描述性和辨別性
——漢字組件化的兩個方面

Descriptiveness and Discriminativeness — the Two Aspects of the Componentization of Chinese Characters

梁崇烈

說文２分碼中國語文系統作者
澳門民政總署顧問高級技術員
sonnet@shuowen.net

2007年1月

摘要

據說，胡適先生留學西洋時，首次看到英文打字機，驚歎在文字輸入方面，中國比英美落後一百年。其實，不獨在輸入方面，連在排序索引方面，中文也比英文落後，不如英文之按字母順序：as easy as ABC。

本文作者認為，一國自有一國的文化特質，不能單獨從一二個項目決定優劣。問題只在於能否順應中文為二維的方塊文字的本質，而尋求最有效的排檢方法。

中國歷來有部首、拼音等排序索引方法，但都各有嚴重的缺點。本文試因應中文輸入法和排檢方法的既有共通點，而又有不同的要求和適用對象，利用漢字組件化的描述性和辨別性兩種功能，在說文２分碼以外，提出相關的說文組件碼，希望以最簡單的方式為普羅大眾提供一種最快捷的檢索中文的方法。

壹、　排序索引——中文輸入法以外

在第三屆兩岸四地中文數字(位)化合作論壇上本文作者提及漢字組件化有三大目的：

1. 輸入
2. 教育
3. 排序索引

上屆已就說文 2 分碼作出示範，並指出輸入法對學童認識中國語文方面影響深重，獲得與會者熱烈討論。今屆擬就排序索引方面補充闡述。

註：上次專題論文見《中文輸入法大勢評議》http://shuowen2.com/documents/SWcdf.pdf

貳、　排序索引關係重大

「吾生也有涯，而知也無涯」(莊子．養生主)，在現今知識爆炸的年代，懸梁刺股式的苦讀，已追不上一日千里的知識增長，我們更著重的是，在有需要的時候，如何搜尋相關資料以組織成有系統的知識。

電腦的出現固然起著組織資訊的作用，但縱使是傳統的工具書，例如字典辭書、百科全書等，都對知識的搜尋起著關鍵作用。進而至於圖書館的書目檢索，我們都需要一個簡單而又快捷的方法。

知識不是憑空出現的，而是經過古今中外不同的人的經驗和研究成果累積而成。能否有效地搜尋知識，成了個人以至一國一族之能否興盛的重要因素。

退一步說，就算是一般升斗市民，在日常生活中，亦須經常面對排序索引的問題，諸如電話簿親友的排列，以及貨物清單等，都必須快速地檢索和增刪。否則，如以隨機的方式加入或僅附加在名單的最後，在日後檢索時便只好從頭到尾逐項比對，其費時失事可想而知了。

相反的，進一步說，是不是有了電腦，就可以由電腦以每秒億萬次運算的速度搜尋，就可以用順序檢索(sequential search)解決問題呢？懂電腦的人都知道，電腦的功能也是有時而盡的，而且電腦的運算如果不能透過接口或介面，讓用戶輸入指示，和將結果有序地顯示給用戶看，它的工作能力再高也是徒然的。因此，有關數據的排序索引仍然是一個重要的攻關項目。

英文的排檢就有如 ABC 般容易，那麼中文呢？

參、　歷來排序索引方法之不足

自古以來中文當然也有其搜尋知識的方法，重要的有以下幾種：

1. 內容分類：

   例如美國人發明的杜威十進圖書分類法（Dewey Decimal Classification）也可以用於中文書目，而中國文化悠久，歷代以來，如唐《北堂書鈔》、宋《太平廣記》、明《永樂大典》、清《古今圖書集成》等，也是類似方法，都是用某種既定的分類方式，將不同的知識，分層歸類到天文、地理、君臣、山川等類別去。

用這個方法來排檢的缺點是分類過於廣泛，一類之中項目太多。正如「杜威十進圖書分類法」這個名稱所示，只適宜用於圖書館的書目排檢。

2. 韻書：
例如隋《切韻》、宋《廣韻》、金《新刊韻略》(平水韻)、清《佩文韻府》、黃錫凌《粵音韻彙》等，將漢字按其韻目分別歸到一東二冬等類別去。通常用於寫詩填詞，一般人未能掌握，而且同韻字太多。

3. 部首：
為中文排序索引最重要的方法，下面有較詳細討論。

4. 拼音：
為目前中國大陸最廣泛採用的檢索方式，依據漢語拼音方案用普通話拼音排檢漢字。詳情亦在稍後討論。
廣東話及其他方言，也有各自的拼音方案，但是因為沒有權威機構以行政或立法手段強制推行，所以各家各派沒有共通的標準，未能形成排檢的有力工具。本文作者提出一套「說文粵音碼」，可作參考。
（見 http://shuowen2.com/ShuoWen2/SWcantonese.pdf）

