

寓知於字—談中文構字中蘊含的知識

科學月刊

第十八卷第十期

謝清俊

「字」由「文」生

古時候，「文」和「字」有不同的定義。東漢許慎「說文解字敘」中云：「倉頡之初作書，蓋依類象形，故謂之文；其後形聲相益，即謂之字。」由此可知，所謂「文」是指漢字在最初造字時所產生的一群最原始的基本符號，它們是中國文字最基本的元素（**primitiv-es**），亦具有「字母」的功能，可用以複合而孳生出許許多多的「字」。

「依類象形」的方式有兩種：一是象「具體」之形，謂之「象形」；一是象「抽象」之形，謂之「指事」。換言之，六書中「象形」和「指事」是造「文」的法則。當「文」緣事物之形象構形之後，再以「音」、「義」相附，便完整地擁有了文字「形、音、義」的要件。漢字中的「文」的數目不多，根據「說文解字」記載的只有489個而已。這些造字之初的文，我們有時特別名之曰「初文」以資識別。

在表一和表二中，分別舉了些象形和指事的例子。”（圖表省略）”舉這些例子的目的，是讓讀者體會體會先賢是如何從事這麼富於智慧的工作：對自然界事物詳實的觀察，精闢地體會並掌握它們的「特徵」（**features**），然後再用幾筆簡單的圖繪，將事物「特徵」傳神地轉變為文字。而這個過程和目前人工智慧中，做識別（**recognition**）類工作時的第一步驟「特徵抽取」

（**feature abstraction**）的程序和要旨，是不謀而合的。若是對事物主要的特徵不能把握住，那麼識別的效果就不會很好。讀者不妨把我們的文字暫時拋開，

自己造造這些字試試，是不是能造得更好？若能這樣的嘗試，當會更深切的體認先民造文字之睿智與優美。

由這些例子裡，我們也可發現近代認知科學中歸納「聯想」的四個主要法則：接近律、類似律、（圖表省略）反對律和因果律，亦在此全部呈現。在象形的例子中多合類似律，而「旦」可解釋為因果律，「叵」則為依反對律所造之形。

到了宋朝，鄭樵對「文」下了另一個定義，也說：「獨體為文，合體為字。」這個定義的精神和許慎的並無不同，都是持「字」由「文」生的觀點，然而做法和結果上卻截然不同。根據相關資料對「象類書」（該書已失傳）之描述，鄭樵定義了870個以聲為主的「文」和330個以形為表的「文」，合計共1200「文」，是為所有「字」孳生的基礎。顯然，鄭樵將重點放在字的孳乳上，才得到這樣個結果。當然，文字在字體演變過程中經歷筆劃的嬗變亦可能是導致做法不同的原因之一。很可惜的是我們已無法一睹此1200「文」的面目了。然而，「字」是怎麼產生的呢？

根據許慎的說法是「形聲相益」，其實就是六書中「形聲」和「會意」二者。會意是形與形的相益；形聲是形與聲的組合。關於文字孳乳的過程，鄭樵在「六書略」序中云：「象形為本；形不可象，則屬諸事（指事）；事不可知，則屬諸意（會意和形聲）；意不可會，則屬諸聲（形聲、轉注和假借），聲則無不諧矣。」這段話把文字孳乳綿延寫得十分明晰。

在表三和表四中，分別舉了些會意和形聲的例子。由表可知，抽象含義的字越來越多，這說明了文字孳乳由實而虛的發展方向。在文字學中對六書有四體二用之說：「象形」、「指事」、「會意」和「形聲」是造文字的法則，是體；而「轉注」和「假借」是文字運的法則，是用。「轉注」是指互訓的字，約略可視為一群近義字之間，其義「相轉相為」的關係。「假借」是指一字數用的方法，約可視為一個字兼具數音數義時，其「依聲記事」而引申其義的關係。由表三與表四觀之，在中文裡對於一些不認識的文字，可以「望文生意」

是有道理的，因為中文構字之中，已經蘊含了許多知識，由此可略窺字義，而字與字之間也頗有形、音、義交互的關係可尋。

構字的發展

為了彌補古書失傳的遺憾和因應中文資訊處理的需要，自民國六十年以降，近人對中文構字的問題上也花費不少心血研究。民國六十二年謝清俊等人發表了496個「字根」，這組字根可組成4萬8千餘漢字。這個工作並未考量到文字的源流與孳乳，它純是在為了能根據字形檢索文字和能產生中文字形的需求下，用制式系統的技術所推導出漢字字形構成的基本元素。之後，這種做法引起了許多人的興趣，發展出許多「字根集合」（雖然所用之名稱各異）。這些理論與基礎上的探討，是今日許多新的輸入方法和設計字形產生器的基礎，同時亦引導著以標準鍵盤取代了已往大鍵盤的發展歷程。

