

漢字構形資料庫的建置與應用

莊德明 中央研究院資訊科學研究所

derming@gate.sinica.edu.tw

謝清俊 玄奘大學講座教授

hsieh@sinica.edu.tw

摘要

目前電腦處理漢字的諸多缺失，例如缺字、異體字等問題，主要的原因在於電腦裡的漢字知識嚴重不足。有鑑於此，中央研究院資訊所文獻處理實驗室自 1993 年起，即先由字形著手，建置漢字構形資料庫。

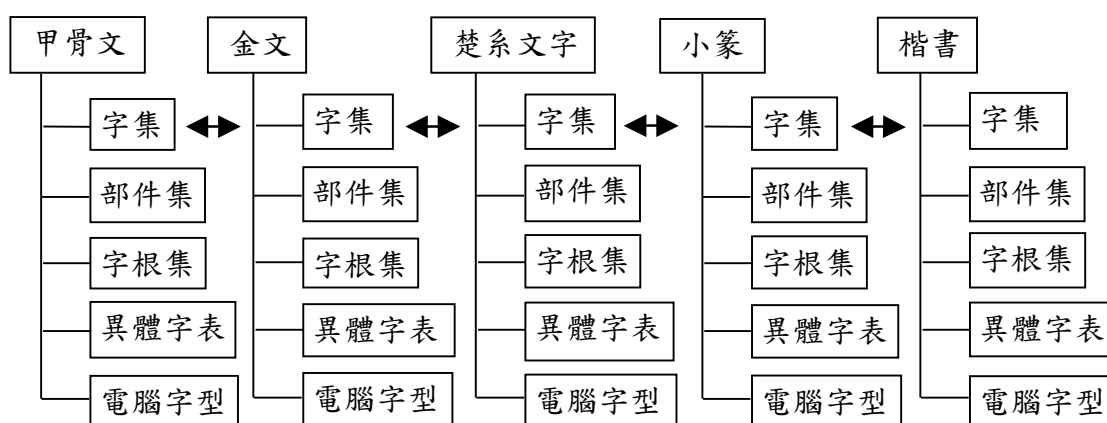
漢字構形資料庫早期收錄的字形是以楷書的現代印刷字體為主，其後陸續增加小篆、金文、甲骨文、楚系文字等古漢字。收錄這些不同歷史時期的漢字，除了要作字形、字義銜接外，還要依據不同的形體來作構形分析。字形的銜接是依據字形的演變，在電腦中使用相同的編碼位置，編入不同的字型。字義的銜接是參考字義的隸屬，在電腦中使用不同的編碼位置，編入異體字表。不同歷史時期的漢字構形分析，雖然不見得合乎構形理據，但是在字形的檢索上也有一定的功能，這也符合漢字『以義構形』的使用心理。

漢字構形資料庫目前最主要的應用是用來解決缺字問題。缺字的問題在於漢字的字集是一個開放性質的，它的字數根本不適合作固定數量的限定；這與數量已定的西方語言的『字母集』，是不可以一概而論的。然而，現行漢字交換碼的結構，卻仿照西方語言的字母集的結構來設計，這不能不說是『削足適履』。相對於現行漢字交換碼利用字碼來區別漢字，以致缺字問題層出不窮，我們認為漢字的差異在於字形，缺字問題的解決應當由字形結構著手，才是根本解決之道。

壹、漢字構形資料庫簡介

目前電腦處理漢字的諸多缺失，例如缺字、異體字等問題，主要的原因在於電腦裡的漢字知識嚴重不足。有鑑於此，中央研究院資訊所文獻處理實驗室自1993年起，即先由字形著手，建置漢字構形資料庫。漢字構形資料庫早期收錄的字形是以楷書的現代印刷字體為主，其後陸續增加小篆、金文、甲骨文、楚系文字等古漢字。在2004年12月份推出的漢字構形資料庫2.2版中，收錄楷體字形59,220個、小篆及重文11,100個、金文3,459個、甲骨文177個、楚系文字372個，另外還包含了異體字12,681組及相關應用程式。這個系統目前只能安裝在微軟中文視窗(繁體版)，有興趣的讀者可由文獻處理實驗室的網址下載：<http://www.sinica.edu.tw/~cdp/>。

目前漢字構形資料庫是由甲骨文、金文、楚系文字、小篆及楷書構形資料庫組合而成，如圖一。從圖一可看到每個構形資料庫都有各自的字集、部件集、字根集、異體字表及電腦字型，各個字集間彼此也有銜接。簡單的說，漢字構形資料庫的主要目標有以下兩點：一、銜接古今文字。二、提供不同時期漢字的構形分析。這兩點將分別在第二、三節說明。



圖一、漢字構形資料庫的組成

早期楷書構形資料庫的建置主要是延續交通大學在1972年發展的字根系統，並以《中文電腦基本用字》的8,532個字為對象，隨後陸續擴充到《中文大辭典》的49,905個字、《漢語大字典》的54,678個字。在古漢字的構形資料庫中，最早完成的為小篆構形資料庫，這個資料庫能順利完成有以下三個原因：一、楷書構形資料庫在1999年初已收錄《漢語大字典》的字頭及異體字表，而小篆和楷體的銜接可由《漢語大字典》字頭下的字形源流演變及異體字表中得知。二、1999年2月適時取得北京師範大學小篆字型，經過我們重新編碼後，並於2002年底取得北師大的授權，可使用在Big5的中文電腦系統。三、2001年初和台灣師範大學合作，依據《說文解字》建立小篆的構形分析。當小篆構形資料庫於2003年3月完成後，我們在中研院史語所的協助下，陸續建置金文、甲骨文、楚系文字構形資料庫。表一分別列出這些資料庫建置的相關資料，包含參考的字集、預定建置的字數、建置日期及預定完成日期、部件及異體字的個數、採用的電腦字型、提供的索引、合作的單位等。

