

“十三五”国家重点出版物出版规划项目

藏文信息处理技术

西藏大学珠峰学者计划-高原学者-珠杰项目资助

ཅུས་འཁོར་ཐོག་ནས་བོད་ཡིག་རང་འགྲུལ་སྒྲིག་གཙོད་ཀྱི་ཐབས་ཤེས་ལ་དབྱད་པ།

藏文文本自动处理方法研究

珠杰 著

西南交通大学出版社

· 成 都 ·

前言

藏文作为人类语言的一个典型例子，具有人类共同的思维方式和语言组织形式，具有自身悠久的历史 and 完备的理论体系，同样受到现代科学技术进步的影响，也不断适应现代社会日新月异的变化；藏文虽然是一个小语种，同样受到自然语言处理领域研究者的关注。自计算机诞生之日起，人们就开始了藏文在计算机上的表示、显示、输入和输出的研究。目前，人们开始探索藏文自然语言处理问题，以不断提升藏文自身适应现代社会的能力。

随着藏文信息技术的不断发展，经过科研院所、高等学校和企业众多研究者的努力，藏文信息技术研究已经取得了丰硕成果，使得藏文字处理技术趋于成熟。随着互联网的普及和大数据时代的到来，藏文电子资源数据得到了迅速增长，这些数据成为藏文信息处理进一步发展的基石。由此，研究人员广泛开展藏文字处理、词处理、短语处理和语句处理等相关研究工作。目前，在藏文字处理、词处理、短语处理和语句处理等领域上取得了不少成绩，但也存在很多尚未解决的问题。本书从目前亟待解决的几个关键问题出发，研究其解决方案和相应的实现算法，这也是本人从事藏文信息处理技术研究的相关成果，大部分成果已经发表在国内中文核心期刊上。在本书编写过程中，作者得到了多方的大力支持。在此，感谢我的导师李天瑞教授，西藏大学欧珠教授、格桑多吉教授、仁青诺布副教授等；感谢我的学生郑亚楠的辛勤努力。另外，本书还得到了“西藏大学珠峰学者计划-高原学者-珠杰”项目、国家自然科学基金项目（61751216）的资助。

本书总共分三个部分，第一部分以藏文字处理为研究对象，讨论了藏文排序方法、藏文音节规则和自动拼写算法的内容；第二部分以藏文词处理为研究对象，讨论了藏文停用词自动处理方法、藏文人名识别方法，研究了条件随机场（CRF）和深度学习的藏文人名识别技术；第三部分以藏文自动校对为研究对象，讨论了

基于音节规则的藏文拼写检查算法、藏文自动校对系统框架和接续关系检查算法。

本书可以作为高等院校藏文信息处理技术、计算机科学与技术、藏语言文学等相关专业研究生的参考书，也可以作为从事藏文信息处理技术、藏语计算语言学、藏语言文学研究相关人员的参考书。

由于本人水平有限，加之时间仓促，书中难免存在疏漏和不妥之处，恳请广大读者批评指正。

珠 杰

2018年3月

目 录

第一篇 藏文字处理技术

第一章 藏文基础理论	3
第一节 藏文字符	3
第二节 藏文音节	3
第三节 藏文词语	4
第四节 藏语句子	5
第二章 基于藏文编码 GB 的藏文排序方法研究	7
第一节 概 述	7
第二节 藏文字排序规则	7
第三节 藏文字排序算法	8
第四节 结 论	11
第三章 藏文音节规则库的建立与应用分析	12
第一节 概 述	12
第二节 藏文的结构	12
第三节 藏文规则库的建立	13
第四节 规则库的应用	23
第五节 结 论	26
第四章 藏文音节规则模型及应用	28
第一节 概 述	28
第二节 藏文音节结构	28
第三节 藏文音节规则模型	30
第四节 规则方法的应用研究	34
第五节 结论与展望	37

第二篇 藏文词处理方法

第五章 藏文停用词选取与自动处理方法研究	41
第一节 概 述	41
第二节 相关研究工作	41
第三节 藏文停用词选取方法	42
第四节 停用词处理实验	44
第五节 结论与展望	51

