

《五代史平話》(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
2012/08/09 丁致伶

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 1616 的「運」字，Unicode 編碼是 284E6，由於 XP 並不支援，仍需使用構字式「辶△里」。

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：

《五代史平話》(標記語料)使用舊版造字 16 個，字頻 22 次，這 16 個造字中，10 個可轉成 Windows XP 能顯示的字，字頻 13 次；另外 6 個字必須轉成構字式，字頻 9 次。

轉碼完成製作轉碼分析表，請參考附件一《五代史平話》(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. 構字式部件：由於漢字構形資料庫目前無法支援 Unicode，因此無法判讀構字式中的 Unicode 部件，構字式的 Unicode 部件應改為 Big5 部件。例：造字編號 3021 的「𠂇」字，為 Unicode 字形，因其為構字式組合部件，需改為 Big5 部件之「𠂇」取代。

附件一、《五代史平話》(標記語料)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
296	厶	2	FBEC			厶	
297	厶	3	FBED			厶	
432	業	1	FCD7			業	
717	足	1	FEBA			足	
899	馳	1	8ED3	99DE	馳		
1491	擲	4	92AF	6527	擲		
1616	運	1	936D	284E6		這厶里	
2999	咆	1	9C4F	5523	咆		
3021	𠂇	1	9C65	5FC4	𠂇		𠂇，構字式部件。
3840	坂	1	81A9	5742	坂		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4000	疴	1	82AC	75B4	疴		
4035	穽	1	82CF	7A7D	穽		
4232	隲	1	83F7	96B2	隲		
5008	幪	1	88EE	5E5E	幪		
5074	𠂔	1	8971	5C2B	𠂔		
5707	棧	1	8D76	6830	棧		

二、手動取代表說明：

《五代史平話》(標記語料)檔案中需手動修改的情況分為以下幾種：

(一) 《五代史平話》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。請參考附件二《五代史平話》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《五代史平話》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《五代史平話》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。

附件二、《五代史平話》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. 42	迫脇天子	迫脅(VF) 天子	
p. 43	鐘簾如故	鐘虞(Na) 如故	
p. 46	賀瓌	賀瑰(Nb)	有 9 處
p. 57	李鷗兒	李鴉兒(Nb)	有 2 處
p. 60	夜夢黑鷗	夜(Nd) 夢(VE) 黑 (VH) 鴉	
p. 60	有一鷗	有(V_2) 一(Neu) 鴉	
p. 60	鷗軍到	鴉(Na) 軍(Na) 到	
p. 69	賜鷗鵠酒卮	賜(VD) 鷗鵠酒卮	
p. 70	所有脇乘輿	所有(Neqa) 脅(VK) 乘輿	
p. 75	稔惡召衅	稔惡(VH) 召(VF) 釁	
p. 76	脇服其眾	脅(VK) 服(VK) 其 (Nh) 眾	
p. 76	逼脇德倫	逼脅(VJ) 德倫	
p. 91	舟△見，彥章退保	艦(Na)	
p. 95	觀衅而動	觀(VE) 釁(Na) 而 (Cbb) 動	
p. 95	鄭珽	鄭珽(Nb)	

頁碼	素語料原文摘錄	標記語料原文摘錄	備註
p. 97	李繼 葬	李繼 麟 (Nb)	
p. 100	巫 尪	巫 尪 (Na)	有 2 處
p. 108	為人輕 雋	為人(Na) 輕 雋	
P. 108	周 瓌	周 瑰 (Nb)	有 3 處
p. 111	堪笑 鴟 兒	堪(VL) 笑(VE) 鴟 兒	
p. 123	隱 矛 稍	隱(Nb) 矛(Na) 槩 (Na)	
p. 135	卮 酒	卮 (Nf) 酒(Na)	有 3 處
p. 146	自啟 衅 端	自(Nh) 啟(VC) 釁 端	
p. 147	衅 而動	釁 (Na) 而(Cbb) 動	
p. 147	浼 卿一行	浼 (VF) 卿(Nh) 一 (Neu) 行	
p. 155	連年兵 衅	連(VJ) 年(Nd) 兵 釁	
p. 169	伺 衅 隙	伺(VJ) 釁 隙(Na)	
p. 177	迫 脇	迫 脅 (VF)	有 2 處
p. 197	以 蒞 武事	以(Cbb) 蒞 (VC) 武事	
p. 219	席 藁	席 稿 (Na)	
p. 224	才 冘 屍首歸葬	冘 (VC) 屍首(Na) 歸 葬	
p. . 238	一大 卮	一(Neu) 大(VH) 卮 (Na)	

頁碼	素語料原文摘錄	標記語料原文摘錄	備註
p. 248	汜 <u>△</u> 公流而北	沿(P) 流(Na) 而 (Cbb) 北	

(二)《五代史平話》(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，請參考附件三。欄位說明如下：

1. 檔案原文摘錄：摘錄檔案《五代史平話》(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。

附件三、《五代史平話》(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
橫(VHC) 移(VC) <u>殆</u> (Dd) 宿(Nd)	橫(VHC) 移(VC) <u>星</u> (Dd) 宿(Nd)	錯字後之標記符號 應修改