《國語》(標記)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
2008/3/28　　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 964 的「隣」字，Unicode 編碼是 28EF3，由於 XP 並不支援，仍需使用構字式「阝𠂤馬」。

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：

《國語》(標記)使用舊版造字 55 個，字頻 136 次，這 55 個造字中，50 個可轉成 Windows XP 能顯示的字，字頻 128 次；另外 5 個字必須轉成構字式，字頻 8 次。

轉碼完成製作轉碼分析表，請參考附件一《國語》(標記)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，如：人名、地名等或特殊情形。例：造字編號 981 的「豺」字，是「豺」的異體字，以標準字「豺」取代。
2. 部份使用不同字形，手動修改：檔案中為同一舊版造字，校對時卻發現原書使用兩個以上不同的字形，可能部份舊版造字為錯字或原書使用兩種異體字形，導致同一舊版造字轉碼時對應到兩個以上的字形；將一部份手動修改，剩餘的頻次則用程式取代成對應的 Unicode 或構字式。例：造字編號 5591 的「杞」字頻次為 3，依書將國名的 2 字手動修改為「杞」，剩餘 1 字轉為「木⿰己」。手動修改部份請參考附件二《國語》手動取代表。
3. 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 5505 的「耆」字，而原書使用之字為「耆」，查詢教育部異體字字典確認「耆」並非「耆」之異體字，因此歸為錯字，以程式取代為正確字形「耆」。

附件一、《國語》(標記)轉碼分析表

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
827	鉉	8	8E69	9B8C	鉉		
938	燮	2	8EFA	7215	燮		
964	隅	1	8F55	28EF3		阝⿰禹	

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
980	犨	4	8F65	72AB	犨		
981	豺	3	8F66	72B2	豺		豺，異體字問題。
988	羋	1	8F6D	7F8B	羋		
1040	聑	1	8FC3	803C	聑		
1044	儤	1	8FC7	511B	儤		
1158	晲	4	907A	7561	晲		
1297	晲	1	9168	6650	晲		
1565	鯀	1	92F9	7DDC	鯀		
1642	脉	3	93A9	8109	脉		脈，異體字問題。
1857	釜	1	94E3	86A0	釜		
2043	踁	1	9641	8E01	踁		
2146	蜗	1	96CA	8739	蜗		
2524	饎	1	994B	994D	饎		
2546	鉻	1	9961	8849	鉻		
2765	塼	1	9AC1	7AF1	塼		
2828	蜩	1	9B41	8744	蜩		
3770	个	2	8141	4E2A	个		
3787	冢	2	8152	51A2	冢		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
3849	獎	1	81B2	596C	獎		
3883	徧	12	81D4	5FA7	徧		
3889	慤	1	81DA	6164	慤		
3891	慙	1	81DC	6159	慙		
3911	攩	1	81F0	651F	攩		
3921	晢	3	81FA	6673	晢		
3924	朞	1	81FD	671E	朞		期，異體字問題。
3974	牀	1	8270	7240	牀		
3976	犁	2	8272	7282	犁		犁，異體字問題。
3987	毅	1	827D	7474	毅		
4035	穽	1	82CF	7A7D	穽		
4038	竈	3	82D2	7AC8	竈		
4045	篬	1	82D9	7B92	篬		
4063	糵	1	82EB	7CF5	糵		蘗，錯字，以程式全部取代。
4079	徛	1	82FB	38E5		亻⿰寺	
4080	羣	29	82FC	7FA3	羣		

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
4160	謫	2	83AF	8B81	謫		
4184	迹	1	83C7	8FF9	迹		
4203	隣	1	83DA	96A3	隣		鄰，異體字問題。
4217	韠	1	83E8	97E0	韠		
4242	鬬	5	8442	9B2D	鬬		
4267	鼇	7	845B	9F08	鼇		
4469	匀	2	85AA	5304	匀		
4913	墢	1	886D	58A2	墢		
5089		1	89A2			冗⿰畢	罼，異體字問題。
5249	蛹	1	8AA5	873D	蛹		
5335	虫	1	8AFB	459F		亡⿰虫	
5437	幷	1	8BC4	5E77	幷		
5456		2	8BD7			韋⿰未	韖，異體字問題。
5505	耆	2	8C49	264BE		老⿰占	考⿰口，錯字，以程式全部取代。
5564	槑	1	8CA6	6ABF	槑		
5583	飧	2	8CB9	98F1	飧		飱，異體字問