由文字學方面著手研究這個問題的有：民國七十三年周何的「中文字根孳乳表稿」為代表。周教授依據「中文資訊交換碼」第一、二集中共2萬2千餘字，歸納出文字孳乳的「字根」共計1134個，其中「聲母」869個，「形母」265個。這個數字和失傳的鄭樵「象類書」中記載的頗為一致。

以上的討論說明了一件很重要的事情：漢字的數目雖然很多（目前已蒐集到的8萬個，尚不包括書法之變異），而且字數與日俱增〔此字之集合是開放的（open set）〕，然而漢字字形構成的基本元素（無論從「初文」、「文」，或「字根」的角度來分析），卻是有限且恆定的（closed set）。這個現象對中文資訊處理技術的發展有極其深遠的影響。

由以上的討論和舉例，我們已經瀏覽了中國文字的孳乳和構字的法則，在構字中所蘊含的知識究竟可不可以用既有的人工智慧技術來表達呢？在下文中且讓我們用些題例加以說明。

構字中的知識表達

由於文字和語言是依約、定、俗、成的法則演進的，文字的含義自然亦隨時間的步伐而轉變。因此，我們將以造字之初，文字孳乳和近代字書等三種資料為例，說明構字與文字間知識表達的結構。此所指的知識表達的結構是由人工智慧的角度來看的，和知識論、認知科學等所討論者不同（有心一探人工智慧中知識表達法的讀者，請參閱參考資料7）。

冰字由「冫」和「水」組成，是水凝凍狀（見表二）。冰這個字含有表示由水凍結成冰的過程的意思。這是先民對自然現象觀察的心得所造的字。水凍結成冰的現象可以用下列的語意網路表示：讀者將發現表示冰「字根式」的（1）和表示水和冰關係的語意網路（2），實是極為相似。換言之，若將構字中含的知識結構以語意網路（**semantic net**）表達，則該字的字根式至少可作為它語意網路之索引，或作為該知識庫控制功能之用。圖二表示一個較複雜的字，「愛」的語意網路，或更可助於讀者了解。若是我們用制式系統中的產生系統（**production system, rule based**）作為知識表達的工具，則可寫一條規則：若（**if**）水在0°C以下則（**then**）凝固（3）來表示水結冰的現象，而（1）式亦可為（3）式之索引。讀者試比較此（1）、（2）、（3）式間的關係與英文中 **ice** 和 **water** 之關係，自不難發現我們文字中蘊含之知識，實較西方拼音文字不可同日而語。由些例亦可知，若當我們有一如（3）

◇ c 成的物理知識系統時，（1）式或（2）式仍然可以與之配合使用。

〔例二〕字根「缶」之孳乳字系

現在，讓我們看一個更複雜的例子。在圖三中是由「缶」孳乳出一系列的示意圖，缶之本義是上面有蓋、有頸、有把手，下端為容體的容器，是一個象形字。由於缶是瓦器，因此缶、瓦、和●是屬可通義之字。

缶有六個主要的性質，其一是●酒「漿」用，此「漿」之意被借用而孳生出「菊」字，它是「●之含漿者」之意。由於「缶」需開蓋使用，故覆以●（見表三），而「●」成為「缶」之重文。又缶可「握持」，「握持」之意借於「●」字則為持刀之意。

缶之顏色有「灰」有「黃」，借這些顏色而產生「●」字，表示色如「缶」之馬也。

又缶為瓦器，故「●」象燒缶之「窯」，因而「●」亦與「缶」通。在窯中缶多「堆積」而燒，故生「陶」字，表示色如「缶」之馬也。

又缶之形狀渾圓宏張，囊之大者如缶，故有「囊」字。再者蟻封形狀如缶，合「囊」與「虫」產生「囊」字，即通「蟻」字。

缶中空，因而擊之有「聲」，借「聲」之意而生「●」字，彭●之聲也。又因有「聲」故可擊節為「歌」，借「歌」之意而生「謠」字。為歌則需「張口舒氣」，借此意生「●●」之詞表示出氣貌。又聞歌則「喜」，借而生「●」字，喜也。又歌則●譁吵人，故生「●」字。

缶在古時甚為貴重，故「寶」字中亦有「缶」焉。

這個缶字孳乳的字系是依據杜學知「文字孳乳考」而引來的，雖然杜教授的論點在文字學界尚有爭論，即使圖三中有些關連不盡符合文字孳生之真實過程，它還是很有用的，適合人工智慧的處理。

這個孳乳之字系，可用人工智慧中的語意網路，或是框架結構（frame）來表示。除了字的意義所含的知識以外，像燒●時坯是成堆放置的、中空之容器擊之有聲、唱歌時須張口舒氣...等等知識，盡含於此結構之中。若能設計有效之譯出程序（interpretive procedure），則可在此資料結構中尋出上述之知識。

