认识 UniCode

□湖北 金菁

曾见过一篇文章，谈到中文简体的 Word97 可以打开 BIG5 码的 Word97 文件，而且没有乱码，作者断言：简体 Word97 是台湾人做的本地化。

本人做软件本地化时，接触了许多繁体中文版的软件，这些软件在简体中文 Windows98 上均可以正常工作(只是界面显示为繁体，乱码很少，甚至没有)。这时才明白那位作者错了——没有 BIG5 码的 Word97 文件，也没有 GB 码的 Word97 文件，Word97 文件内码为 UniCode。因为 Windows 95/98 支持 GBK 大字库，且大字库中包含了所有的繁体字，GBK 编码又与 UniCode 中相应的编码一一对应，所以在简体 Word97 中也可以正常显示港台的 Word97 文档。至于在字体框中显示所谓“华康体”，是因为字体名称也用 UniCode 保存在文档之中。

鉴于许多人对 UniCode 不太了解，本文先用 Word 97 做一个小小的测试，然后详细介绍 UniCode 及其与 GB、GBK、BIG5 之间的转换。

一、在 Word97 中测试 UniCode

1. 打开 Word97，在新文档中输入“革命”，设置为“宋体”，存为 1.doc；再另存为 Unicode 文本 1.txt。

2. 用十六进制编辑器 Hex Workshop 打开 1.doc 和 1.txt。

3. “革命”的 Unicode 编码为“9769547D”，二进制文本中显示为“69977D54”，在 1.txt 和 1.doc 中均可找到。

4. 在 1.doc 中查找“8B5B534F”(宋体)，将“8B5B534F”改为“534F8B5B”(体宋)，存盘退出。

5. 在简体 Word97 中打开 1.doc，可以见到字体框中显示“体宋”，但是系统中并不存在这样的字体。对于不存在的字体，Word97 用系统字体自动替换。

二、什么是 UniCode

根据《通用多八位编码字符集 (UCS)》及国际标准 ISO/IEC10646.1－1993，UniCode 用于世界上各种语言的书面形式以及附加符号的表示、传输、交换、处理、存储、输入及显现。因此，用 UniCode 编码制作的软件或文档，在任何支持 UniCode 编码的操作系统中使用均不会产生乱码。可以预见，如果所有的主页均采用 UniCode 作为 HTML 字符集，那么 Internet 就成为真正意义上的地球村了。

1. UCS 的总体结构

UCS 编码字符集的总体结构是一个四维编码空间，它包含 00～7F 共 128 个三维组，每个三维组中包含 00～FF 共 256 个二维平面，每个二维平面包含 00～FF 共 256 个一维行，每行共 256 个字位 (00～FF)，每个字位用一个字节二进制数表示。因此在 UCS 中每一个字符用 4 个二进制数编码，以确定每个字符在编码空间的组、平面、行和字位。上述四个 8 位二进制数编码形式称为 UCS 的四八位正则形式，记作 UCS－4。

2. 基本多文种平面

在 UCS 编码空间中 00 组的 00 平面称为基本多文种平面。在此平面包含了字母文字、音节文字和表意文字中通常使用的字符以及各种符号和数字。

基本多文种平面的组编码为 00H。UCS 规定当正则形式的组、平面编码为 00H 时可以省略，因此安排在基本多文种平面上的字符可用两个字节的二进制数来表示，形成双八位编码字符集，记作 UCS－2。

我们通常谈论的 UniCode 即指 UCS－2，本文中如果没有特殊声明，UniCode 均指 UCS－2。

基本多文种平面分成 A、I、O、R 四个区。

A 区：代码位置从 0000～4DFF，共 19903 个字位。此区用于字母文字、音节文字以及各种符号的编码，其中 0000～001F 和 007E～009F 保留用于控制字符。

I 区：代码位置从 4E00～9FFF，共 20992 个字位。此区用于中、日、韩(CJK)统一的表意文字，即中国、日本、韩国等三国汉字的编码。通常谈论的 GBK、BIG5、UniCode 编码之间的转换都是针对此区。

O 区：代码位置从 A000～DFFF，共 16384 个字位。此区目前未用，留作未来的标准化。

R 区：代码位置从 E000～FFFD，共 8190 个字位。此区是限制使用区，用于专用字符、变形显现形式和兼容字符的编码。

3. 双字节字符集(DBCS)

DBCS 是有别于 Unicode 的另一类字符编码，它支持很多不同的东亚语言字符，如汉语、日语和朝鲜语。DBCS 使用数字 0－128 表示 ASCII 字符集。其它大于 128 的数字作为前导字节字符，它并不是真正的字符，只是简单的表明下一个字符属于非拉丁字符集。在 DBCS 中，ASCII 字符的长度是一个字节，而汉语、日语和朝鲜语字符的长度是 2 个字节。GBK、GB、BIG5 字符集是 DBCS 编码，尽管一个中文字符要用两个字节进行 DBCS 编码，但使用半角英文字符时仍然保持单字节 ANSI 编码。在 UCS－2 中，即使半角英文字符也用两个字节编码，如“A”的编码为“0041”，经过调换高低位，二进制文件中显示为“4100”。所以，在用 UniCode 编写的软件中，“OK”与“确定”都占用 4 个字节，大可不必害怕汉化后汉字字串长于原英文字串而将“OK”翻译为“好”。

三、GB、GBK、BIG5、UniCode 内码互换

在中文(简体)本地化一个软件时，往往希望将繁体中文版本直接本地化成简体中文版。首先取出可执行文件的资源部分，比如用 VisualC＋＋提取资源时，系统自动将 UniCode 编码转换为 GBK 编码存为 RC 文件，将 RC 文件从 GBK 编码转换为 GB 编码，回存至可执行文件中（此时系统又将 GB 编码转换为 UniCode 编码），便完成了 UniCode 编码部分的中文（简体)本地化过程。

1. UniCode ⇔ GBK 和 UniCode ⇔ BIG5

Windows NT 4.0 自带了一个“中文转码器”，可以在 GBK、BIG5、UniCode 编码间互相转换。

2. GBK ⇔ BIG5 和 GBK ⇒ GB

UCWIN 附带的 “Text Converter” 可以在 GBK、GB、BIG5、CNS、TCA、ETEN、IBM 5550、Shift JIS、JIS、KSC 编码间互相转换，也可以将以上编码转换为 HZ、ISO2022－GB、ISO2022－CNS、ISO2022－JIS、ISO2022－KSC 编码，而且可以选择自动添加空格或智能处理简繁汉字“一对多”的问题。

3. 制作码表

以制作 UniCode ⇒ GBK 的码表为例。

(1) 创建一个二进制文件 unicode.tab，内容为 4E00～9FFF 的所有 I 区编码；

(2) 用“中文转码器”将 unicode.tab 内码转换为 GBK；

(3) 新 unicode.tab 文件中 (xxH－4eH) ＊100H＋(yy－9fH)偏移处的字即为 UniCode 编码 xxyy 对应的 GBK 编码。

注：本文所指操作系统为 Win98 简体中文版和 WinNT 4.0 简体中文版，Word97 也是简体中文版。