# 藏文字笔画编码排序的设想①

刘 城, 黄鹤鸣, 李继文

(青海师范大学 计算机学院, 西宁 810008)

**摘 要**: 藏文字符排序将被广泛应用于藏文文字信息处理的各个方面, 包括字、词典的排序、系统软件和其他应用软件. 试图对藏文的书写笔画排序规则做出较为正确、合理的归纳和富有逻辑性的描述, 目的是为了找到一种在计算机里自动实现藏文笔画排序的算法模型, 并打破了藏文字符仅依赖于音节部首结构排序的传统思维定式和框架.

**关键词**: 藏文; 笔画编码; 笔画; 模式识别; 音节

## Tibetan Strokes of Computer Codes Sorting

LIU Cheng, HUANG He-Ming, LI Ji-Wen

(Computer College, Qinghai Normal University, Xining 810008, China)

**Abstract**: Tibetan sort will be widely used in every aspect of Tibetan language text information processing, including word, dictionary sequence, system software and other application software. This paper attempts to describe Tibetan writing stroke sorting rules which make more correct, reasonable induction, the purpose is to find a sorting algorithm model that could realize automatically the Tibetan strokes character. And which break the framework in Tibetan character syllable sequence depends only on the syllable key radical structure sorting traditional of thinking.

**Key words**: Tibetan; stroke coding; stroke; pattern recognition; syllable

藏文的字母和其它符号有一定的笔顺, 藏文 1 的笔顺有些和汉文字的笔顺一样, 按照笔顺写, 字才写得漂亮. 不过, 笔顺作为规则是活的, 各种教材里的英文字母笔顺体样式就多种多样, 汉字笔顺也时有调整, 所以对藏文的标准不唯一, 是正常现象.

## 1 常见藏文字体的笔画样式概述

首先我们来分析以下列出的 3 套体系的藏文印刷体笔顺, 各有一些差别(包括声调符号). 不难看出他们的字体样式会有所不同, 但字的笔画是近似的. 三类分别为: 一是带有箭头笔画指向的版本藏文字书写, 见图 1; 二是藏族朋友们初学时, 教师常教的一种藏文字书写笔画顺序, 见图 2; 三是藏文报刊、教科书的上的书写版本, 见图 3.

图 1 带有箭头笔画指向的藏文字书写方式

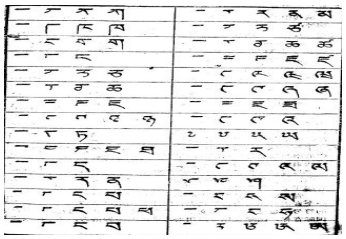

图 2 教师常教的一种藏文字书写笔画

① 基金项目:国家自然科学基金(60963016);藏文信息处理省部共建重点实验室开放课题
收稿时间:2012-10-15;收到修改稿时间:2012-11-14

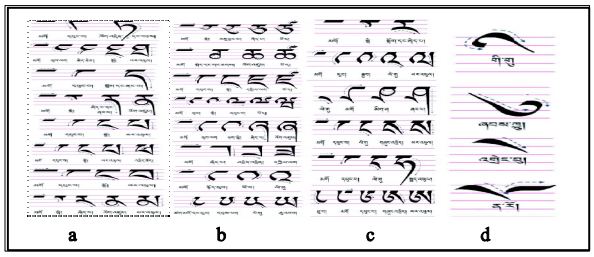


图 3 藏文报刊、教科书的上的书写版本

通过上面公认的三套体系的藏文书写版本，因人而异的会喜欢自己的一套书写藏文的笔画顺序，作为写惯了中韩文日文的学者朋友们，肯定会将其与其他文字的书写笔画或方式进行对比，不难看出，藏文的笔画字体有其独特的特点是：每个字母最上一笔是横直的，字母排列时．上端必须在一条直线上，形似平顶帽．由于这种字体多用作刊印书籍、录、写文章的字体，也成出版字体，另一种笔顺书写则很随意．

对于每一个了解和熟悉藏文字的人，笔画和书写是习得文字的前提，因此采取笔画编码形成的输入方法不但规范而且易于掌握，无需强记，真正可以做到计算机汉字输入如同写字，得心应手．

## 2 计算机藏文字笔画编码具有它的科学性

计算机藏文字编码包括字库编码和输入法编码两类[1]；其中输入法编码分为，键盘输入法编码和非键盘输入法编码；在键盘输入法编码中主要有音节编码和字型编码两大类．经过长期深入研究和实践，我们认为计算机藏文输入法采取笔画编码具有其它方法难以逾越的优势，它能最有效地解决无法用藏文音符编码输入所有藏文的问题，它能最有效地解决以往字型类编码难学难记的问题．

