

国家自然科学基金：基于依存关系的藏文语义角色标注研究[61363057]项目资助
甘肃省民族语言智能处理重点实验室资助

藏文分词与标注研究

Research on Tibetan Segmentation and Tagging

祁坤钰 著



甘肃民族出版社

祁坤钰，男，甘肃天祝人，副教授，硕士生导师，所学专业中国少数民族语言文学，现从事专业为语言学、藏文信息处理及计算语言学。兼任藏文国家标准工作组成员。

1992 年 7 月本科毕业于西北民族学院藏语言文学专业，1996 年 6 月至 1999 年 6 月攻读藏语言文学系藏汉翻译硕士学位。1999 年至今主要从事藏文信息处理及藏语语言研究工作。2002 年在清华大学做访问学者，与模式识别重点实验室合作研发了藏文 OCR 文字识别系统。近年来参与 863 等多项国家级科研项目，2001 年《藏文视窗平台、字处理软件和藏文网站》项目获得国家科技进步二等奖，主持国家自然科学基金项目《基于依存关系的藏文语义角色标注研究》一项，承担《计算语言学》、《统计语言学》、《面向对象技术》和《藏语语法》等硕士研究生专业课程的教学任务。先后发表了《基于语料库的藏语名词分类与统计研究》、《信息处理用藏文自动分词研究》、《基于国际标准编码系统的藏语分词词典机制研究》、《基于依存关系的藏文语义角色标注研究》、《基于词汇语料的白马藏语语音分析研究》和《面向信息处理的藏语语义角色研究》等学术论文 30 多篇。

国家自然科学基金：基于依存关系的藏文语义角色标注研究[61363057]项目资助

甘肃省民族语言智能处理重点实验室资助

藏文分词与标注研究

Research on Tibetan Segmentation and Tagging

祁坤钰 著



甘肃民族出版社

图书在版编目 (CIP) 数据

藏文分词与标注研究 / 祁坤钰著. --兰州: 甘肃
民族出版社, 2015.3
ISBN 978-7-5421-2839-3

I. ①藏… II. ①祁… III. ①藏语—分词—研究
IV. ①H214.4

中国版本图书馆CIP数据核字 (2015) 第 059405 号

书 名: 藏文分词与标注研究

作 者: 祁坤钰 著

责任编辑: 瞿广业

装帧设计: 扬忠仁青

出 版: 甘肃民族出版社(730030 兰州市城关区读者大道 568 号)

发 行: 甘肃民族出版社发行部(730030 兰州市城关区读者大道 568 号)

印 刷: 甘肃北辰印务有限公司

开 本: 710 毫米×1020 毫米 1/16 印张:13 插页:1

字 数: 230 千

版 次: 2015 年 4 月第 1 版 2015 年 4 月第 1 次印刷

印 数: 1~1 000

书 号: ISBN 978-7-5421-2839-3

定 价: 30.00 元

甘肃民族出版社图书若有破损、缺页或无文字现象,可直接与本社调换。

邮编:730030 地址:兰州市城关区读者大道568号 网址:<http://www.gsminzu.com>

投稿邮箱:448925720@qq.com

发行部:屈滢 联系电话:0931-8773312 8773264(传真) E-mail:1130375943@qq.com

版权所有 翻印必究

目 录

第一章 绪 论	1
一、藏文信息处理概述	1
(一) 藏文文字处理	1
(二) 藏文编码标准	3
(三) 藏文操作系统	9
(四) 藏语自然语言处理	10
二、藏文自动分词与词性标注的概念	11
(一) 基本概念	11
(二) 藏文自动分词与标注的重要性	13
(三) 藏文自动分词与标注的目标	15
三、藏文自动分词的应用领域	16
(一) 藏外电子翻译	16
(二) 藏文智能信息检索	17
(三) 语音处理技术	17
(四) 自动文摘	18
第二章 分词与标注的原理	19
一、语言与词法分析	19
二、英语的词法分析	20
(一) 形态学的基本概念	20
(二) 英语形态分析	23
(三) 英语的词类划分与标注集	27

