

構字式的處理技巧

中研院資訊所 莊德明 96/04/10

一、構字符號

構字符號共有 13 個，分別為 Δ 、 \triangle 、 \triangleleft 、 \circ 、 $\circ\circ$ 、 $\circ\circ\circ$ 、 $\circ\circ\circ$ 、 $\circ\circ\circ$ 、 $\circ\circ\circ$ 、 $\circ\circ\circ$ 、 \square 、 \square ，說明如表一。這 13 個構字符號在 Big5 的編碼是連續的，分別為 8DF1(\triangle)-8DFD(\square)。

表一、構字符號

類別	符號	Big5	說明	構字式範例
連接	Δ	8DF2	當部件的連接順序由左至右	順 = 川 Δ 頁
	\triangle	8DF1	當部件的連接順序由上至下	含 = 今 \triangle 口
	\triangleleft	8DF3	當部件的連接順序由外至內	圍 = 口 \triangleleft 韋
部件序	\square	8DFC	按部件書寫順序輸入，前後以起始符號 (\square) 和終止符號 (\square) 包夾。	解 = \square 角 刀 牛 \square
	\square	8DFD		
方便符號	\circ	8DF4	二個相同部件直連	炎 = \circ 火
	$\circ\circ$	8DF7	三個相同部件直連	
	$\circ\circ$	8DF5	二個相同部件橫連	朋 = $\circ\circ$ 月
	$\circ\circ\circ$	8DF8	三個相同部件橫連	
	$\circ\circ\circ$	8DF6	三個相同部件呈三角狀排列	焱 = $\circ\circ\circ$ 火
	$\circ\circ\circ\circ$	8DFB	四個相同部件橫連	
	$\circ\circ\circ\circ$	8DFA	四個相同部件直連	
	$\circ\circ\circ\circ$	8DF9	四個相同部件呈四角狀排列	燚 = $\circ\circ\circ\circ$ 火

二、構字式的類型

構字式可分成五個類型，詳如表二。

表二、構字式的類型

類型	說明	字例	構字式	備註
一	由部件、橫連符號或方便符號構成	鉢	金△本	只要出現橫連符號即屬於此型
		剌	⊖水△刂	
		位	亻△竝	
		紉	纟△力△欠	
		瀾	氵△○屏△刂	
二	由部件、直連符號或方便符號構成	憑	馮△几	只要出現直連符號即屬於此型
		甕	○甬△花	
		杰	本△○○口	
		甕	宀△明△死	
		崇	品山△石△木	
三	由部件、包含符號或方便符號構成	磚	瓦△專	只要出現包含符號即屬於此型
		閨	門△品下	
四	由部件、起始標示、終結標示或方便符號構成	戡	形山灵戈□	只要出現起始標示及終結標示即屬於此型
		漚	形氵大品牛□	
		儂	形亻西域哲□	
		羈	形韭井孔女□	
		霽	形雲彡○○火鬲□	
五	只由方便符號構成	吉	○○吉	只有出現方便符號
		喜	⊖吉	
		喆	○○吉	
		雷	品雷	
		龍	○○龍	

三、構字式的擷取

構字式的五個類型，除了型四有明顯的起始和終結標示外，其餘的四個類型必須先找出起始和終結字元，其擷取類型詳如表三。

表三、構字式的擷取類型

類型	符號	字例	構字式	起始	終結	備註
一	△	鉢	金△本	金	本	當第一個構字符號為連接符時，則起始字元為△、△、△的前一個；終結字元為連續兩個非構字符號的第一個，例如「持金△本入舍衛大城」之「本」。
		位	亻△竝	亻	竝	
		紉	糸△力△欠	糸	欠	
		瀾	氵△〇屏△丿	氵	丿	
	△	馮	馮△几	馮	几	
		杰	本△〇〇口	本	口	
		甍	宀△明△死	宀	死	
△	磚	瓦△專	瓦	專		
	閑	門△品下	門	下		
二	⊠	戡	⊠山灵戈⊠	⊠	⊠	當第一個構字符號為⊠時，則起始字元為⊠，終結字元為⊠。
		漣	⊠氵大品牛⊠	⊠	⊠	
		僂	⊠亻西域哲⊠	⊠	⊠	
		羈	⊠韭井孔女⊠	⊠	⊠	
		羸	⊠雲彡〇火鬲⊠	⊠	⊠	
三	〇	剌	〇水△丿	〇	丿	當第一個構字符號為方便符號時，則起始字元為方便符號，終結字元為連續兩個非構字符號的第一個。
	〇〇	甍	〇〇甬△花	〇〇	花	
	品	崇	品山△石△木	品	木	
	〇〇	吉	〇〇吉	〇〇	吉	
	〇	喜	〇吉	〇	吉	
	〇〇	喆	〇〇吉	〇〇	吉	
	品	雷	品雷	品	雷	
	〇〇	龍	〇〇龍	〇〇	龍	