表一、漢字構形資料庫的建置

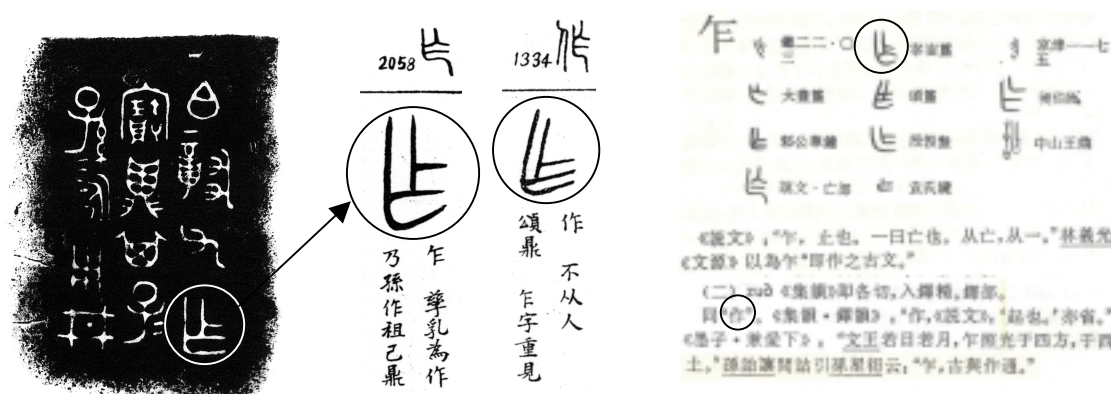
| | 甲骨文 | 金文 | 楚系文字 | 小篆 | 楷書 |
|---------------|------------------------|-------------------------|-----------------|---------------|-------------------------------|
| 主要參考字書或字集 | 殷墟甲骨刻辭類纂 | 金文編 | 楚系簡帛文字編 | 說文解字詁林 | 中文電腦基本用字、中文大辭典、漢語大字典、殷周金文集成引得 |
| 預定建置字頭數 | 4,488 | 3,771 | 2,228 | 9,831 | |
| 預定建置字形數 | (未定) | 20,489 | 19,250 | 11,100 | (未定) |
| 預定建置異構字數 | 約 4,500 | 約 5,000 | 約 3,000 | | |
| 已建置字數 | 177 | 3,459 | 372 | 11,100 | 59,220 |
| 已建置異體字組數/字數 | | 875/2,059 | | 1,081/2,350 | 12,681/38,902 |
| 開始建置日期(年/月) | 2004/7 | 2003/10 | 2004/10 | 1999/2 | 1993/12 |
| 預定或已完成日期(年/月) | (未定) | 2005/12 | 2005/12 | 2003/3 | |
| 已分析字形的部件數/字根數 | | 804/469 | | 2,004/367 | (整理中) |
| 電腦字型 | 中研院甲骨文 | 中研院金文 | 中研院楚系文字 | 北師大說文小篆 | 標楷體及細明體外字集 |
| 內建索引 | 殷墟甲骨刻辭類纂、甲骨文字詁林、甲骨文字集釋 | 金文編、金文詁林、殷周金文集成引得、器號及器名 | 楚系簡帛文字編、出土墓號及簡號 | 說文解字詁林 | 漢語大字典、中文大辭典、Unicode、Big5 |
| 合作單位 | 中研院史語所 | 中研院史語所 | 中研院史語所 | 北京師範大學、台灣師範大學 | 中研院史語所 |

除了資料庫的建置外，隨著 Unicode 的逐漸普及，我們也同時在開發 Unicode 版的漢字構形資料庫。累積了幾年古漢字構形資料庫建置的經驗後，我們發覺目

前也是回頭修正楷書構形資料庫的大好時機。楷書構形資料庫可在 2005 年 3 月底前修正，而 Unicode 版的漢字構形資料庫希望能在 2005 年 6 月底前推出。

貳、古今文字的銜接

本節說明古今文字的銜接。圖二是伯斝父鼎的拓片，我們以拓片中的「𠄎」為例，說明漢字構形資料庫中金文、小篆和楷體間的銜接機制。圖三是《金文編》和「𠄎」相關的小篆字頭和楷定字，圖四為《漢語大字典》字頭「乍」的字形源流演變和部分釋義。圖三的《金文編》指出「𠄎」對映的小篆字頭為「𠄎」或「𠄎」，而「𠄎」或「𠄎」隸定成「乍」或「作」。圖四由《漢語大字典》「乍」的釋義可得知「乍」、「作」為古今字。



圖二、伯斝父鼎拓片

圖三、金文編的「𠄎」

圖四、漢語大字典的「乍」

在漢字構形資料庫中，古今文字的銜接分成字形的銜接及字義的銜接兩種。字形的銜接是依據字形的演變，在電腦中使用相同的編碼位置，編入不同的字型；字義的銜接是參考字義的隸屬，在電腦中使用不同的編碼位置，編入異體字表。例如「𠄎」、「𠄎」、「乍」的 Big5 碼同樣是 A545，但是隸屬的字型分別是中研院金文、北師大說文小篆、標楷體；而「作」的 Big5 碼為 A740，在《漢語大字典》的異體字表中，「作」為主體字，「乍」為主體字。北師大說文字型原為北京師範大學研發，而中研院資訊所重新編碼；至於中研院金文字型為中研院史語所和資訊所合作開發，是將《金文編》的字形掃描切割後再轉成 Truetype 字型。

表二列出伯斝父鼎中的每個金文在《金文編》對映的小篆字頭及楷定字形，釋文記成『白（伯）斝父乍（作）／寶鼎，其子子／孫孫永用。』其中金文「斝」為《說文》所無。

表二、伯斝父鼎的金文和小篆、楷體的銜接

| | | | | | | | | | | | | |
|----|---|---|---|---|---|---|---|---|---|---|---|---|
| 金文 | 白 | 斝 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 |
| 小篆 | 白 | 伯 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 | 𠄎 |
| 楷體 | 白 | 伯 | 斝 | 乍 | 作 | 寶 | 鼎 | 其 | 子 | 孫 | 永 | 用 |

事實上在處理字形銜接時，也並不總是如表二那麼順利，常見的一個問題就

是無法在字書裡找到對映的楷定字。例如《金文編》字頭「寶」下的金文有 273 個，其中異構字有 20 個如表三。這些金文的楷定字常散布在相關字書中，有時還得自行楷化。

表三、金文「寶」的異構字及對映的楷化字

| | | | | | | | | | | | | | | | | | | | | |
|----|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 金文 | | | | | | | | | | | | | | | | | | | | |
| 小篆 | 寶 | | | | | | | | 寶 | | | | | | | | | | | |
| 楷體 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 | 寶 |

接著說明異體字表的處理。1999 年初漢字構形資料庫已納入《漢語大字典》的異體字表，這個異體字表是採用由主體字統領異體字的編排方法，將同一主體字統領的簡化字、古今字、全同異體字(指音義全同而形體不同的字)和非全同異體字(指音義部分相同的異體字)，集中在該主體字下編為一組，共收 12,208 組，包含 36,309 個字形。然而這個異體字表除了標明簡化字外，其餘的異體字並未作說明。

