《醋葫蘆》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中央研究院語言所文獻語料小組
2008/7/18　　丁玟伶

轉碼主要工作是把檔案中的舊版造字轉換成 Windows XP 能支援的 Unicode 字形，Unicode 目前共收錄漢字 70194 個字，而 XP 只能支援 20902 個字（詳如表一），不支援之字將以構字式表達。例：造字編號 14 的「徹」字，Unicode 編碼是 38E7，由於 XP 並不支援，仍需使用構字式「⿰彳⿰么月⿺乚⊙」。

表一、Unicode 的字數及編碼區段

Unicode	新增字數	新增編碼區段	總字數	WinXP
1.1 版	20902	4E00-9FFF	20902	支援
3.0 版	6582	3400-4DFF	27484	不支援
3.1 版	42710	20000-2A6D6	70194	不支援

一、舊版造字轉碼分析：

《醋葫蘆》使用舊版造字 68 個，字頻 1036 次，這 68 個造字中，53 個可轉成 Windows XP 能顯示的字，字頻 1001 次；另外 15 個字必須轉成構字式，字頻 35 次。

轉碼完成製作轉碼分析表，請參考附件一《醋葫蘆》轉碼分析表，欄位說明如下：

（一）編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

（二）造字：舊版造字。

（三）頻次：舊版造字在文件的出現次數。

（四）Big5：造字的 Big5 碼。

（五）Unicode：造字所對應的 Unicode 碼。

（六）WinXP：造字在 Windows XP 的對應字形。

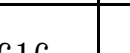
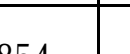
（七）構字式：Windows XP 無法對應字形改用構字式。

（八）備註凡例：備註欄中記錄轉碼後字形及修改原因，凡例如下：

1. 異體字問題：為了使用者查詢和使用的方便，在處理異體字時最主要的方式是以標準字取代，除非是專有名詞或特殊情形，如：人名、地名等。例：造字編號 2252 的「乱」字，是「亂」的異體字，以標準字「亂」取代。
2. 錯字，以程式全部取代：與原書字形不符，並查詢教育部異體字字典確認非異體字後，皆歸類為錯字，以程式取代為正確字形。例：檔案中使用造字編號 4750 的「廾」字，而原書使用之字為「艹」,「廾」為錯字，以程式取代為正確字形「艹」。
3. Unicode 字型呈現差異：Unicode 字型與舊版造字有些微差異，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 3098 的「揸」字，Unicode 字型呈現為「揸」，實際上仍為同一字。
4. 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 3265 的「王曼」。

附件一、《醋葫蘆》轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
14		1	FA4D	38E7			
296		5	FBEC				
297		9	FBED				

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
298		5	FBEE				
639		1	FE4A				
725		4	FEC2				
770		1	FEEF				
1213	睃	1	90D3	7743	睃		
1616		1	936D	284E6			
1854		1	94E0	20E95			
1905	褌	3	9554	88E9	褌		
2046	踪	6	9644	8E2A	踪		
2252	乱	1	9775	4E71	乱		亂，異體字問題。
2602	鰥	1	99BB	9C10	鰥		
2650	鶵	1	99EB	9E18	鶵		
3098	揸	1	9CD4	63F8	揸		Unicode 字型呈現差異。
3265		1	9DDE	3EF4			待造字。
3279		1	9DEC				待造字。
3789	冲	8	8154	51B2	冲		沖，異體字問題。
3791	况	78	8156	51B5	况		況，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3824	叙	12	8177	53D9	叙		
3825	叠	2	8178	53E0	叠		疊，異體字問題。
3826	叶	1	8179	53F6	叶		
3829	咏	2	817C	548F	咏		
3840	坂	1	81A9	5742	坂		
3886	悞	10	81D7	609E	悞		
3899	担	1	81E4	62C5	担		
3909	擡	1	81EE	64E1	擡		
3966	烟	12	8268	70DF	烟		
3969	煊	80	826B	714A	煊		
3974	牀	6	8270	7240	牀		
4000	疴	1	82AC	75B4	疴		
4016	着	398	82BC	7740	着		
4038	竈	5	82D2	7AC8	竈		
4042	笋	4	82D6	7B0B	笋		
4044	筋	1	82D8	7B6F	筋		
4063	糵	1	82EB	7CF5	糵		
4067	綉	8	82EF	7D89	綉		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4068	綫	12	82F0	7DAB	綫		
4073	纊	2	82F5	7E6E	纊		
4085	耻	2	8342	803B	耻		
4107	葱	2	8358	8471	葱		
4109	蓟	1	835A	84AD	蓟		
4115	藁	1	8360	85C1	藁		
4129	虬	1	836E	866C	虬		
4132		120	8371			血𠆢巫	眾，異體字問題。
4139	袴	1	8378	88B4	袴		
4167	賮	14	83B6	8CEB	賮		
4169	趁	1	83B8	8DA6	趁		
4184	迹	14	83C7	8FF9	迹		
4190	郤	1	83CD	90C4	郤		
4193	鈎	10	83D0	920E	鈎		鉤，異體字問題。
4195	鉢	1	83D2	9262	鉢		
4252	鰥	1	844C			魚𠙵眾	鰥，異體字問題。
4256	鷄	20	8450	9DC4	鷄		雞，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4307	罵	46	84A5	99E1	罵		罵，異體字問題。
4347	蟨	1	84CD	27445		虫⿰走越	
4750	卄	1	8767	5344	卄		卄，錯字，以程式全部取代。
4768	妬	85	8779	59AC	妬		妒，異體字問題。
4940	韵	2	88AA	97F5	韵		
5034	弃	1	8949	5F03	弃		棄，異體字問題。
5075	犭	2	8972			犭	
5347	慼	1	8B48	617D	慼		
5404	鴪	1	8BA3	4CED		即⿰走鳥	
5543	鰲	1	8C6F	9B9D	鰲		
5621	泪	13	8CDF	6CEA	泪		淚，異體字問題。
5753	潛	1	8DC6	6F5C	潛		
5797	灩	1	8DF2	7054	灩		

二、手動取代表說明：

《醋葫蘆》檔案中需手動修改的情況分為以下幾種：

（一） 《醋葫蘆》檔案中出現「●」字形，共計 1 字，比對原書後以 Unicod 字形或構字式手動取代，並製作「●」字手動取代表，請參考附件二。欄位說明如下：

1. 頁碼：「●」字形所在位置之原書頁碼。
2. 檔案原文摘錄：摘錄《醋葫蘆》檔案有「●」字形的文句。
3. 原書字形：「●」在原書中之字形，以此字取代檔案原文的「●」。
4. 備註：記錄修改相同詞句的次數或其他事項。

附件二、《醋葫蘆》「●」字手動取代表

頁碼	檔案原文摘錄	原書字形	備註
p.829	盡作●衣之色	染	