5. 筆劃數：
以繁體中文版微軟視窗為例，操作系統預設的排檢機制為「筆劃數」。用家對於要在包含數百個檔案的文件夾中，找尋某個檔案是如何的費時，應該有深切體會。點算筆劃過程繁複易錯，而同筆劃數含字太多，是一種效率較低的排檢方法。它通常和其他排序方法配合使用。

6. 筆劃順序：
如大陸之用 12345 代表橫豎撇點折。作為輸入法，它的好處是會寫字就會輸入，幾乎不用學。雖然速度很低，仍然有很多手機使用。但是畢竟多筆劃字須使用一大串數目字，而同碼的字仍然極多，這是它的缺點，尤其不適合作為排檢方法，通常不單獨使用而只是作為其他排序的輔助。

7. 以常用程度(字頻)分級：
如果只分成常用字、通用字、備用字、罕用字等，未足以鎖定個別漢字；如果以字頻統計表排序，亦不是普通人可以掌握。譬如說，第 1234 號常用漢字是哪一個？「碼」字算多常用，應排在第幾位？

8. 內碼：
如電報碼、Big-5、Unicode 等，只宜由電腦處理，一般人不可能記得數以萬計漢字的內碼。

9. 輸入法：
如倉頡、五筆、行列、縱橫等。上屆論文已討論各種輸入法，此處不贅。只就排檢方法而言，簡單易學的則重碼字太多，重碼字少的又複雜難學，因此有些字典只以輸入法排序作為附錄，而不直接用作正文的排列，但卻令查字典變成「兩輪」的程序。

肆、　部首源流

部首是中國自古至今最重要的排序索引方法，因此有必要進一步研究它的發展

源流，從而判斷它的優劣：

■東漢・許慎《說文解字》

《說文解字》共有部首 540 個，是中國第一部以部首排列的漢語字典。其中部首的排列始於「一」，終於「亥」，受天干地支、陰陽五行影響，其排列順序沒有嚴格的規律，檢索起來並不容易。

■遼《龍龕手鑑》、金《五音篇海》

這些字書開始調整部首數量，在同一部首內的漢字，部分採用筆劃排列。但這些字書的部首本身仍非採用筆劃數排列，因此查核部首仍然相當費時。

■明・梅膺祚《字彙》

首訂現行 214 部首，並且是第一本部首本身和部內漢字均採用筆劃數排列的字書。它對於《說文解字》的部首作了大刀闊斧的修改，減少部首數量，例如《說文》「男部」有「男、甥、舅」三個字，而《字彙》取消了「男部」，分別歸入「力、生、臼」三部。這種修改也有它的缺點，就是開始背離以意符作部首的原則，不利於字義的分析和歸類。

■清《康熙字典》

清康熙帝勅修，翰林院張玉書、陳廷敬主編，承襲《字彙》214 部首。由於清法規定科舉考試書寫字體必須以《康熙字典》為標準，因此它成為士人研習的寶典。

乾隆時王錫侯著《字貫》指出《康熙字典》錯處，冒犯皇帝威嚴，遭受滿門抄斬的厄運。我們對專制皇朝的殘酷欺壓士人感到憤慨之餘，亦可見出《康熙字典》在客觀事實上有不可動搖的地位。至今臺港澳字典辭書基本上仍採用《康熙字典》編纂體例。

伍、　部首檢索的不足

用部首檢索漢字，有近二千年的歷史，加上封建皇朝的積極推行，已成為中文排序索引的主流方法。但它不是最理想的檢字方法，它的缺點有以下幾點：

1. 同筆劃的部首的排列順序，沒有嚴格規律，例如四筆的部首有 35 個，連變形共 46 個，尋找費時。
2. 有些常用的意符沒有列入，例如從「光」引伸出來的字有很多個，但《康熙字典》竟然沒有「光部」，於是「光」歸「儿部」，「耀」歸「羽部」，「輝」歸「車部」，毫無準則，檢索只能靠死記或逐個部分嘗試。
3. 有些字的部首形狀不明顯，例如難以看出「與、興、豐」歸「臼部」，而「學」卻又歸「子部」，「覺」歸「見部」。
4. 獨體字的部首更不易掌握，例如：

- 「丈、世、丑、並」同屬「一部」；
- 「一」屬「一」，「二」屬「二」，但「三」卻屬「一」，「五」又屬「二」；
- 「之」一般寫三筆，但《康熙字典》算四筆，並且不以第一筆的「丶」而以第三筆的「丿」作部首。

5. 有些字的意符不作部首而反以聲符作部首，例如前述的「甥」和「舅」。如因此就以為這是一種改良，總之先行的部分就是名符其實的「部首」，則顯然又非事實，因為亦有很多例子違反這個原則，例如同為形聲字的「視」

