《新書》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
2008/2/14 陳建安 製作

一、 《新書》一書(新書.xml)使用舊版造字 57 個，字頻 180 次，詳如附件一。這 57 個造字中，45 個可轉成 Windows XP 能顯示的字，字頻 156 次；另外 12 個字必須轉成構字式，字頻 24 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 2103 的「⿰形辶⿱口⿱乇⊡」字，原字為「還」。

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 4878 的「宂」字，Unicode 字型呈現為「宂」，實際上仍為同一字。

3、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 2036 的「石⿰⿰⿰⿰⿰⿰⿰⿰⿰⿰萬」。

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 4855 的「⿰氵⿱⿰⿰⿰⿰⿰⿰⿰⿰⿰⿰

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、 《新書》中未確定字形「●」共 41 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《新書》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(新書.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、《新書》造字分析表

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
841	礲	1	1	8E77	7932	礲		
890	𦬊	1	1	8ECA	71AD	𦬊		
1024	虵	2	2	8FB3	8675	虵		
1040	聑	1	1	8FC3	803C	聑		
1981	讐	5	5	95C2	8B90	讐		
2036		1	1	95F9			石⿰⿰萬	待造字。
2103		1	1	967D			⿰形辶⿱口丐⿴⿰⿱口丐丶	逞，校對問題，舊版漢籍錯字，可用程式全部取代。
2415	鬪	4	4	987B	95D8	鬪		鬬，校對問題，舊版漢籍錯字，可用程式全部取代。

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
2817	穷	1	1	9AF5	3AC4		穷	
3791	况	4	4	8156	51B5	况		
3804	効	3	3	8163	52B9	効		
3835	噍	1	1	81A4	564D	噍		
3855	孽	4	4	81B8	5B7C	孽		
3860	尠	1	1	81BD	5C20	尠		
3883	徧	3	3	81D4	5FA7	徧		
3891	慙	1	1	81DC	6159	慙		
3910	擥	1	1	81EF	64E5	擥		
3924	朞	1	1	81FD	671E	朞		
3931	椀	1	1	8245	6900	椀		
3974	牀	1	1	8270	7240	牀		
3996	疎	5	5	82A8	758E	疎		
4015	眦	1	1	82BB	7726	眦		
4040	竝	12	12	82D4	7ADD	竝		
4045	箒	1	1	82D9	7B92	箒		
4055	粃	8	8	82E3	7C83	粃		
4072	綵	1	1	82F4	7DE5	綵		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
4104	莅	1	1	8355	8385	莅		
4134	衞	27	27	8373	885E	衞		
4146	覩	1	1	83A1	89A9	覩		
4155	謌	2	2	83AA	8B0C	謌		
4172	蹙	1	1	83BB	8E77	蹙		Unicode 呈現差異。
4184	迹	12	12	83C7	8FF9	迹		
4190	郄	12	12	83CD	90C4	郄		
4250	鰌	2	2	844A	9C0C	鰌		
4267	鼈	3	3	845B	9F08	鼈		
4271	齩	1	1	845F	9F69	齩		
4371	僩	2	2	84E5	50F4	僩		
4455	劒	1	1	857A			僉⿰刀	
4533	唫	1	1	85EA	552B	唫		
4855	汗	4	4	87F2	3CC5		氵⿰不	
4869	孋	1	1	8841	5B4A	孋		
4878	宂	1	1	884A	5B82	宂		Unicode 呈現差異。
5034	弃	9	9	8949	5F03	弃		Unicode 呈現差異。
5048	婚	3	3	8957	36F0		女⿰昏	

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
5091	鉶	1	1	89A4	92E3	鉶		
5108	惛	3	3	89B5	60FD	惛		
5109	樐	1	1	89B6	6A10	樐		
5306	簡	6	6	8ADE			竹⿱亼冂	
5325	踆	2	2	8AF1	28082		足⿰⿱亼人灸	
5335	𧊒	1	1	8AFB	459F		亡⿱亼虫	
5355	𡗗	1	1	8B50	20AE4		𡗗	
5371	冐	9	9	8B60	808E	冐		
5398	湻	1	1	8B7B	6E7B	湻		
5651	難	1	1	8CFD	3E10		難⿱亼灬	
5682	蠚	1	1	8D5D	881A	蠚		
5739		2	2	8DB8			冗⿱亼古	罟，校對問題，舊版漢籍錯字，可用程式全部取代。
5752	揔	1	1	8DC5	63D4	揔		

附件二、《新書》未確定字形「●」手動取代表

檔案原文摘錄	手動取代結果
氓●之人	氓隸之人
韓王信●胡	韓王信倚胡
望其章而知其勢●	望其章而知其勢山⿱亼人⿱亼子
●過黑	慟過黑
●自完足矣	慟自完足矣

檔案原文摘錄	手動取代結果
衛侯更其名曰●	衛侯更其名曰火⿰為
周襄王出逃伯●	周襄王出逃伯門⿺卦
縛、●、笞	縛、木⿰穷、笞
罵詈而●笞之	罵詈而木⿰穷笞之
不同禽獸●焉耳	不同禽獸⿰勞焉耳
吟●而堅控	吟齒⿰休而堅控
孰敢●然	孰敢忄⿰分然
●炙肉	⿸形月者火⿰炙肉
垂●而相告	垂氵⿰羡而相告
適●拘也	適糹⿰冓拘也
飯羹啗膹●●	飯羹啗膹夕⿱肉肉嘽
●多飲酒	氵⿰脊多飲酒
王●不踏蹴	王⿰勞不踏蹴
●問小誦之	⿰間問小誦之
●帷而入	扌⿰寨帷而入
過則●	過則言⿰間
●牲而	亻⿰貳牲而
漻然●然	漻然氵⿰愁然
廢首●肘曰卑坐	廢首低肘曰卑坐
如屋之●	如屋之氐
食●餕	食豊⿰葬餕
●蚍生	虫⿰兀蚍生
食●菹而蛭口	食⿸形艹禾盧⿰菹而蛭口
●，重裘而立	酉⿰旦，重裘而立
不●於鄉曲	不作⿱心心於鄉曲
不●於朝廷	不作⿱心心於朝廷
嘗試傅白●黑	嘗試傅白⿰黑黑
●。弟敬愛兄	癶⿱齒齒。弟敬愛兄
反端為●	反端為⿰防
義為●	義為⿰夢
欣●可安	欣忄⿰熏可安
反嚴為●	反嚴為車⿰辰
●測之也	冫⿱四⿱勹測之也
夫玉者真德●也	夫玉者真德冫⿱四⿱勹也
猶景之●形也	猶景之冫⿱四⿱勹形也