文章编号: 1003-0077(2008)04-0105-04

在通用字符集中藏文编码模式的研究与应用

欧 珠

(西藏大学 工学院,西藏 拉萨 850000)

摘 要: 藏文软件开发者在现代计算机系统中处理藏文数据时必须所具备的知识之一是藏文在通用字符集(Universal Character Set, UCS)中是如何进行编码。在设计藏文网页内容时 UCS 藏文数据的整理、设计藏文应用软件时藏文文本的处理操作或者在设计藏文 OpenType 或 AAT 字库时、UCS 藏文编码模式应用等都要首先去理解 UCS 藏文编码模式。因此,理解和掌握 UCS 藏文编码模式是软件制作商首选目标。详细介绍了 UCS 藏文编码模式的组织结构和设计方法,以便于使用 OpenType 来支持复杂藏文文本的显示。

关键字: 计算机应用;中文信息处理;UCS;藏文编码;组合;排序;重排

中图分类号: TP391 **文献标识码**: A

A Study on Tibetan Script Encoding in the UCS

Ngodrup

(Engineering School of Tibet University, Lhasa, Tibet 85000, China)

Abstract: It's necessary for the developers of Tibetan software to know how do encode Tibetan in UCS when processing Tibetan data. Understanding the UCS Tibetan encoding system must come before reorganizing UCS Tibetan data when designing Tibetan websites, processing Tibetan text, developing Tibetan application software, or designing OpenType or AAT fonts. To facilitating the understanding of the UCS Tibetan encoding system, this article explains in detail the organizational structure and design methods of UCS Tibetan encoding system, so that the OpenType can be applied to display complex Tibetan documents.

Key words: computer application; Chinese information processing; UCS; Tibetan encoding; combining; order; reordering

1 引言

为推动藏语文规范化、标准化和信息处理现代化,弘扬藏族优秀文化,使藏语文适应现代信息技术的发展需要,在国家有关部门的大力支持和帮助下,西藏从 1993 年开始开展了藏文信息技术标准化工作,着手起草和制定藏文编码国际标准和国内标准的方案。藏文编码国际标准最终方案于 1997 年获得国际标准组织顺利通过,使藏文在中国少数民族文字中成为第一个具有国际标准的文字。这为藏语文步入现代信息媒体领域,在网络媒体中实现信息处理和交换,建立了良好的基础。但由于藏语言本身的特点,其处理方法不同于拉丁文字、汉字的处理方法更复杂,这给开发藏文版本的软件带来了很大的困难。当前国内所使用的藏文软件,几乎都支持国际编码标准——ISO/IEC 10646 中藏文编码字符集国际标准(基本集),因此结合藏文本身的基本结构,正确理解其编码结构是一项非常重要的基础概念。

2 藏文具备了作为一种复杂文本语言的基本特征

藏文可以被视为基本字符和基本字符通过纵向

收稿日期: 2007-06-29 **定稿日期**: 2007-11-20

基金项目: 信息产业部电子信息产业发展基金资助项目(信部运[2005]425 号)

作者简介: 欧珠 (1964—),男,硕士,教授,研究方向为计算机软件与理论、藏文信息处理。

叠加而成的字符串,构成一个完整藏文词素的基本单位是由藏文中的"音节分割符 tsheg bar"来确定。一个藏文词由一个或多个音节构成。每一个音节包含着"基字(Root letter)(Ming gzhi)"和可能跟随的如前加字(Prefix)、上加字(Head letter)、元音符号(Vowel)、后加字(Suffix)、再后加字(Post suffix)。音节,通常是由音节分割符 tsheg bar 或者其他标点符号来划分的。图 1 给出了一个藏文字的各组成构建。

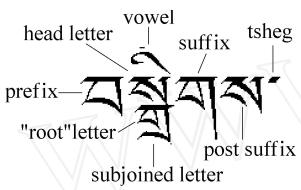

图 1 藏文字的各组成构建

在 ISO 10646/ Unicode 标准编码中,像拉丁语、汉语,它们文字的显现形式与 ISO 10646/ Unicode 中编码字符是一一对应的,即字符(Character)与它的显现字型(Glyph)是一一对应的,而且显示的顺序和在内存中存放的顺序是一样的,但藏文却有着比这更复杂的特性,即一个藏文字符则需要用几个编码来表示,长度不定,给藏文在信息系统的实现带来极大的麻烦。具体而言,藏文有如下一些特性:

(1) 字符置位性:虽然藏文书写方向是从左向右,但是在一个藏文文本中纵向叠加的辅音加上元音的组合字是经常存在和使用的。然而,无论是两个相邻的字符将要纵向地堆叠或者从左到右形式地拼写,后一个字符不能简单地由上下文或语法规则来确定。不管是什么文字,在计算机中,Unicode 字符串都是以逻辑顺序存储的,即它们的存储方向都是从左到右。在处理多语言文字的 Unicode 字符串时,系统就必须识别出各种文字的方向。

(2) 与上下文内容相关的显现形式:字符在词的不同位置有着不同的显现形式,如藏文字 0F62 在作为藏文的基字和上加字时有不同的显现形式。这里引出了两个概念:名义字符及其变形显现形式,名义字符指在 Unicode 中已编码的字符;变形显现形式指在语义上还是那个字符,但它却有着和那个字符完全不同的显现形式,它们在 Unicode 中没有码位、没有编码。

3 ISO/ IEC 10646 和 Unicode 国际标准中藏文的编码模式及应用

藏文由于叠加字符的各构件变形和跨度都较大,特别是叠加层数较多的字符,各字母在不同层的高度和宽度都有不同的要求,因此,藏文字符的纵向叠加是藏文信息处理的一大难点。正因为如此,在 UCS 藏文编码中所使用的编码模式是一个基于藏文正字法或字布局而不是基于藏文语法规则的确切的叠加模式。

3.1 藏文辅音和组合用辅音字符

藏文编码中被采纳的编码模式是一个基于藏文正字法或字布局而不是基于藏文语法规则的确切的叠加模式。在 UCS 中安排了两个完整的辅音集合:一个是从 U0F40 到 U0F69 的主辅音字符,被用于单一的辅音或者是任何的组合叠加中出现在最上层位置的辅音字符,即藏文的最基本辅音字符和六个来自印度文的预组合好的辅音字符。另一个是从 U0F90 到 U0FBC 的组合用辅音字符,用于在叠加时出现的附加辅音。

3.2 藏文及梵音转写藏文元音符号

对于藏文元音,经常是作为标记与组合辅音或者辅音的叠加相结合,在 UCS 中其编码安排在从 U0F71 到 U0FB1 之间。其中,藏文中常用的四个元音编码分别被定义为 U+0F72、U+F74、U+0F7A,和 U+0F7C。为了保证处理和兼容梵音转写藏文用元音之需要,还提供了编码点从 U+0F71 到 U+0F7D 之间的元音符号。大多数藏文用户不会把这些复合元音作为一个单独的元音字符,并且它们在梵文中的应用也是被限制的。同样,用户也可以把这些组合元音认为由其他相似的元音经过一系列的组合而成。元音的拼写顺序是与藏文书写时的顺序一致,即总把元音字符放在它所应用到的主辅音或组合辅音的后面。

3.3 藏文断行符

在藏语言中有两个可断行的藏文字符。其主要的字符为标准的 TSHEG,在一些语言中我们把它翻译为"音节符",其 UCS 编码被定义为 U+0F0B。第二个断行字符为"空格"字符。虽然在 UCS 中这

两个被定义为可断行的字符,但在具体应用中根据藏文构字规律和语言规则来正确地定义和使用断行规则。

对于主断行符 TSHEG,除非出现在藏文辅音字符 NGA (U + 0F44)之后和藏文符号 SHAY (U + 0F0D)之前时,根据藏文语法规则不允许断行外,其他任何情况下允许作为换行符。当它出现在藏文辅音字符 NGA (U + 0F44)之后和藏文符号 SHAY (U + 0F0D)之前时,通常由不断行的音节符 Tshe(U + 0F0C)来代替 TSHEG(U + 0F0B)。

3.4　上加字为完整形式的"RA"(具有固定格式的 RA)