字不歸先行的聲符「示部」，而歸意符「見部」。

6. 非形聲字(象形、指事、會意)的部首歸類更沒有統一標準，例如前述的「男」字(《說文》：用力於田)，《康熙字典》改歸「田部」，取前不取後，但是「相」字(《說文》：地可觀者，莫可觀於木)，卻又歸「目部」，取後不取前。
7. 部首的筆劃點算不易，筆劃多者固然點算耗時且易數錯，而筆劃少者亦無絕對標準，例如：
   - 「氏」算四筆，「瓦」算五筆；
   - 「𠂇」一般寫兩筆，但《康熙字典》分成三筆；
   - 「辶」無論當是三筆或四筆都找不到，而要找它的原形「辵」，算七筆；
   - 「阝」還要看是左邊八筆的「阜」還是右邊七筆的「邑」；
   - 類似的還有「人、刀、心、手、水、犬、火、肉、艸……」等皆以多種形式出現。

其他沒有規律的部首舉例：

- 「戀」屬「心部」，「巒」屬「山部」，但「變」屬「言部」；
  (《說文》「變」屬「攴(攵)部」，反較合理)
- 「犛」屬「牛部」，「氂」屬「毛部」，但「斄」屬「攴(攵)部」；
- 「意」屬「心部」，但「章」屬「立部」；

如此者多不勝數，造成中文學習及排序索引的沉重負擔。

即使在部首的判別和筆劃的點算上都沒有出錯，部首檢字仍然是一個效率很低的排檢方法：

- 數筆劃費時：筆劃多者至二三十筆，只能用手指憑空模擬寫字，一筆一劃，沒有捷徑；
- 部首檢索是一種「兩輪」的程序：先在「部首索引」搜尋所需的部首，再按頁碼搜尋字典的正文。若果像大陸出版的字典般，將部首檢索當成附錄，則更是「三輪」的程序。

陸、　大陸使用的部首

正是由於部首檢字有以上缺點，於是有心人便研究如何改良。新的部首見於大陸版的《辭海》、《漢語大字典》等。

大部分漢字簡化是整個部首調整，例如「言」作左偏旁時變成「讠」，對部首判斷的影響尚不至於很大。但是簡化字中有不少意符被取消，漢字成為純粹的「符號」，例如「發」(癶部)和「髮」(髟部)簡化成「发」，表意的特性減弱(如非消失)，只好歸入意義無關的「又部」；又如「內」本屬「入部」，「簡化」成「内」後既和「人」無關，只好取書寫順序較前的「冂」作部首。

此外，簡化字造成大量獨體字，無從分析，只好以其中的部分或個別筆劃歸部，例如「電」本和「雨」有關，簡化成「电」後只好歸入「田部」；又如「樂」本象絃線繃在木上之形，歸入「木部」，簡化成「乐」不可解，只好用首筆「丿」作部首。

再以「東」字為例，讓我們看看部首的演變：

- 《說文解字》歸入「東部」(「動也。从木。官溥說：从日在木中。凡東之屬皆从東。」)

註：「東」本來是一個兩頭束起來的囊的象形，《說文》的解釋只適用於當時見到的小篆。

- 《康熙字典》歸入「木部」(「.....按《說文》東自爲部，今併入。 」)
- 大陸簡化字「东」為「草書楷化」，失去「會意」的作用，只好歸入「一部」。

以首筆劃作部首看似簡單，其實造成這些「單筆部首」負擔過重，部內包含的字過多，失去以部查字的作用。

以 1999 年版上海辭書出版社【缩印本】《辞海》為例：

- 不計繁體字部首，共有部首 250 個
- 「一部」在正文共佔 71 頁，共收字 72 個
- 刪去一些部首(丿)，合併一些部首(人、入→人)
- 新改一些部首(辶→辶、艸→艹)
- 確立部首筆劃數和一般寫法一致
  - 「辶」算三筆
  - 草頭採用三筆的「艹」而不用四筆的「艹」，亦不用六筆的「艸」
- 分立一些部首(人→人、亻)，新增一些部首(冖、亡)
- 訂立一些取部首的規則，如取左不取右、取上不取下、取外不取內、取大不取小、取複筆不取單筆等，因此解決前述一些問題，例如「相」取「木」不再取「目」，「意」和「章」皆屬「音」等
- 力避在字的中間取部首：「囊」字捨「口」而取「一」
- 形聲字不問形符聲符，例如「闻」取外不取內，所以用聲符「门」做部首而不用形符(意符)「耳」

柒、 誰定至尊？

大陸的 250 個部首新編排除了因為簡體字的出現而有其必要性外，如單以繁體字而論，是不是就比《康熙字典》214 個部首優勝？

事實上，臺港澳繁體字地區仍然採用康熙部首。《康熙字典》為封建皇朝的欽定產物，集古代近二千年文字學之大成，自有其權威性。訂立規則無可避免有主觀成分，反正都是主觀，並不見得新的主觀一定比舊的主觀好。因此，250 部首未能建立唯我獨尊的地位，不能與《康熙字典》相比。試看徐中書主編的《漢語大字典》有部首 200 個，就並不以 250 部首為標準。