(四) 英语词性标注方法.....	28
三、汉语的词法分析	32
(一) 汉文分词的概念.....	32
(二) 汉文分词规范.....	33
(三) 汉文词类与标注规范.....	34
(四) 汉文分词方案.....	35
(五) 汉文标注方案.....	52
(六) 汉文分词与标注系统介绍.....	53
第三章 藏文与藏文文法.....	55
一、藏文概述	55
(一) 藏文字的产生.....	55
(二) 藏文字的厘定和改革.....	57
二、藏文文法概述	62
(一) 古印度声明学体系.....	62
(二) 文法理论的藏译.....	63
(三) 藏文文法史.....	64
(四) 藏文文法的主要内容.....	74
第四章 藏文自动分词及标注规范	81
一、藏文分词与标注规范说明	81
二、使用范围与参考规范	81
(一) 使用范围.....	81
(二) 参考规范.....	82
三、基本术语	82

四、分词与标注原则	84
五、词类划分与标注集	84
六、切分与标注细则	87
第五章 匹配与统计相结合的分词与标注	108
一、藏文分词标注现状	109
(一) 分词与标注规范	109
(二) 分词与标注系统	109
二、藏文分词标注语料库	110
(一) 语料库的属性	110
(二) 语料分类	112
(三) 语料分布表	118
(四) 标注语料实例	119
三、字符串匹配与统计相结合的藏文分词算法	120
(一) 正向最大匹配方法	121
(二) 逆向最大匹配方法	122
(三) 双向最大匹配方法	123
(四) 切分错误与切分歧义	123
四、分词词典机制	129
(一) 分词词典数据	130
(二) 词典的数据结构与搜索算法	132
五、基于隐马尔科夫模型的藏文词性标注	136
(一) 统计与概率的基础概念	136
(二) 统计语言模型	138

(三) 隐马尔科夫模型 (HMM)	140
(四) 藏文兼类词	147
(五) 藏文标注模型	149
(六) HMM 模型参数训练	151
(七) 韦特比算法	153
六、藏文分词与标注测评	158
(一) 黄金标准	159
(二) 评价指标	159
(三) 中文测评	160
(四) 藏文分词与标注测试	162
七、本方案存在的问题与改进的思路	167
(一) 存在的问题	168
(二) 今后的改进思路	168
八、本软件的使用	171
附录	172
1. 藏文国际标准编码	172
2. 英语词性标注集	180
3. 信息处理用现代汉语分词规范	185
4. 信息处理用现代汉语词类标记规范	194
参考文献	199

第一章 绪 论

一、藏文信息处理概述

藏民族是中华民族大家庭中的一员，具有悠久的历史 and 灿烂的文化。藏族文化是中华民族文化的重要组成部分，也是世界文化的瑰宝。世界各国人民都对藏民族的传统文化、宗教文化和风俗习惯怀有浓厚的兴趣。文化的融合与传播是继承和弘扬民族文化的重要途径之一，以大藏经为代表的藏族传统文化本来就是苯教文化、古印度文化和内地中原文化交融的结晶。在信息时代，计算机网络成为各民族文化传播的重要渠道。要想弘扬藏族优秀文化，使藏语文适应现代化的发展，首要的问题就是必须不失时机地发展藏文信息处理技术，使藏语文能够紧跟世界信息技术的大潮，使得藏语文的学习、使用和发展真正步入现代化的轨道。藏文信息处理是伴随着信息技术的发展和计算机的普及而产生的。最早采用计算机程序来研究藏文的是张连生。他于 1981 年采用于道泉先生提出的藏文拉丁转写方案，在英文 DOS 系统下用 COBOL 语言实现了一个简单的藏文排序软件^[1]。之后更多的科研单位、高校和个人都从事过藏文系统技术开发。到目前为止，藏文信息处理走过了三十多年的历程，其研究内容也从单一的办公文字处理扩展到了藏语自然语言处理领域。为了更清楚地了解藏文信息处理的发展历程，我们从文字处理、标准编码、藏文操作系统和自然语言理解等几个方面来做简要介绍。

