藏文字符集基本集的修订方案

黄鹤鸣[1],契嘎·德熙嘉措(赵晨星)[2]
HUANG He-ming[1],ZHAO Chen-xing[2]

1.青海师范大学 物理系,西宁 810008
2.青海藏文信息技术研究所,西宁 810008
1.Physics Department,Qinghai Normal University,Xining 810008,China
2.Qinghai Institute of Tibetan Information and Technology,Xining 810008,China

HUANG He-ming,ZHAO Chen-xing.Some proposes on modification of Tibetan encoded character set—basic set.Computer Engineering and Applications,2007,43(20):187-189.

Abstract: This paper proposes some suggestions on modification of Tibetan encoded character set—basic set:Modify the property value of the character element (U+0FB2) and (U+0FBC);Modify the value of Canonical_Combining_Class field of character elements U+0F90~U+0FBC;Add a zero consonant;Add 36 compound vowels and add a graphical symbol .With these modifications and additions,basic set of Tibetan encoded character will perform better in Tibetan's computer input,output,storage etc.It will accelerate the process of Tibetan informationization.

Key words: Tibetan;character set;basic set;Unicode database;modification

摘 要:制作藏文字符集扩展集 A 和扩展集 B 的 Unico 数据库工作中,发现藏文编码字符集基本集有不完善的地方,现提出了几点修改建议:修改字元(U+0F62)的名称和字元(U+0F6A)的图形;修改字元(U+0FB2)与字元(U+0FBC)的属性值;修改字元(U+0F6A)的属性值;修改组合用下加字元 U+0F90~U+0FBC 的组合定位字段的属性值;增加一个辅音和对应的不占位形式;增加一个空辅音;增加 36 个复合元音;增加一个图形符号。通过这些修改,藏文编码字符集基本集的 Unicode 数据库将更趋完善,实现基本集的"利用基本集中的基本字符通过垂直组合形成藏文(叠字)字符,从而实现所有藏文字符的计算机处理"的目的。

关键词:藏文;字符集;基本集;Unicode 数据库;修订方案
文章编号:1002-8331(2007)20-0187-03 **文献标识码**:A **中图分类号**:TP391.1

1 藏文字符集和 Unico 简介

藏文字符集的建设是实现藏文信息化的基础工作之一。在国家信息产业部和信标委的大力支持下,在藏文信息技术专家的共同合作和不懈努力下,相继制定了《信息技术 信息交换用藏文编码字符集 基本集》(1997 年 9 月颁布),《信息技术 藏文编码字符集 扩充集 A》(送审),《信息技术 藏文编码字符集 扩充集 B》(送审),对藏文编码字符集的国际标准有了配套的标准。

《信息技术 信息交换用藏文编码字符集 基本集》于 1997 年 9 月由国家技术监督局批准。该字符集共收集了 193 个藏文基本字符和字元等,这些基本字符包括:辅音字母(包括变形辅音字母)、作为上加字或下加字的辅音字母、单元音、语音符、藏文标点符号、藏文数字(包括 10 个藏文中独有的半值数字)、藏文图形符号等。建立基本集的目的是利用基本集中的基本字元通过垂直组合形成藏文(叠字)字符,从而实现所有藏文字符的计算机处理。《信息技术 藏文编码字符集 扩充集 A》是基本集的扩充集,共收录了 1 536 个垂直组合字符,其中 574 个字符是现代藏文字符,其余 962 个字符是梵音藏文字符。扩充集 A 和基本集结合能表示和交换以现代藏文为载体的文字信息,满足现代藏文的信息处理需求。《信息技术 藏文编码字符集 扩充集 B》是扩充集 A 的补充,共收录了 5 669 个垂直组合梵音藏文字符。藏文字符集基本集、扩展集 A 和扩展集 B 三者结合起来能实现 99.99%的藏文字符(包括梵音藏文字符)的计算机表示、传输、交换、处理、存储及显现。

Unico(万国代码组织)是字符编码的非官方国际权威组织,它所制定的字符编码标准以及提出的许多国家、民族的字符编码数据库得到国际标准化组织(ISO)的认可。Unico 数据库用 15 个参数来描述每个字符,从而对每个字符的属性描述较全面,因此国际标准化组织(ISO)要求世界各国各民族的文字都严格按照 Unico 标准来制定字符集的编码标准。作为中华人民共和国国家标准的《信息技术 信息交换用藏文编码字符集 基本集》得到了 Unico 的承认,它是藏文字符集的国际标准。可以

