《東軒筆錄》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
2008/5/8 葉冠孜 製作

一、 《東軒筆錄》一書(東軒筆錄.xml)使用舊版造字 108 個，字頻 465 次，詳如附件一。這 108 個造字中，93 個可轉成 Windows XP 能顯示的字，字頻 381 次；另外 15 個字必須轉成構字式，字頻 84 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、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 4080 的「群」字，xml 檔在 181 頁漏列了一個註釋，因而造成頻次不同，處理方式為將注釋補上，並將所有頻次皆修改為「群」。

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 1226 的「罰」係「罰」字之異體，故以「罰」字取代。又編號 4016 的「着」字，係「著」字之簡體字，亦以標準字的「著」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 3846 的「墖」字，Unicode 字型呈現為「墖」，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 2564 的「剌」字，Unicode 編碼為 34E8，由於 Windows XP 並不支援，仍須使用構字式「夾 ⿰ 刂」。

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:舊漢籍造字。
丙、 取代：使用字形與造字欄不同時，於此欄列出。
丁、 檔案原文摘錄：摘錄檔案(東軒筆錄.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。
戊、 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、《東軒筆錄》造字分析表

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
74	髣	1	1	FAAB	29B06		髟⿰仿丐	髣，異體字問題
296	⿰仿	5	5	FBEC			⿰仿	

編號	造字	頻次（txt）	頻次（xml）	Big5	Unicode	WinXP	構字式	備註
297		7	7	FBED				
768		1	1	FEED				
821	脅	6	6	8E63	7718	脅		
1076	鳥	4	4	8FE7	5368	鳥		Unicode 呈現差異
1226	罸	2	2	90E0	7F78	罸		罰，異體字問題
1416	埃	1	1	9242	7AE2	埃		
1591	羌	1	1	9354	7F97	羌		
1664	腸	1	1	93BF	8193	腸		腸，異體字問題
1668	膕	1	1	93C3	81C8	膕		
1893	衮	1	1	9548	886E	衮		
1994	達	1	1	95CF	6FBE	達		
2044	踈	3	3	9642	8E08	踈		
2162	醎	2	2	96DA	918E	醎		
2213	鉊	1	1	974E	9277	鉊		
2395	鑾	2	2	9867	947B	鑾		
2502	颱	1	1	98F4	98C8	颱		
2564	剌	1	1	9973	34E8		夾 刂	

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
2651	鸝	3	3	99EC	9E1C	鸝		
2770	絳	2	2	9AC6	7D98	絳		
2885	吋	2	2	9B7A	544C	吋		
3028	礻	2	2	9C6C			礻	
3789	冲	4	4	8154	51B2	冲		沖，異體字問題
3791	况	1	1	8156	51B5	况		況，異體字問題
3792	凉	9	9	8157	51C9	凉		涼，異體字問題
3814	却	10	10	816D	5374	却		卻，異體字問題
3819	厮	5	5	8172	53AE	厮		廝，異體字問題
3826	叶	1	1	8179	53F6	叶		
3843	堦	1	1	81AC	5826	堦		
3846	堺	13	13	81AF	58FB	堺		Unicode 呈現差異
3858	寃	4	4	81BB	5BC3	寃		冤，異體字問題
3868	巵	3	3	81C5	5DF5	巵		
3876	廻	3	3	81CD	5EFB	廻		迴，異體字問題
3886	悞	2	2	81D7	609E	悞		
3891	慙	4	4	81DC	6159	慙		慚，異體字問題