作为"RA"的一个例子当其出现在上加字时,这个辅音以不变形的情况下完整地出现在上加字的位置,我们需要用到编码为 U + 0F6A 的"RA"。很显然,这个不是藏文书写规则中的一个标准,但有时在梵音转写时出现这种形式。为此,当且仅当遇到这种情况时,其编码由 U + 0F6A 来代替,即固定格式的藏文字符 RA。例如,组合字符"RA + YA"出现在梵音转写藏文字符时,其上加字 RA 的编码应该为 U + 0F6A, YA 的编码即可以是组合用辅音字符 YA(U + 0FBB),也可以是变形下加辅音字符 YA (U + 0FB1)。

字库开发商需要注意的是,"RA + 下加辅音字符 NYA"时,没有要求使用固定格式的 RA(U + 0F6A)。

3.5　下加位置上的辅音字符"WA","YA"和"RA"

根据藏文的语法规则,"WA","YA", 和"RA"三个辅音字符出现在主辅音字符的下方时,它们的形状会发生变化。当辅音字符"WA"出现在下加字的位置时通常不是完整形式的"WA",而是被剪掉在该主辅音字符右下角的位置,通常被称为 wazur (wa zur),并且其编码为独立的 U + 0FAD。同样,辅音字符"YA"和"RA"出现在该主辅音字符右下角的位置分别称为 ya-ta (ya_btags) 和 ra-ta (ra_btags),其编码分别是 U + 0FB1 和 U + 0FB2。

在梵音转写藏文中也经常需要不变形的组合用辅音字符"WA","YA"和"RA"。为了满足这个要求,在 UCS 中也安排了不变形的完整形式的对应编码,分别是 U + 0FBA 、U + 0FBB 和 U + 0FBC 。"RA + YA"的组合是一个很好的例子,因为它既有变形的 YA (ya_btags) 形也有不变形的完整式"YA"。

3.6　下加字符-A (a-chung) (U + 0FB0):

下加字符-A (a-chung)即完整大小格式的辅音字符很少被用在主辅音字母的下加位置上,但为了构成一个完整的编码体系之需要,在 UCS 中也安排了完整格式的下加辅音字符(a-chung)。通常在藏文中辅音字符 U + 0F60 出现在其主辅音字符的下加位置上时,经常是使用短格式的下加辅音 AA(U + 0F71)。

4　字符的排序

虽然藏文书写方向是从左向右,但是在一个藏文文本中自顶而下顺序的纵向叠加的辅音标上自底而上顺序的元音的组合字是经常存在和使用的。正是由于在藏文文本中这种纵向叠加字符所出现的频率和复杂性,在 UCS 藏文编码中所采纳的编码模式不同于象天城体(Devanagari)或者印度文中所采纳的编码模式。所以,藏文中对组合字符的叠加有一定的规则,但藏字定型引擎允许任何辅音字符与任意的组合用字符叠加。根据 UCS 藏文编码模式和藏文书写顺序,在藏文字定型引擎处理的顺序是:

(1) 头层辅音字符(主辅音字符): (U + 0F40-U + 0F6A, U + 0F88, U + 0F89) tsa 'phru (U + 0F39) [if any]

(2) 组合用辅音字符 [如果有]: U + 0F90-U + 0FBC [if any]

(3) 下加字符 [如果有]: a-chung U + 0F71 [if any]

(4) 下加元音符 [如果有]: zhabs-skyu U + 0F74 [if any]

(5) 在梵音转写藏文中偶尔出现的 halant [如果有]: U + 0F84 [if any]

(6) 上加元音符 [如果有]: U + 0f72, U + 0F7A-U + 0F7E [if any]

(7) 上加符号: U + 0F82 和 U + 0F83

(8) 其他上加符号: U + 0F86 或 U + 0F87

说明: (1) 现代藏文(藏文)中经常出现的三个上加字 ra mgo, la mgo 和 sa mgo ,根据 UCS 藏文编码模式与之相对应的主辅音字符相同,即为 U + 0F62, U + 0F63 和 U + 0F66。(2) 下加元音符优先处理于上加元音符,例如,藏文字符ཀྲི的组成顺序如图 2 所示。