隨後開始建置小篆構形資料庫，由於小篆重文都已收錄在《漢語大字典》的異體字表，所以並無增補異體字表的需求。直到 2003 年底建置金文構形資料庫時，部分金文的異體字《漢語大字典》並未收錄，才有了增補異體字表的想法。例如《金文編》字頭「乍」下有「𠄎」、「𠄏」、「𠄐」、「𠄑」四個字，楷定作「乍」、「𠄎」、「詐」、「𠄏」，其中「𠄏」《漢語大字典》並未收錄，而「乍」、「𠄎」、「詐」並未編入「乍」的異體字。增補後的異體字表共含異體字 12,681 組，包含 38,902 個字形。增補異體字表的同時，我們也對古漢字加了標示，這些標示、已標示字數及說明如表四。

表四、異體字標示、已標示字數及說明

| 標示 | 字數 | 字例 | | |
|------|-------|----|----|---------------------------|
| | | 異體 | 主體 | 說明 |
| 說文小篆 | 2,585 | 然 | 燃 | 「然」在《說文》的本義為「燃燒」 |
| | | 𠄎 | 便 | 「𠄎」為小篆「𠄎」的楷化，而「便」為隸定字。 |
| 說文或體 | 468 | 灾 | 災 | 「灾」為《說文》「裁(災)」的或體字 |
| 說文古文 | 570 | 求 | 裘 | 「求」為《說文》「裘」的古文 |
| 說文籀文 | 263 | 鵬 | 雕 | 「鵬」為《說文》「雕」的籀文 |
| 說文奇字 | 3 | 无 | 無 | 「无」為《說文》「無」的奇字 |
| 金文 | 2619 | 乍 | 作 | 「乍」、「𠄎」為《金文編》「作」的重文 |
| | | 𠄎 | | |
| 甲骨文 | 12 | 帝 | 禘 | 甲骨文「禘」作「帝」 |
| 楚系文字 | 80 | 𠄎 | 秋 | 「𠄎」、「𠄎」為《楚系文簡帛文字編》「秋」的異體字 |
| | | 𠄎 | | |

表五、「災」的異體字標記及釋義

| 說文字頭 | 小篆 | 楷體 | 標記 | 《漢語大字典》釋義 | 備註 |
|------|----|------|----------|-----------|-----------|
| 𤇑 | 𤇑 | 𤇑 | 說文小篆 | 同「災」。 | 「𤇑」的隸定字 |
| | | 𤇑 | 說文小篆 | 同「𤇑(災)」。 | 「𤇑」的楷化字 |
| | 𤇑 | 災 | 說文或體 | 同「𤇑(災)」。 | 也是「災」的簡化字 |
| | 𤇑 | 𤇑 | 說文古文 | 同「災」。 | |
| | 𤇑 | 災 | 說文籀文 | 自然發生的火災 | 「災」的隸定字 |
| 災 | | 說文籀文 | 同「𤇑(災)」。 | 「災」的楷化字 | |
| 𤇑 | 𤇑 | 𤇑 | 說文小篆 | 災害。後作「災」。 | |

異體字表可充分反映出漢字一字多形的特色。例如「災」的異體字有 18 個：「灾」、「𤇑」、「𤇑」、「𤇑」、「𤇑」、「𤇑」、「𤇑」、「災」、「災」、「𤇑」、「𤇑」、「燭」、「燭」、「𤇑」、「𤇑」、「𤇑」、「𤇑」、「𤇑」，其中 6 個和《說文》相關。這些異體字在《說文》原為兩組，字頭分別是「𤇑」、「𤇑」，而「𤇑」、「𤇑」、「災」為「𤇑」的或體、古文、籀文。這些異體字依《說文》重新整理如表五。

參、漢字構形分析

字形分析是漢字構形資料庫的核心，也是最花時間的一部分。不同歷史時期的漢字由於字形的遞變，分析方式也不盡相同。例如金文「𤇑」字，本義為『螢火或鬼火』，从「大」、「𤇑」，字形中的四點像螢光或鬼火閃爍；然而「𤇑」對映的小篆作「𤇑」，《說文》分析成从「災」、「𤇑」，已不見閃爍之意；後來「𤇑」隸定成「𤇑」，「災」變成「米」，連螢火都不見了。雖然「𤇑」分析成从「災」、「𤇑」，「𤇑」分析成从「米」、「𤇑」，都不合乎構形理據，但是在字形的檢索上也有一定的功能。例如「𤇑」《說文》歸在「災」部，「𤇑」《漢語大字典》歸在「米」部。在保存各個歷史時期漢字形體的同時，同時記錄不同形體的構形分析，這也符合漢字『以義構形』的使用心理。

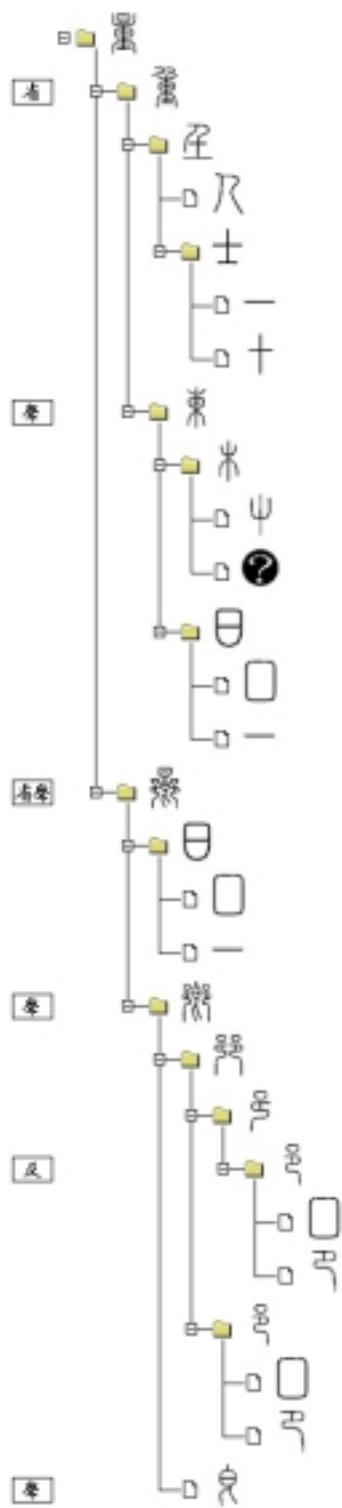
在漢字構形資料庫中，小篆構形資料庫毋寧是最重要的，而小篆的構形分析則如實反映許慎在《說文解字》的釋形。例如圖五為「量」字在《說文》的結構，表六則為「量」字部件在《說文》的釋義，我們同時附上《漢語大字典》的引用資料及補充說明。小篆的構形分析可分以下幾點說明：

一、《說文解字詁林》收錄 9,831 個小篆及 1,269 個重文，合計 11,100 個字形。這些字形依據《說文》的釋形來分析，部件總數為 2,030 個，其中有 51 個為《說文解字詁林》未收錄的成字部件。例如「笛」字釋形為『从竹、由聲』，而《說文解字詁林》並未收錄「由」字。

