《全相平話五種》(標記語料)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
2012/08/17　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 1808 的「槴」字，Unicode 編碼是 3BB6，由於 XP 並不支援，仍需使用構字式「木⿰朔」。

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：

《全相平話五種》(標記語料)使用舊版造字 79 個，字頻 410 次，這 79 個造字中，62 個可轉成 Windows XP 能顯示的字，字頻 359 次；另外 17 個字必須轉成構字式，字頻 51 次。

轉碼完成製作轉碼分析表，請參考附件一《全相平話五種》(標記語料)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「衛」字，是「衛」的異體字，以標準字「衛」取代。
2. 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 3935 的「槩」字，Unicode 字型呈現為「槩」，實際上仍為同一字。

附件一、《全相平話五種》(標記語料)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
296		2	FBEC				
297		24	FBED				
298		3	FBEE				
759	广	2	FEE4			广	
768		1	FEED	8279			
770		1	FEEF	624C			
780		1	FEF9				
784		3	FEFD				
1067	棊	1	8FDE	68CA	棊		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
1073		8	8FE4			箕𠆢么	篡，異體字問題。
1278	碍	1	9155	788D	碍		
1491	擷	2	92AF	6527	擷		
1520	曒	1	92CC	3B20		日𠙵煞	
1808	槰	1	94B2	3BB6		木𠙵朔	
1885	鑾	1	9540	882D	鑾		
1905	褍	1	9554	88E9	褍		
2157	酎	2	96D5	9167	酎		
2343	靼	1	97F2	4A5E		革𠙵占	
2432	鞮	1	98AE	29350		革𠙵登	
2796	鍕	1	9AE0	295E9		金𠙵風	
2833	刃	4	9B46	5204	刃		
3069	彡	4	9CB7			彡	
3071	彡	2	9CB9			彡	
3072	彡	2	9CBA			彡	
3789	冲	1	8154	51B2	冲		沖，異體字問題。
3792	凉	2	8157	51C9	凉		涼，異體字問

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
							題。
3797	凭	1	815C	51ED	凭		
3801	剋	1	8160	5257	剋		
3826	叶	2	8179	53F6	叶		
3830	咤	2	817D	54A4	咤		
3834	嗶	1	81A3	5637	嗶		Unicode 字型呈現差異。
3840	坂	18	81A9	5742	坂		
3843	堦	8	81AC	5826	堦		
3846	堺	4	81AF	58FB	堺		
3864	峯	3	81C1	5CEF	峯		
3876	廻	2	81CD	5EFB	廻		
3886	悞	1	81D7	609E	悞		
3909	擡	1	81EE	64E1	擡		
3931	椀	2	8245	6900	椀		
3935	槩	2	8249	3BA3	槩		Unicode 字型呈現差異。
3966	烟	10	8268	70DF	烟		
3977	狥	1	8273	72E5	狥		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3978	猪	2	8274	732A	猪		Unicode 字型呈現差異。
4000	疴	1	82AC	75B4	疴		
4021	磞	2	82C1	78AF	磞		
4042	笋	1	82D6	7B0B	笋		
4067	綉	22	82EF	7D89	綉		
4069	綳	1	82F1	7DB3	綳		
4080	羣	43	82FC	7FA3	羣		
4093	舐	1	834A	445B		舌厶氏	
4094	舘	1	834B	8218	舘		
4105	菓	1	8356	83D3	菓		
4107	葱	3	8358	8471	葱		
4115	藁	1	8360	85C1	藁		
4124	蝨	1	8369	8771	蝨		
4129	虬	6	836E	866C	虬		
4134	衛	17	8373	885E	衛		衛，異體字問題。
4139	袴	1	8378	88B4	袴		
4146	覩	18	83A1	89A9	覩		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4167	賮	20	83B6	8CEB	賮		
4173	躭	1	83BC	8EAD	躭		
4184	迹	3	83C7	8FF9	迹		
4195	鉢	4	83D2	9262	鉢		
4223	顒	1	83EE	984B	顒		
4226	飡	1	83F1	98E1	飡		
4231	騌	1	83F6	9A0C	騌		
4256	鷄	2	8450	9DC4	鷄		
4267	鼇	1	845B	9F08	鼇		
4309	刱	4	84A7	523C	刱		
4460	勅	57	85A1	52D1	勅		
4515	呪	3	85D8	546A	呪		
4768	妬	4	8779	59AC	妬		
5008	幙	2	88EE	5E5E	幙		
5034	弃	3	8949	5F03	弃		
5112	园	1	89B9	56ED	园		
5300	鬪	34	8AD8	9B2A	鬪		
5551	賮	11	8C77	8CF7	賮		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
5558	龕	2	8C7E	9E84	龕		
5860	无	1	C6D3	65E0	无		

二、手動取代表說明：

《全相平話五種》(標記語料)檔案中需手動修改的情況分為以下幾種：

（一）《全相平話五種》(標記語料)檔案中出現「●」字形，共計 1 字，比對原書後以 Unicode 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. 標記行碼：語言所標記檔裡設定的行碼。
3. 檔案原文摘錄：摘錄《全相平話五種》(標記語料)檔案有「●」字形的文句。
4. 原書字形：「●」在原書中之字形，以此字取代檔案原文的「●」。
5. 備註：記錄修改相同詞句的次數或其他事項。

附件二、《全相平話五種》(標記語料)「●」字手動取代表

頁碼	標記行碼	檔案原文摘錄	原書字形	備註
p.227	24645	臂(Na)　忙●(U)　蹄(Na)		

（二）《全相平話五種》素語料中之「●」字，為日後線上檢索方便，在標記語料多修正為古今通用字，與文獻處理實驗室較以原書為準修改其轉碼後素語料檔案的「●」字，兩者用字略有

不同，其中多為異體字差異，特製作用字差異記錄表，以供未來參考。請參考附件三《全相平話五種》素語料與標記語料「●」字差異記錄表。欄位說明如下：

1. 頁碼：修改字形所在位置之原書頁碼。
2. 素語料原文摘錄：摘錄《全相平話五種》素語料檔案的文句段落，其中紅字部分，是經由文獻處理實驗室轉碼修改後的「●」字。
3. 標記語料原文摘錄：摘錄《全相平話五種》標記語料檔案的文句段落，其中藍字的部分，為標記語料修正素語料中之「●」字形。
4. 備註：記錄修改相同辭句的次數或其他事項。

附件三、《全相平話五種》素語料與標記語料「●」字差異記錄表

頁碼	素語料原文摘錄	標記語料原文摘錄	備註
p.227	叫牙朱火东	叫牙朱煠	

（三）《全相平話五種》(標記語料)檔案中，對於舊版造字以外發現的錯字，以手動修改，並製作錯字修改表，請參考附件四。欄位說明如下：

1. 檔案原文摘錄：摘錄檔案《全相平話五種》(標記語料)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。
2. 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。
3. 備註：記錄修改相同詞句的次數或其他事項。

附件四、《全相平話五種》(標記語料)缺字外錯字修改

檔案原文摘錄	手動取代結果	備註
蕭古達(Nb)　橫(VHC)	蕭古達(Nb)　橫(VHC)	漢語大字典該字為

檔案原文摘錄	手動取代結果	備註
丈(Nf)　四(Neu)　紫金(Na)　穴𠔃戌(Na)	丈(Nf)　四(Neu)　紫金(Na)　穴𠔃戌(Na)	穴𠔃戌