《韓詩外傳》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
2011/02/24　　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 961 的「枏」字，Unicode 編碼是 3B4A，由於 XP 並不支援，仍需使用構字式「木⿰冉片」。

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：

《韓詩外傳》使用舊版造字 45 個，字頻 101 次，這 45 個造字中，43 個可轉成 Windows XP 能顯示的字，字頻 97 次；另外 2 個字必須轉成構字式，字頻 4 次。

轉碼完成製作轉碼分析表，請參考附件一《韓詩外傳》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 4134 的「衞」字，是「衛」的異體字，以標準字「衛」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 5308 的「檅」字，而原書使用之字為「檅」，查詢教育部異體字字典確認「檅」並非「檅」之異體字，因此歸為錯字，以程式取代為正確字形「檅」。
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 4442 的「潛」字，Unicode 字型呈現為「潛」，實際上仍為同一字。

附件一、《韓詩外傳》轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
826	癱	1	8E68	7655	癱		
910	窾	1	8EDE	5D78	窾		
961	枎	3	8F52	3B4A		木⿰厶片	
982	胄	1	8F67	80F7	胄		
1024	虵	1	8FB3	8675	虵		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
1871	𧘇	1	94F1	87C1	𧘇		
1981	𪚥	1	95C2	8B90	𪚥		
2044	跾	2	9642	8E08	跾		
2053	躚	2	964B	8EA7	躚		
2152	鞬	1	96D0	979F	鞬		
2244	虫	1	976D			ノ𠆢虫	
2572	衒	1	997B	8858	衒		
3804	効	3	8163	52B9	効		
3813	卬	2	816C	536D	卬		
3850	姏	1	81B3	59D9	姏		
3883	徧	3	81D4	5FA7	徧		
3891	慙	6	81DC	6159	慙		
3929	棃	1	8243	68C3	棃		
3930	椁	1	8244	6901	椁		
3994	毕	1	82A6	754A	毕		
3996	跾	5	82A8	758E	跾		
4010	皋	5	82B6	7690	皋		皋，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4035	穻	1	82CF	7A7D	穻		
4038	竈	1	82D2	7AC8	竈		
4103	茔	1	8354	831D	茔		
4109	葛	2	835A	84AD	葛		
4130	衅	1	836F	8845	衅		
4134	衛	20	8373	885E	衛		衛，異體字問題。
4140	衹	2	8379	88B5	衹		
4167	賮	2	83B6	8CEB	賮		
4184	迹	2	83C7	8FF9	迹		
4193	鈎	3	83D0	920E	鈎		
4242	鬬	3	8442	9B2D	鬬		Unicode 字型呈現差異。
4267	鼇	4	845B	9F08	鼇		
4342	倮	1	84C8	502E	倮		
4442	潛	1	856D	6F98	潛		Unicode 字型呈現差異。
4519	咲	1	85DC	54B2	咲		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4768	妬	1	8779	59AC	妬		
4996	斷	1	88E2	65B5	斷		
5147	恆	1	89DC	6060	恆		
5308	機	1	8AE0			棹⿰木戈	機，錯字，以程式全部取代。
5482	蠻	1	8BF1	8812	蠻		
5680		1	8D5B			⿰形殼、心⊙	慤，錯字，以程式全部取代。
5684	滓	2	8D5F	6E12	滓		
5722	暮	4	8DA7	671E	暮		