二、值得一提的是，2,030 個部件中只有 367 個成字字根，這和李佳信在《說文小篆字根研究》一書指出的 546 個字根相差甚多，主要原因在於凡是《說文》釋形有「从某」的字，我們就把「某」作部件拆分，以致於《說文小篆字根研究》一書中的很多字根都拆分了。例如「木」字，《說文》釋形『从中，下象其根。』於是在圖五的「木」字結構下出現「中」，而『下象其根』這部分僅以❶表示。

表六、「量」字部件的相關釋義

| 號 | 篆 | 楷 | 釋形 |
|----|---|---|---|
| 1 | 量 | 量 | 《說文》：稱輕重也。从重省，鄉省聲。 |
| 2 | 重 | 重 | 《說文》：厚也。从壬，東聲。 林義光《文源》：人挺立於地，為厚重象。 |
| 3 | 壬 | 壬 | 《說文》：善也。从人士。士，事也。一曰象物出地挺生也。 《漢語大字典》：甲骨文象人挺立土上之形。 |
| 4 | 人 | 人 | 《說文》：天地之性最貴者也。此籀文，像臂脛之形。 《漢語大字典》：甲骨文象人側面站立形。 |
| 5 | 士 | 士 | 《說文》：事也，數始於一，終於十。从一，从十。 |
| 6 | 一 | 一 | 《說文》：惟初太始，道立於一，造分天地，化成萬物。 《漢語大字典》：古文字一至四橫畫表示數字一至四，是原始記數符號。 |
| 7 | 十 | 十 | 《說文》：數之具也。一為東西，丨為南北，則四方中央備矣。 于省吾《甲骨文字釋林》：字初形本為直畫，繼而中間加肥，後則加點為飾，又由點孳化為小橫。 |
| 8 | 東 | 東 | 《說文》：動也。从木，官溥說，从日在木中。 《漢語大字典》：甲骨文象實物囊中括其兩端之形，為「橐」的初文。後世借為『東方』之東。 |
| 9 | 木 | 木 | 《說文》：冒也，冒地而生。東方之行。从中，下象其根。 王筠釋例：木固全體象形字也。 |
| 10 | 日 | 日 | 《說文》：實也，太陽之精不虧。从口、一，象形。 |
| 11 | 口 | 口 | 《說文》：回也，象回市之形。 |
| 12 | 鄉 | 鄉 | 《說文》：不久也。从日，鄉聲。 |
| 13 | 鄉 | 鄉 | 《說文》：國離邑，民所封鄉也。嗇夫別治，封圻之內六鄉，六鄉治之，从罷，邑聲。 |
| 14 | 罷 | 罷 | 《說文》：鄰道也。从邑，从邑。 《甲骨文編》：象二人相向之形。 |
| 15 | 邑 | 邑 | 《說文》：从反邑，罷字从此。 |
| 16 | 邑 | 邑 | 《說文》：國也，从口。先王之制，尊卑有大小，从卩。 《漢語大字典》：从口象疆域。(下面)象人跽形，乃人之變體，即指人民。有土有人，斯成一域。 |
| 17 | 卩 | 卩 | 《說文》：瑞信也，守國者用玉卩，……，象相合之形。 《漢語大字典》：甲骨文像人跪坐之形。後作『符節』之『節』。 |
| 18 | 良 | 良 | 《說文》：穀之馨香也，象嘉穀在裹中之形，匕所以扱之。 |



圖五、「量」字在《說文》的結構

三、小篆的部件功能區分可依《說文》的釋形加上『省』、『省聲』等標示，這些標示及包含標示的字數、字例、釋形如表七。

表七、小篆的部件標示、包含標示字數及字例釋形

| 標示 | 字數 | 字例 | 《說文》釋形 |
|-----|------|----|-----------------|
| 聲 | 8083 | 重 | 厚也。从壬，東聲。 |
| 省 | 159 | 量 | 稱輕重也。从重省，鄉省聲。 |
| 省聲 | 327 | 量 | 稱輕重也。从重省，鄉省聲。 |
| 亦聲 | 220 | 政 | 正也。从支，从正，正亦聲。 |
| 省亦聲 | 1 | 季 | 少侑也。从子，从稚省，稚亦聲。 |
| 反 | 28 | 𠂔 | 步止也。从反彳。 |
| 倒 | 6 | 尾 | 微也，从到毛在尸後。 |
| 二 | 64 | 艸 | 百艸也。从二中。 |
| 三 | 28 | 品 | 眾庶也。从三口。 |
| 四 | 4 | 𦰇 | 眾艸也。从四中。 |
| 半 | 6 | 支 | 去竹之枝也。从手持半竹。 |

其他古漢字的分析是由中研院史語所文字組協助進行，目前以金文的進度最快，《金文編》正文的字形大約已分析完畢。相對於圖五「量」字在《說文》的結構，表八列出金文「量」的構形分析。相對於小篆，金文的構形分析說明如下：

表八、金文「量」的構形分析

| 小篆 | 金文(楷體) | 金文(楷體)字根 |
|----|--------|----------|
| 量 | 量(量) | 日東土 |
| | 量(景) | 日東 |

一、《金文編》正文的異構字共計 3,459 個字形，分析後的部件數為 804 個，字根為 469 個。

二、可用的合成部件不如小篆豐富，常須直接分析成更小的字根。例如「鬻」字《說文》分析成从「鬻」、「𠂔」，然而《金文編》無「𠂔」字，「鬻」字只好分析成从「鬻」、「𠂔」，「𠂔」。

三、未收錄的成字部件較多，可能是沒有單獨出現，因此《金文編》並未收錄。這類部件有 82 個，有些還是常用的部件，如「刀」、「火」、「宀」等。

四、未能有一部如同《說文解字》一般可供字形拆分依據的字書，因此分析方式若和大家認知的不同，還待各方不吝指正。

接著說明楷體字形的分析。《漢語大字典》收錄的的 54,678 個字，有小篆等古漢字構形可參考的有 13,056 個，其他的 41,622 個字則無從參考。另外，可參考的 13,056 個字的形體也可能和古文不同，不能採用一致的分析。楷體字形分

析分以下幾點說明：

一、單字歸部(首)為字形拆分的重要依據。例如「量」字在《漢語大字典》為「里」部，因此可拆分成「旦」、「里」；「旦」為「日」部，再拆分成「日」、「一」；「里」、「日」、「一」皆為部首，可不拆分。只是利用部首來拆分字形，《漢語大字典》並不是一個好的選擇，雖然《漢語大字典》在單字歸部基本上與《康熙字典》相同，但對其中原歸部難於查檢的字，已略加調整。例如「條」、「穀」、「穎」、「贏」等字的部首在《漢語大字典》的歸部和《康熙字典》不同，詳見表九。雖然目前《康熙字典》的索引尚未加入，但是《中文大辭典》在單字歸部方面和《康熙字典》一致，可先採用。