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
3921	晢	1	1	81FA	6673	晢		
3930	椁	1	1	8244	6901	椁		
3935	槩	1	1	8249	69E9	槩		概，異體字問題
3937	檅	2	2	824B	6A9D	檅		
3940	欝	3	3	824E	6B1D	欝		
3944	氷	4	4	8252	6C37	氷		冰，異體字問題
3968	煅	3	3	826A	7145	煅		
3996	疎	3	3	82A8	758E	疎		
4012	皷	10	10	82B8	76B7	皷		
4013	盌	2	2	82B9	76CC	盌		
4015	眦	4	4	82BB	7726	眦		
4016	着	5	5	82BC	7740	着		著，異體字問題
4028	稟	2	2	82C8	7980	稟		稟，異體字問題
4034	穐	1	1	82CE	7A49	穐		
4038	竈	1	1	82D2	7AC8	竈		
4043	笄	1	1	82D7	7B53	笄		
4045	篳	2	2	82D9	7B92	篳		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
4077	鱒	1	1	82F9	7F47	鱒		
4080	羣	25	24	82FC	7FA3	羣		群，標記問題
4085	耻	6	6	8342	803B	耻		恥，異體字問題
4090	脚	13	13	8347	811A	脚		腳，異體字問題
4091	皋	3	3	8348	81EF	皋		
4094	舘	3	3	834B	8218	舘		館，異體字問題
4105	菓	2	2	8356	83D3	菓		
4109	葛	1	1	835A	84AD	葛		
4131	劒	3	3	8370	8842	劒		
4132		41	41	8371			血㑒巫	眾，異體字問題
4146	覩	6	6	83A1	89A9	覩		
4156	詔	1	1	83AB	8B1F	詔		
4160	謫	1	1	83AF	8B81	謫		
4167	賮	5	5	83B6	8CEB	賮		
4176	輙	2	2	83BF	8F19	輙		
4184	迹	14	14	83C7	8FF9	迹		
4193	鈎	2	2	83D0	920E	鈎		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
4198	鑛	1	1	83D5	945B	鑛		
4203	隣	9	9	83DA	96A3	隣		鄰，異體字問題
4242	鬬	10	10	8442	9B2D	鬬		Unicode 呈現差異
4256	鶏	5	5	8450	9DC4	鶏		雞，異體字問題
4308	刧	1	1	84A6	5227	刧		劫，異體字問題
4423	臀	3	3	855A	81CB	臀		臀，異體字問題
4437	廩	1	1	8568	5EEA	廩		廩，異體字問題
4455	劒	4	4	857A			僉⿰刃	劍，異體字問題
4460	勅	18	18	85A1	52D1	勅		敕，異體字問題
4507	吴	54	54	85D0	5434	吴		吳，異體字問題
4515	呪	4	4	85D8	546A	呪		咒，異體字問題
4546	鵐	2	2	85F7	9D76	鵐		
4603	憇	1	1	8671	6187	憇		憩，異體字問題
4702	塲	2	2	86F6	5872	塲		場，異體字問題
4768	妬	1	1	8779	59AC	妬		妒，異體字問題
4806	囓	1	1	87C1	56B1	囓		
4814	徤	2	2	87C9	5FA4	徤		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
5019	罌	2	2	88F9			罌	
5043	殰	1	1	8952	5F4D	殰		
5147	恮	5	5	89DC	6060	恮		
5176		8	8	89F9			氵⿰冗	沉，異體字問題
5218		4	4	8A64			⿹宀莧、⊡	寬，異體字問題
5225	卹	1	1	8A6B	460F		血⿰卩	卹，異體字問題
5290	携	4	4	8ACE	643A	携		
5300	鬪	1	1	8AD8	9B2A	鬪		
5360	鄺	2	2	8B55	913D	鄺		
5439	栢	1	1	8BC6	6822	栢		
5442	栟	1	1	8BC9	67BF	栟		
5478	飠	3	3	8BED	4D77	飠		
5542	鑼	1	1	8C6E	93C1	鑼		
5549	藁	2	2	8C75	85F3	藁		
5578	弛	1	1	8CB4	38AE		弓⿰也	弛，異體字問題
5591	杞	5	5	8CC1	233CC		木⿰巳	杞，異體字問題
5697	煑	1	1	8D6C	7151	煑		煮，異體字問題

編號	造字	頻次（txt）	頻次（xml）	Big5	Unicode	WinXP	構字式	備註
5706	敍	14	14	8D75	654D	敍		敘，異體字問題
5723	牂	1	1	8DA8	4367		牜 ⺶羊	
5742	澂	2	2	8DBB	6F81	澂		澄，異體字問題
5753	潛	1	1	8DC6	6F5C	潛		

附件二、《東軒筆錄》手動取代表

編號	造字	取代	檔案原文摘錄	手動取代結果
4080	羣	群	姑匿之重橑上［四］，出血數升，	姑匿之重橑上<qe>重橑上　群書類編故事卷六同，卷一七引作「重傷腦上」。</qe>，出血數升，
			羣（全部取代）	群（全部取代）