《謝疊山全集校注》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
2008/5/20 陳建安 製作

一、 《謝疊山全集校注》一書(謝疊山全集校注.xml)使用舊版造字 102 個，字頻 383 次，詳如附件一。這 102 個造字中，84 個可轉成 Windows XP 能顯示的字，字頻 301 次；另外 18 個字必須轉成構字式，字頻 82 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、 構字式：Windows XP 若無對應字形，則改採用構字式
壬、 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 4032 的「禾  未」字，原字為「秣」。

2、標記問題：標記問題主要是漢籍電子文獻在舊轉新的過程中，採用了新的標記語言，而在修改標記時，對於某些標題句做了增刪的動作，因為這些標題句的內容包含了缺字,所以這些更改標記的動作造成 txt 檔與 xml 檔缺字頻次不符的情形。如編號 3780 的「僎」字，xml 檔在 170 頁新增了一個標題，因而造成頻次不同，處理方式為將所有頻次皆修改為「僎」。

3、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 4132 的「血  巫」係「眾」字之異體，故以「眾」字取代。又編號 1641 的「胆」字，係「膽」字之簡體字，亦以標準字的「膽」字取代。

4、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 4158 的「譕」字，Unicode 字型呈現為「譞」，實際上仍為同一字。

5、待造字：Unicode 及漢字構形資料庫皆未收錄的舊漢籍造字，正在等待補造字中，所以「造字」欄空白無法看到字形。如編號 1131 的「米  糞」。

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 5795 的「婕」字，Unicode 編碼為 3715，由於 Windows XP 並不支援，仍須使用構字式「女連」。

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、 txt 檔與 xml 檔頻次不同且造字分析表「備註」欄有備註問題情況者，詳細取代內容可參考附件二的手動取代表。附件二的手動取代表的欄位說明如下：

甲、 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

乙、 造字：舊版造字字形。

丙、 檔案原文摘錄：摘錄檔案(謝疊山全集校注.xml)中包含該造字的文句段落，其中紅字底線的部分，為需要手動取代的原文。文句後方標示（全部取代）者，表示執行全部取代的動作，不一一列舉所有原文。

丁、 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。文句後方標示（全部取代）者，表示執行全部取代的動作，不一一列舉所有原文。

附件一、《謝疊山全集校注》造字分析表

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
296		2	2	FBEC				
297		9	9	FBED				
299		2	2	FBEF				
300		2	2	FBF0				
493	冒	1	1	FD55	5190	冒		冒，異體字問題。
608	夬	2	2	FDEA			夬	
656	音	1	1	FE5B			音	
675	飠	2	2	FE6E			飠	
780	氵	1	1	FEF9			氵	
814	蟆	1	1	8E5C	87C7	蟆		蟆，異體字問題。
881	飇	1	1	8EC1	98C7	飇		
1024	虵	2	2	8FB3	8675	虵		
1131		1	1	905F			米  冓	待造字。
1490	籱	3	3	92AE	7C5D	籱		
1565	鯀	1	1	92F9	7DDC	鯀		
1641	胆	1	1	93A8	80C6	胆		膽，異體字問題。
1673	茜	1	1	93C8	34BC		茜	