表九、單字歸部及釋形

| 歸部 \ 單字 | 條 | 穀 | 穎 | 贏 |
|-----------|--------|--------|--------|--------|
| 《說文》釋形 | 从木、攸聲。 | 从禾、彳聲。 | 从禾、頃聲。 | 从貝、羸聲。 |
| 《說文》歸部 | 木 | 禾 | 禾 | 貝 |
| 《康熙字典》歸部 | 木 | 禾 | 禾 | 貝 |
| 《漢語大字典》歸部 | 人 | 彳 | 頁 | 月 |
| 《中文大辭典》歸部 | 木 | 禾 | 禾 | 貝 |

二、字根(或稱基礎部件)沿用相關標準。大陸信息處理用的漢字部件規範已於1997年推出，參考字集為GB 13000.1字符集的20,902個漢字，基礎部件為560個。台灣的中文字基礎部件標準草案CNS 11643-2仍在審定中，參考字集為CNS 11643第一、二字面的13,051個字，基礎部件有517個。先前我們基於『《康熙字典》部首不予拆分、其他字形按單字歸部拆分』的原則，CNS 11643第一、二字面的13,051個字拆分後的字根數為479，若再加上簡化字的字根則為526個。待中文字基礎部件標準制定後，我們再重新修正相關字形的拆分。

三、合成部件以小篆的成字部件或異體部件為主要來源，而字形當作偏旁的次數也是重要參考。小篆的2,030個隸定部件中有1,588個為合成部件，經由異體字表找到的異體部件也將近有560個。例如表十可由小篆的部件「更」、「便」、「恩」、「蔥」找到合成部件「叟」、「叟」、「忽」、「忿」、「蔥」、「葱」。扣除上述的合成部件個數外，出現在偏旁的可能部件個數可參考表十一，出現一次的部件雖有1,860個，但不見得就是合成部件。

表十、由異體字表尋找合成部件

| | | | | | | | | |
|-----|---|---|---|----|---|----|---|---|
| 小篆 | 𠄎 | 𠄎 | 𠄎 | | 𠄎 | 𠄎 | 𠄎 | |
| 主體字 | 更 | 便 | 鞭 | 匆 | 恩 | 葱 | 總 | 檇 |
| 異體字 | 叟 | 叟 | 鞭 | 忽恩 | 忿 | 蔥葱 | 總 | 檇 |

表十一、由偏旁字數尋找可能的合成部件

| | | | | | | | | | | |
|--------|-------|-----|-----|----|----|----|----|---|---|----|
| 偏旁字數 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | >9 |
| 可能的部件數 | 1,860 | 358 | 114 | 52 | 28 | 27 | 11 | 7 | 6 | 31 |

綜合上述，雖然對於不同歷史時期的漢字，我們採用不同的構形分析，但是不同的只是部件的選擇及個數，整個架構是一樣的，相關的應用程式都可共享。

肆、漢字構形資料庫與缺字處理

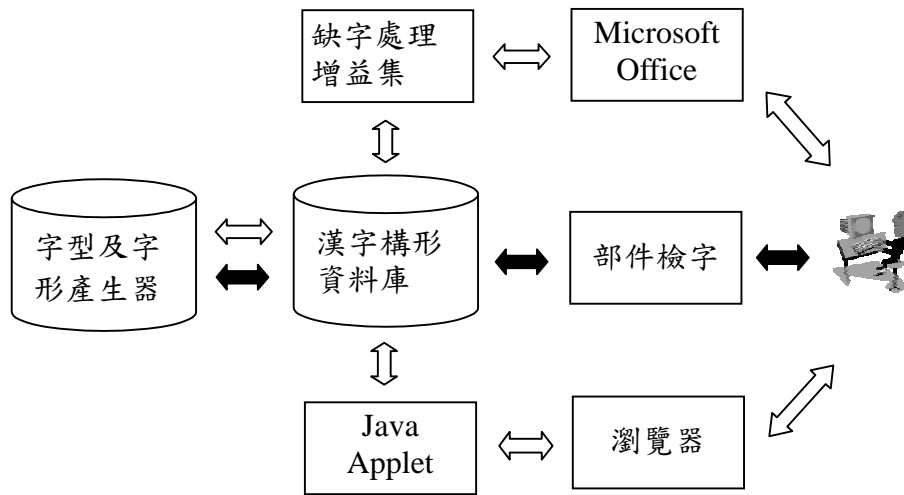
缺字指的是電腦交換碼沒有的字形。雖然不斷的增加交換碼的字形對於缺字問題有紓解的作用，但仍無法真正擺脫缺字問題的夢魘。缺字的問題在於漢字的字集是一個開放性質的，它的字數根本不適合作固定數量的限定；這與數量已定的西方語言的『字母集』，是不可以一概而論的。然而，現行漢字交換碼的結構，卻仿照西方語言的字母集的結構來設計，這不能不說是『削足適履』。相對於現行漢字交換碼利用字碼來區別漢字，以致缺字問題層出不窮，我們認為漢字的差異在於字形，缺字問題的解決應當由字形結構著手。

由第三節的漢字構形分析可以看出，字形結構的差異主要在於部件的選擇，其次才是部件的位置。例如「輝」、「暉」的差別在於部件「光」、「日」的不同，而「暈」、「暉」的差別在於部件「日」、「軍」位置的不同。絕大多數因部件位置不同而差異的字形，它的部件都只有兩個，而部件的相對位置為左右、上下或內外。於是我們定義表十二的構字符號，並且用構字符號和部件來表達字形結構，這樣一個結構式稱為構字式。相較於現行的交換碼，構字式更適合來表達缺字。

