《鬼谷子》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中研院史語所漢籍電子文獻工作小組
2008/2/20 陳建安 製作

一、《鬼谷子》一書(鬼谷子.xml)使用舊版造字 14 個，字頻 43 次，詳如附件一。這 14 個造字中，11 個可轉成 Windows XP 能顯示的字，字頻 31 次；另外 3 個字必須轉成構字式，字頻 12 次。

二、附件一的造字分析表說明如下：

甲、編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

乙、造字：舊版造字

丙、字頻(txt)：造字在「.txt」文件的出現次數

丁、字頻(xml)：造字在「.xml」文件的出現次數

戊、Big5：造字的 Big5 碼

己、Unicode：造字所對應的 Unicode 碼

庚、WinXP：造字在 Windows XP 的對應字形

辛、備註凡例：

1、異體字問題：新版漢籍考量到使用者檢索及使用時的便利性，將用字原則改為除專詞等特殊情形之外，一律改用標準字呈現。如編號 2042 的「足 Δ 充」係「疏」字之異體，故以「疏」字取代。

2、Unicode 字型呈現差異：Unicode 字型與舊漢籍造字有些微差異，但只是字體風格差異，實際上仍為同一個字，因此仍取 Unicode 字型。如編號 4081 的「翺」字，Unicode 字型呈現為「翺」，實際上仍為同一字。

三、Unicode 目前收錄的漢字總數為 70194，分屬於三個不同區段，詳如表一。目前 Windows XP 只支援 CJK 認同表意文字區的 20902 個字，內碼為 4E00-9FFF。所以造字編號 4551 的「嶠」字，Unicode 編碼為 21F0F，由於 Windows XP 並不支援，仍須使用構字式「山 Δ 間」。

表一、Unicode 的字數及編碼區段

Unicode	字集子集合	新增字數	新增編碼區段	總字數	WinXP
1.1 版	CJK 認同表意文字區	20902	4E00-9FFF	20902	支援
3.0 版	CJK 認同表意文字擴充 A 區	6582	3400-4DFF	27484	不支援

3.1 版	CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	70194	不支援
-------	------------------	-------	-------------	-------	-----

四、《鬼谷子》中未確定字形「●」共 1 處，詳細處理內容可參考附件二的未確定字形「●」手動取代表。附件二《鬼谷子》未確定字形「●」手動取代表欄位說明如下：

- 甲、檔案原文摘錄：摘錄檔案(鬼谷子.xml)中包含未確定字形「●」的文句段落，其中紅字底線的部分，為需要手動取代的字形。
- 乙、手動取代結果：變更過後的檔案內容，其中藍字底線的部分，為修改後的結果。

附件一、《鬼谷子》造字分析表

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
2042	疏	4	4	9640	47FD		足厶充	疏，異體字問題。
2044	踈	3	3	9642	8E08	踈		
2947	聰	1	1	9BDA	8066	聰		
3804	効	1	1	8163	52B9	効		
3814	却	4	4	816D	5374	却		
3996	踈	4	4	82A8	758E	踈		
4080	羣	1	1	82FC	7FA3	羣		Unicode 呈現差異
4081	翱	1	1	82FD	7FFA	翱		Unicode 呈現差異。
4132		5	5	8371			血厶丞	眾，異體字問題。
4190	郟	6	6	83CD	90C4	郟		
4240	髯	2	2	8440	9AF4	髯		

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
4242	鬪	6	6	8442	9B2D	鬪		Unicode 呈現差異
4551	嶠	3	3	85FC	21F0F		山厶間	
5860	无	2	2	C6D3	65E0	无		

附件二、《鬼谷子》未確定字形「●」手動取代表

檔案原文摘錄	手動取代結果
熒●之處安在	熒火厶惑之處安在