同样是拼音文字的现代藏文，也有其科学的、明确的、传统的排序规则，只是由于现代藏文在字符结构构成、拼写方法和书写走向等方面与英文有所不同，使得对现代藏文排序规则的描述相比英文字符排序要多些步骤．

任何一种语言文字都有自己的排序规则，人们在使用该语言文字时都习惯性的有种共识和规范．人们在使用这熟知和共同遵守的规则，可以对各种字符、词典和字、词表以及查找其中的字符进行编排．例如：英文作为在世界范围内最普遍实用的语言文字，有其自身的排序规则要求．在众多程序设计语言的库函数中也都有其字符(串)比较函数(模块)，这样对程序员编程来说很方便，也更有利于广大用户的使用．

## 3 藏文字的笔画规范

### 3.1 藏文字笔画顺序具有很强的规范性

汉字的笔画编码在上世纪八十年代风行一时，此文的思路基本上按照这个思路进行编码，因为汉藏文字属于同一体系，所以从原则上这是讲得通的：由于国家语言文字委员会对藏文字定有《现代藏文通用字笔顺规范》，小学藏语文教学必教笔画顺序，因此采取笔画编码是书写藏文字的人都可接受和掌握，它不要求什么特殊的记忆，只要会写的人就能使用，这点与汉字的笔画编码的理由一致．汉字的笔画编码已经在被人们广泛使用，所以藏文字笔画顺序具有很强的规范性．

在实际编码中，由于藏文字结构具有较强的规律性，但是藏文字很少有完全相同的，个别笔画的使用常常凝聚在部分组合上，而且藏文字的单音节字符笔画多的达到 6 画以内，按照标准四键编码，如不进行有效合理的处理，重码率[1]难以降低．因此在制定编码技术过程中，必须着重根据藏文字的笔画分布规律，结构规律，普通人群的识字规律对所有藏文字笔画和结构以及词组进行充分整理和分析，确定符合人机行为学的编码规则．

藏文字本身由简单的藏文基本音节构成，也可以看成是由基本笔画通过不同顺序和笔画数组合构成，笔画和笔画顺序完全相同的藏文字极少，这就使得我们有针对性地采取一定的措施之后，笔画编码是能够有效地控制编码重码率的．

### 3.2 藏文字笔画的结构

藏文字笔画分布规律要求最合理的定义使用的笔画，在汉字笔画中包括“横、竖、撇、捺、点、折、弯钩、提”，而在藏文字中这些笔画的使用率各不相同，有些使用率很低，以 30 个基字单音节为例研究藏文文字笔画的使用频率分别为：横的使用率为 96.67%、撇的使用率为 20.02%、捺的使用率为 6.67%竖的使用率为 33.3%、弯钩的使用率为 10.03%、提的使用率为 3.33%．因此需要科学的归类和组合笔画才能最合理的代表藏文字组字信息，以均匀分布笔画编码．

其中结构规律要求最直观的是定义藏文字字型，并科学的确定笔画分配，藏文字字型包括“左右型、上下型、左中右型、上中下型、独体型”等等，由于藏文字字符串笔画数多，笔画编码必须根据字型分配笔画，此时必须考虑编码的重码率，这一点很重要，任何编码如果重码率太高，就意味着输入时需要更多的选择，输入效率不高是一个方案被淘汰的主要原因之一，因此在藏文字字型选择越多则编码重码率就易于下降，但规则就相对复杂，因此应尽可能的简单化．设计补

充藏文字笔画编码是由于部分藏文字笔画较少，不足以编辑四码；采取两种方法共同补充，一是按标准藏文音节字母顺序补充，二是重新按藏文字书写笔画补充，从而确保了会使用藏文字汉化音的和会藏文字笔画的人均可以方便灵活使用.