編號	造字	頻次	Big5	Unicode	WinXP	構字式	備註
							題。
5591	杞	3	8CC1	233CC		木  巳	部份使用不同字形「杞」，手動修改。
5703	亂	1	8D72	9F53	亂		

二、手動取代表說明：

《國語》(標記)檔案中需手動修改的情況分為以下幾種：

（一）《國語》(標記)檔案中之舊版造字，因字形部份不同而無法全部以程式取代，皆手動修改成Windows XP能顯示的字或構字式，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. 取代：實際使用字形與舊版造字不同時，於此欄列出。
4. 標記行碼：語言所標記檔裡設定的行碼。
5. 檔案原文摘錄：摘錄檔案《國語》(標記)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。
6. 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。
7. 備註：記錄修改相同辭句的次數或其他事項。

附件二、《國語》(標記)手動取代表

編號	造字	取代	標記行碼	檔案原文摘錄	手動取代結果	備註

編號	造字	取代	標記行碼	檔案原文摘錄	手動取代結果	備註
5591	杞	杞	837	唐(NB3)[+prop]	杞(NB3)[+prop]	
			1935	唐(NB3)[+prop]	杞(NB3)[+prop]	

（二）《國語》(標記)檔案中出現「●」字形，共計 22 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. 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《國語》(標記)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。

附件三、《國語》(標記)「●」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. 74	1281	不(DC)　亦(DL)　●(VH1N)　姓(NI)	嬻	
p. 109	1980	佐(VF)　●(NA1)　者(NH)	食 从雍	
p. 178	3303	禁(VE)　罝(NA2)　罝(NA2)　●(NA2)	罒 公鹿	
p. 178	3310	獸(NA1)　長(VP)　麑●	鹿 夭	
p. 205	3888	懼(VH2S)　●(VC1)　季孫	忏	
p. 209	3974	●(VP)　門(NA2)	門 為	

頁碼	標記行碼	檔案原文摘錄	原書字形	備註
		與(P)　之(NH)　言		
p.223	4245	遊車(NA2)[+attr] 之(T)　●(NI)	前𠔃衣	
p.228	4356	擊(VC1)　●(NA2) 除(VAN)　田	某	
p.228	4366	身(NA1)　衣 (VC1)[+nv]　襏● (NA2)[+co]	褫	
p.239	4708	贖(VJO)　以(P) ●盾(NA2)[+attr]	韇	
p.241	4748	地(NA2)　南(NG) 至(VA)　於(P)　● 陰	𠂤𠙵匋	
p.472	10340	今(NA5)　忨(VH1N) 日(NA5)　而(C) ●	氵𠙵歇	
p.507	11153	依(NB2)[+prop]　、 ●(NB2)[+prop]	黑𠙵柔	
p.514	11221	羋(NB3)[+prop] 姓(NI)　●越 (NB3)[+prop]	首𠔃𠂉𠔃歧	
p.516	11281	行(VP)　●極 (NI)[+co]	姟	
p.538	11668	若(C)　合(VH1) 而(C)　●(VAN)	刍	
p.549	11891	亡𠔃虫●	維𠔃虫	

頁碼	標記行碼	檔案原文摘錄	原書字形	備註
		(NA1)[+co]　之(T)		
p.567	12268	會(VC10)　于(P)　龍●(NA1)[+attr]	豕⿰龸尨	
p.597	12953	於(P)　其(NH)　心(NI)　也(T)　●然(VI)[+poly]	伐	
p.602	13084	盛(VC10)　以(P)　鴟●	鷅	
p.605	13122	將(DD)　夾(VC1)　溝(NA2)　而(C)　●(VC1)	广⿰夂侈	
p.612	13255	自(DH)　●(VA)　於(P)　客(NA1)　前	亞⿰龸刂	

（三）《國語》(標記)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「?」字形手動取代表，請參考附件四。欄位說明如下：

1. 標記行碼：語言所標記檔裡設定的行碼。
2. 檔案原文：摘錄《國語》(標記)檔案有「?」字形的文句。
3. 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。