四、構字式在漢字構形資料庫中的結構

構字式在漢字構形資料庫中的結構，詳如表四。

表四、構字式在資料庫中的結構

類型	說明	字例	構字符號		部件序
一	由部件、橫連符號或方便符號構成	鉢	△	1	金本
		剌	△	1	∪水丿
		位	△	1	亻竝
		紉	△	1	纟力欠
		瀾	△	1	氵∪屏丿
二	由部件、直連符號或方便符號構成	憑	△	2	馮几
		甕	△	2	∪甬花
		杰	△	2	本∪∪口
		覓	△	2	宀目死
		崇	△	2	∪山石木
三	由部件、包含符號或方便符號構成	甌	△	3	瓦專
		閑	△	3	門∪下
四	由部件、起始標示、終結標示或方便符號構成	戡	形□	4	山𠃉戈
		漣	形□	4	氵大∪牛
		僂	形□	4	亻西域哲
		𦍋	形□	4	韭井卂女
		霽	形□	4	雲效∪火鬲
五	只由方便符號構成	喆		5	∪吉
		喜		5	∪吉
		喆		5	∪∪吉
		雷雷		5	∪雷
		龍龍龍		5	∪∪龍

漢字構形資料庫中有關字形結構的欄位，共有七個，分別是「連接符號」、「部件序」、「字根序」、「字根組」、「部件序二」、「字根序二」、「字根組二」。連接符號、部件序、字根序、字根組的使用舉例如表五。

表五、「灩」字在漢字構形資料庫中的相關結構(一)

編號	字例	連接符號	部件序	字根序	字根組
1	一	0	一	一	一
51	尸	0	尸	尸	尸
60	卅	0	卅	卅	卅
376	并	2	丿开	丿一卅	一丿卅
1206	屏	3	尸并	尸丿一卅	一丿卅尸
18454	丿	0	丿	丿	丿
21859	卂	0	卂	卂	卂
23463	开	2	一卅	一卅	一卅
33121	灩	1	氵 〇〇屏 卂	氵 尸 丿 一 卅 尸 丿 一 卅 卂	一一 卂 丿 丿 卅 卅 氵 尸 尸
63471	氵	0	氵	氵	氵

以下是表五的補充說明：

1. 字根即是不再拆分的部件，例如表五中的「一」、「尸」、「卅」、「丿」、「卂」、「氵」。
2. 表二列出的五個構字式類型，「連接符號」欄位對應的值分別是1-5，而0表示字根。
3. 字根序可經由部件序推得，例如「屏」的字根序可推導如下：
屏→尸并→尸丿开→尸丿一卅
4. 字根組可由字根序重新排序而得，例如依照字根的筆畫及筆順，「屏」的字根序「尸丿一卅」重新排序成「一丿卅尸」，即為字根組。

至於部件序二、字根序二、字根組二的使用舉例如表六。

表六、「瀾」字在漢字構形資料庫中的相關結構(二)

編號	字例	連接符號	部件序二	字根序二	字根組二
1	一	0	一	一	一
51	尸	0	尸	尸	尸
60	卅	0	卅	卅	卅
376	并	2	八开	八一卅	一八卅
1206	屏	3	尸并	尸八一卅	一八卅尸
18454	丷	0	八	八	八
21859	刂	0	刀	刀	刀
23463	开	2	一卅	一卅	一卅
33121	瀾	1	水〇〇屏刀	水尸八一卅尸 八一卅刀	一一八八刀卅卅尸 尸水
63471	氵	0	水	水	水

「部件序」、「字根序」、「字根組」及「部件序二」、「字根序二」、「字根組二」等欄位的差異在於「異體字根」。異體字根的例子，如表六中的「水」及「氵」、「刀」及「刂」、「八」及「丷」為異體字根。表六的補充說明如下：

1. 部件序可經由異體字根的取代轉成部件序二，例如表六中「瀾」的部件序二「水〇〇屏刀」是表五部件序「氵〇〇屏刂」取代異體字根「氵」、「刂」成「水」、「刀」而得。
2. 字根序可經由異體字根的取代轉成字根序二，例如表六中「屏」的字根序二「尸八一卅」是表五字根序「尸丷一卅」取代異體字根「丷」成「八」而得。
3. 字根組二可由字根序二重新排序而得，例如依照字根的筆畫及首筆順序，「屏」的字根序二「尸八一卅」重新排序成「一八卅尸」，即為字根組二。