另一個相似的例子是「不」字的字系（見圖四）。因限於篇幅，不再贅述（若有不明可參考資料6）。（圖表省略）。

〔例三〕缶部的字（國語日報辭典）

根據國語日報辭典缶部共10字。其中2字，「缺」與「●」表示「狀態」或「性質」。其餘8字為器物名稱，其間關如圖五所示。圖五亦可用語意網路或框架結構來表達。

根據圖五，我們可一眼看出，此圖與圖三（同是缶）所含之知識差異甚大，在此圖中，應可推導出瓦器皆有易「缺」及易生「●」之性質，若不然此二現象（狀態）之字不會從「缶」孳生。同理，瓦器可盛酒漿和固體，亦可為樂器，此等性質可由缶部器物類之8字中導出。

在表五中列有中華大辭典中木字部336字之分類表。此分類表雖現甚為粗略，然而木材之主要功用卻不言可知。此外，根據字義，關於木製器物之中，有關建築者39字，有關交通者18字（其中13字與舟楫有關），此外家具13字、容器12字、文具9字、敲打工具9字、棍棒7字、刑具及棺具各6字及武器4字等為較大宗者。由此可推知木材之主要用途。

在木字部中有「柴」字，其意為作為燃料之木材，（圖表省略），是否由此可推知木之可燃性呢？是否可由此推知「有些家具可燃」的知識呢？當然，這些知識均已蘊含在構字和字典的部首結構中，我們人可以明瞭，是不是電腦也可以聰明地推導出這些知識呢？則是尚得努力的研究題目。

總之，由上面的例子，我們可確信中國文字中蘊含的知識成分遠較拼音的文字為多，而且它有一套系統化的架構，若能將之電腦化，則其蘊藏之知識實可組成一極佳的常識知識庫（common sense knowlege base），而常識知識

庫，正是國外學者為解決其他人工智慧問題：如自然語文的處理及機器翻譯等研究工作，而急於想獲得的。

最佳的知識表達系統

先賢造字，肇始於對自然現象之觀察，此中涉及人體、天文、物理、化學、動物、植物等等科學領域。以現代科學的水準來觀察千百年前文字的演進，我們幾乎找不到違反科學方痕跡。這是多麼令人驚訝的事！由此可知，先賢造字均能屏除巫迷，在文字的結構裡表現出對事物真象客觀正確的體認。這種具體而微，詳盡周延的觀察功夫，深深契合科學方法的要領。對於人倫、文化與社會方面的觀察，也充分表現出持正不阿、善體天心的心本文化精神。由是觀之，亦無怪乎春秋時期有諸子百家的蔚蔚風流。在文字孳乳方面，則思路明晰，舉凡推論之邏輯、觀念之延申、聯想之法則、重點之把握，以及由實物進而引申到抽象意義的創意過程等等，精闢絕倫，無不表現出卓越的睿智和極其科學化的內涵。這種成就，發生在數千年前的上古時代，實在令人難以想像！

有些語言學者認為「音」與「義」是語言最主要的兩大要素，至於文字「形」則是次要的，它只是記錄語言的符號罷了。文盲亦可流利的使用語言，就是常用的佐證。這種說法，在西方拼音語系中尤顯得有力，因為拼音文字的「形」與「音」的關係是直接的，而「形」與「義」的關係就比較疏遠，「形」常須經由「音」的媒介與「義」相關連。然而，我們文字「形」與「義」的關係就不盡然如此。在中國文字結構中蘊藏了極豐富的知識，這些知識和字「義」緊密的結合在一起。換言之，我們的文字，「形」、「音」、「義」三者是密切相關無分親疏之分。其實，世界上任何一種語言都有豐富的知識內涵，然而中西語言表達知識的層次卻不相同。在我們的語言裡，即使在

構字的層面上，就已經把知識和語言緊密地結合在一起，更不用說構詞、構句和文章了。

由人工智慧的角度來看，我國文字的結構是一極佳的知識表達系統。雖然它是給人用的，然而它的結構卻與目前人工智慧中所了解的知識表達技術，一呼應。這是多麼令人拍案驚奇的因果關係！莊子與「得魚得筌」之語，他以「筌」喻語言，以「魚」喻語言要傳達的意念。我們今天來看看「筌」，卻有意料之外的收獲：若能善用我們文字的特質，對於計算機處理中國語文的研究，或將有獨樹一幟的發展。

參考資料

1. 謝清俊談中國文字中的知識結構民國七十六年五月三十一日於中研院史語所語言組例會之講稿
2. 林尹 文字學概說 正中書局印行 民國六十年初版
3. 齊佩□ 訓詁學概論 漢京公司印行 民國七十四年台初版
4. 王初慶 中國文字結構析論 文史哲出版社印行民國七十二年再版
5. 周何等 中文字根孳乳表稿 文建會國字整理小組印行 民國七十三年初版
6. 杜學知 文字孳乳考 世界書局印行 民國六十八年再版
7. A Barr & E.A.Feigenbaum, The Handbook of Artificial Intelligence, Vol.1, William Kaufman, Inc.,1981
8. 謝清俊等 中文字根之研究 交大學刊 民國六十二年
9. 王申倍 Intelligence in chinese character 中文計算機學會年會論文集 June,1987. (作者保留著作權)