與其說新的和舊的哪一個更能定於一尊，毋寧謂兩者都不是好的排序索引方法。因為部首檢字有其先天不足：既希望以意符作為部首，而偏偏有些意符的位置並不明顯，在字的開頭部分較明顯的卻又是聲符，此其矛盾一；又要減少部首的數量，又希望部內字數不可太多，兩者實難兼得，此其矛盾二；又要照顧字源，而一般人寫字的習慣偏偏又和字源相違，此其矛盾三。這些矛盾在本質上不可能解決。因此，本文作者認為部首檢字法已經完成了它的歷史任務，不適宜再作為新的字典辭書以至任何排序索引的方法。

捌、 拼音排序

既然部首檢字法有其不可解決的矛盾，新的排檢方法遂應運而生，尤其是大陸

方面，棄用部首而改用漢語拼音，已成為主流方案。

它有以下的特點：

- 適用於百科全書、人物清單等的排序；
- 優點為簡單易學，但是大前提是須先懂普通話，而且熟悉拼音方案，這對一些人，例如澳門人來說，並不是必然的；
- 最大缺點為同音字太多，例如發[yi]音而一般人認識的字近百個，如作為字典須包括罕用字，則更可能超過二百個，縱使加上聲調，同音字亦有數十個之多；
- 漢字為形系文字【按：不應稱為「表意文字」，因為任何文字都是表意的】，有別於西洋的拼音文字，用音來檢索不符合漢字的先天特性；
- 拼音不適宜作為漢語詞典，更不適宜作為漢語字典的排檢方法——查字典往往正為索其讀音，以音查字為適得其反。

玖、 漢字組件化逆波蘭表達式

鑑於上述種種排檢方法的不足，本文作者提出「描述性和辨別性——漢字組件化的兩個方面」。所謂描述性，指從不同角度表達事物的某方面，尤其是較有代表性的特徵，不一定求其全面；所謂辨別性，指表達方式足以全面定義某種事物，目標唯一而不會引起任何歧義。

作者上屆發表《漢字組件化逆波蘭表達式》(見《中文輸入法大勢評議》http://shuowen2.com/documents/SWcdf.pdf 附錄)，針對「辨別性」說明如何從結合位置和結合方式對一個漢字作全面表達。所表達的漢字甚至可以是一個新造字，閱讀表達式的人或電腦，可以據此重塑完整漢字。例如「樂」字的表達式為：

白幺＞%幺＜%木𠂉！

意為：先寫「白」，然後在左方寫「幺」向右靠攏，以相離方式結合，再在右方寫「幺」向左靠攏，同樣以相離方式結合，最後在下方寫「木」向上靠攏，以相觸方式結合。

詳細的漢字組件化逆波蘭表達式(詳細式)，得出來的結果是獨一無二的，是漢字組件化的最具「辨別性」的表達。

「辨別性」和「描述性」並不是互斥的概念。「詳細式」同時具有對漢字的高度描述作用。但是除了用電腦作向量組字，或對漢字進行理論分析之外，一般人並不一定需要絕對的辨別性，反而更重視字和字之間的相對關係。亦即更有價值的表達式，可能要相對減低辨別性，而更重視描述性的作用，例如「樂」字的「簡單式」省去結合方式，成為：

白幺＞幺＜木𠂉

但是如果不要求對漢字的唯一辨別性，那麼即使是簡單式也時常過於複雜，例如要將「鬱」字的組件全部臚列出來，將會是一大串：八個組件加上多個關係符號。

研究新的排序索引方法，所要求的並不是「無中生有」利用表達式創造漢字，而只是「有中生有」，利用表達式的描述作用，辨別表達式相對應的漢字而已。

壹拾、 適度的辨別性

對輸入法而言，好的表達式，亦即「編碼」，應該盡可能是唯一的，亦即「無重碼」；對排序索引而言，對漢字的描述只須保持「適度」的辨別性即可。因此，漢字組件化逆波蘭表達式過於詳細，並不是最佳的排檢方法；但相反的，以部首檢字在同一部首的同一部外筆劃數之中，動輒要翻閱多頁才能找到，而以拼音檢字，同音字亦多達數十個，顯見兩者的辨別性皆有不足。

何謂「適度」是相對而言的，作為輸入法，常用字應該零重碼，亦即百分之百的辨別性；罕用字則容許選字。要做到零重碼，又要求編碼短，縱使是常用字亦不易達到，無可避免要使用「例外編碼」，增加學習困難。更有甚者，對字典來說，查罕用字的機會可能比查常用字的機會更大，對辨別性有另一種要求，因此更不能做到零重碼。