## 4 藏文字符音节构成部件的基本字符

Unicode 现代藏文编码字符集为现代藏文分配的编码空间是从 OFOO 至 OFFF. 其中，对现代藏文字符的音节是按类连续编码的，字符的分类规则和排放顺序完全依照现代藏文传统的排序方法. 针对藏文字符的 30 个辅音字母连续分配于 OF4O 至 0F69 之间，4 个元音符号分配于 OF71 至 OF7D 之间. 因此，比较现代藏文字符音节间各构件时，不需要再为辅音字母和元音字符编码进行编码，直接比较其字符音节的编码值. 根据 Unicode 国际字符集编码原则，现代藏文的上加字音节必定都为非组合用字符区的字符，下加字音节则都为组合用字符区的字符，通过代码值就可以正确反映它们的排序顺序. 藏文字符音节构成部件的基本字符具体如下表 1:

表 1 藏文字符音节基本字符表

| 序号 | 音节部件 | 藏文音节字符 | 个数 |
|---|---|---|---|
| 1 | 基字字符 | ཀ ཁ ག ང ཅ ཆ ཇ ཉ ཏ ཐ ད ན པ ཕ བ མ ཙ ཚ ཛ ཝ ཞ ཟ འ ཡ ར ལ ཤ ས ཧ ཨ | 30 |
| 2 | 前加字 | ག,ང,ད,བ,མ | 5 |
| 3 | 上加字 | ར,ལ,ས | 3 |
| 4 | 下加字 | ཡ,ར,ལ,ཝ | 4 |
| 5 | 元音 | ི ,ུ ,ེ ,ོ | 4 |
| 6 | 后加字 | ག,ང ,ད,ན,བ,མ,འ,ར,ལ,ས | 10 |
| 7 | 再后加字 | ད,ས | 2 |

因此，要设计符合规范的现代藏文字排序模块，必须使之完全基于 Unicode 现代藏文编码字符集及其编码原则，本文就是基于 Unicode 现代藏文编码字符集标准进行应用设计构思的[2-5]. 在藏文字中“横竖撇捺点折弯钩提”对其笔画部首进行编码，将会很省计算机的内存空间，处理数据的时间度将会降低. 这样更方便、更有效的字符编码排序的趋势也将会更快捷和简单化.

## 5 对现代藏文字笔画编码排序顺序的构思

现代藏文排序将被广泛应用于现代藏文字符的字处理的各个方面，包括字、词典的排序、系统软件和其他应用软件. 目的是在 Unicode 国际字符编码的标准下，实现一种在计算机里自动分割现代藏文单音节排序的算法模型. 采用藏文笔画编码的排序前提还是要依赖于藏文基本字符音节的科学排序规则.

现代藏文字笔画分部有如下表 2:

表 2 现代藏文字笔画分部

| 笔画数 | 藏文字 | 个数 |
|---|---|---|
| 一画 | | 0 |
| 二画 | ད, ན, ི ,ུ ,ེ ,ོ | 6 |
| 三画 | ཅ ,ཆ, ཇ, ཉ, ཏ ,ཐ, ད ,ན, པ, ཕ, བ | 11 |
| 四画 | ཀ, ཁ, ཀ, ག ཚ ,ཛ, ཝ, ཞ, ཟ ཡ ,ར ,ལ | 15 |
| 五画 | ཤ, | 1 |
| 六画 | ས | 1 |

掌握藏语知识的人，他们对现代藏文语言文字的传统人工排序规则很熟悉，他们能很快的进行编排和查找现代藏文词典. 可是想要在计算机中实现对这一传统的排序规则做一个简单明了而又富有逻辑性的描述却具有一定的挑战性. 现代藏文字符音节的排序是指依照一定的规则，对输入单个现代藏文音节结构部件进行优先级的排序，从而指出现代藏文单音节的各个结构部件，为进一步的多音节现代藏文排序提供很好的参考与实现. 对现代藏文字笔画编码排序顺序则需要把藏文字的基本音节字符掌握清楚，在其基础上作进一步的细化，这样藏文字笔画编码排序规则就易于程序设计及算法优化了.

## 参考文献


1 张学涛.汉字的笔画、部件、偏旁和基本字四大组成部分—中文信息名词标准化概念探讨.计算机中文环境,1995,(7).

2 黄鹤鸣,达飞鹏.基于排序的现代藏文音节判定.计算机应用,2009,29(7).

3 黄鹤鸣,赵晨星.对 DUCET 中部分藏文字母排序码的商榷.计算机工程与应用,2008,44(29).

4 扎西次仁.藏文的排序规则及其计算机自动排序的实现.中国藏学,1999,4.

5 春燕,曲珍.藏文文本编码识别方法研究.计算机工程与应用,2011,10.

6 珠杰,欧珠.基于藏文编码 GB 的藏文排序方法研究.西藏大学学报(自然学科报版),2008,5.

7 江荻,康才晙.书面藏语排序的数学模型及算法.计算机学报,2004,4.

8 江荻,燕海雄.藏文字符的分类与功能描述.西藏研究,2010,10(5).

9 王华,丁晓青.一种多字体印刷藏文字符识别方法计算机工程,2004,30(13).