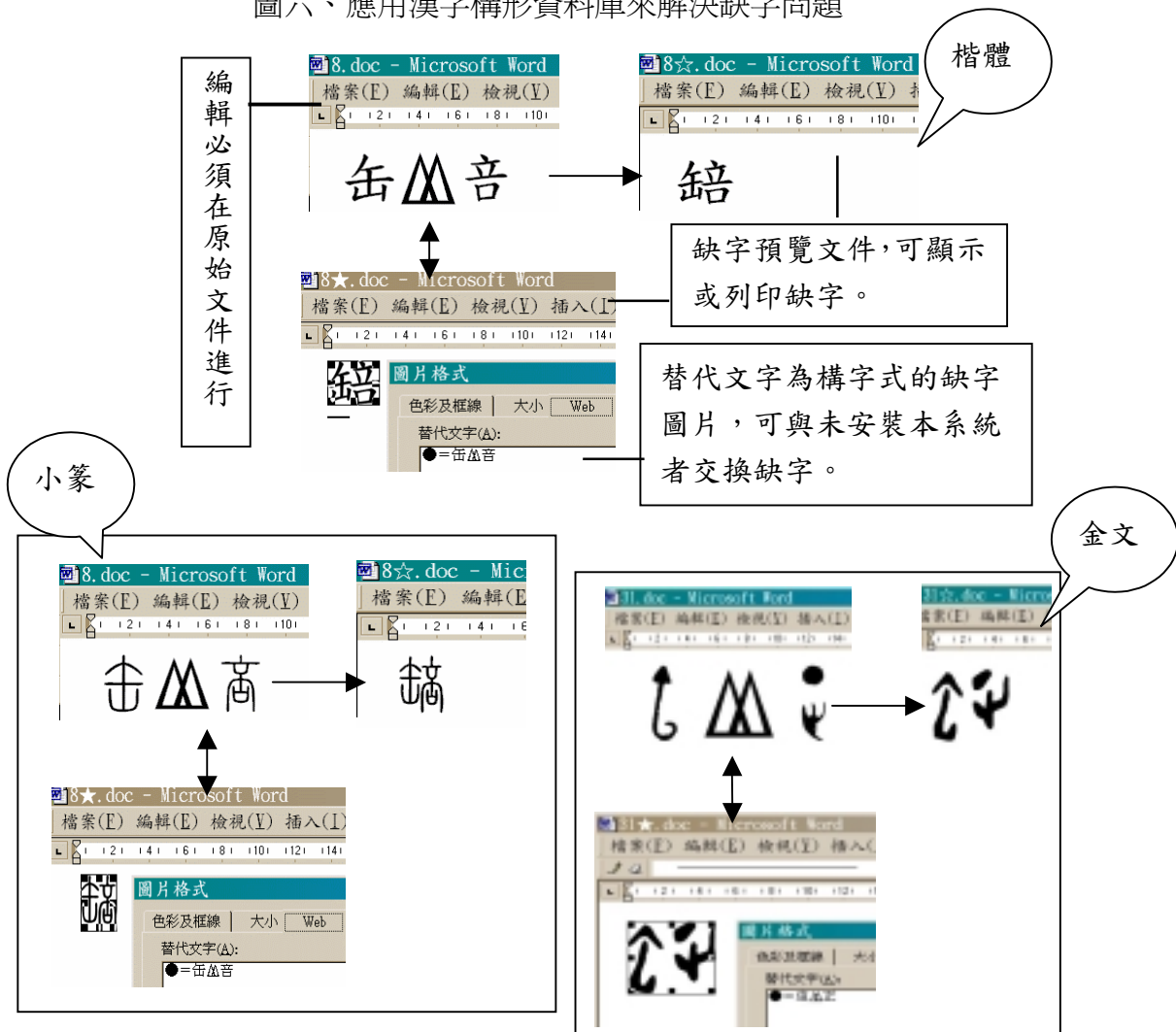
表十二、構字符號及構字式

| 類別 | 符號 | 說明 | 構字式範例 |
|------|----|---------------------------------|-----------|
| 連接 | △ | 當部件的連接順序由左至右 | 順 = 川△頁 |
| | △ | 當部件的連接順序由上至下 | 含 = 今△口 |
| | △ | 當部件的連接順序由外至內 | 圍 = 口△韋 |
| 部件序 | ⊠ | 按部件書寫順序輸入，前後以起始符號（⊠）和終止符號（⊡）包夾。 | 牖 = ⊠片戶甫⊡ |
| | ⊡ | | |
| 方便符號 | ∞ | 二個相同部件直連 | 炎 = ∞火 |
| | ∞ | 三個相同部件直連 | |
| | ∞ | 二個相同部件橫連 | 朋 = ∞月 |
| | ∞ | 三個相同部件橫連 | |
| | ∞ | 三個相同部件呈三角狀排列 | 焱 = ∞火 |
| | ∞ | 四個相同部件橫連 | |
| | ∞ | 四個相同部件直連 | |
| | ∞ | 四個相同部件呈四角狀排列 | 燚 = ∞火 |

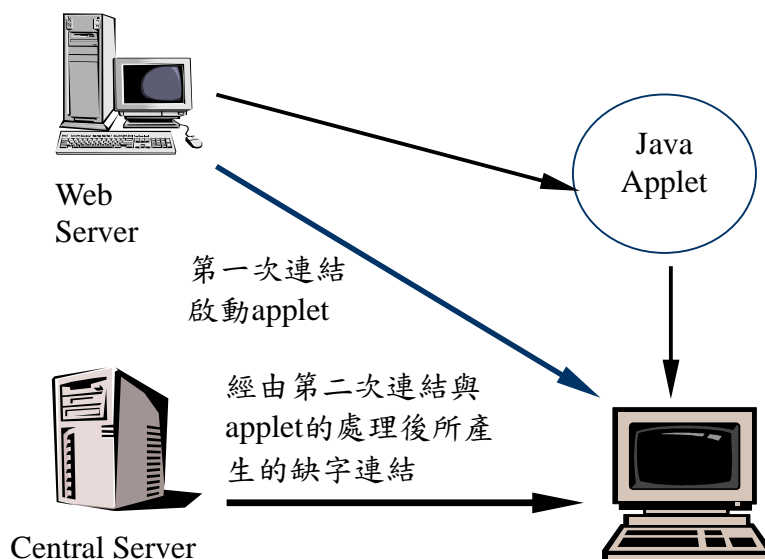
除了缺字表達外，缺字處理還需包括字型、字形產生器及應用程式，而這個系統的核心即為漢字構形資料庫，系統架構參見圖六。使用 Microsoft Word 處理構字式見圖七，網頁缺字處理見圖八，詳細用法可見漢字構形資料庫使用手冊。