（一）藏文文字处理

藏文文字处理是藏文字信息化的基本需求，也是催生藏文标准编码的制定和加速藏文操作系统研发的根本动力。根据藏文字符编码与操作系统的不同，把以前的文字处理软件分为如下两类：

1. 基于 DOS 系统和非标准编码的文字处理

计算机操作系统源自美国，DOS 操作系统从 1.0 到最后的 7.22 版本都是西文版本。CCDOS 及 WPS 系统均为汉文字处理软件，不能算作严格意义上的操作系统。在 DOS 系统下，所有的用户接口命令都是英文字符，不支持汉字，更不能支持藏文。1980 年，中国国家标准总局发布了汉字 GB2312 编码标准，即：信息交换用汉字编码字符集，共收入 6763 个汉字。

整个字符集分成 94 个区，每区有 94 个位，每个区位上只有一个字符。因此，可用所在的区和位来对汉字进行编码，称为区位码。当时藏文编码还没有国家标准，所有尝试研发的藏文字处理软件都各自为政，编码不统一，加之藏文叠加组合复杂，起初的软件都是采用拉丁字母或数字来代表藏文字母。根据相关资料显示，1984 年上海教育学院物理系俞乐等人在 VICTOR9000 微机上利用 BASIC 语言实现了一个具有输入、显示和打印功能的藏文字处理系统，并用 BASIC 编写了藏文报表软件。1985 年甘肃省计算中心胡彦发等人 and 西北民族学院合作，也在 WANGVS/80 机上用扩展 BASIC 语言实现了一个藏文的字处理系统 ZWCL，他们还配套编写了藏文文献联机检索系统。航天部 710 所罗圣仪等人在微机 PC-8001 和 IBM-PC 上实现了一个藏文字处理系统。该系统利用 TLLP 字符写入程序实现，并采用拉丁转写和数字对藏文字符进行了编码。

随着 CCDOS 汉字字处理系统的产生，很多藏文字处理软件也在该系统基础上进行了扩展开发。1986 年青海师范大学赵晨星、青海民族学院毛继祖和熊涛等在 CCDOS2.13 下开发了与汉英文兼容的藏文操作系统 TCDOS。后来，熊涛与西北民族学院于洪志等人合作开发了可挂接在 WPS 下的兰海藏文系统。1988 年中国藏学研究中心和航天部 710 所联合推出了藏文字处理及激光编辑排版印刷系统，1989 年该系统与潍坊华光合作开发了华光藏文排版系统。1990 年北京大学计算机研究所、民族印刷厂、中国民族语文翻译中心等单位在华光藏文系统的基础上，联合推出了北大方正藏文书报排版系统。1992 年西藏大学尼玛扎西等人研究开发了 TCE 藏汉英文信息处理系统，通过了西藏自治区的鉴定。其中北大方正藏文和华光藏文在藏区应用非常普及，青海、甘肃、四川、西藏等地的民族出版社都采用方正或华光藏文，对藏文古籍整理和藏文图书出版起到了重要的作用。在藏文信息处理领域产生了深远影响。这些系统有一个共同的特点是：都在 DOS 环境下使用，并且藏文字符编码都不统一，软件互补兼容。

2. 基于 Window 系统和非标准编码的文字处理

1995 年以后，随着图形界面 Windows 操作系统的迅速普及，Windows 系统很快取代了 DOS 系统成为主流的操作系统，以文档为中心的资源管理模式给用户带来了极大的方便。但是，由于微软公司未公开操作系统的内核技术，要开发基于 Windows 系统的民族语言本地化技术几乎没有可能。因而，也只有采用外挂模块技术才能实现 Windows 上的藏文字处理。这样的局面直到 2007 年微软推出 Windows Vista 和藏文复杂文本引