第六章 基于词向量的藏文词性标注方法研究	53
第一节 概 述	53
第二节 相关研究工作	53
第三节 词性标注算法	54
第四节 实验及数据分析	59
第五节 结论与展望	62
第七章 基于条件随机场的藏文人名识别技术研究	64
第一节 概 述	64
第二节 相关研究工作	65
第三节 藏文人名特征	66
第四节 模型及人名识别	67
第五节 实 验	72
第六节 结论与展望	77
第八章 基于深度学习模型的藏文人名识别方法研究	80
第一节 概 述	80
第二节 相关研究工作	80
第三节 深度学习模型	81
第四节 训练模型	84
第五节 词向量训练	85
第六节 实验及数据分析	88

第三篇 藏文自动校对方法

第九章 TSRM 藏文拼写检查算法	97
第一节 概 述	97
第二节 藏文音节规则模型	98
第三节 拼写检查算法	100
第四节 实 验	102
第五节 结论与展望	105
第十章 藏文文本自动校对方法及系统设计	107
第一节 概 述	107
第二节 藏文文本自动校对系统	107
第三节 藏文文本校对方法	111
第四节 结论与展望	116
第十一章 文本自动处理技术比较	118
第一节 字层面的处理	118
第二节 词层面的处理	119
第三节 序列标注层面的处理	119
第四节 句子层面的处理	120

第一篇

藏文字处理技术

从 20 世纪 80 年代初于道泉先生的论文《藏文数码代字》开始，藏文信息处理技术研究的大门被打开了，从此，许多研究人员致力于藏文信息处理技术的研究。藏文字处理技术包含藏字构件的属性统计、藏文编码体系研究、编码制定、编码转换、输入输出、字型与字库设计、藏文排序等众多内容，历经 30 多年的努力，藏文信息处理技术已经进入藏文字处理技术逐步成熟、词处理技术蓬勃发展、应用技术逐步完善的阶段。本篇内容由四章构成。第一章是藏文基础理论，后三章分别介绍了基于藏文编码 GB 的藏文排序方法研究，藏文音节规则库的建立与应用分析，藏文音节规则模型及应用。主要讨论了藏文字符排序问题，提出了藏文基字定位算法；讨论了藏文音节规则模型，建立了藏文音节规则库和应用范围，提出了藏文音节自动拼写算法。

第一章 藏文基础理论

传统的藏文语法主要由“文法根本三十颂”(简称“三十颂”)和“字性组织法”组成^[1, 2]。早在 1300 多年前吐蕃王朝的松赞干布时期,由文臣图弥桑布扎始创,经过噶玛司都·却吉迥乃、欧曲·法贤、欧曲·央金珠白多吉等历代语言学家的注释和扩充,形成了以虚词和动词为核心,具有显性语言组织结构特征的传统文法体系^[3]。其中“三十颂”主要描述了藏文音节拼写结构、格助词和各类虚词的用法;“字性组织法”描述了以动词为中心的形态变化、时态变化、施受关系、能所关系、单句句式 and 复句句式等有关动词的一系列语法规则和功能。

第一节 藏文字符

按语法理论，藏文有显性的三十个字母和四个元音，但在藏文字符国家标准^[4]中，藏文字符有 211 个，包含了语法理论中的字母、藏文数字以及藏文文本中出现的各类符号。

三十个字母：ཀ་ཁ་ག་ང་ ཅ་ཆ་ཇ་ཉ་ ཏ་ཐ་ད་ན་ བ་མ་མ་མ་ ཙ་ཚ་ཛ་ལ་ ཞ་ཟ་འ་ཡ་ ར་ལ་ཤ་ས་
ཏ་ཨ་

四个元音：ḥi·ḥu·ḥe·ḥo·

藏文数字：༠ ༡ ༢ ༣ ༤ ༥ ༦ ༧ ༨ ༩ ཁ ག ཇ ཉ ཊ ཋ ཌ ཌྷ ཎ ཏ ཐ ད དྷ

部分符号：|| ◌ x x x ॐ ॐ ॐ ॐ ॐ ॐ ॐ

第二节 藏文音节

藏文音节（有些文献称为藏字）是由藏文三十个字母和元音组成的字符序列。藏文音节结构是二维的字符序列组织结构，如图 1-1 所示。根据音节中基字为中心的位置关系，分为前加字、基字、后加字、再后加字、上加字、下加字和元音。