附件四、《國語》(標記)「?」字形手動取代表

標記行碼	檔案原文	字形	備註
1053	師(NA3)　必(VM)　有	謫	

標記行碼	檔案原文	字形	備註
	(VG) ?(NI)		
1064	秦(NB3)[+prop] 師(NA3) 無(VG) ?	謫	
6219	死(VH1) 又(DL) 不(DC) 敢(VM) ? (VAN)	泣	
11115	沈(VP) 竈(NA2) 產(VC1) ?(NA1)	黿	
14430	而(C) ?黿(NA1)[+co] 之(T) 與	黿	

（四）《國語》(標記)檔案中出現「☆」字形，共計 9 字，比對原書後以 Unicode 字形或構字式手動取代，並製作「☆」字形手動取代表，請參考附件五。欄位說明如下：

1. 標記行碼：語言所標記檔裡設定的行碼。
2. 檔案原文：摘錄《國語》(標記)檔案有「☆」字形的文句。
3. 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。

附件五、《國語》(標記)「☆」字形手動取代表

標記行碼	檔案原文摘錄	原書字形	備註
1840	洛(NB2)[+prop] ☆(VA)	鬬	
1981	佐(VF) ☆(VA) 者(NH) 傷	鬬	
1999	王(NA1) 將(DD) 防(VC1) ☆川	鬬	
2000	飾(VC1) 亂(NI) 而(C) 佐(VF) ☆(NI) 也(T)	鬬	

標記行碼	檔案原文摘錄	原書字形	備註
4901	遂(DL)　☆(VA)　而(C)　死	鬬	
6762	☆士(NA1)[+attr]　眾(VH1)	鬬	
6774	☆士(NA1)[+attr]　是故	鬬	
8238	戰☆(NI)[+co]	鬬	
8439	則(C)　濟(NI)　可(VM)　☆(VKW)	鬬	

（五）《國語》(標記)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，請參考附件六。欄位說明如下：

1. 標記行碼：語言所標記檔裡設定的行碼。
2. 檔案原文摘錄：摘錄檔案《國語》(標記)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。
3. 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。
4. 備註：記錄修改相同詞句的次數或其他事項。

附件六、《國語》(標記)缺字外錯字修改

標記行碼	檔案原文摘錄	手動取代結果	備註
13362	戮(VC1)　力(NI)　同(VP)　德	勠(VC1)　力(NI)　同(VP)　德	
13738	能(VM)　下(VAN)[+nv]　其(NH)　臣(NA1)	能(VM)　下(VAN)[+nv]　其(NH)　羣臣(NA1)	
712	太宰(NA1)[+attr]　(VAN)　之	太宰(NA1)[+attr]　(VAN)　莅之	

標記行碼	檔案原文摘錄	手動取代結果	備註
1262	官正(NA1)[+attr] (VAN) _事(NI)	官正(NA1)[+attr] (VAN) 莅事(NI)	
5195	使(VF) 奚齊 (NB1)[+prop] _(VAN) 事	使(VF) 奚齊 (NB1)[+prop] 莅 (VAN) 事	
4980	有-{(NB3)[+prop]	有褒(NB3)[+prop]	
4981	-(NB3)[+prop] 人 (NA1) 以(P) -姒 (NB1)[+prop]	褒(NB3)[+prop] 人 (NA1) 以(P) 褒姒 (NB1)[+prop]	
4982	-姒(NB1)[+prop] 有(VG) 寵(NI)	褒姒(NB1)[+prop] 有(VG) 寵(NI)	
10698	閻明 (NB1)[+prop] 、 叔-(NB1)[+prop]	閻明 (NB1)[+prop] 、 叔褒(NB1)[+prop]	
11325	收(VC20) 以(P) 奔(VAN) -(NB4)[+prop]	收(VC20) 以(P) 奔(VAN) 褒 (NB4)[+prop]	
11330	-(NB3)[+prop] 人 (NA1) 之(T) 神 (NI)	褒(NB3)[+prop] 人 (NA1) 之(T) 神 (NI)	
11334	- (NB3)[+prop] 之 (T) 二(S) 君	褒(NB3)[+prop] 之 (T) 二(S) 君	
11360	逃(VA) 于(P) -(NB4)[+prop]	逃(VA) 于(P) 褒 (NB4)[+prop]	
11361	-(NB3)[+prop] 人 (NA1) -姁(NB1)	褒(NB3)[+prop] 人 (NA1) 褒姁(NB1)	

標記行碼	檔案原文摘錄	手動取代結果	備註
10167	候遮(NA2)[+co] 扞%(NA2)	候遮(NA2)[+co] 扞衛(NA2)	