基金项目:信息产业部(信部运[2002]393 号)项目。
作者简介:黄鹤鸣(1969-),男,藏族,硕士,副教授,研究方向:藏文信息处理,模式识别;契嘎·德熙嘉措(赵晨星)(1946-),男,藏族,教授,研究方向:藏文信息技术,计算机理论和软件。

说:藏文字符集的建设领先于国内其他少数民族语言字符集的建设,因为藏文是我国少数民族语言中第一个建立国际标准的语言。

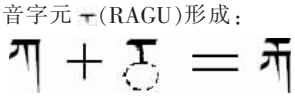
藏文字符集扩展集 A 中的字符基本确定后,作者参照 Unico 标准建立了该字符集的 Unico 数据库,在建立该数据库的过程中发现有必要对藏文字符集的基本集做一些局部的修改和适当的补充。在《对藏文字符基本集 UNICODE 数据库的商榷》一文中曾提出了部分的修改、补充建议;近期作者在完成藏文扩展集 B 中的 Unico 数据库后发现对藏文字符基本集还需要更进一步的补充和完善。作者归纳整理了所有的这些修改建议后,提出了这个修改方案。为了使修改方案更具系统性作者仍然引用了《对藏文字符基本集 UNICODE 数据库的商榷》的部分论点。

这个修改方案是在完成扩充集 A 和扩充集 B 的 Unico 数据库后提出的,所以按此方案修改后的基本集中的基本字符确实能组合出扩充集 A 和扩充集 B 中的共 7 205 个组合藏文字符,能实现 99.99%的藏文字符的信息处理。从而实现建立基本集的目的即利用基本集中的基本字元通过垂直组合形成藏文(叠字)字符,从而实现几乎所有藏文字符的计算机处理。

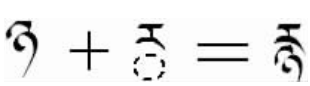
2 修改部分字符的属性值

2.1 藏文语法和信息处理对构成组合字符的字元的分解差异

对于组合辅音字符,藏文语法分解其构成元素时,有上加辅音的概念。例如,组合辅音字符 是在基本字元 的上方添加变形的上加辅音字元 (RAGU)形成:



而组合字符 是在基本字元 的上方添加不变形的上加辅音字元 形成:



因此辅音字元根据它在组合字符中出现的位置可以分为基本辅音字元、上加辅音字元、下加辅音字元三种形式,但是部分辅音做上加字元或者下加字元时可能需要变形。例如 既可以做上加字元又可以做下加字元,并且在这两种情况下都有可能发生变形。因此从藏文语法的角度来看藏文字符 应该有五种形式:基本字元、上加变形字元、上加不变形字元、下加变形字元、下加不变形字元等。

但是在藏文信息处理中没有上加辅音的概念,组合辅音字符的第一层辅音被认为是基本辅音,以下各层的辅音字元都是下加辅音。因此在藏文信息处理中,前面提到组合辅音 认为:基本辅音是 而 是下加辅音。

2.2 修改字元 (U+0F62)的名称和字元 (U+0F6A)的图形

上节谈到在计算机处理藏文时,组合辅音字符的第一层辅音被认为是基本辅音,以下各层的辅音字元都是下加辅音,没有上加辅音的概念。但是辅音 作为上加字时可能会发生变形,因此有两个字形: 和 。并且组合时变形的 (RAGU)出现的频率比没有变形的 出现频率高。

考察基本集发现有两个相同的图形 (U+0F62)和字元 (U+0F6A)。这两个字元的 Unico 记录如下:

0F62;TIBETAN LETTER RA;Lo;0;L;;;;;N;;*;;;

0F6A;TIBETAN LETTER FIXED-FORM RA;Lo;0;L;;;;;N;;*;;;

从字元 0F6A 的名称 LETTER FIXED-FORM RA 可以知道,这表示没有变形的辅音 ,而 U+0F62 表示的是变形了的辅音 ,因此应将 U+0F62 的字形修改 。

藏文中对 U+0F62 表示的字形有专门的名称:RAGU,但是如果名称字段填写为 TIBETAN LETTER RAGU,则有可能将其误认为一个新的辅音。为了指明它是 的变形,应填写 TIBETAN LETTER TRANSFORMED RA。而 U+0F6A 表示的是没有变形的辅音 ,因此它的名称字段应和其他没有变形的辅音的名称字段一致,填写为 TIBETAN LETTER RA。因此作者认为这两个字的 Unico 记录应改为:

0F62;TIBETAN LETTER TRANSFORMED RA;Lo;0;L;;;;;N;;*;;;

