《一片情》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
2008/7/04　　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 5094 的「翉」字，Unicode 編碼是 26407，由於 XP 並不支援，仍需使用構字式部件「翉」。

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：

《一片情》使用舊版造字 44 個，字頻 816 次，這 44 個造字中，37 個可轉成 Windows XP 能顯示的字，字頻 606 次；另外 7 個字必須轉成構字式，字頻 210 次。

轉碼完成製作轉碼分析表，請參考附件一《一片情》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 3791 的「况」字，是「況」的異體字，以標準字「況」取代。
2. 部份使用不同字形，手動修改：檔案中為同一舊版造字，因異體字問題，將專有名詞或特殊情形之舊版造字保留，剩餘的頻次則用標準字取代。例：造字編號 3789 的「冲」字頻次為 4，依書將人名的 2 字轉為「冲」，其他 2 字手動修改為標準字「沖」。手動修改部份請參考附件二：《一片情》手動取代表。
3. 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 2589 的「笪」字，而原書使用之字為「笪」,「笪」為錯字，以程式取代為正確字形「笪」。
4. 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。

附件一、《一片情》轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
256	叟	1	FBC4			叟	
296	㑩	59	FBEC			㑩	

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
297	𠙵	42	FBED			𠙵	
298	𠙶	80	FBEE			𠙶	
678	𠂤	2	FE71			𠂤	
770	扌	25	FEEF			扌	
1213	睃	3	90D3	7743	睃		
2046	踪	3	9644	8E2A	踪		
2524	饎	1	994B	994D	饎		
2589	笪	1	99AE	7B21	笪		笪，錯字，以程式全部取代。
2999	唿	2	9C4F	5523	唿		
3001	蜡	1	9C51	87A5	蜡		
3098	揸	1	9CD4	63F8	揸		Unicode 字型呈現差異。
3161	壜	3	9D54	58DC	壜		
3789	冲	4	8154	51B2	冲		異體字問題。部份使用不同字形「沖」，手動修改。
3791	况	23	8156	51B5	况		況，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3815	嗇	1	816E	537A	嗇		
3824	叙	2	8177	53D9	叙		
3825	叠	24	8178	53E0	叠		疊，異體字問題。
3829	咏	1	817C	548F	咏		
3830	咤	1	817D	54A4	咤		
3938	橈	8	824C	6AC8	橈		
3966	烟	11	8268	70DF	烟		
3974	牀	2	8270	7240	牀		
4016	着	342	82BC	7740	着		
4018	矚	1	82BE	77AF	矚		
4038	竈	1	82D2	7AC8	竈		
4042	笋	6	82D6	7B0B	笋		
4044	筋	2	82D8	7B6F	筋		
4068	綫	2	82F0	7DAB	綫		
4073	纊	1	82F5	7E6E	纊		
4085	耻	5	8342	803B	耻		
4104	荢	1	8355	8385	荢		
4105	菓	1	8356	83D3	菓		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4107	葱	2	8358	8471	葱		
4132		56	8371			血⿰巫	眾，異體字問題。
4193	鈎	4	83D0	920E	鈎		鉤，異體字問題。
4256	鷄	13	8450	9DC4	鷄		雞，異體字問題。
4307	罵	26	84A5	99E1	罵		罵，異體字問題。
4768	妬	3	8779	59AC	妬		妒，異體字問題。
4798		39	87B9			囗⿰夕	窗，錯字，以程式全部取代。
5094	翉	1	89A7	26407		翉	
5543	鴦	1	8C6F	9B9D	鴦		
5753	潛	8	8DC6	6F5C	潛		

二、手動取代表說明：

《一片情》檔案中之舊版造字，因字形部份不同而無法全部以程式取代，皆手動修改成Windows XP能顯示的字或構字式，並製作手動取代表，請參考附件二。欄位說明如下：

1. 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。

2. 造字：舊版造字字形。

3. 取代：實際使用字形與舊版造字不同時，於此欄列出。
4. 檔案原文摘錄：摘錄檔案《一片情》中的文句段落，其中紅字底線的部分，為需要手動取代的原文。
5. 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。
6. 備註：記錄修改相同辭句的次數或其他事項。

附件二、《一片情》手動取代表

編號	造字	取代	檔案原文摘錄	手動取代結果	備註
3789	冲	沖	怒髮庱冠	怒髮沖冠	
			一庱行去	一沖行去	