五、構字式的正規化

構字式可用來表達缺字，而其最大的優點是不需要查表，每個人所表達的通常是一致的。例如「鉢」、「憑」、「甕」的構字式分別為「金△本」、「馮△几」、「瓦△專」，幾乎不會有不同的表達式。但是部分字形由於結構較為複雜，或是允許使用者有更大的表達彈性，仍可出現不同的構字式，詳見表七。

表七、構字式及可能的構字式

類型	字例	構字式	可能的構字式
一	剌	⊗ 水 △ 刂	⊗ 水 水 刀 ⊙
	伫	亻 △ 立	人 △ ⊙ 立
	紉	纟 △ 力 △ 欠	纟 △ 力 △ 欠
	瀾	氵 △ ⊙ 屏 △ 刂	水 △ 屏 △ 屏 △ 刀、⊗ 水 屏 屏 刀 ⊙
二	甬	⊙ 甬 △ 花	⊗ 甬 甬 花 ⊙
	杰	本 △ ⊙ 口	⊗ 本 口 口 口 ⊙
	甕	宀 △ 目 △ 死	宀 △ ⊙ 目 △ 死
	樂	⊙ 山 △ 石 △ 木	⊗ 山 山 山 石 木 ⊙
三	閤	門 △ ⊙ 下	⊗ 門 下 下 下 ⊙
四	戡	⊗ 山 灵 戈 ⊙	⊗ 山 日 火 戈 ⊙
	滄	⊗ 氵 大 ⊙ 牛 ⊙	⊗ 水 大 牛 牛 牛 ⊙
	僂	⊗ 亻 西 域 哲 ⊙	⊗ 人 西 域 哲 ⊙、⊗ 西 域 哲 人 ⊙
	羈	⊗ 韭 井 卂 女 ⊙	⊗ 韭 卂 井 女 ⊙
	霽	⊗ 雲 彡 ⊙ 火 鬲 ⊙	⊗ 雨 彡 云 彡 火 鬲 火 ⊙
五	吉	⊙ 吉	吉 △ 吉
	喜	⊗ 吉	吉 △ 吉
	喆	⊙ 吉	吉 △ 吉 △ 吉、⊗ 吉 吉 吉 ⊙
	雷	⊙ 雷	雷 △ ⊙ 雷、⊗ 雷 雷 雷 ⊙
	龍	⊙ 龍	⊗ 龍、⊙ 龍 △ ⊙ 龍、⊗ 龍 △ ⊗ 龍

構字式是由構字符號和部件組成，而其中最重要的仍是部件。漢字構形資料庫中有關字形結構的欄位中，欄位連接符號及部件序是用來表達構字式，而字根序、字根組、部件序二、字根序二、字根組二等欄位可作為構字式正規化之用，尤其以字根序二及字根組二最為常用。表八由表七中各挑幾個字例，說明如何利用欄位字根序二及字根組二來作構字式的正規化。例如「覓」的構字式「宀△○○目△死」及「形宀目目死□」可利用欄位字根序二或字根組二正規化成「宀△目△死」；而「僂」的構字式「形西域哲人□」只能利用欄位字根組二正規化成「形人西域哲□」。

表八、構字式的正規化

字例	構字式	字根序二	字根組二	正規化欄位
瀾	氵△△○○屏△△リ	水尸八一卅 尸八一卅刀	一一八八刀 卅卅尸尸水	字根序二或 字根組二
	水△屏△屏△刀			
	形水屏屏刀□			
覓	宀△△目△死	宀目目歹匕	匕宀歹目目	字根序二或 字根組二
	宀△△○○目△死			
	形宀目目死□			
開	門△△品下	門一卜一卜 一卜	一一一卜卜 卜門	字根序二或 字根組二
	形門下下下□			
戡	形山灵戈□	山日火戈	山戈日火	字根序二或 字根組二
	形山日火戈□			
僂	形人西域哲□	人西土戈口 一手斤口 西土戈口一 手斤口人	一人土口口 戈手斤西	字根序二或 字根組二 字根組二
	形人西域哲□			
	形西域哲人□			
龍	龍	龍龍龍龍	龍龍龍龍	字根序二或 字根組二
	龍			
	龍△龍			
	○○龍△○○龍			
	龍△龍			
	形龍龍龍龍□			

安裝漢字構形資料庫光碟後，可在安裝目錄找到檔案cdphanzi.mdb。cdphanzi.mdb為Microsoft Access格式的資料庫，其中和本文相關的資料表有三，分別是「檢字表」、「字根」及「異寫字根」。

檢字表包含連接符號、部件序、字根序、字根組、部件序二、字根序二、字根組二等欄位。字根資料表包含編號、字形等欄位，其編號是依字形筆畫、筆順排序。異寫字根資料表則列出字根間的異寫關係。