擎技术(USP 技术)才宣告结束。

在这一阶段有代表性的藏文字处理软件有：北大方正的维思彩色印刷系统、西北民族大学的同元藏文字处理软件、青海师范大学的班智达藏文字处理软件、西藏大学的火狐藏文字处理软件，还有国外的 Sambhota 藏文处理软件。由于技术实现的难度，这些软件都未采用藏文国际编码标准，几乎都用 TrueType 字库与输入法相结合的方式实现藏文输入，字符编码都采用纵向叠加字丁形式，未能涵盖所有的藏文字符，软件应用领域非常有限。而且编码不统一，相互不兼容。但是，同元藏文字处理软件和班智达藏文字处理系统在藏文 OCR 文字识别、语音合成和机器翻译等自然语言处理领域具有划时代的意义。

近年来，由于藏文标准系统的出现以及信息技术的更新和计算机的广泛普及，这些非标准的文字处理软件已经逐步退出了人们的视线，已经成为了历史。

（二）藏文编码标准

藏文字符编码是藏文信息标准化研究的根本问题，统一的字符编码标准是实现藏文系统的基本前提。藏文编码标准是指藏文字符编码要通过国际标准化组织（ISO）的授权，使得藏文能够在世界范围内与英文、汉字等其他文字一样统一编码，能够在互联网统一平台上实现藏文信息交换与数据处理。从理论上说，一种文字的标准化首先要完成该语言的文字编码工作，这个环节是电子信息技术与产品开发应用、信息化建设必不可少的技术基础。有了语言文字的编码及标准，计算机网络和通信系统才能实现各民族及各国家之间的经济、文化和社会信息的实时传输与交流。

1. 藏文国家编码标准制定的过程

藏文信息技术标准化的问题首先由国际标准化组织（ISO）于 1992 年提出，其核心任务是藏文编码体系的标准化问题。主要包括藏文编码字符集标准（交换码）、基本集和构建集、藏文字符键盘布局标准（输入码）和藏文字形标准（字形码）等。为了使我国少数民族文字尽快进入国际编码标准体系，1993 年在国家技术监督局的领导下，国家电子工业部与国家民委等部门联合，通过全国信息技术标准化委员会组织少数民族语言文字、信息技术和标准化三方面专家，对藏、蒙、彝、维、哈、柯、傣、朝等民族文字的信息交换用编码，相继开展国内、国际的标准化工作，并对此项工作给予专门经费投入。由于技术实现等各方面的原

因,当时的藏文字符编码出现了 2 种方案。一种是按上下叠加的整字编码,称大字符集方案。另一种是按字丁的构件编码,称小字符集方案。这种情况的出现与藏文字的结构形式息息相关。藏文是拼音文字,拼写的方向是左右上下,基字与上加字、下加字、元音上下叠加构成一个整体字丁。字丁与前加字、后加字、又后加字从左到右组合,组合成一个音节整体。如果按大字符集方案编码,技术实现相对容易,但字丁的数量无法确定,最少也要 600 个字丁,多则几千上万个字丁。若按小字符集编码,藏梵字母可以任意叠加构成想要的字符结构,符合藏文的结构特点,不需要太多的码位,但是技术实现难度较大。因此,从 1994 年到 1997 年在藏文编码方案选择中走过了艰难的历程。还有一个更重要的原因,就是当时在 ISO/IEC10646 UNICODE 基本多文种平面(BMP 平面,也就是第一个平面)已经没有足够的空间容纳上千个藏文字符,但是 BMP 平面是信息交换的基本平面,藏文必须要进入基本多文种平面。鉴于藏文编码进入 BMP 平面的重要性,也考虑到藏文字母拼写结构的特点和梵音藏文叠加结构的任意性,专家们最终选择了藏文小字符集编码作为国际标准编码方案。