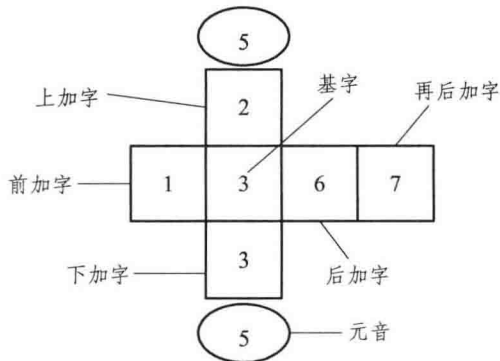


图 1-1 藏文音节结构

藏文音节的拼写规则主要依据藏文“字性组织法”中的音韵理论，将基字、元音、前加字和后加字分为阳性、中性和阴性，阴性又细分为准阴性、纯阴性和极阴性。字符的音韵分类如表 1-1 ~ 1-3 所示。

表 1-1 基字字性

字性	阳性	中性	阴性			无性
			准阴性	纯阴性	极阴性	
基字	ཀ་ཅ་ཉ་ས་ཙ་	ཁ་ཆ་ཐ་ས་ཨ་	ག་ང་ད་བ་ཇ་མ་ཞ་ཟ་འ་ཡ་ག་ས་	ར་ལ་ཉ་ཙ་	ང་ཉ་ན་མ་	ཨ་

表 1-2 前加字字性

字性	阳性	中性	阴性	极阴性
前加字	བ་	ག་ད་	འ་	མ་

表 1-3 后加字字性

字性	阳性	中性	阴性
后加字	ག་ད་བ་ས་	ན་ར་ལ་	ང་མ་འ་

第三节 藏文词语

一个及以上藏文音节构成词语（ཡི་གེའི་ཁོངས་ནས་མིང་དབྱེད་ཟླ།）。传统藏文语法中，关于词语讨论最丰富的应属动词、格助词和各类虚词。“三十颂”中分析了格助词和各类虚词的功能和用法，“字性组织法”中分析了动词的形态变化、功能和意义。在“三十颂”中，对格助词和虚词的分析是从不同视角讨论虚词的两种表象，即“一套是从形式到意义，另一套是从意义到形式”^[3]。

格助词的分析是从意义到形式的论述，各类虚词的分析是从形式到意义的论述。藏语“格”从语法意义来分远不止八“格”（传统语法分为八“格”），如果从语法形式来分，却只有属格、作格、位格和从格。这些形式不仅表示词与词、句子成分之间的结构关系，还可以附加在动词谓语之后，用来表示分句之间的关联关系，具有“格”和虚词的双重功能^[3]。从形态上划分的藏语格助词如表 1-4 所示。

表 1-4 藏语格助词表

格助词	属格助词	作格助词	位格助词	从格助词
意义格	领属格	施动格	受动格、目的格、处所格	来源格
词数	5	5	7	2
形态变化	གི་གྱི་གྱི་འི་ཡི་	གིས་གྱིས་གྱིས་འིས་ཡིས་	སུ་ར་ར་རུ་ན་ལ་ཏུ་	ནས་ལས་

藏语虚词分为自由虚词和不自由虚词。不自由虚词是指有多种形态，受接续规则约束的一类虚词^[3]；自由虚词正好相反，是不受接续规则约束的一类虚词。不自由虚词共有 9 种，除了属格助词、作格助词和位格助词之外，其余如表 1-5 所示。

表 1-5 藏语不自由虚词表

虚词	饰集词	待述词	离合词	终结词	ཞིང་等	གྱིན་等
个数	3	3	10	10	15	4
形态变化	གྲང་ཡང་ འང་	ཏི་ཉི་དེ་	གམ་ངམ་དམ་ནམ་ བམ་མམ་འམ་རམ་ ལམ་སམ་ཏམ་	གོ་ངོ་ནོ་པོ་མོ་འོ་ རོ་ལོ་སོ་ཏོ་	ཞིང་ཞིས་ཞིའོ་ཞི་ན་ཞིག་ ཅིང་ཅིས་ཅིའོ་ཅི་ན་ཅིག་ ཤིང་ཤིས་ཤིའོ་ཤི་ན་ཤིག་	གྱིན་གྱིན་གིན་ ཡིན་

接续规则是指不自由虚词的接续受前一个音节后置字母的语音强弱影响的一种语法规则。自由虚词则不受接续规则的约束。在传统文法“三十颂”中，提到了 6 种自由虚词，具体如表 1-6 所示。

