《毛詩》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
2010/04/28　　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 2042 的「疏」字，Unicode 編碼是 47FD，由於 XP 並不支援，仍需使用構字式「疋⿰亠厷」。

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：

《毛詩》使用舊版造字 76 個，字頻 224 次，這 76 個造字中，68 個可轉成 Windows XP 能顯示的字，字頻 198 次；另外 8 個字必須轉成構字式，字頻 26 次。

轉碼完成製作轉碼分析表，請參考附件一《毛詩》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 4032 的「秣」字，是「秣」的異體字，以標準字「秣」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 4902 的「宪」字，而原書使用之字為「衮」，查詢教育部異體字字典確認「衮」並非「宪」之異體字，因此歸為錯字，以程式取代為正確字形「衮」。
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 5705 的「廐」字，Unicode 字型呈現為「廄」，實際上仍為同一字。
4. 該處無字，手動刪除：檔案中使用之舊版造字，校對原書後發現該處無字，以程式轉碼後手動刪除。例：檔案中使用造字編號 533 的「朮」字，而原書該處無字，以程式轉碼後手動刪除。手動修改部份請參考附件二《毛詩》手動取代表。

附件一、《毛詩》轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
533	朮	1	FD7D			朮	該處無字，手動刪除。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
858	罙	1	8EAA	7F59	罙		
915	遲	2	8EE3	905F	遲		
938	燬	1	8EFA	7215	燬		
947	爆	2	8F44	7217	爆		
1062		2	8FD9			馬  帀	騀，異體字問題。
1079	總	4	8FEA	7DEB	總		
1381	稱	1	91DE	7A3A	稱		
1565	縣	15	92F9	7DDC	縣		
1893	衮	1	9548	886E	衮		
2042	疏	1	9640	47FD		足  流	
2142	景	2	96C6	7758	景		Unicode 字型呈現差異。
2168	醻	6	96E0	91BB	醻		
2472	靐	6	98D6	9741	靐		
2858	潀	2	9B5F	6EFA	潀		
2947	聰	3	9BDA	8066	聰		
2970	貟	1	9BF1	8C9F	貟		
2987	鼜	1	9C43	9B57	鼜		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3787	冢	3	8152	51A2	冢		
3813	卬	2	816C	536D	卬		
3832	嘅	2	81A1	5605	嘅		Unicode 字型呈現差異。
3877	廼	11	81CE	5EFC	廼		
3883	徧	3	81D4	5FA7	徧		
3921	晢	1	81FA	6673	晢		
3922	晳	1	81FB	23293		執𠆢日	
3949	淞	3	8257	6D96	淞		
3964	戚	1	8266	70D5	戚		
3965	戠	1	8267	70D6	戠		
3974	牀	3	8270	7240	牀		
4032	秣	2	82CC	2578A		禾𠘨未	秣，異體字問題。
4034	穉	3	82CE	7A49	穉		
4061	糦	1	82E9	7CE6	糦		
4080	羣	10	82FC	7FA3	羣		
4081	翺	6	82FD	7FFA	翺		
4091	皋	7	8348	81EF	皋		Unicode 字型

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
							呈現差異。
4097	芭	7	834E			⺾巴	
4102	苷	8	8353	82E2	苷		
4107	葱	1	8358	8471	葱		
4113	蕑	2	835E			⺾間	
4114	蘗	1	835F	8616	蘗		
4124	蝨	1	8369	8771	蝨		
4132		4	8371			血巫	眾，異體字問題。
4142	褒	1	837B	8943	褒		褒，異體字問題。
4150	訟	3	83A5	8A29	訟		
4160	謫	1	83AF	8B81	謫		
4185	逿	1	83C8	9037	逿		
4186	遡	8	83C9	9061	遡		
4209	靴	3	83E0	9789	靴		
4212	韠	1	83E3	97B8	韠		
4217	韡	1	83E8	97E0	韡		
4238	髭	1	83FD	9AE2	髭		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4256	鶏	3	8450	9DC4	鶏		雞，異體字問題。
4259	麕	3	8453	9E95	麕		
4267	鼇	1	845B	9F08	鼇		
4371	僩	2	84E5	50F4	僩		
4428	蹔	1	855F	8E70	蹔		
4442	潛	1	856D	6F98	潛		Unicode 字型呈現差異。
4460	勑	1	85A1	52D1	勑		
4603	慼	1	8671	6187	慼		
4732	脣	2	8755	2664D		形ノ幺月⊙	
4733		11	8756			形丘-4⊙	
4878	穴	1	884A	5B82	穴		Unicode 字型呈現差異。
4902	宼	1	8862	5156	宼		衮，錯字，以程式全部取代。

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4965	橄	1	88C3			木  敕	
4996	斷	1	88E2	65B5	斷		
5111	旆	7	89B8	65BE	旆		
5215	毚	1	8A61			昆免	毚，異體字問題。
5302	昏	13	8ADA	662C	昏		
5319	晰	1	8AEB	6663	晰		
5344	晷	2	8B45	55F8	晷		
5505	耆	4	8C49	264BE		老占	耆，錯字，以程式全部取代。
5564	槧	1	8CA6	6ABF	槧		
5583	飧	1	8CB9	98F1	飧		
5586	歙	2	8CBC	6B57	歙		
5655	躅	4	8D42	56BB	躅		
5705	廐	2	8D74	5ED0	廐		Unicode 字型呈現差異。

二、手動取代表說明：

《毛詩》檔案中需手動修改的情況分為以下幾種：

（一）《毛詩》檔案中使用之舊版造字，校對原書後發現該處無字，以程式轉碼後手動刪除，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. 檔案原文摘錄：摘錄檔案《毛詩》中的文句段落，其中紅字底線的部分，為需要手動刪除的舊版造字。
4. 手動取代結果：變更過後的檔案內容。
5. 備註：記錄修改相同辭句的次數或其他事項。

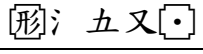
附件二、《毛詩》手動取代表

編號	造字	檔案原文摘錄	手動取代結果	備註
533	朮	狼𧰼__其尾	狼𧰼其尾	

（二）《毛詩》檔案中出現「●」字形，共計 5 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，請參考附件三。欄位說明如下：

1. 頁碼：「●」字形所在位置之原書頁碼。
2. 檔案原文摘錄：摘錄《毛詩》檔案有「●」字形的文句。
3. 原書字形：「●」在原書中之字形，以此字取代檔案原文的「●」。
4. 備註：記錄修改相同詞句的次數或其他事項。

附件三、《毛詩》「●」字手動取代表

頁碼	檔案原文摘錄	原書字形	備註
p. 46	鉤膺●革		
p. 47	●拾既伙		待造字

頁碼	檔案原文摘錄	原書字形	備註
p. 56	祇自●兮	疵	
p. 63	俾我●兮	疵	
p. 71	可以●饎	餚	

（三）《毛詩》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《毛詩》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、《毛詩》缺字外錯字修改

檔案原文摘錄	手動取代結果	備註
庭燎晰哲	庭燎晰晰	