

《歧路燈》(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
2012/08/08 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 1691 的「鬚」字，Unicode 編碼是 4BFC，由於 XP 並不支援，仍需使用構字式「鬚𠄎狄」。

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：

《歧路燈》(標記語料)使用舊版造字 135 個，字頻 2695 次，這 135 個造字中，116 個可轉成 Windows XP 能顯示的字，字頻 2593 次；另外 19 個字必須轉成構字式，字頻 102 次。

轉碼完成製作轉碼分析表，請參考附件一《歧路燈》(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 3978 的「猪」字，Unicode 字型呈現為「猪」，實際上仍為同一字。


附件一、《歧路燈》(標記語料)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
296	𠄎	13	FBEC			𠄎	
297	𠄎	37	FBED			𠄎	
298	𠄎	7	FBEE			𠄎	
639	彳	3	FE4A			彳	
658	疒	2	FE5D	7592		疒	
717	疒	2	FEBA			疒	
752	𠄎	1	FEDD			𠄎	
768	𠄎	2	FEED	8279		𠄎	
770	扌	1	FEEF	624C		扌	

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
1201	昵	1	90C7	7724	昵		
1278	碍	7	9155	788D	碍		
1634	忬	7	93A1	5FEC	忬		
1641	胆	9	93A8	80C6	胆		
1642	脉	8	93A9	8109	脉		
1691	髻	1	93DA	4BFC		髻𠄎狄	
2046	踪	20	9644	8E2A	踪		
2206	祆	1	9747	8884	祆		
2361	鎗	1	9845	93F3	鎗		
2374	冢	1	9852	51A1	冢		
2408	鬪	8	9874	9599	鬪		鬪，異體字問題。
2415	鬪	2	987B	95D8	鬪		鬪，異體字問題。
2448	隼	3	98BE	96BD	隼		
2592	臟	10	99B1	8D1C	臟		
2758	獻	1	9ABA	732E	獻		獻，異體字問題。
2812	舉	4	9AF0	64E7	舉		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
2883	涂	1	9B78	51C3	涂		
2997	鹽	2	9C4D	25081		形土 宀 口皿	
2999	咆	5	9C4F	5523	咆		
3016	闕	1	9C60	28D91		門△臭	
3160	坟	3	9D53	575F	坟		
3168	瘡	1	9D5B	24EA5		疒△答	
3177	蠃	3	9D64	7F4E	蠃		
3251	灯	4	9DD0	706F	灯		
3789	冲	21	8154	51B2	冲		沖，異體字問題。
3791	况	21	8156	51B5	况		況，異體字問題。
3792	凉	5	8157	51C9	凉		涼，異體字問題。
3794	凑	20	8159	51D1	凑		湊，異體字問題。
3797	凭	2	815C	51ED	凭		
3805	勅	1	8164	52C5	勅		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3814	却	594	816D	5374	却		
3815	𡗗	1	816E	537A	𡗗		𡗗，異體字問題。
3817	廁	13	8170	53A0	廁		
3818	𡗗	1	8171	53AB	𡗗		
3819	𡗗	42	8172	53AE	𡗗		𡗗，異體字問題。
3824	叙	11	8177	53D9	叙		
3825	疊	9	8178	53E0	疊		
3829	咏	2	817C	548F	咏		
3849	獎	7	81B2	596C	獎		獎，異體字問題。
3851	姪	1	81B4	59F9	姪		
3858	冤	19	81BB	5BC3	冤		冤，異體字問題。
3883	徧	1	81D4	5FA7	徧		
3890	憑	1	81DB	6142	憑		
3893	倭	1	81DE	4FAB		亻 𡗗 倭	
3899	担	1	81E4	62C5	担		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3909	擡	34	81EE	64E1	擡		
3921	暫	2	81FA	6673	暫		
3930	憵	8	8244	6901	憵		
3934	榻	19	8248	69C5	榻		
3938	欖	8	824C	6AC8	欖		
3942	毡	3	8250	6BE1	毡		
3966	烟	25	8268	70DF	烟		
3975	牕	2	8271	7255	牕		
3976	犁	3	8272	7282	犁		
3978	猪	33	8274	732A	猪		Unicode 字型 呈現差異。
3979	猫	4	8275	732B	猫		
4000	疴	1	82AC	75B4	疴		
4016	着	437	82BC	7740	着		
4028	稟	111	82C8	7980	稟		
4030	秆	2	82CA	79C6	秆		
4032	秣	1	82CC	2578A	禾  未		秣，異體字問 題。
4036	窰	4	82D0	7AB0	窰		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4038	竈	1	82D2	7AC8	竈		
4041	豎	4	82D5	7AEA	豎		
4042	笋	8	82D6	7B0B	笋		
4066	綉	2	82EE	7D5D	綉		
4067	綉	24	82EF	7D89	綉		
4068	綫	9	82F0	7DAB	綫		
4073	纒	2	82F5	7E6E	纒		
4074	縑	2	82F6	7E27	縑		
4085	耻	11	8342	803B	耻		
4090	脚	165	8347	811A	脚		
4107	葱	6	8358	8471	葱		
4111	蕤	1	835C	8493	蕤		
4129	虬	3	836E	866C	虬		
4130	衅	2	836F	8845	衅		
4131	劔	1	8370	8842	劔		
4134	衛	6	8373	885E	衛		衛，異體字問題。
4139	袴	1	8378	88B4	袴		
4140	裱	1	8379	88B5	裱		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4145	羈	1	837E	898A	羈		Unicode 字型 呈現差異。
4146	覩	1	83A1	89A9	覩		
4167	賫	2	83B6	8CEB	賫		
4169	趁	1	83B8	8DA6	趁		
4173	駛	6	83BC	8EAD	駛		
4184	迹	23	83C7	8FF9	迹		
4193	鈎	7	83D0	920E	鈎		
4195	鉢	3	83D2	9262	鉢		
4218	萑	4	83E9	97EE	萑		
4247	鮎	1	8447	9B8E	鮎		
4256	鷄	55	8450	9DC4	鷄		
4262	麪	4	8456	9EAA	麪		Unicode 字型 呈現差異。
4267	鼈	1	845B	9F08	鼈		
4297	躔	1	8479	281E0		𠃉足厂 墨𠃉	
4307	罵	45	84A5	99E1	罵		罵，異體字問 題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4308	劫	3	84A6	5227	劫		
4316	艷	2	84AE	8276	艷		
4345	嗽	3	84CB			口 ㄩ 敕	
4372	廩	12	84E6			形 广 回 禾 口	
4432		55	8563			形 纟 厂 墨 口	纏，異體字問題。
4447	捏	30	8572	63D1	捏		
4448		1	8573			形 纟 厂 墨 口	纏，異體字問題。
4476		1	85B1			良 ㄩ 尸	即，異體字問題。
4518	响	1	85DB	54CD	响		
4531	攢	1	85E8	6505	攢		
4616	壳	1	867E	58F3	壳		
4742		2	875F			衣 ㄩ 執	褻，異體字問題。
4761	標	1	8772	6AA9	標		
4768	妬	1	8779	59AC	妬		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4805	瀆	1	87C0	51DF	瀆		瀆，異體字問題。
4940	韵	12	88AA	97F5	韵		
5005	帮	1	88EB	5E2E	帮		
5008	幪	3	88EE	5E5E	幪		
5075	彡	3	8972	72AD		彡	
5101	馱	9	89AE	4B7E		馬厶犬	
5163	憂	3	89EC	621E	憂		
5290	携	37	8ACE	643A	携		
5299	戟	3	8AD7	39B8		卓厶戈	戟，異體字問題。
5543	養	2	8C6F	9B9D	養		
5590	朶	2	8CC0	6736	朶		
5591	杞	3	8CC1	233CC		木厶巳	杞，異體字問題。
5621	泪	23	8CDF	6CEA	泪		
5702	厨	110	8D71	53A8	厨		
5706	敘	48	8D75	654D	敘		敘，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
5753	潛	270	8DC6	6F5C	潛		
5768	濶	4	8DD5	6FF6	濶		