1997 年在国家民委、国家技术监督局及电子工业部的直接支持和指导下,由西藏自治区藏语文工作委员会牵头,西藏大学、西藏技术监督局、西北民族大学、青海师范大学等单位共同完成了 GB16959-1997《信息技术交换用藏文编码字符集基本集》、GB/T16960.1-1997《信息技术交换用藏文编码字符集(基本集)24×48 点阵字形第一部分:白体》和 GB16959-1997《藏文字符键盘布局》等三项国家标准的研制任务。同年 7 月,由我国提交的藏文编码字符集国际标准在第 33 届 WG2 会议及 SC2 会议上经过国际标准化组织的多道严格程序和数轮投票正式获得通过,同时藏文成为我国少数民族文字中第一个正式进入 ISO/IEC10646 标准编码体系结构的文字,标志着藏文开始步入信息时代,为藏文在以计算机和网络为主体的信息化社会中实现信息处理和交换打下了坚实的基础,使这一古老的民族文字在信息时代焕发出新的活力。

2. 非国际标准编码

由于藏文标准编码选择构件式的基本集方案,当时 OpenType 字库技术还未广泛使用,加之 Windows 内核技术不开放,除微软公司以外,其他研究机构无法在 Windows 系统中植入支持藏文基本集编码的语言接口引擎(复杂文字引擎)。因此,为了在 Windows 系统中实现藏文,选择了折中的办法,即选择一些常用的叠加字丁作为字模,占用了其他字符的

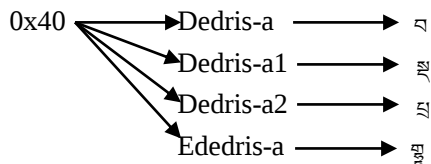
编码位置，开发了 TrueType 藏文字库。同时，根据 IME 技术编写输入法代码，就可以在 Windows 的一些通用软件中，通过选择字库的方式来实现藏文的显示和打印。这些软件都是非标准的文字处理软件，如：SAMBHOTA 1.0 和 2.0 编码、华光藏文编码、北大方正藏文编码、同元藏文编码和班智达藏文编码等。

SAMBHOTA 1.0 编码：其字符占用 ASCII 位置，实现方式是字丁与元音组合，字丁是经过上下叠加的组合，在字库中所有的字丁具有相同的宽度，元音宽度均为零，因而，静态的藏文字丁字模与元音叠加构成藏文组合，前加字和后加均按字丁对待。

如：ཨ + ཨ = ཨ, ཨ + ཨ = ཨ

SAMBHOTA 2.0 编码：SAMBHOTA 2.0 与 1.0 版码位完全相同，但比 1.0 版本增加了大量的梵文。多增加的梵文通过增加字库的方式来实现，把字库文件分为 Dedris-a 和 EDedris-a 两类，共有 30 多个字库文件。在 Word 系统中通过切换字库的方式来实现不同藏文字符的显示，字符的切换通过 Word 动态链接库 WLL 方式实现，即通过 Word 模板加载项来实现。所以该软件只能在微软公司的 Word 软件中使用。

比如：在字库中编码 0x40 的对应不同字库文件的字符：



华光藏文编码：是独立开发的藏文书版系统，字库中共有 4597 个码位，编码占用汉字 GB2312 的位置，即从 0xB0A1 开始的 4597 个码位。该编码方案所有字符都是上下叠加的字丁编码。如：ཨ(0xB6A1)、ཨ(0xB3F3)、ཨ(0xB1FB)、ཨ(0xB6AB)、ཨ(0xB4AE)等。