表 1-6 藏文自由虚词

虚词	ནི་སྒྲ་	དང་སྒྲ་	དེ་སྒྲ་	ཟྱི་སྒྲ་	དགག་སྒྲ་	བདག་སྒྲ་
功能	语气助词	连词	指示代词	疑问代词	否定词	指人构词后置
形态	ནི་	དང་	དེ་	ཅི་ཇི་སྟུ་གང་	མ་མི་མིན་མེད་	པ་བ་མ་པོ་པོ་མོ་མཁན་ཅན་ལྟན་

虚词和动词是传统藏文文法的两根支柱，动词作为“字性组织法”的核心，词法和句法都受制于动词理论，在整个文法体系中占有支配地位^[3]。动词的分类有三种形式：按是否带有宾语，分为及物动词和不及物动词。能带宾语的动词是及物动词，不能带宾语的是不及物动词。按主观能动性，分为自主动词与不自主动词^[3]。动作行为能够由行为主体自由支配的动词是自主动词，不能由行为主体自由支配的动词是不自主动词。按与其他助词的组合形式，分为使动式和兼语式。与英语类似，藏文动词也有时态的划分，分为过去时、现在时、将来时和命令式，动词一般有少于四种的形态变化。

为了适应藏文信息处理的发展，与传统文法不同，已有不少研究成果在讨论藏文词类和词性的问题；同时借鉴英文和汉文的词性分类方法，研究名词、形容词等更多词性内容。

第四节 藏语句子

藏语句子是多个词语或短语组成的字符序列（མིང་གི་ཁྲོད་ས་ནས་ཆོག་ཕུང་ནས། ཆོག་གིས་དོན་ནམས་སྟོན་པར་བྱེད།）。与英语和汉语不同，藏语句子是 SOV 型，是动词后置的语言，谓语成分作为句子的核心，在传统文法中有充分的论述。藏文句子一般分为单句和复句，单句又分陈述句和祈使句，每个句式还可以进行更多的细分；复句也分为联合复句和偏正复句。联合复句可以细分为并列、连贯、选择、对比和递进五种关系，偏正复句可以细分为因果、转折、条件、假设和目的五种关系^[3]。

参考文献

- [1] 土弥三菩札. 藏文文法四种合编[M]. 北京: 民族出版社, 2000.
- [2] 嘎玛司都. 四都文法详解[M]. 西宁: 青海民族出版社, 2003.
- [3] 格桑居冕, 格桑央京. 实用藏文文法教程[M]. 成都: 四川民族出版社, 2004.
- [4] 国家技术监督局. 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, 1997.

第二章 基于藏文编码 GB 的藏文排序方法研究

藏文排序在字、词典排序，计算机中藏文排序等方面有着广泛的应用，本章主要根据藏文编码国家标准（GB）的整字编码方案，研究藏文的排序问题。通过藏文结构的线性化处理，提出基于藏文编码国家标准的基字定位算法和排序算法，并将其应用于藏文电子词典的排序中。

第一节 概 述

藏文拼写是既有横向组合又有纵向组合的非线性结构，是一种二维的独特的文字构成形式。不少专家对藏文字的字典序性进行过深入的研究，并对其在计算机中的实现进行过有益的探讨，为藏文在计算机中的处理奠定了良好的理论基础。藏文音节构件（构造级^[1]）排序顺序^[2]：基字→前加字→上加字→下加字→元音字符→后加字→再后加字，这符合藏文语法的排序规则。根据藏文编码国家标准，本章通过藏文非线性的结构变成线性化的思想，使其按线性规则进行比较、排序。

第二节 藏文字排序规则

一、一级排序规则

在藏文排序时，总体上需要遵守的排序规则定为一级排序规则，其基本原则就是藏文编码国家标准基本集中的 0F40 到 0F68 字符序列，符合传统约定藏文字母次序排列规则，即基本辅音序^[1]，还包含从梵文转写来的基本的几个叠置字符，如表 2-1 所示。

表 2-1 藏文一级排序规则

字符	ཀ	ཁ	ག	གྲ	ང	ཅ	ཆ	ཇ	ཉ	ཊ	ཋ
编码	0F40	0F41	0F42	0F43	0F44	0F45	0F46	0F47	0F49	0F4A	0F4B
字符	ཌ	ཌྲ	ཎ	ཏ	ཐ	ཏྲ	ཏླ	ཏྵ	ཏྶ	ཏྷ	ཏྸ
编码	0F4C	0F4D	0F4E	0F4F	0F50	0F51	0F52	0F53	0F54	0F55	0F56
字符	ཏྐྵ	ཏྺ	ཏྻ	ཏྼ	ཏ྽	ཏ྿	ཏྺྲ	ཏྺླ	ཏྺྴ	ཏྺྵ	ཏྺྶ
编码	0F57	0F58	0F59	0F5A	0F5B	0F5C	0F5D	0F5E	0F5F	0F60	0F61
字符	ཏྺྷ	ཏྺྸ	ཏྺྐྵ	ཏྺྺ	ཏྺྻ	ཏྺྼ	ཏྺ྽				
编码	0F62	0F63	0F64	0F65	0F66	0F67	0F68				