二、手動取代表說明：

《歧路燈》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《歧路燈》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《歧路燈》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《歧路燈》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、《歧路燈》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. 35	門△龜一韵	鬪(VC) 一(Neu) 韵	
p. 224	台公冢孫	台公(Na) 冢孫	
p. 290	璀璨奪目	璀灿奪目	
p. 302	茅拔茹賴箱	茅拔茹賴箱	

頁碼	素語料原文摘錄	標記語料原文摘錄	備註
p. 441	小字滙兒	小(VH) 字[羽]; 口佳[口]	標記語料改為滙
p. 442	一總算	一總(Dab) 算(VE)	
p. 798	燦耀奪目	燦耀(Dh) 奪目	
p. 809	與相公淩	與(P) 相公(Na) ; ㄩ ㄩ	標記語料改為淩
p. 809	吃淩柿	吃(VC) ; ㄩ 淩(VC)	標記語料改為淩
p. 1060	必定口ㄩ么喝你	必定(Dba) 吆喝	
p. 1068	千斤白砵	千(Neu) 斤(Nf) 白礬	
p. 1068	紙上加砵	紙(Na) 上(Ng) 加(VC) 礬	
p. 1068	一分白砵	一(Neu) 分(Nf) 白礬	
p. 1075	月ㄩ脅從誨	脅從(Na) 誨	
p. 1079	口ㄩ么喝官府	吆喝(VE) 官府(Na)	
p. 1080	撫藩自公本的旗	督撫(Na) 藩臬	