圖六、應用漢字構形資料庫來解決缺字問題

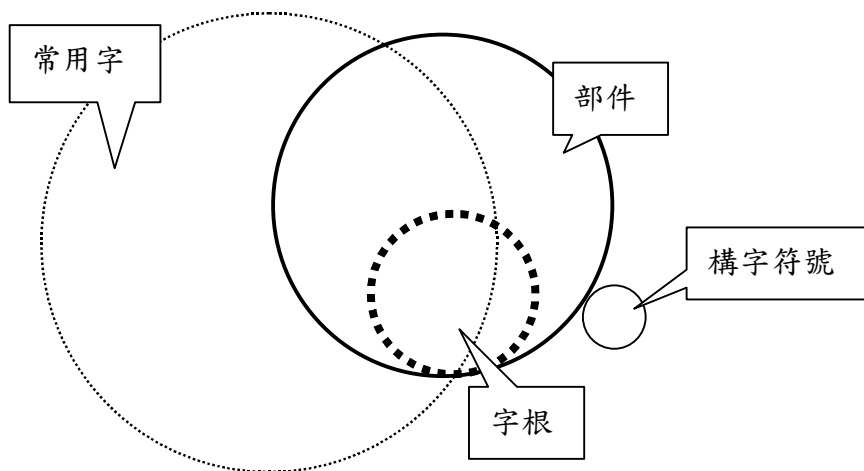


圖七、使用 Microsoft Word 處理構字式



圖八、網頁的缺字處理

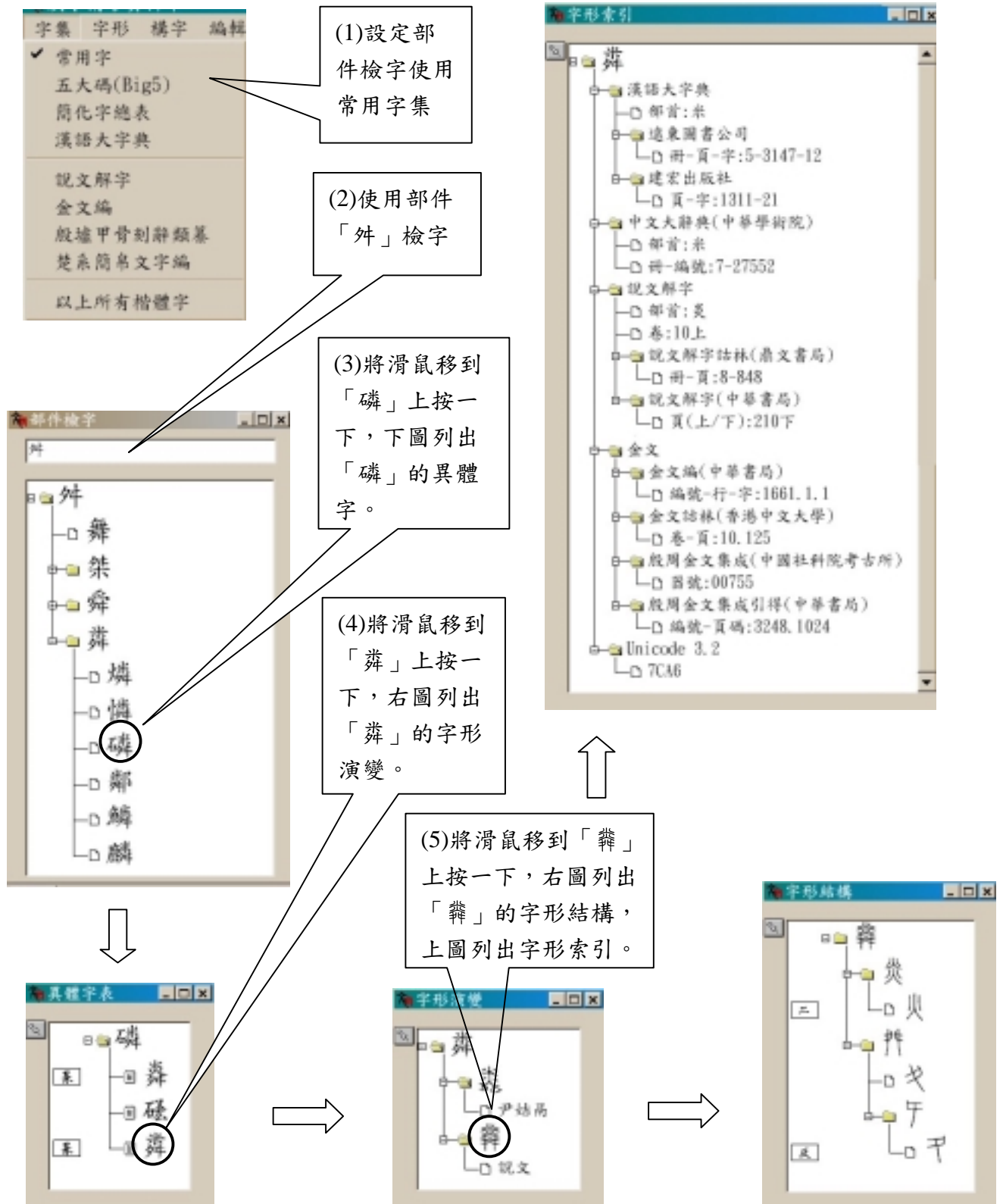
漢字構形資料庫目前仍有些部件未收錄在 Big5 或 Unicode 2.0 中，相較於這些交換碼，圖九指出電腦系統內碼除了收錄常用字外，部件及構字符號實不可或缺。對於缺字處理而言，增收幾百個部件遠比增收幾萬個字形來得實際。