四、基字定位算法

(1) 判断藏文单音节是否是一个音节字符组成。如果是, 则该字符为基字; 否则, 转到第 2 步。

(2) 判断藏文单音节的一个音节中, 第一个字符是否包含了 TPSET 集合中的字符。如果否的话, 判定该字符为基字; 否则, 转到第 3 步。

(3) 判断藏文单音节是否由两个音节字符组成。如果是, 则第一个字符为基字; 否则, 转到第 4 步。

(4) 这里第一个字符可能是前加字, 也可能是基字。可以通过第二个字符来判定, 所以应查找第二个字符。根据藏文的语法规则, 判断第二个字符是否为叠加字符。如果是, 第二个字符为基字, 语法规则说明 1^[6]和说明 2^[7]如图 2-1 和图 2-2 所示; 否则, 转到第 5 步。

མོན་འཇུག་པོ་ཉི་པ་ཡིག་ནི།
བཅེགས་པོ་རྒྱ་ཆ་མུ་དང་། ཉ་བཅེགས་གསུམ་བཅས་བདུན་ལ་འབྲུལ།
མོ་ཡིག་ར་མགོ་ས་མགོ་ཡི། ག་ང་ཉ་ན་བཞི་གཉིས་བརྒྱད།
རྩ་རྩ་ཉ་བཅེགས་གསུམ་དང་སྒྲ། བརྒྱ་བཞི་མོན་རུ་འཇུག་པ་འོ།

མོན་འཇུག་པ་ནི་མགོ་ཅན་གྱི་ཡི་གེ་ཉི་པ་གཅིག་ལ་འཇུག

图 2-1 语法规则说明 1

པོ་ནི་མ་ནིང་ལ་མི་འཇུག མོ་ནི་པོ་ལ་འཇུག་པ་མིན།
མ་ནིང་རང་ལ་རང་མི་འཇུག རང་གི་ཕྱེ་དང་ཕྱད་པ་ནི།
པོ་ཡང་མོ་ལ་འཇུག་མི་འགྱུར། མ་ནིང་མོ་ཡང་དེ་བཞིན་ནོ།

图 2-2 语法规则说明 2

(5) 藏文前加字文法^[7]有一定的匹配规则, 语法规则说明 3 如图 2-3 所示。

པོ་ནི་པོ་དང་མོ་ལ་འཇུག མོ་ནི་མོ་དང་མ་ནིང་ལ།
མ་ནིང་ཡང་ནི་པོ་མོ་ལ་འོ། ཤིན་རུ་མོ་ནི་མ་ནིང་དང་།
མོ་དང་ཤིན་རུ་མོ་ཉིད་ལ་འོ།

图 2-3 语法规则说明 3

根据该规则, 可以得到藏文文法匹配表, 表中的左边一列是前加字, 右边一列是相应基字匹配字符, 如表 2-3 所示。

表 2-3 前加字匹配表

前加字	匹配基字
ག	ཅ་ ཉ་ ཅ་ ཉ་ ད་ ན་ ར་ ཟ་ ཡ་ ཤ་ ས་
ད་	ཀ་ པ་ ག་ ང་ བ་ མ་
བ་	ཀ་ ཅ་ ཉ་ ཅ་ ག་ ང་ ར་ ཉ་ ད་ ན་ ར་ ར་ ཟ་ ར་ ཤ་ ས་
མ	ག་ ར་ ད་ ར་ ང་ ཉ་ ན་ མ་
འ	ག་ ར་ ད་ བ་ ར་ ལ་ ཆ་ ཐ་ པ་ བ་ ཆ་