綜合來說，排檢方法有如下的要求：

- 使用排序索引的對象為普羅大眾而不一定是使用電腦的人(雖然現今世界兩者漸趨於一致)；
- 排檢方法應該極之容易學習，其易學程度的要求比輸入法更高；
- 不可有任何例外編碼；
- 只要求適度的辨別性，同碼字一般三四個，最多不應超過十個，而不要求唯一；
- 應該充分利用已有的中文知識；
- 以形索字而非以音索字，避免本末倒置；
- 照顧常用字以至罕用字；
- 異於輸入法之利用詞組輸入以提高效率及減少重碼，排序索引應以字為單位。

壹拾壹、 說文組件碼

基於上述構想，本文作者提出「說文組件碼」作為排序索引的依據方法。它的總則是：

一個組件一個碼，每字最多三個碼

以下先提出幾個例字，在其後的章節才進一步詳加解釋：

口(o)吕(oo)呂(oo)品(ooo)

區(eoo)靐(eoo)

鷗(eo5)鑪(5eo)

壹拾貳、 「組件」的探討

海峽兩岸對部件的研究各得出的部件數量，由五百餘至近千，數量過多，不利學習。針對部件拆和不拆的問題，本文作者在上屆專題論文中提出：

『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。』

對於這些問題的新認知是：作為輸入法要避免重碼，拆與不拆是相對的，如果堅持「交連不拆」就會引起大量重碼。但是，作為排檢方法容許適度重碼，拆與不拆要求絕對、沒有例外，才能滿足普羅大眾對易學易用的要求。

從以下例子，可以看出說文組件碼(用於排檢)和說文首選碼(用於輸入)的不同(括號中前為組件碼，後為首選碼，大小寫相通)：

中(I - i)　申(I - i.)　巾(I - ni)
田(U - u0)　由(U - u.)　曲(U - uh)
忠(IV - oiv)　沖(3I – 3oi)
農(UF - uf)　粬(WU - wuh)

壹拾參、　組件的基本定義

組件大致上是字中「相離」、「相接」或「相觸」(時離時接，離接之間)的各部分：

相離的字例：

初(礻刀)、每(𠂉母)、因(口大)、坐(人人土)

相接的字例：

玄(亠幺)、南(十冂相接，另加羊)

相觸的字例：

受(爫冖相觸，另加又)、南(十，另加冂羊相觸)

然而，並不是所有相離、相接、相觸的都一定要分拆。這裏可以回過頭來說一說康熙部首。

部首檢字法誠如前述有相當多不足之處，但也有它的價值：

- 二千年來文化的傳承；
- 高度的權威性；
- 普遍流通，縱使是現代，臺港澳學童仍然自小學習，而且字典辭書廣泛採用；
- 利用字形分析而非利用拼音，從漢字為形系文字的本質上去處理漢字；
- 部首多數為意符，有助字義的了解(雖然也有一些例外，如前述「甥」之歸入「生部」等)。

我們正可以利用前人的智慧成果，而不是一筆抹殺：所有 214 個康熙部首都是「組件」。因為部首已廣泛使用，故不致造成學習過大的負擔，但是仍須特別留意較罕用的部首，例如「玄、生、谷、辛、首、齊」等。不論其內部結構是否相離，例如「言、龍」等皆視為組件；也不論該部分在字中是否當成部首，減少判斷的困難，例如「生」在「甥」中為聲符，在「產」中為意符，皆為部首，而在「牲」中為聲符，不是部首，但在說文組件碼中一律視為組件。此外，判斷組件時無須數筆劃，不論其位置，例如「阝」無論當兩筆、三筆、七筆或八筆，無論在左或在右，都當成一個組件。

於是，現在可以對上屆提出的其中一個問題作出解答：

- 「因」要拆，因為不是康熙部首，而其外內相離；
- 「田」作為輸入法首選碼要拆，因為整個字當成一個組件的編碼「u」已有重碼字(不拆也可以，但是要選字)；作為排檢方法不拆，因為是康熙部首，整體外內「相屬」，而排檢容許重碼，所以寧願犧牲輸入法的效率，而更重視它簡單易用的一致性。

壹拾肆、　　相交不拆

我們還可進一步討論拆與不拆的問題。

正如大陸的學者專家提出，字中所有筆劃相交的部分都不分拆。例如「事」字全部筆劃都相交，不再分拆；「東」雖解釋成「日在木中」，在輸入法中拆成「木」和「日」兩部分，但作為排檢方法的說文組件碼當成一個組件，不再分拆；「朿、東」等亦然。

於是，現在可以對上屆提出的另一個問題作出解答：

- 作為輸入法首選碼，「中」不拆，因為是高頻字只打一碼，可以大幅度提升效率(拆亦可，但不能首選輸出)。「中」在「忠、沖」等字中要拆，因為分拆可提高辨別性避免重碼(不拆亦可，但不能首選輸出)。
- 作為排檢方法，「中」無論作為單字或是字中的一部分，都永遠不要分拆，因為筆劃相交，構成組件整體。