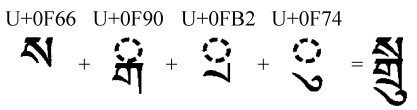

图 2 藏文字符的组成顺序

在藏文文献中出现的层数较高的叠加方式,也可以按这种顺序录入。如梵音转写藏文字符的叠加方式。如图 3。

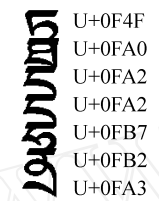

图 3 梵音转写藏文字符的叠加方式

5 字符的重排

在一个串中规正化(Normalized)处理也许影响藏文字符的顺序。在 Unicode 标准中为了实现不同的目标,很多字符标都标上一系列特征值,特别是所有组合用字符都被标上正则组合类值(CCCV: Canonical Combining Class Value)。当一个字符串被规正化时,组合字符根据它们的正则组合类值而重新排序: 大值字符应重排在小值字符之后。CCCV 值为 0 的字符不能被重排。

尽管有一些藏文字符的 CCCV 被赋上了不正确的值,但 Unicode 技术委员会(UTC)为了保持标准的稳定性而制定的严格政策,那些不正确的值将不能被改变。如:

(1) 字符 TSA 'PHRU (U + 0F39),从藏文的正常写入顺序而言,应该是立即组合或输入一个组合用辅音字符之后。但由于字符的 CCCV 值为 216,大于 CCCV 赋给一个元音的值,因此,根据其原则,进行组合后重排在元音之后。因为 TSA 'PHRU 应该组合在任意一个元音之前,而且元音的定位依赖于不管 TSA 'PHRU 是否已经加入到基本辅音字符。所以,TSA 'PHRU 的 CCCV 值应该小于任意一个组合的元音。

(2) 至少有一个元音以上的堆栈里的下加基本元音(U + 0F74) 应该在任意一个上元音之前组合或输入完成,它的 CCCV 值应小于任何一个上元音字符的 CCCV 值。而在标准中,的 CCCV 值为 132,上元音的 CCCV 值为 130。同样,下加元音的 CCCV 值为 129,也不符合藏文的书写规则。

因此,在开发基于 UCS 中的藏文编码模式的藏文软件,尤其是利用藏文字体的 OpenType 特征在分析音节、重排字符以及在字库中使用 GPOS (Glyph Glyph Positioning Table)、GSUB (Glyph Substitution Table) 特征和查找链表(Lookups Table)对显现字符进行定位和替换时,必须考虑以上的编码模式方式、不同的组合方式、字符的 Unicode 特征数据以及藏文字符的排序和重排特征。

6 总结

藏文的 UCS 编码模式是把藏文完全当作拼音文字处理,更符合藏文属于拼音文字的本质。因此,本文就应用这一编码模式在藏文计算机网络产品中其藏文数据的通讯和存储、藏文应用软件中藏文文本的处理操作以及 OpenType 藏文字库中藏文字的正确组合和排版等设计时,结合藏文自身的一些特性,分析了 UCS 藏文编码模式以及它的应用,对缺乏藏文知识的专家开发藏文软件时遇到的困难提供指导。

参考文献:

[1] 国家质量技术监督局. GB16959-1997 信息技术 —信息交换用藏文编码字符集 —基本集[S]. 北京: 中国标准出版社,1998.

[2] 芮建武,吴健,孙玉芳. 基于 ISO/IEC 10646 标准的藏文操作系统若干问题研究[J]. 中文信息学报,2005,19(5): 50-66.

[3] 江荻,周季文. 论藏文的序性及排序方法[J]. 中文信息学报 2000,14(1),56-64.

[4] The Unicode Consortium. The Unicode Standard 5.0 [S]. Boston, MA, Addison Wesley, 2006. ISBN 0-321-49091-0.

[5] 陈玉忠,俞士汶. 藏文信息处理的研究现状与展望[M]. 中国藏学,2004.04

[6] 李晋有. 中国少数民族语言文字现代化文集[M]. 北京: 民族出版社,1999.

[7] 朱巧明. 中文信息处理技术[M]. 北京: 清华大学出版社. 2005.

[8] Micro Software. OpenType specification [EB/OL]. http://www.microsoft.com/typography/otspec/.