根据文法规则说明 3 和说明 4, 如图 2-3 和 2-4 所示, 如果有两个以上字符, 并具有如下文法特征, 可以进行以下判断。

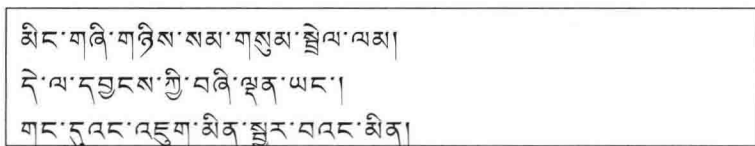


图 2-4 文法规则说明 4

① 当第一个字符是ག, 第二个字符判定为ཅ་ཏ་ཙ་ཉ་ཞ་ཟ་ཡ་ཤ单音基字时, 可以直接确定ག是前加字, 第二个字符为基字。

② 当第一个字符是ད, 第二个字符判定为ཀ་པ单音基字时, 可以直接确定ད是前加字, 第二个字符为基字。

③ 当第一个字符是བ, 第二个字符判定为ཀ་ཅ་ཏ་ཙ་ཇ་ཉ་ེ་ཞ་ཟ་ཤ单音基字时, 可以直接确定བ是前加字, 第二个字符为基字。

④ 当第一个字符是མ, 第二个字符判定为ཇ་ེ་ཉ单音基字时, 可以直接确定མ是前加字, 第二个字符为基字。

⑤ 当第一个字符是འ, 第二个字符判定为ཇ་ེ་ཁ་ཆ་ཐ་པ་ཨ་མ་མ单音基字时, 可以直接确定འ是前加字, 第二个字符为基字; 否则转到第 6 步。

(6) 如果是三个以上字符, 判断第三个字符是否为{ད, ས}^[7]的其中之一。如果是, 则第一个字符必定是基字; 否则为语法错误。文法规则说明 5 如图 2-5 所示。

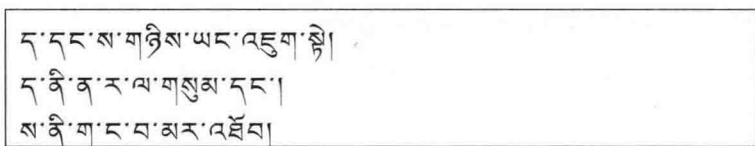


图 2-5 文法规则说明 5

五、排序算法

根据线性化的构造级: 基字(字丁)→前加字→后加字→再后加字, 可以建立如下排序算法。

(1) 通过基字定位算法确定基字, 然后依据藏文文法规则确定的字丁集合 TBSET 和一级排序规则, 判定基字属于哪个“系”。

① 先在基本集合中查找, 若未找到, 按顺序在扩充集 A 和扩充集 B 中查找。示例如图 2-6 所示。即确定属于哪个“系”级, 然后在相应的基本集, 扩充集 A、B 中查找。