壹拾伍、　　相連基本上不拆

大陸學者提出「交重不拆」，相交要不要拆已如上述，而「相重」實際上應是「相連」，例如「串」字可不可以拆成兩個「中」，這是另一個廣受爭議的問題。

繁體字地區最流行的中文輸入法是倉頡，但是它最為人所詬病的一個問題是「割裂漢字」，例如將「目」拆成「月」和「⺆」。

（詳見上屆專題論文《中文輸入法大勢評議》http://shuowen2.com/documents/SWcdf.pdf)

相連不拆符合中國人寫字以筆劃為單位的習慣，說文組件碼亦以此為基礎，但是提出額外的兩條規則：

- 「拆十不拆一」，例如「未」拆成「十木」，「本」拆成「木十」；
- 「田上不出頭」，例如「里」拆成「田土」，「果」拆成「田木」，但「申」上有出頭不再分拆。

為保持「交連不拆」的一貫性，拆和不拆的編碼可以並存，例如「未、里」皆可當成一個組件或兩個組件。

壹拾陸、　　附筆不計

本文一直強調組件的「辨別性」和「描述性」兩個方面，那麼對一個漢字的描述，應該著重其中各部分的「特徵」，才能彰顯其與別不同之處，將它從眾多漢字中辨別出來。這就正如我們描述一個人的樣貌時，我們通常會說「方面隆準」(國字臉、高鼻子)、「身裁高大」之類，但卻很少會用「頭髮很幼」、「腳趾很短」等來形容。

「附筆不計」可說是說文２分碼，以至現在發表的說文組件碼，對漢字分析的最大突破，前人從來未有類似的討論。

附筆指除折筆之外的單筆(橫豎撇點) ，而相交或相接、相觸於另一主體之前或後者，不獨立當成組件，而辨別一個組件的依據主要落在主體上。舉例如下：

- 「合、共」各為兩個組件，「一」為附筆不計
- 「存、在」各為兩個組件，「丨」為附筆不計
- 「白、才」各為一個組件，「丿」為附筆不計
- 「主、太」各為一個組件，「丶」為附筆不計
- 「亡」為「亠」「乚」兩個組件，「乚」為折筆不算附筆
- 「吳」拆成「口、𠃌、大」三個組件，「𠃌」為折筆不算附筆
- 「旦」為兩個組件，「一」和「日」相離不是附筆
- 「引」為兩個組件，「丨」和「弓」相離不是附筆
- 「糸」大陸寫法「丿」和「糸」相接成附筆，但臺港寫法相離為兩個組件
- 「勺」為兩個組件，「丶」和「勹」相離不是附筆

應要注意的是：如果兩個單筆可以合成一個組件，那就不算附筆，例如「亡」中的「亠」、「反」中的「厂」等。

壹拾柒、　取碼元件

將一個漢字分析成若干個組件是第一步。接下來，要為組件編碼，就取決於組件內的「元件」。一個組件由一個或多個元件構成。元件依照〈取碼元件〉表。
(見說文２分碼中國語文系統之《說明文件》http://shuowen2.com/ShuoWen2/SWdocument.pdf 或本文附錄)

一個組件只取一個元件編碼——取「描述性」最強、最有特徵的元件。如果組件開頭無附筆，則按書寫順序以首元件取碼，例如：

- 「美」可拆成「羊(羊)」「大」兩個組件，「羊」只以第一個元件「丷」取碼，「王」為組件中第二個元件不作取碼；(「羊(羊)」為康熙部首，故不拆成兩個組件)
- 「英」可拆成「艹」和「央」兩個組件，「央」只以第一個元件「冂」取碼，「大」不作取碼；(「冂」和「大」因為相交故不拆)

如果組件開頭的元件為附筆，不足以代表該組件，則以其後的元件取碼，例如：

- 「皆」以「𠂤日」取碼而非「一丿」(「比」為康熙部首，故不拆成兩個組件)
- 「泛」由「氵」和「乏」組成，而「乏」首兩筆不構成元件，當作附筆，故以較具特徵作用的折筆「𠂉」取碼。

元件有時自成組件，亦即一個組件只由一個沒有附筆的元件構成，例如「為」內部相離，亦非康熙部首，但因整個字屬取碼元件，所以不再分拆。

壹拾捌、　獨體字的編碼

將一個漢字拆分成不同的組件後，又知道每個組件以什麼元件作代表，現在便可以為漢字編碼了。

首先看獨體字。獨體字由一個組件構成，根據「一個組件一個碼」的原則，可

按首元件只取一碼，例如：

乙(z) 日(q) 三(3) 九(9) 羊(y) 言(')

作為輔助方法，亦可按書寫順序取最多兩個元件，以加強其辨別性，例如：

羊(yi) 言('m)

這種兩碼的獨體字可以**不避附筆**，例如：

白(q/8q) 我(c/8c) 丸(9/9')

而且**可拆相交**，例如：

丰(i/3i) 事(0/0o)