0F6A;TIBETAN LETTER RA;Lo;0;L;;;;;N;;*;;;

2.3 修改字元 、、以及 、、的名称

下加辅音字符 (U+0FBA)、(U+0FBB)、(U+0FBC)有变形和不变形两种情况,在梵音藏文中它们一般以原形出现,而在现代藏文中它们分别变形为: (U+0FAD)、(U+0FB1)、(U+0FB2)。这 3 个变形的下加字元有专门的名称,分别是:WADA、YADA、RADA。如果将它们名称字段填写为 TIBETAN SUBJOINED LETTER WADA、TIBETAN SUBJOINED LETTER YADA 和 TIBETAN SUBJOINED LETTER RADA,则会使人误认为它们是新的辅音。为了指明它们是下加辅音 、、的变形,可以将它们命名为:TIBETAN SUBJOINED LETTER TRANSFORMED WA、TIBETAN SUBJOINED LETTER TRANSFORMED YA、TIBETAN SUBJOINED LETTER TRANSFORMED RA。

字元 在 Unico 中的名称为:TIBETAN SUBJOINED LETTER FIXED-FORM WA,这里应该去掉修饰 FIXED-FORM 修改为 TIBETAN SUBJOINED LETTER WA,以便和其他辅音的命名规律一致。另外两个也应分别改为 TIBETAN SUBJOINED LETTER YA,TIBETAN SUBJOINED LETTER RA。

2.4 修改字元 (U+0FB2)与 (U+0FBC)的属性值,严格区分下加变形字元 (U+0FB2)与下加不变形字元 (U+0FBC)

藏文编码字符基本集对下加变形字元 与下加不变形字元 在字形上作了明确的区分,但在属性方面特别是对它们的组合定位(Canonical_Combining_Class)字段的属性值没有作严格区分,将这两个字元的组合定位字段的值都填为 0。这样做没有反映出这两个字符的组合特性,应该进行如下修改:由于下加变形字元 (RADA)是下接于基本字元,所以它的组合定位(Canonical_Combining_Class)字段的值应该是 202;而下加不变形字元 (U+0FBC)出现在基本字元的下部,所以它的组合定位(Canonical_Combining_Class)字段的值应该是 220。即在 Unicode 字符数据库中字元 (U+0FB2)与字元 (U+0FBC)相应的记录应为:

0FB2;TIBETAN SUBJOINED LETTER RA;Mn;202;NSM;;;;;N;;;;;

0FBC;TIBETAN SUBJOINED LETTER FIXED-FORM RA;Mn;220;NSM;;;;;N;;;;;

2.5 修改部分组合用下加字元的属性值

字元 U+0F90~U+0FAC、U+0FAE~U+0FB0、U+0FB4~U+0FBC 为组合用下加字元,当它们与基本字元组合时出现在基本字元的下部,因此它们的组合定位字段的值应为 220;而字

元(WATA)(U+0FAD)、字元(YATA)(U+0FB1)、字元(RATA)(U+0FB2)以及字元(LATA)(U+0FB3)与基本字元组合时下接于基本字元，因此它们的组合定位字段的值应为202。但在Unico数据库中所有这些字符的组合定位字段值均为0，这没有反映出这些字符的特性，因此建议按照前面的讨论修改这些字符的组合定位字段的值。

3 建议增加部分基本字元

3.1 增加一个辅音和对应的不占位形式

在藏文编码字符集基本集中，收集了辅音字元和对应的不占位形式。这是因为，虽然在吴坚白体中可以将它们认为是和组合而成，但是在藏文的其他字体中将它们看成整体，因此将其作为一个独立的辅音收集。同理应该将辅音收集到基本集中来，并且将它的不占位形式也收集进来，分别放在U+0F6B和U+0FBD的位置。

其中的不占位形式出现在基本辅音的下部，因此根据Unico的规定，它的组合定位(Canonical_Combining_Class)字段的值应该是220。所以这两个字元的Unico记录为：

0F6B;TIBETAN LETTER KSS;Lo;0;L;;;;;N;;*;;;

0FBD;TIBETAN SUBJOINED LETTER KSS;Mn;220;NSM;;;;;N;;;;;

3.2 建议增加一个空辅音的编码

藏文字符集中有许多不占位字符：元音、上加字符(包括变形的上加字符)、下加字符(包括下加的变形字符)。有时需要独立显示这些不占位字符(例如当有人要编写有关藏文的语法书或语音书时)，但是如果将所有这些不占位字符对应的占位字符都添加到基本集中去将造成基本集中有限编码资源的浪费。可以通过添加一个空辅音的方式来解决这个问题。例如，如果想显示上加字符，就可以先敲一次空辅音对应的键，然后敲不占位的上加字符对应的键就可以独立显示该字符。即

