《孔叢子》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
2008/3/3 陳建安 製作

一、 《孔叢子》一書(孔叢子.xml)使用舊版造字 58 個，字頻 161 次，詳如附件一。這 58 個造字中，52 個可轉成 Windows XP 能顯示的字，字頻 125 次；另外 6 個字必須轉成構字式，字頻 36 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、校對問題，舊版漢籍錯字，可用程式全部取代：在舊版漢籍電子文獻中即存在的錯字，因校對時的疏漏而未更正，持續留存在新版漢籍電子文獻中；若該造字的所有頻次，皆屬於錯誤使用的錯字情形，可以用程式全部取代為正確字形。如編號 4275 的「义」字，原字為「叉」。

2、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 2811 的「」係「留」字之異體，故以「留」字取代。

3、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 1925 的「」字，Unicode 字型呈現為「  」，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 4493 的「  」字，Unicode 編碼為 3E40，由於 Windows XP 並不支援，仍須使用構字式「  」。

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、 《孔叢子》中未確定字形「●」共 6 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《孔叢子》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(孔叢子.xml)中包含「●」的文句段落，其中紅字底線的部分，為需要手動取代的原文。

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、《孔叢子》造字分析表

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
74	髺	1	1	FAAB	29B06		髟𠂤丐	鬢，異體字問題。
980	犨	1	1	8F65	72AB	犨		
1014	珧	1	1	8FA9	73CE	珧		
1238	庑	15	15	90EC	77E6	庑		
1925	羴	1	1	9568	8989	羴		Unicode 呈現差異。
1981	讋	2	2	95C2	8B90	讋		
2044	跠	2	2	9642	8E08	跠		
2415	鬪	1	1	987B	95D8	鬪		
2505	飜	1	1	98F7	98DC	飜		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
2516	餚	1	1	9943	9919	餚		
2811		2	2	9AEF			丽亼田	留，異體字問題。
2935	穱	1	1	9BCE	7A6A	穱		
2977	却	1	1	9BF8			去厶卩	卻，異體字問題。
3035	賨	10	10	9C73	8CD4	賨		賓，異體字問題。
3048	叅	1	1	9CA2	53C5	叅		參，異體字問題。
3791	况	3	3	8156	51B5	况		況，異體字問題。
3796	决	1	1	815B	51B3	决		決，異體字問題。
3804	効	2	2	8163	52B9	効		
3817	厠	1	1	8170	53A0	厠		廁，異體字問題。
3876	廻	2	2	81CD	5EFB	廻		迴，異體字問題。
3883	徧	3	3	81D4	5FA7	徧		
3935	槩	1	1	8249	69E9	槩		概，異體字問題。
3940	欝	1	1	824E	6B1D	欝		
3974	牀	2	2	8270	7240	牀		
3979	猫	1	1	8275	732B	猫		
3996	疎	2	2	82A8	758E	疎		
4010	皐	1	1	82B6	7690	皐		Unicode 呈現差異。

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
4014	葢	4	4	82BA	8462	葢		蓋，異體字問題。
4028	稟	1	1	82C8	7980	稟		稟，異體字問題。
4038	竈	1	1	82D2	7AC8	竈		
4057	莖	8	8	82E5	585F	莖		
4080	羣	9	9	82FC	7FA3	羣		群，異體字問題。
4085	耻	4	4	8342	803B	耻		
4091	皋	4	4	8348	81EF	皋		Unicode 呈現差異。
4132		30	30	8371			血  巫	眾，異體字問題。
4145	羈	1	1	837E	898A	羈		Unicode 呈現差異。
4146	覩	5	5	83A1	89A9	覩		
4156	謟	2	2	83AB	8B1F	謟		諂，異體字問題。
4184	迹	1	1	83C7	8FF9	迹		
4203	隣	4	4	83DA	96A3	隣		鄰，異體字問題。
4256	鷄	1	1	8450	9DC4	鷄		
4259	麕	1	1	8453	9E95	麕		
4275	乂	1	1	8463	4E49	乂		叉，校對問題，舊版漢籍錯字，可用程式全部取代。
4437	廩	1	1	8568	5EEA	廩		廩，異體字問題。

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
4460	勑	1	1	85A1	52D1	勑		
4493	牮	1	1	85C2	3E40		牜 ⿰李	
4801	媊	1	1	87BC	5A2C	媊		
5046	彝	1	1	8955	5F5C	彝		
5066	瑣	1	1	8969	24A0F		王⿰貝	瑣，異體字問題。
5074	厊	3	3	8971	5C2B	厊		
5098	寤	1	1	89AB	7AB9	寤		
5147	恮	1	1	89DC	6060	恮		
5290	携	2	2	8ACE	643A	携		
5300	鬪	6	6	8AD8	9B2A	鬪		
5408	紝	1	1	8BA7	7D4D	紝		
5439	栢	1	1	8BC6	6822	栢		
5558	麁	1	1	8C7E	9E84	麁		
5753	潜	1	1	8DC6	6F5C	潜		潛，異體字問題。

附件二、《孔叢子》未確定字形「●」手動取代表

檔案原文摘錄	手動取代結果
●者不可生	死者不可生
梁甫迴●枳棘充路	梁甫迴迍枳棘充路
麟出而●	麟出而死
●牛元武	牜⿰麗牛元武
●嬖寵之官	口⿰攵嬖寵之官

檔案原文摘錄	手動取代結果
手搏  獸	手搏足  獸