如果這個組件是康熙成字部首，亦可按相離或相接的各部分最多取三碼，亦即這個字可視為合體字而非由一個組件構成，例如：

气(9/am9) 辛('/'y0) 雨(m/myy) 言('/'mo)

鹿(f/fev) 鼻(m/mom) 龍('/'5l)

有些獨體字可以取一碼、兩碼或三碼。這種一字對應多碼的現象，在字典學界稱為「多開門」。

壹拾玖、 合體字的編碼

接著看看合體字的編碼方法。

首先將漢字一分為二作第一層分拆，如果兩部分都是單組件，則不再分拆，整個字就只取兩碼，例如：

- 康熙部首不拆：信(a') 討('[) 秋(tw) 烋(wt) 到(]j) 隻(ax) 因(uk) 房(pf)
- 附筆不拆：吞(ko) 共(h8) 在(9g) 伯(aq) 住(ae)
- 說文元件不拆：偽(a5) 妹(xt) 曷(qu) 吧(ol)

如果分拆出來的兩部分還可再分拆，則再將其中一份一分為二。其中如果後組件為複組件，則優先拆後不拆前，不論前組件為單或複，例如：

- 前單組件、後複組件：後(cs;) 萌(hqn) 國(uco) 夾(kaa)
- 前複組件、後複組件：歸(ber) 豔(vg4) 器(oko) 鬱(trx)

但是如果後組件為單組件，則分拆推移到前組件，拆前不拆後，例如：

- 前複組件、後單組件：對(wy[ ) 頡(gor) 然(;kw) 賢(exr) 坐(aag)

貳拾、 分拆順序

一般來說，怎樣描述就怎樣拆，例如：

- 「憶」為「忄音心」，非「忄立日(心)」
- 「陰」為「阝今云」，非「阝人(一)厶(云)」
- 「暮」為「莫日→艹日(大)日」，非「艹日大(日)」

另一方面，分拆的順序考慮三種組件結合的方式，即 1. 相離、2. 相觸、3.相接：

- 相離先於相接，例如「盍」拆成「土厶皿」，但「蓋」不拆成「艹土厶」而拆成「艹土皿」；
- 相離先於相觸，例如「受」拆成「爫冖又」，但「授」不拆成「扌爫冖」而拆成「扌爫又」；

- 相觸先於相接，例如「南」拆成「十冂羊」，但「楠」不拆成「木十冂」而拆成「木十羊」。

貳拾壹、　最多三碼

「每字最多三個碼」，只取分拆部分的首碼，例如：

- 「句」拆成兩碼(勹口)
- 「苟」的後組件再拆成兩碼，合共三碼(艹勹口)
- 「敬」的前組件拆成兩碼之後，不作第三層分拆，全字只取三碼(艹勹𠂉)
- 「警」第一層分拆成「敬」和「言」,「敬」第二層分拆成「苟」和「攵」,「苟」的第一個組件(亦是元件)是「艹」,「攵」的第一個元件是「𠂉」,「言」的第一個元件是「丶」,所以「警」只取三碼成為「艹𠂉丶」

如果各部分並無特殊的小組結合，又屬於同級，例如同屬「相離」，則按先後次序取碼，例如「襄」拆成相離的上中下三部分，而「讓」只取「言」和「襄」其中的前兩部分而成為「丶亠口」。

貳拾貳、　元件與字符的對應

說文組件碼的學習關鍵是記憶元件與拉丁字母及符號的對應，亦是與鍵盤按鍵的對應。〈取碼元件〉盡量利用形狀的聯想，例如「人」的形狀似英文字母Ａ，進而聯想到立人旁「亻」和臥人兒「𠂉」，以及形似的「入」和「乀」(捺)，不難記憶。(見說文２分碼中國語文系統之《說明文件》http://shuowen2.com/ShuoWen2/SWdocument.pdf 或本文附錄)

約 200 個說文元件對應到 40 個拉丁字母及符號上，漢字的排序就依據其對應的拉丁字符排列：

'34567890;ABCDEFGHIJKLMNOPQRSTUVWXYZ[]

無論以 ASCII 或 Unicode 來排列都是一致的，檢索漢字就有如檢索英文一樣。

論檢索效率，雖然每個編碼可能有三數個重碼，但因每字最多三個碼，可作為補償，所以檢索效率應與英文相差不遠，而這是「一輪」的檢索，遠遠優於目前任何一種漢字排序索引方法

貳拾參、　結語

語云：「知不足，然後能自反也」(禮記．學記第十八)，我們必須先知道中國歷來的排序索引方法，和英文相比明顯落後，有所不足，然後才能創造和使用一個更好的方法。這不但影響學術研究，日常生活亦無不與之息息相關。