藏文字符基本集可以粗略地分为两部分：占位字符区和不占位字符区。其中占位字符区到U+0F6A结束，因此建议将编码U+0F6B分配给空辅音，由于空辅音虽然不显示任何图形，但它是占位字符。

下面简要讨论一下空辅音在Unico数据库中记录的填写。藏文中的空辅音像英语键盘中的空格键一样，可以将其看作是空间占位符，因此它的类型(General_Category)字段的值应该是Zs；虽然空辅音不显示任何图形，但它是占位字符，因此组合定位(Canonical_ Combining_Class)字段的值应为0；而书写方向(Bidi_Class)字段的值应为WS。因此在藏文基本集中空辅音的Unico数据库中对应的记录应该是：

0F6B;TIBETAN ZERO CONSONANT;Zs;0;WS;;;;;N;;;;;

3.3 建议增加36个复合元音的编码

藏文中除基本集中出现的单元音外，还有36个复合元音：

复合元音是由多个单元音组成的一个整体。这些复合元音中有些只叠加在辅音的上面(例如：)，而有些则叠加在辅音的上下两部(例如：)。

有些人可能认为既然复合元音是由单元音组合而成，那么何必在基本集中还要添加这些复合元音，用相应的单元音组合表示不行吗?这样做不妥，因为单元音在组合成相应的复合元音的过程中它们的位置发生了改变。例如，复合元音是由两个单元音和复合而成，在这个复合元音中单元音出现在辅音的左上部而单元音在辅音的右上部。但是当单独使用时单元音和都出现在辅音的上部，因此严格地从这两个字符的组合定位的属性来讲，它们无法构成复合元音。因此建议在基本集中增加这36个复合元音。

在藏文编码字符集B(审定稿)中把这36个复合元音的32个已经收集进去，给了编码，但这样产生了几个问题：

(1)复合元音不全应补全，没有将复合元音、、、收集进来。理由是在藏文字符扩展集B中收集了以它们为元音的字符。例如：以为元音的字符有：(U+0F0017)、(U+0F081A)；以为元音的字符有：(U+0F0005)，以为元音的字符有(U+0F0007)，以为元音的字符有(U+0F0009)等，所以这4个复合元音也应该收集到藏文字符集中。

(2)复合元音是作为组合藏文字符的一个基本元音，用它上下叠加而构成藏文字符的。基本集收集了辅音字母和单元音的组合用的形式，也应该把组合用的字元放在基本集中，不应放在扩展集B中，扩展集A和扩展集B是已组合的字符集，不是组合用的字元。所以组合用的字元归基本集，字符归到扩展集A和B才是有序的。

(3)用Opentype技术构成藏文字符时，构造规则是由组合用的藏文字元作为构件来描述的，因此字元都放在基本集中是方便的。

(4)用Opentype技术构成藏文字符时，辅音字元和单元音字元为双字节编码，而复合元音字元放在扩展集B中，则成了三字节编码，这样处理不等长编码是非常麻烦的。

(5)这36个复合元音相对而言构成了一个整体，基本集中有空位，因此最好将它们放在一起。可以将U+0FD7到U+0FFA之间的36个编码分配给它们。

下面简要讨论每一个复合元音的Unico记录中各个字段的值。其中名称字段采用藏文的拉丁转写，藏文字符基本集中采用的是由青海师大物理系的赵晨星教授设计的藏文拉丁转写方案，作者在填写这些复合元音的名称字段时仍然采用该方案；因为它们都是不占位字符，所以类型字段的值为Mn；有些复合元音(例如：)出现在辅音的上部，所以组合定位字段的值应该填写230，而有些复合元音(例如：)同时出现在辅音的上部和下部，所以组合定位字段的值应填写0；对于书写方向字段的值则统一填写为NSM，因为它们都是不占位字符；其他字段则可以保留缺省值。

3.4 建议增加一个图形符号

藏文中的图形符号较多，这些图形符号大致可以分为佛教仪轨符号和天文历算符号两类。对于这两类图形符号在基本集中都有收集，但在天文历算符号类中缺了一个图形符号：，希望能将该符号补充进去，从而使藏文的图形符号更完整。图形符号和图形符号都属于天文历算符，所以可以放在之

(下转193页)