

《東坡志林》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

2008/4/10 陳建安 製作

一、《東坡志林》一書(東坡志林.xml)使用舊版造字 69 個，字頻 200 次，詳如附件一。這 69 個造字中，55 個可轉成 Windows XP 能顯示的字，字頻 146 次；另外 14 個字必須轉成構字式，字頻 54 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 74 的「髟𠂔」字，原字為「髻」。
- 2、校對問題，新舊版缺字並存：漢籍在舊轉新的過程中，以人工修改方式取代舊漢籍造字，雖然修改的結果是正確的，但是只修改到部分造字，沒有全部修改到，導致舊漢籍造字在新版漢籍檔案中，一部分已被取代成 Unicode 或構字式，另一部分仍為舊漢籍造字的並存情形。處理方式為配合漢籍工作室的修改，將沒被修改到的舊漢籍造字，以對應的 Unicode 或構字式取代。如編號 715，舊漢籍造字為「𠂔」，比對 txt 檔及 xml 檔後，發現原先在 txt 檔使用舊漢籍造字「𠂔」者，在 xml 檔中部分已人工修改為構字式的「𠂔」字，故配合修改，將編號 715「𠂔」字，手動取代為構字式的「𠂔」字。
- 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 3891 的「慙」係「慚」字之異體，故以「慚」字取代。又編號 3966 的「烟」字，係「煙」字之簡體字，亦以標準字的「煙」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 3978 的「猪」字，Unicode 字型呈現為「猪」，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 5653 的「𠂔」字，Unicode 編碼為 3B09，由於 Windows XP 並不支援，仍須使用構字式「日𠂔」。

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、檔案原文摘錄：摘錄檔案(東坡志林.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。

丙、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、《東坡志林》造字分析表

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
74	髻	1	1	FAAB	29B06		髻𠂔𠂔	髻，校對問題，舊版漢籍錯字，可用程式全部取代。
297	𡵓	13	13	FBED			𡵓	
298	𡵔	1	1	FBEE			𡵔	
299	𠂔	6	3	FBEF			𠂔	𠂔，校對問題，新舊版缺字並存。
300	𠂔	6	3	FBF0			𠂔	𠂔，校對問題，新舊版缺字並存。
580	𡵕	1	1	FDCE			𡵕	
658	𡵖	1	1	FE5D			𡵖	
715	𡵗	4	1	FEB8			𡵗	𡵗，校對問題，新舊版缺字並存。
770	𡵘	1	1	FEEF			𡵘	
820	𡵙	1	1	8E62	70F1	𡵙		
1024	𡵚	1	1	8FB3	8675	𡵚		
1040	𡵛	1	1	8FC3	803C	𡵛		
1056	𡵜	1	1	8FD3	9641	𡵜		
1205	𡵝	1	1	90CB	770E	𡵝		
1354	𡵞	1	1	91C3	79A9	𡵞		
1944	𡵟	4	4	957B	8A67	𡵟		

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
1981	讐	2	2	95C2	8B90	讐		
2131	弼	1	1	96BB	5F3B	弼		弼，異體字問題。
2162	醎	1	1	96DA	918E	醎		
3816	厓	1	1	816F	5393	厓		
3819	廐	1	1	8172	53AE	廐		廐，異體字問題。
3834	嗃	2	2	81A3	5637	嗃		
3858	寃	2	2	81BB	5BC3	寃		寃，異體字問題。
3864	峯	1	1	81C1	5CEF	峯		峰，異體字問題。
3891	慙	1	1	81DC	6159	慙		慙，異體字問題。
3957	凜	1	1	825F	6F9F	凜		凜，異體字問題。
3966	烟	1	1	8268	70DF	烟		煙，異體字問題。
3974	牀	4	4	8270	7240	牀		
3978	猪	2	2	8274	732A	猪		Unicode 呈現差異。
3981	珉	1	1	8277	73C9	珉		
3996	疎	6	6	82A8	758E	疎		
4008	癯	1	1	82B4	766F	癯		
4010	皐	3	3	82B6	7690	皐		
4012	鼓	2	2	82B8	76B7	鼓		

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
4013	盃	1	1	82B9	76CC	盃		
4023	碇	1	1	82C3	78D3	碇		
4038	竈	3	3	82D2	7AC8	竈		
4055	粃	1	1	82E3	7C83	粃		
4080	羣	12	12	82FC	7FA3	羣		群，異體字問題。
4091	皐	3	3	8348	81EF	皐		
4093	舐	1	1	834A	445B		舌𠂔氏	舐，異體字問題。
4107	葱	1	1	8358	8471	葱		
4132		22	22	8371			血𠂔𠂔	眾，異體字問題。
4134	衛	9	9	8373	885E	衛		衛，異體字問題。
4146	覩	3	3	83A1	89A9	覩		
4169	趲	2	2	83B8	8DA6	趲		
4184	迹	7	7	83C7	8FF9	迹		
4186	遡	1	1	83C9	9061	遡		
4190	邳	3	3	83CD	90C4	邳		
4193	鈎	1	1	83D0	920E	鈎		
4226	𠂔	1	1	83F1	98E1	𠂔		
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	2	2	8442	9B2D	鬪		Unicode 呈現差異。
4256	鷄	2	2	8450	9DC4	鷄		
4267	鼈	1	1	845B	9F08	鼈		
4270	鼾	1	1	845E	9F45	鼾		
4460	勅	2	2	85A1	52D1	勅		
4507	吳	26	26	85D0	5434	吳		吳，異體字問題。
4515	呪	8	8	85D8	546A	呪		
4575	縊	1	1	8655	7E7F	縊		
4735	壠	1	1	8758	58E0	壠		
4768	妬	1	1	8779	59AC	妬		妒，異體字問題。
4778	鞞	3	3	87A5	97B4	鞞		
4816	亻	1	1	87CB			亻	
4957	崑	1	1	88BB	5D53	崑		
5302	昏	2	2	8ADA	662C	昏		
5582	鞣	1	1	8CB8	97B5	鞣		
5653	暎	4	4	8D40	3B09		日 𠂔 𠂔	
5713	臧	1	1	8D7C	481E		足 𠂔 臧	

附件二、《東坡志林》手動取代表

編號	檔案原文摘錄	手動取代結果
299	<u> </u> （全部取代）	<u>形</u> （全部取代）
300	<u> </u> （全部取代）	<u>□</u> （全部取代）
715	<u>形</u> <u> </u> 止頁 <u>□</u> 王	<u>形</u> <u> </u> 止頁 <u>□</u> 王