本文作者不揣淺薄，就描述性和辨別性兩方面，提出漢字組件化與排序索引的關係和解決方法，謹就教於諸君子！

說文2分碼中國語文系統　　取碼元件

'	丶亠讠	（以 apostrophe 作點）
;	ク㇉力	（圓點伸長，逗號彎曲伸長）
[	寸	（形狀無關；跨組件搶先：特 0[、樹 t[[）
]	至	（形狀無關；跨組件搶先：臺 g]、握 j]）
3	三彡氵水氺巛川	（後面五個均與水有關，但「川」屬 N）
4	四罒覀皿目且目身	（「四」平放或直放，加一劃或筆劃伸長）
5	五丑岦五韦丟金钅与鸟乌為为	（金為五金之首；「考」之下及「鳥」之上似 5）
6	六食飠饣	（匙羹[勺子]用以進食）
7	七毛世屯也	（橫劃在 L 形的左右穿出）
8	八癶丿千	（「八」在下時末筆變點；「八」之左撇作 8）
9	九儿𠂇乁⺄气瓦	（截取「九」之不同部位）
0	十車车	（拆十不拆一）
A	人亻𠂉入㇏	（立人旁及臥人兒均作 A，「人」之右捺亦作 A）
B	乃㠯耳	（人耳形似 B）
C	㇛㇂彳	（逆時針方向旋轉；「彳」為 C 的三條切線；組件內搶先：戊 c.、識'c、藏 hc、虤 cu.）
D	㇇㇆㇉勹豸豸豕馬马	（順時針方向旋轉；狗貓豬馬均以 D 歸類，但羊屬 Y，牛屬 I，牛旁則屬 0）
E	匚巨𠄌𠂆𠂆㐄コ王壬	（向右開口或向左開口或向兩邊開口；注意豎鉤）
F	厂𠂆𠂆𠂆干方	（左上包圍；「干」為 F 向兩邊望；「方」除去曲筆有 F 之形，且發音屬 F）
G	土士主	（形狀無關，或可以 Ground 聯想）
H	艹廾丌	（一橫劃加兩直豎）
I	丨巾中申丰𦍌工㇗⺊⺕⺊	（「工」似 I 連稜線；「非」字台港以撇起筆，大陸規定作直豎）
J	亅刂寸手扌𠂠	（直豎向左鈎；最後一個是兩橫加斜撇，與「丰」作 I 不同）
K	夂𠂉衣衤片大夫矢长	（「衣旁」有兩點，與「示旁」作 W 不同；「大」似 K 向右傾斜）
L	乚乚己巳巴㔾电辶廴	（最後兩個按筆順規範後寫，但說文 2 分碼規定左下包圍要先打 L）
M	一㇀冂冂門门鬥币⻗	（將 M 拉直變成橫劃；除 N 型、9 型之外的上包圍）
N	月冂川冂用巾	（左邊為撇，包括上面不封口；月之變形；用之變形；上部穿出）
O	口	（空心，與 U 不同；相接讓先：葆 hat、涼 3'w、銳 589、園 ugk）
P	卩㔾㔾尸尹	（左邊為豎或撇，有鈎或無鈎，上部中央有筆畫或無筆畫）
Q	日曰毌㔿足⻊	（「口」中間或下面多一點東西）
R	貝贝頁页冖	（R 中間加兩橫；約定作寶蓋兒，上有點或無點）
S	幺弓㇉	（絲線糾纏成彎彎曲曲之形）
T	丅⺊罒木未	（T 正放或橫放；樹木亭亭如蓋之形，或可以 Tree 聯想）
U	凵臼囗由𠂤匃酉	（下包圍，向上的兜，或加蓋成全包圍，或有物穿出）
V	𠄌山屮心忄	（直豎向右鈎，和 J 不同；山峰之形倒放；心形）
W	丷⺍小⺌示礻灬业氺	（三點或四點向上或向下作輻射狀散出，和 Y 不同）
X	乂又女	（交叉；「女」像交腿而立）
Y	丷冫⺀𡿨㇃丬⺊丆亻𠄌丬丬	（兩點向上、向左及向右；直線有斜線分岔）
Z	乙乛𠃋𠃊	（Z 的全部、上半截及下半截）

.	（完成）	（重碼字後加句號：to to、杏 to.、束 to..、占 to...、稱 twn、稍 twn.、梢 twn..）
,	（省略）	（「會 a,」「同 m,」等高頻字用「,」表示省略）
;	（特殊）	（「半 w;」「山 v;」等無論作單字或第二組件時皆附以「;」）

（重碼字的處理：一字如取三碼而重碼，則可續取第四碼，如仍有重碼則取第五碼，以利首選出字；如一字已用盡取碼元件，則以句號表示重碼。例如：

蕾 hmu、薑 hmu/hmum、韁 hmu/hmum/hmumu

暮 hqq、昔日 hqq/hqq.、菖 hqq/hqq./hqq..

以上避重碼規則以及「搶先」和「讓先」只適用於說文首選碼，說文組件碼則容許重碼）