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
1718	菪	1	1	93F5	83ED	菪		
1809	囷	3	3	94B3	56E6	囷		
2046	踪	1	1	9644	8E2A	踪		
2157	酘	1	1	96D5	9167	酘		
2197	鉄	5	5	96FD	9244	鉄		鐵，異體字問題。
2502	颱	1	1	98F4	98C8	颱		
3780	僞	5	6	814B	5101	僞		僞，標記問題，增加標題，人工取代。
3789	冲	1	1	8154	51B2	冲		沖，異體字問題。
3791	况	18	18	8156	51B5	况		況，異體字問題。
3796	决	7	7	815B	51B3	决		決，異體字問題。
3801	剋	1	1	8160	5257	剋		
3802	劇	1	1	8161	529A	劇		
3805	勅	2	2	8164	52C5	勅		
3814	却	9	9	816D	5374	却		卻，異體字問題。
3817	厠	1	1	8170	53A0	厠		廁，異體字問題。
3819	厮	1	1	8172	53AE	厮		廝，異體字問題。
3829	咏	2	2	817C	548F	咏		
3830	咤	1	1	817D	54A4	咤		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
3840	坂	5	5	81A9	5742	坂		
3852	媚	3	3	81B5	5A63	媚		
3858	冤	10	10	81BB	5BC3	冤		冤，異體字問題。
3864	峯	42	42	81C1	5CEF	峯		峰，異體字問題。
3883	徧	4	4	81D4	5FA7	徧		
3930	椁	1	1	8244	6901	椁		
3942	毡	1	1	8250	6BE1	毡		
3948	沭	1	1	8256	6CAD	沭		沐，校對問題，舊版漢籍錯字，可用程式全部取代。
3949	泣	1	1	8257	6D96	泣		
3957	凜	8	8	825F	6F9F	凜		凜，異體字問題。
3963	炁	2	2	8265	7081	炁		
3966	烟	7	7	8268	70DF	烟		煙，異體字問題。
3974	牀	5	5	8270	7240	牀		
3978	猪	1	1	8274	732A	猪		Unicode 呈現差異。
3979	猫	1	1	8275	732B	猫		
3996	疎	2	2	82A8	758E	疎		
4006	癗	1	1	82B2	764F	癗		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
4016	着	4	4	82BC	7740	着		著，異體字問題。
4032	秣	4	4	82CC	2578A		禾⿰未	秣，校對問題，舊版漢籍錯字，可用程式全部取代。
4040	竝	1	1	82D4	7ADD	竝		
4041	竪	1	1	82D5	7AEA	竪		
4043	笄	2	2	82D7	7B53	笄		
4067	綉	4	4	82EF	7D89	綉		繡，異體字問題。
4068	綫	3	3	82F0	7DAB	綫		線，異體字問題。
4080	羣	17	17	82FC	7FA3	羣		群，異體字問題。
4090	脚	3	3	8347	811A	脚		腳，異體字問題。
4112	蘇	1	1	835D	8534	蘇		
4115	藁	2	2	8360	85C1	藁		
4129	虬	4	4	836E	866C	虬		
4132		42	42	8371			血⿱巫	眾，異體字問題。
4134	衛	21	21	8373	885E	衛		衛，異體字問題。
4140	衹	1	1	8379	88B5	衹		
4146	覩	4	4	83A1	89A9	覩		
4158	譌	1	1	83AD	8B4C	譌		Unicode 呈現差異。

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
4167	賮	1	1	83B6	8CEB	賮		
4176	輙	2	2	83BF	8F19	輙		輒，異體字問題。
4184	迹	4	4	83C7	8FF9	迹		
4186	遡	2	2	83C9	9061	遡		
4193	鈎	3	3	83D0	920E	鈎		
4195	鉢	2	2	83D2	9262	鉢		
4198	鑛	1	1	83D5	945B	鑛		
4218	韮	1	1	83E9	97EE	韮		
4226	凔	6	6	83F1	98E1	凔		
4242	鬪	4	4	8442	9B2D	鬪		Unicode 呈現差異。
4256	鶏	4	4	8450	9DC4	鶏		
4262	麪	5	5	8456	9EAA	麪		
4266	黽	1	1	845A	9F02	黽		
4267	鼈	1	1	845B	9F08	鼈		
4307	罵	6	6	84A5	99E1	罵		罵，異體字問題。
4309	刼	3	3	84A7	523C	刼		劫，異體字問題。
4702	塲	4	4	86F6	5872	塲		場，異體字問題。
4735	壠	1	1	8758	58E0	壠		

編號	造字	字頻(txt)	字頻(xml)	Big5	Unicode	WinXP	構字式	備註
4955	龜	6	6	88B9	9F9C		龜	龜，異體字問題。
5013	氵	2	2	88F3			氵	
5169	戶	1	1	89F2	6239	戶		
5219	妃	1	1	8A65			女⿰巳	妃，異體字問題。
5289	慤	1	1	8ACD	6164	慤		慤，異體字問題。
5303	昰	1	1	8ADB	6630	昰		
5306		1	1	8ADE			𥫱⿰公閒	簡，異體字問題。
5466	倏	1	1	8BE1	5010	倏		Unicode 呈現差異。
5542	鑲	1	1	8C6E	93C1	鑲		
5559	欵	2	2	8CA1	6B35	欵		款，異體字問題。
5591	杞	2	2	8CC1	233CC		木⿰巳	杞，異體字問題。
5666	汚	3	3	8D4D	6C5A	汚		污，異體字問題。
5704	穩	1	1	8D73	6C0A	穩		
5706	敘	3	3	8D75	654D	敘		敘，異體字問題。
5795	婕	1	1	8DF0	3715		女⿰連	

附件二、《謝疊山全集校注》手動取代表

編號	造字	檔案原文摘錄	手動取代結果
3780	僊	鹵（全部取代）	僊（全部取代）