② 在同一“系”级中, 先查找基本集, 再查找扩充集 A, 最后查找扩充集 B。同一级别中, 具有多个基字的时候, 排序按照编码的顺序进行。

(2) 当基字相同时, 查看前加字。排序过程根据 TPSET 的字符顺序进行。

(3) 当前加字相同时, 查看后加字, 排序过程根据 TSSET 的字符顺序进行。

(4) 当后加字相同时, 查看再后加字, 排序过程根据 TSSET1 的字符顺序进行。

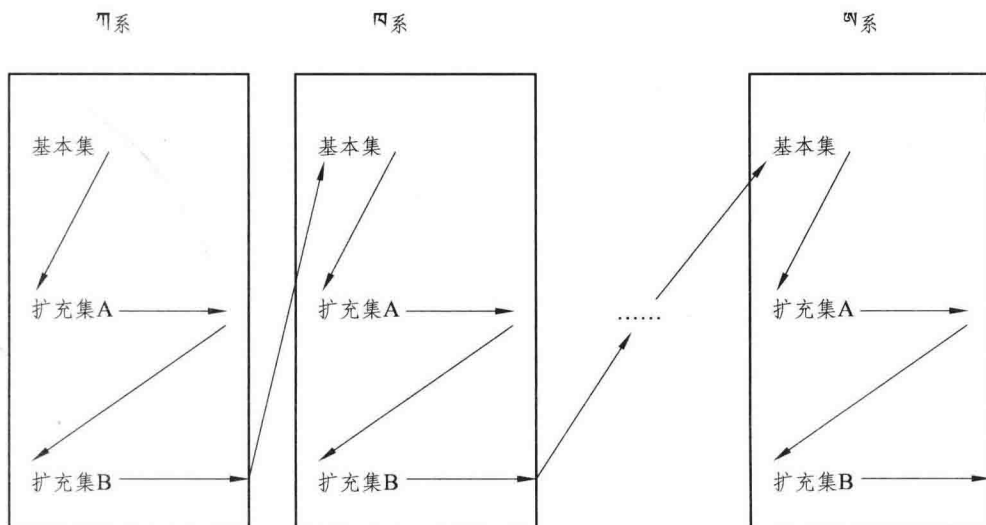


图 2-6 藏文排序过程图

第四节 结 论

本章的排序顺序和排序规则，依据藏文编码国家标准和藏文的文法特点，具有很强的易实现特点，已在藏文电子词典中得到实现。

参考文献

- [1] 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.
- [2] 扎西次仁. 藏文的排序规则及其计算机自动排序的实现[J]. 中国藏学，1999（4）：128-134.
- [3] 林河水，程伟，曹晖，等. 一种符合 ISO14651 语义的藏文排序实现方法[J]. 中文信息学报，2004，05（18）：35-41.
- [4] 西藏自治区藏语言工作委员会办公室，等. 信息技术 藏文编码字符集（基本集及扩充集 A）[S]. 北京：中国标准出版社，2009.
- [5] ISO/IEC.ISO/IEC FCD14651-International String Ordering-Method for Comparing Character and Description of the Common Template Tailorable Ordering[S].1997.
- [6] དཔལ་རིས་མངས་ཀྱིས་བཅུམས་ བརྒྱུད་གསལ་བྱེད་པའི་སློན་མཆོ་སློན་མི་རིགས་དཔེ་སློན་ཁང་ནས་
- [7] ཐོན་མི་མཆོ་ཏྲ་དང་དབུལ་ཁྱེད་ཀྱིས་བཅུམས་ ལུམ་ཉགས་ཅ་འབྲེལ་མི་ཏུའི་ཞལ་ལྷུང་[M]. མི་རིགས་དཔེ་སློན་ཁང་ནས་

第三章 藏文音节规则库的建立与应用分析

藏文音节具有独特的构造方法，不同的构造位上有不同的藏文字符，根据不同的组合，构成了千变万化的藏文音节。由于字符的语音特性，藏文组合形式上有很多的限制。本章借助藏文文法规则和藏汉大词典，建立了现代藏文音节规则库，并分析了可能的应用领域。

第一节 概 述

随着现代信息技术的发展和互联网的普及，藏文信息处理技术有了较快的进步。从藏文的属性统计工作开始^[1]，许多专家通过几十年的努力，从多格局编码状况^[2]到统一编码的时代^[3]，从多键盘布局的设计到统一键盘布局的出台^[4]，解决了藏文在计算机中的输入、输出问题，并在现代互联网上实现了藏文信息的共享。现在不少高校和科研机构，在前人的研究基础上不断探索，开始在藏文语音识别^[5]、文字识别^[6, 7]、分词、词类标记^[8]、机器翻译^[9]等领域着手研究，并取得了一些研究进展。

随着藏文信息处理技术的进一步发展，藏文文本处理成为藏语自然语言处理的研究内容。藏文音节作为文本组成的重要成分，对其进行分析是一个基础性工作。从书面藏文的信源属性来看，藏文文本中的音节有 72% 的冗余度，这说明 3/4 的藏文字母是保证依据语法规则来组合藏文音节的，只有 28% 是可自由选择的^[10]。根据此特点，本章以“预组合”的形式建立一个规则库，并分析了在藏文信息处理研究领域中的应用的可能性。

第二节 藏文的结构

藏文音节结构是以基字为核心，既有横向拼写又有纵向拼写。前加字、基字、后加字、再后加字是横向拼写；上加字、基字、下加字和元音符是纵向拼写。藏文音节结构十分复杂，字符在音节中的特定位置可以称为“构造位”。根据藏文的文法，各个构造位上出现的字符的性质与数量均有一定限制，相互之间也形成一种约束关系。

不包括梵音转写藏文，藏文音节结构中的构造位共有 7 个，如图 3-1 所示。

每个构造位在藏文字中的表示为：1 是前加位，2 是上加位，3 是基字位，4 是下加位，5 是元音位，6 是后加位，7 是再后加位，分别由前加字、上加字、基字、下加字、元音、后加字、再后加字来表示，在字中的位置如图 3-2 所示。