方正藏文编码：是北大藏文书版系统编码，藏文有 4500 个字丁，编码方式与华光藏文相同，只是码位点对应的藏文字形不同，不能与华光藏文系统兼容。

同元藏文编码：是西北民族大学同元字处理软件的藏文编码方案。共有 4100 个藏文字丁，为了在文本方式下与简体汉字保持兼容，编码选择了 GBK 的扩展区域，即汉文繁体字的编码位置，与方正和华光码位不兼容。

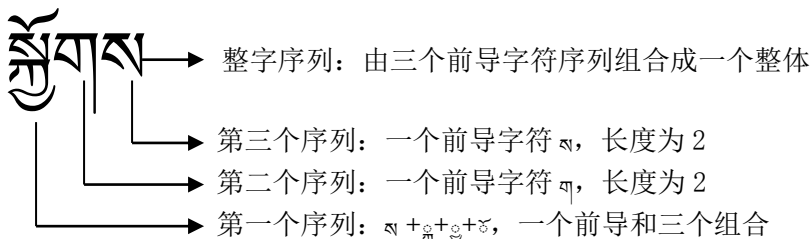
班智达藏文编码：是青海师范大学藏文字处理软件的藏文编码方案。共有 4000 多个藏梵文字丁，编码占用了 GB2312 简体汉字的位置。与华

光、方正和同元编码都不兼容。

3. 藏文国际标准编码

ISO/IEC10646《通用多八位编码字符集》藏文编码国际标准在基本多文种平面（Basic Multilingual Plane）分配给藏文的编码空间是 U+0F00-U+0x0FFF 共 256 个位置。最新 UNICODE 6.3 标准确定的藏文编码字符为 211 个。包括藏文字母 30 个、梵音藏文字母 11 个、藏文组合用字符 58 个、藏文数字符 20 个、其他图形字符 91 个、控制字符 1 个。

藏文编码国际标准基本集是小字符集，称小字符集就是与之前的大字符集相比较而言的。国际标准编码把藏文上下叠加的整字部分当做一个变长序列码，这个变长码序列的第一个辅音字母称为前导字符（head consonants），除前导辅音外的其他组合用的辅音都称为组合用辅音（sub-joined consonants）。比如：



- 变形显示字符都有独立的编码，如：

འ 作为藏文上加字编码为 U+0F62，如字符 ཀཀའའའའའའའའའའའ 等，ཀ 不变形；作为梵文字母的编码为 U+0F6A，如上加字符 ཀྐ 等，下加字符 ཀྒ 等；作为藏文下加字编码为 U+0FB2，如字符 ཀཀཀཀཀ 等。

ཡ 作为基字的编码为 U+0F61，如字符 ཡ 等；作下加字的编码的 U+0FB1，如 ཡཱ 等。

མ 作为基字的编码为 U+0F5D，如字符 མ 等；作下加字的编码的 U+0FAD，如 མཱ 等。

- 在藏文编码中，为了便于前导字符和与其对应的组合用字符间的转换，一个前导字符与其对应的组合用字符间相差 0x50。

- 梵音藏文 6 个叠字 ཀྐྐྐྐྐྐ, 具有独立的整天编码，不能上下分开输入。

- 梵音藏文的 ཀྐྐྐྐྐྐྐྐྐ 等元音分别是独立编码，在与梵文辅音组合时，元音必须作为一个整体，不能拆开来输入。

• 《藏文国际标准编码基本集》将藏文辅音字母及梵音藏文辅音字母有机地融为一体，包含了藏文文献中所有的字母和符号。前导字符、组

合用字符及元音三者结合能够表示所有的藏梵文组合结构。

其他符号的功能和说明，请看附录 1《藏文国际标准编码表》。

4. 藏文字符编码国家标准

藏文国际标准编码 ISO/IEC 10646 《信息技术通用多八位编码字符集》于 1997 年发布。按国际藏文字符编码方案，藏文属于复杂文字，字符叠加及变形显现技术给操作系统开发带来了一定难度，加之微软操作系统内核技术封闭，Linux 等其他操作系统未能广泛普及等原因，一度阻碍了藏文操作系统研发的进程。但是在藏族地区对藏文信息技术的需求刻不容缓。在这样的大环境下产生了藏文字符编码国家标准。