利用上述的資料表，構字式正規化的程序如下：

1. 利用檢字表的連接符號及部件序欄位，查核目前的構字式是否已收錄在檢字表中；若尚未收錄，則利用以下程序進行構字式的正規化。
2. 利用檢字表的字根序二欄位，將構字式中的每個部件換成字根序，再利用字根序二欄位查核此構字式的字根序是否已收錄在檢字表中；若尚未收錄，則利用程序三繼續進行構字式的正規化。
3. 利用字根資料表將程序二的字根序重新排序，再利用字根組二欄位查核此構字式的字根組是否已收錄在檢字表中；若尚未收錄，則該字應為新的缺字。



上述的正規化程序，可先以構字式「宀𠂇〇〇目𠂇死」為例來說明：

1. 構字式「宀𠂇〇〇目𠂇死」的連接符號為2（直連），部件序為「宀〇〇目死」，核對檢字表的連接符號及部件序兩個欄位，此一組合並未出現。
2. 在檢字表的字根序二欄位中，「宀」、「目」、「死」三個字的值分別為「宀」、「目」、「歹匕」，所以部件序「宀〇〇目死」可轉成字根序「宀目目歹匕」。核對檢字表的字根序二欄位，「宀目目歹匕」對應的連接符號為2，部件序為「宀目死」，即「覷」的構字式「宀𠂇𠂇死」。

以下再以構字式「形西域哲人◻」為例繼續說明正規化的程序：

1. 構字式「形西域哲人◻」的連接符號為4，部件序為「西域哲人」，核對檢字表的連接符號及部件序兩個欄位，此一組合並未出現。
2. 在檢字表的字根序二欄位中，「西」、「域」、「哲」、「人」四個字的值分別為「西」、「土戈口一」、「手斤口」、「人」，所以部件序「西域哲人」可轉成字根序「西土戈口一手斤口人」。核對檢字

表的字根序二欄位，此字根序並未出現。

- 表九為字根資料表中的字根排序，依此排序，字根序「西土戈口一手斤口人」可再轉成字根組「一人土口口戈手斤西」。核對檢字表的字根組二欄位，「一人土口口戈手斤西」對應的連接符號為4，部件序為「亻 西 域 哲」，即「儼」的構字式「亻 西 域 哲」。

表九、字根排序

字根	一	人	土	口	戈	手	斤	西
編號	1	69	147	197	393	453	482	1057

六、缺字的呈現

安裝漢字構形資料庫光碟後，可在安裝目錄找到表九的缺字字型檔。每套缺字字型只用到五大碼的外字區，最多只能編到 6,217 個字。

表九、缺字字型檔

標楷體缺字字型	檔案名稱	細明體缺字字型	檔案名稱	字體
標楷體部件外字	cdpeudck.tte	細明體部件外字	cdpeudc.tte	0
標楷體外字集一	hzcdp01k.ttf	細明體外字集一	hzcdp01m.ttf	1
標楷體外字集二	hzcdp02k.ttf	細明體外字集二	hzcdp02m.ttf	2
標楷體外字集三	hzcdp03k.ttf	細明體外字集三	hzcdp03m.ttf	3
標楷體外字集四	hzcdp04k.ttf	細明體外字集四	hzcdp04m.ttf	4
標楷體外字集五	hzcdp05k.ttf	細明體外字集五	hzcdp05m.ttf	5
標楷體外字集六	hzcdp06k.ttf	細明體外字集六	hzcdp06m.ttf	6
標楷體外字集七	hzcdp07k.ttf	細明體外字集七	hzcdp07m.ttf	7
標楷體外字集八	hzcdp08k.ttf	細明體外字集八	hzcdp08m.ttf	8
標楷體外字集九	hzcdp09k.ttf	細明體外字集九	hzcdp09m.ttf	9

cdphanzi.mdb 的檢字表中，和缺字字型相關的欄位為「字體」、「字碼」，欄位字體所對應的字型可參考 cdphanzi.mdb 的字體資料表。依表九，若將該字碼的字型換成對應的字體，即可顯示該字的字形。例如「覓」對應的字體為標楷體外字集一，字碼為 94C9；「儼」對應的字體為標楷體外字集一，字碼為 9273。

表九中的標楷體及細明體部件外字，即是一般所謂的造字檔，其中細明體部件外字為系統預設外字集，而標楷體部件外字目前只對應到標楷體。檢字表字體欄位為 0 者，不見得就是部件外字，而可能是五大碼的標準字，兩者可透過欄位字碼來區分。五大碼的外字區分成

四段，分別為第一段 FA40-FEFE 共 785 個，第二段 8E40-A0FE 共 2983 個，第三段 8140-8DFE 共 2041 個，第四段 C6A1-C8FE 共 408 個，合計 6217 個。