《儒林外史》(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
2012/09/07　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「髟⿰公狄」。

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：

《儒林外史》(標記語料)使用舊版造字 107 個，字頻 3927 次，這 107 個造字中，90 個可轉成 Windows XP 能顯示的字，字頻 3879 次；另外 17 個字必須轉成構字式，字頻 48 次。

轉碼完成製作轉碼分析表，請參考附件一《儒林外史》(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. 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 4750 的「廾」字，而原書使用之字為「艹」，查詢教育部異體字字典確認「艹」並非「廾」之異體字，因此歸為錯字，以程式取代為正確字形「艹」。
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 3978 的「猪」字，Unicode 字型呈現為「猪」，實際上仍為同一字。
4. 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「  」字，為 Unicode 字形，因其為構字式組合部件，需改為 Big5 部件之「」取代。

附件一、《儒林外史》(標記語料)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
296		5	FBEC				
297		23	FBED				

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
299		1	FBEF				
300		1	FBF0				
304		1	FBF4				
755		1	FEE0				心，錯字，以程式全部取代。
770		1	FEEF	624C			
1153		2	9075	362D			
1213	睃	1	90D3	7743	睃		
1278	碍	3	9155	788D	碍		
1332		8	91AD				攜，異體字問題。
1359		1	91C8	25725			
1513	粮	1	92C5	7CAE	粮		糧，異體字問題。
1691		2	93DA	4BFC			
1981	讐	1	95C2	8B90	讐		
2046	踪	10	9644	8E2A	踪		
2157	酎	1	96D5	9167	酎		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
2346	鐧	1	97F5	9427	鐧		
2592	贓	1	99B1	8D1C	贓		
2749	柏	1	9AB1	67B1	柏		
3016	闐	1	9C60	28D91		門⿰臭	
3021	忄	1	9C65	5FC4	忄		忄，構字式部件。
3097	倖	1	9CD3			亻⿰幸	
3168	瘩	1	9D5B	24EA5		疒⿰答	
3177	鱷	14	9D64	7F4E	鱷		
3789	冲	6	8154	51B2	冲		沖，異體字問題。
3797	凭	2	815C	51ED	凭		
3799	刦	2	815E	5226	刦		
3804	効	13	8163	52B9	効		
3805	勅	2	8164	52C5	勅		
3814	却	316	816D	5374	却		
3815	卺	1	816E	537A	卺		
3819	厮	1	8172	53AE	厮		廝，異體字問題。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3820	厦	4	8173	53A6	厦		
3825	叠	1	8178	53E0	叠		
3829	咏	2	817C	548F	咏		
3843	堦	4	81AC	5826	堦		
3846	堺	38	81AF	58FB	堺		
3849	奬	3	81B2	596C	奬		奬，異體字問題。
3858	寃	22	81BB	5BC3	寃		冤，異體字問題。
3864	峯	36	81C1	5CEF	峯		
3876	廻	4	81CD	5EFB	廻		
3883	徧	3	81D4	5FA7	徧		
3884	忽	8	81D5	6031	忽		
3909	擡	71	81EE	64E1	擡		
3934	椢	6	8248	69C5	椢		
3942	毡	2	8250	6BE1	毡		
3966	烟	6	8268	70DF	烟		
3974	牀	78	8270	7240	牀		
3976	犂	1	8272	7282	犂		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3978	猪	53	8274	732A	猪		Unicode 字型呈現差異。
3996	踈	8	82A8	758E	踈		
4008	癵	1	82B4	766F	癵		
4016	着	2439	82BC	7740	着		
4036	窯	2	82D0	7AB0	窯		
4038	竈	1	82D2	7AC8	竈		
4040	竝	1	82D4	7ADD	竝		
4041	竪	1	82D5	7AEA	竪		
4042	笋	5	82D6	7B0B	笋		
4043	笩	1	82D7	7B53	笩		笩，異體字問題。
4044	筋	6	82D8	7B6F	筋		
4074	縧	6	82F6	7E27	縧		
4077	罇	3	82F9	7F47	罇		
4080	羣	11	82FC	7FA3	羣		
4090	脚	155	8347	811A	脚		
4093	舐	4	834A	445B		舌  氏	
4105	菓	8	8356	83D3	菓		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4112	蘇	7	835D	8534	蘇		
4132		247	8371			血⿰巫	眾，異體字問題。
4134	衛	24	8373	885E	衛		衛，異體字問題。
4139	袴	2	8378	88B4	袴		
4173	躭	17	83BC	8EAD	躭		
4184	迹	1	83C7	8FF9	迹		
4193	鈎	2	83D0	920E	鈎		
4195	鉢	2	83D2	9262	鉢		
4232	隲	1	83F7	96B2	隲		
4256	鷄	15	8450	9DC4	鷄		
4267	鼈	2	845B	9F08	鼈		
4268	鼓	3	845C	9F17	鼓		
4307	罵	52	84A5	99E1	罵		罵，異體字問題。
4308	刧	3	84A6	5227	刧		
4428	蹰	3	855F	8E70	蹰		
4432		1	8563			⿴形糸 厂	纏，異體字問

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
						墨⊙	題。
4447	揑	14	8572	63D1	揑		
4518	响	3	85DB	54CD	响		
4684		1	86E4			礻⩓受	祲，錯字，以程式全部取代。
4750	卄	1	8767	5344	卄		艹，錯字，以程式全部取代。
4805	瀆	1	87C0	51DF	瀆		
4816	亻	1	87CB			亻	
4904	尅	2	8864	5C05	尅		
4996	斷	1	88E2	65B5	斷		
5008	幙	1	88EE	5E5E	幙		
5015	竄	7	88F5	5BD7	竄		
5075	犭	1	8972	72AD		犭	
5176		14	89F9			氵⩓冗	沉，異體字問題。
5195		2	8A4D			形木亩	檀，異體字問

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
						且⊙	題。
5229	撐	1	8A6F	6490	撐		
5230		1	8A70			⿰土亘 且⊙	壇，異體字問題。
5290	携	9	8ACE	643A	携		
5300	鬪	1	8AD8	9B2A	鬪		
5559	欵	1	8CA1	6B35	欵		
5590	朵	8	8CC0	6736	朵		
5595	垛	2	8CC5	579C	垛		
5666	汚	2	8D4D	6C5A	汚		污，異體字問題。
5702	厨	8	8D71	53A8	厨		
5706	敍	41	8D75	654D	敍		敘，異體字問題。
5742	澁	1	8DBB	6F81	澁		

二、手動取代表說明：

《儒林外史》(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，請參考附件二。欄位說明如下：

1. 檔案原文摘錄：摘錄檔案《儒林外史》(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。

附件二、《儒林外史》(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
烏雲(Na)　髟𠂤委髟𠂤隨	烏雲(Na)　髟𠂤委髟𠂤隋	根據漢語大字典修正。