(1) 藏文字符编码国家标准的产生

藏族地区由于受自然条件和经济发展水平的制约，信息化技术大大落后于全国平均水平。藏文国际标准编码 ISO/IEC 10646 《信息技术 通用多八位编码字符集》的发布，迈出了藏文信息标准化建设的第一步。但是，由于微软公司视窗系统的内核技术研发上的限制和国内藏文信息处理相对落后等原因，使得 Windows、Linux 等系统的藏文国际标准编码支持模块研发一度推迟。国内很多单位、个人都在使用非标准的大字符集文字处理系统，如：方正藏文、华光藏文、同元藏文及班智达藏文等等。这些软件编码既不是国家标准，也不是行业标准，编码互不兼容，给藏族地区的信息化建设带来了很大的影响。

正是在这样的背景下，“2002 年国家民委、信息产业部、教育部、科技部、国家标准化管理委员会和全国信息技术标准化委员会等单位的有关领导和专家组成的少数民族文字软件联合调研组的调研结果显示，我国的藏文信息建设处于刚刚起步阶段，标准化建设和服务体系很不完善，藏文信息技术研发单位各自为战，缺乏沟通与合作，存在大量的重复建设，技术水平一般，藏文信息产业缺少后续发展力和竞争力，我国的藏文信息技术水平与国际先进水平存在很大差距，我国对藏文信息处理国际标准研究因遇到了技术困难而暂时处于停滞状态。”^[2] 为了解决这个问题，我国向国际标准化组织 ISO/IEC 10646 提交了“藏文大字符集方案”，但遭到了拒绝。因此，我国信息产业部电子工业标准化研究所组织西藏自治区藏语言工作委员会办公室、西藏大学、青海师范大学、西北民族大学、北京北大方正电子有限公司、潍坊北大青鸟华光科技股份有限公司、中国科学院软件研究所、民族出版社、国家民委文宣司语文室、中国标准化研究院、中央民族大学、中国民族语文翻译局的技术人员，组建了标准研制组。从 2003 年开始，经过多次专家论证和会议讨论，2006

年藏文国家标准 GB/T 20542-2006《信息技术 藏文编码字符集 扩充集 A》正式发布。之后,《信息技术藏文编码字符集 扩充集 B》也相继发布。

(2)藏文字符编码国家标准《扩充集 A》和《扩充集 B》的编码原理

根据 2003 年 12 月成都会议的决议和 2004 年的两次藏文编码标准编制工作组会议精神,确定了藏文编码国家标准制定的指导纲要:“在国际标准框架下制定藏文大字符集编码国家标准,定义垂直预组合的藏文字符,应作为我国藏文信息处理发展的策略;同时,不排斥小字符集的技术方案,并积极跟踪研究动态组合技术。”

1)扩充集 A(Extension set A)

由基本字符纵向叠加而成的结构稳定的藏文字符和最常用梵音转写字符的集合。扩充集 A 共有 1536 个垂直预组合字符,由 962 个藏文字符和 576 个常用梵音藏文字符组成。其中藏文字符包括现代藏文(三次厘定后的规范的藏文书写形式)、古藏文(规范之前藏文书写形式,主要指碑文及敦煌文献中出现的字符等)和已固化为藏文的梵音转写字符等。常用梵音藏文转写字符是以西藏收集的大字符集、藏学中心提供的高频梵音转写藏文字符和其他佛教经典中出现的常用梵音转写藏文字符为主要依据,同时参考了青海师大、西北民大提供的字频统计信息,确定了 574 个最常用梵音转写藏文字符。该字符集包含了所有现代藏文、各种符号及 98%的梵音转写藏文,能满足藏族地区各级政府及企事业单位从事藏文文字编辑的需要。