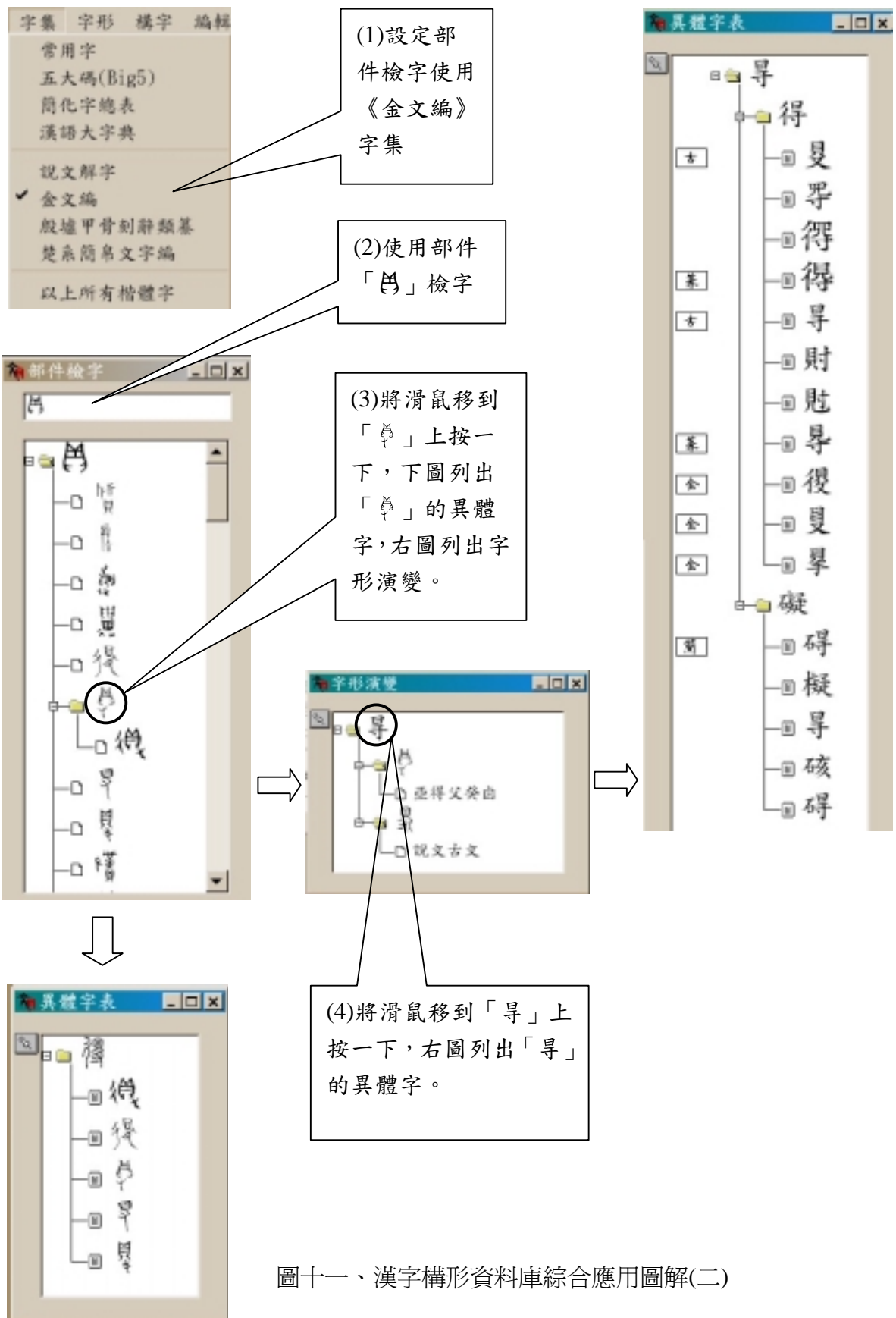
圖九、中文電腦的基本字集

伍、漢字構形資料庫綜合應用圖解

圖十、圖十一以圖來說如何利用漢字構形資料庫查得相關文字知識。



圖十、漢字構形資料庫綜合應用圖解(一)



圖十一、漢字構形資料庫綜合應用圖解(二)

陸、結語

在建置漢字構形資料庫的十一年中，剛好處於個人電腦蓬勃發展的階段。電腦的速度變快了，儲存容量變大了，但是處理漢字的技術卻沒有提昇多少。在一個漢字知識貧乏的電腦系統上建構漢字知識庫，雖然知識的表達不會有問題，但是在應用上總不是那麼如人意。這十一年來，幸好我們一直堅守在這個崗位上，才能累積這麼多的漢字構形知識，才不會因電腦系統的更換而使我們的缺字系統後繼無力。

從小篆構形資料庫完成後的這兩年來，使用漢字構形資料庫的人似乎變多了，網路下載的統計資料顯示，平均每天都有兩次下載。對於這些使用者，我們既感謝，又抱歉。抱歉的是我們始終投注絕大多數的心力在自己的研究上，而無法兼顧到他們的需求；感謝的是他們仍繼續使用我們的系統，相信我們遲早會解決他們的問題。

新的一年是可以期待的，今年我們將完成金文和楚系文字的構形資料庫，徹底的解決金文及楚系文字的缺字問題。金文的主要工作在於處理圖形文字及異寫字，楚系文字則是先處理異構字後，再來處理異寫字。最近也有使用者在反映使用聲韻來檢字的問題，今年過後，漢字構形資料庫不可避免的要開始觸及「音」、「義」，而不能只侷限於「形」，在此也期待這會是一個『漢字知識庫』，而不只是『漢字構形資料庫』。

參考文獻

| 編號 | 作者 | 書(目)/論文 | 出版社/研討會 | 出版年月 |
|----|------|-----------------------|-------------------------|----------|
| 1 | 于省吾 | 甲骨文字詁林 | 中華書局·北京 | 1996年 |
| 2 | 王寧等 | 漢字漢語基礎 | 北京科學出版社 | 1996年7月 |
| 3 | 李孝定 | 甲骨文字集釋 | 中央研究院歷史語言研究所·台北 | 1965年 |
| 4 | 李佳信 | 《說文》小篆字根研究 | 國立台灣師範大學國文研究所碩士論文 | 2000年7月 |
| 5 | 周法高等 | 金文詁林 | 香港中文大學 | 1974年 |
| 6 | 林尹等 | 中文大辭典 | 中國文化大學出版部 | 1973年10月 |
| 7 | 林樹 | 中文電腦基本用字 | 交通大學·新竹 | 1972年3月 |
| 8 | 姚孝遂 | 殷墟甲骨刻辭類纂 | 中華書局·北京 | 1989年 |
| 9 | 容庚 | 金文編 | 中華書局·北京 | 1985年7月 |
| 10 | 徐中舒等 | 遠東·漢語大字典 | 遠東圖書公司 | 1991年9月 |
| 11 | 張亞初 | 殷周金文集成引得 | 中華書局·北京 | 2001年7月 |
| 12 | 莊德明 | 漢字印刷字形的整理 | 電子古籍中的文字問題研討會·台北 | 1999年6月 |
| 13 | 莊德明 | 中文電腦缺字解決方案 | 全國技專院校圖書館自動化規劃第七屆研討會·屏東 | 2001年12月 |
| 14 | 莊德明等 | 漢字構形資料庫使用手冊 | 中研院資訊所·台北 | 2002年7月 |
| 15 | 莊德明等 | 如何使用電腦處理古今文字的銜接—以小篆為例 | 第十四屆中國文字學全國學術研討會·高雄 | 2003年3月 |
| 16 | 陳昭容等 | 金文資料庫字詞檢索系統的設計與應用 | 中央研究院第三屆國際漢學會議·台北 | 2000年6月 |
| 17 | 楊家駱 | 說文解字詁林 | 鼎文書局 | 1994年3月 |
| 18 | 滕壬生 | 楚系簡帛文字編 | 湖北教育出版社 | 1995年7月 |
| 19 | 謝清俊等 | 中文字形資料庫的設計與運用 | 第六屆中國文字學全國學術研討會·台中 | 1995年4月 |
| 20 | 謝清俊 | 電子古籍中的缺字問題 | 第一屆中國文字學會學術討論會·天津 | 1996年8月 |
| 21 | 謝清俊等 | 中央研究院古籍全文資料庫解決缺字問題的方法 | 第二次兩岸古籍整理研究學術研討會·北京 | 1998年5月 |