很大的幫助。

各組件的結合先後次序對部件拆分是非常重要的，亦即「鬳」字要拆應該先拆成「單」和「展」，而不是將「口」從其他部分拆分，這可從逆波蘭表達式的最後一個「＜」表示出來。「單」字(口口＜田十∧∧)如要拆也是應由逆波蘭表達式的後面看起，亦即不是在「田」和「十」間分拆，而應該是將上面兩個「口」和下部分離。

再舉一個例來說明無論多麼複雜的字都可以用逆波蘭表達式表示得很清楚：

鬱(臺灣)：　缶木＞木＜冖𠂉(乂連四點) 凵@(比之左)𠂉彡＜𠂉

鬱(大陸)：　木缶＜木＜冖𠂉……

是不是輸入「ㄖㄖ＜田十𠂉𠂉尸(共之上)(喪之下)𠂉@＜」，電腦就自動砌出一個「鬯」字來——即使是字庫所無的新字？

以現今電腦向量組字的技術來說，基本上是的，但漢字的構造太「千變萬化」，還最少要加上結合方式。

大陸的研究提出部件的結合方式只有「相離」、「相接」和「交重」三種，並主張「相離、相接可拆，交重不拆」，已如上述，但我也提出八個字：「離觸接屬，交連重掛」：

先說前一句：

相離(初旦因)	相觸(合受南)
相接(丁失光)	相屬(氣必斥)

拆還是不拆不是絕對而是相對的，而以相離先拆，相接後拆。相觸是我提出的一種結合方式，介乎離接之間，通常發生在點和撇等長短可稍為伸縮的筆劃，有時寫成相離，有時寫成相接，但都不能算寫錯。反之，如果有人將「初」字兩邊黏連在一起，或將「丁」字的兩筆分離，都可以當成寫錯。相觸的「可拆度」小於相離，但卻大於相接。例如「合」字中的橫劃和「人」相觸而和「口」相離(在相離處分拆)，「受」中的「冖」和「爫」相觸而和「又」相離(在相離處分拆)，「南」中的「冂」和上部相接而和內部相觸(在相觸處分拆)。

相屬為雖相離而不拆，所以根本可以視為同一組件的內部，例如「气」、「心」和「斤」。相屬既可當成同一組件，當中各部分的「可拆度」極低。「氣」字中的「气」和「米」相離可拆固然沒什麼問題，但「必」字中的「心」和撇相交，如要拆寧拆相交而不拆相離。「斥」字中的「斤」和點的關係亦然。康熙部首是判斷是否相屬的其中一個重要參考，但也不是絕對的標準，例如「龍」是一個象形字，畫的就是一條龍的形狀，但我傾向把它拆成三個組件而不是只作為一個「獨體」。即是說：還要看康熙部首的內部結構的緊密程度和該部首的使用頻度、在一般人心目中的認受程度。這涉及人的主觀判斷，要得到多數人的共識，而不能由少數人決定。

再說後一句：

相交(中史束東)	相連(果里未本)
相重(火)	相掛(內確孝寿)

兩個組件相交、相連、相重時盡可能不拆，但如這些組件本身已是單字，在輸入時因無以顯現同系列字的特徵，所以如只打一碼便難以保證一定首選出字，例如「果」、「里」和「畢」如果只根據「田」來編碼，便會都是一樣，所以便不能不將相連的下部分拆另取一個編碼了。

大陸《部件規範》中「交重不拆」的「重」按所舉例字其實應該叫做「相連」，例如「果」和「里」中「田」和下部因直豎而相連。至於真正的「相重」應是指兩個組件之間有某些筆劃重疊，例如「火」字上部成三點輻射部分的中間筆劃和下部「人」的上半截共用，但這只影響編碼的拆分，而且數量極少，一般相重的兩組件可合併當成一個部件看待。

相掛為雖相交而可拆，例如「內」可以拆成「入」和「冂」。「必」字的「心」和撇亦可作如是觀。

在「離觸接屬，交連重掛」中，和逆波蘭表達式有關的是以下幾個：

離：%　　觸：!　　接：&　　交：*　　連：|

上舉「囅」字若加上組件的結合方式，可以如下表達式表示：

囅：　口口＜%田十へ|へ%尸(共之上)(喪之下)へ&@%＜%

只有結合位置(＜＞へ@＄)而沒有結合方式(%!&*|)的漢字組件化逆波蘭表達式，可視為「簡單式」，而有結合方式的則可視為「詳細式」。現再舉幾個詳細式的例子：

賢：　臣又＜%貝へ%
咒：　口口＜%几へ%
呪：　口口儿へ&＜%
受：　爫冖へ!又へ%
陳：　阝木日＄*＜%
函：　乛氷へ&凵@%
贏：　亠(乚無鉤)へ&口へ%貝月＞%几丶@!＜%へ%

詳細表達式通常已可將漢字描述得很清楚。更進一步，相交是在何處相交，各部分的比例如何，如果全部資訊放齊，恐怕會令表達式過分複雜。不過，如果不要求「砌」出來的字很美觀的話，如只要求「認得出」，相信有結合位置加結合方式已經足夠。例如「中」字的「口」和直豎相交，並不是在直豎的正中間，但就算放在正中間，雖然在美觀上打了折扣，但總認得出是這個字，對意義的辨識並不構成太大影響。

漢字組件化逆波蘭表達式這個構想只是拋磚引玉，要做的工作還有很多，例如要將已有統一碼的全部漢字(大概七萬多個)的表達式存入電腦，再進行統計分析，非一人可以做到。另外，這種表達式先要建立在「部件」上，而正如上述，這還須兩岸四地進一步諮詢討論。

還有一個問題是有些部件還未造字型和編碼，如(共之上)、(北之左)，也沒有名稱，不利於處理和教學。這些都是要大家共同努力的。