《新語》舊版造字轉碼說明

中研院資訊所文獻處理實驗室
中研院史語所漢籍電子文獻工作小組
2008/2/13 陳建安 製作

一、 《新語》一書(新語.xml)使用舊版造字 30 個，字頻 80 次，詳如附件一。這 30 個造字中，26 個可轉成 Windows XP 能顯示的字，字頻 44 次；另外 4 個字必須轉成構字式，字頻 36 次。

二、 附件一的造字分析表說明如下：

甲、 編號：Big5 造字空間為 6217 個，編號由 1 到 6217。
乙、 造字：舊版造字
丙、 字頻(txt)：造字在「.txt」文件的出現次數
丁、 字頻(xml)：造字在「.xml」文件的出現次數
戊、 Big5：造字的 Big5 碼
己、 Unicode：造字所對應的 Unicode 碼
庚、 WinXP：造字在 Windows XP 的對應字形
辛、 備註凡例：

1、校對問題，舊版漢籍錯字，可用程式全部取代：在舊版漢籍電子文獻中即存在的錯字，因校對時的疏漏而未更正，持續留存在新版漢籍電子文獻中；若該造字的所有頻次，皆屬於錯誤使用的錯字情形，可以用程式全部取代為正確字形。如編號 4910 的「氵亼䍃」字，原字為「滛」。

2、異體字問題：新版漢籍考量到使用者檢索及使用時的便利性，將用字原則改為除專詞等特殊情形之外，一律改用標準字呈現。如編號 4132 的「血亼乑」係「眾」字之異體，故以「眾」字取代。又編號 4507 的「吴」，係「吳」之簡體字，亦以「吳」字取代。

3、Unicode 字型呈現差異：Unicode 字型與舊漢籍造字有些微差異，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 3935 的「槩」字，Unicode 字型呈現為「槩」，實際上仍為同一字。

三、 Unicode 目前收錄的漢字總數為 70194，分屬於三個不同區段，詳如表一。目前 Windows XP 只支援 CJK 認同表意文字區的 20902 個字，內碼為 4E00-9FFF。所以造字編號 3894 的「䈥」字，Unicode 編碼為 4225，由於 Windows XP 並不支援，仍須使用構字式「⺮

亼勅」。

表一、Unicode 的字數及編碼區段

Unicode	字集子集合	新增字數	新增編碼區段	總字數	WinXP
1.1 版	CJK 認同表意文字區	20902	4E00-9FFF	20902	支援
3.0 版	CJK 認同表意文字擴充 A 區	6582	3400-4DFF	27484	不支援
3.1 版	CJK 認同表意文字擴充 B 區	42710	20000-2A6D6	70194	不支援

附件一、《新語》造字分析表

編號	造字	頻次 (txt)	頻次 (xml)	Big5	Unicode	WinXP	構字式	備註
899	駞	1	1	8ED3	99DE	駞		
1044	儛	1	1	8FC7	511B	儛		
1226	罸	4	4	90E0	7F78	罸		
1565	緜	2	2	92F9	7DDC	緜		
1906	裦	1	1	9555	88E6	裦		
1981	讐	2	2	95C2	8B90	讐		
3791	况	1	1	8156	51B5	况		
3804	効	1	1	8163	52B9	効		
3810	匵	1	1	8169	5335	匵		
3894	䈥	1	1	81DF	4225		⺮亼勅	
3935	槩	1	1	8249	69E9	槩		Unicode 呈現差異。
3937	檝	1	1	824B	6A9D	檝		
3976	犂	1	1	8272	7282	犂		

編號	造字	頻次(txt)	頻次(xml)	Big5	Unicode	WinXP	構字式	備註
3985	璹	1	1	827B	7447	璹		Unicode呈現差異。
3996	疎	2	2	82A8	758E	疎		
4040	竝	1	1	82D4	7ADD	竝		
4077	罇	1	1	82F9	7F47	罇		
4080	羣	9	9	82FC	7FA3	羣		Unicode呈現差異。
4132		33	33	8371			血⿱乑	眾，異體字問題。
4146	覩	4	4	83A1	89A9	覩		
4178	輭	1	1	83C1	8F2D	輭		
4437	廩	1	1	8568	5EEA	廩		
4507	吳	1	1	85D0	5434	吳		吳，異體字問題。
4904	尅	1	1	8864	5C05	尅		
4910		1	1	886A			氵⿰䍃	滛，校對問題，舊版漢籍錯字，可用程式全部取代。
5031	閴	1	1	8946	95B4	閴		
5034	弃	2	2	8949	5F03	弃		Unicode呈現差異。
5434	湼	1	1	8BC1	5E7C		[形]氵 白工[•]	
5705	廐	1	1	8D74	5ED0	廐		Unicode呈現差異。
5706	敍	1	1	8D75	654D	敍		