在 UNICODE 基本多文种平面中有一块系统专用保留区域,编码区间是从 U+E000 到 U+F8FF 之间,有 6400 个码位。藏文扩充集 A 的编码就放在了这块保留区域的后面,即:U+F3000-U+F8FF 间的 1536 个码位。从 UNICODE 规范讲,占用这些保留区域是不应该的。

2)扩充集 B(Extension set B)

扩充集 B 共有 5669 个垂直组合字符。以西藏收集的大字符集、藏学中心提供的梵音转写藏文字符和其他佛教经典中出现的梵音转写藏文字符为主要依据,确定了 5 669 个常用梵文字符。除扩充集 A 收录的部分字符外,其余都收录于扩充集 B 之中。

扩充集 B 在 GB 13000 专用平面 0F 平面上编码,共占用从 U-000F0000~U-000F1645 位置的 5702 个编码位置。

扩充集 A 是 GB 16959-1997《信息技术 藏文编码字符集 基本集》的扩充集;扩充集 A 和基本集能表示和交换以现代藏文为载体的所有信息,满足现代藏文信息处理的需要。扩充集 B 是基本集 GB 16959-1997 的

扩充,也是扩充集 A 的补充;基本集、扩充集 A 和扩充集 B 能表示和交换所有现代藏文和 99.99% 古藏文为载体的信息,能满足藏文信息处理的需要^[3]。

目前,在国内基于国家标准的应用软件有方正飞腾藏文排版系统,中标普华藏文办公软件、方正藏文输入法和西藏大学岗杰藏文输入法等软件。

由于藏文编码国际标准在 PC 及手机移动终端的 Windows、Linux、OS2、Android 等多种操作系统中已广泛应用,新开发的藏文应用系统绝大多数均基于国际标准编码。因此,2014 年 4 月 29 日在北京召开的藏文信息技术国家标准工作组会议上宣布,藏文编码国家标准已完成了它的过渡使命,以后不再使用。

(三)藏文操作系统

2007 年微软公司向全球市场同步发布了最新操作系统 Windows vista 和 Office2007 办公软件,全面支持藏文编码国际标准基本集,采用复杂文本引擎 Uniscribe APIs 与 OpenType 技术相结合的方式完美解决了在 Windows 系统中藏文的语言支持引擎难题。之后, Linux、MacOSX、Android 等 PC 及移动终端操作系统新版本均支持藏文处理。使得藏文能够在世界统一的多语言平台上实现信息交换与数据共享。

藏文是拼音文字,按照藏文的行文习惯,在拼读书写时,不但要从左到右横向组合,而且还要从上到下纵向叠加,才能构成一个完整的拼读音节。藏文中把这样的拼读音节叫“擦青”(ཆེན་ཁྱེད།)。藏文字丁(也叫字元)是上下纵向叠加后形成的结构,最多可以叠加 4 层,有基字、上加字、下加字和元音构成,字丁最少由一个基字构成。梵音藏文的叠加比较自由,可以任意叠加组合,为了计算机处理上的方便,系统默认最多是 5 层。“擦青”是藏文中的一个拼读单位,也是一个语素。在传统藏文书写中把“擦青”当做一个整体看待,行尾断行时“擦青”不能拆分,要么行内紧缩,要么整体换行。

按照藏文编码国际标准基本集的原理来开发藏文操作系统,则涉及两个关键技术:一个是藏文字符上下动态叠加;另一个是文本换行与光标移动。在 Windows vista 及以后的新版本系统中采用 OpenType 技术和复杂文本引擎技术(Uniscribe APIs)来解决藏文字符的组合显示。OpenType 是开放技术,而复杂文本引擎技术 Uniscribe APIs 技术是保密的。OpenType 技术解决藏文动态叠加问题,Uniscribe APIs 解决藏文的