北京大学学报(自然科学版)
Acta Scientiarum Naturalium Universitatis Pekinensis

藏文音节规则模型及应用

珠杰[1, 2, †]　李天瑞[1]　格桑多吉[2]　仁青诺布[2]　乔少杰[1]

1 西南交通大学信息科学与技术学院，成都 610031; 2 西藏大学计算机科学系，藏文信息工程研究中心，拉萨 850000;
† E-mail: trocky.jie@gmail.com

摘要　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。根据藏文字符不同的组合，构成了千变万化的藏文音节。由于字母的语音特性，藏文组合形式上有诸多限制。作者以藏文音节为研究对象，借助藏文语法规则，建立了现代藏文音节的简化模型和相应的规则库，介绍了其应用领域，提出了一种基于音节模型的的藏文音节自动拼写算法，实验验证了规则方法的有效性。

关键词　藏文音节；藏文规则；规则库；音节拼写

中图分类号　TP391

Tibetan Syllable Rule Model and Applications

ZHU Jie[1,2, †], LI Tianrui[1], GE Sangduoji[2], REN Qingnuobu[2], QIAO Shaojie[1]

1. School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031;
2. Department of Computer Science, Tibetan University, Lhasa 850000; † E-mail: trocky.jie@gmail.com

Abstract　Tibetan syllable has a unique configuration approach. Different positions of configuration compose different Tibetan characters. Based on the combination of Tibetan characters, there are a large number of ever-changing Tibetan syllables. Because of the pronunciation features of the letters, there are several limitations of the styles for Tibetan combination. This paper focuses on the study of Tibetan characters. A simplified model for modern Tibetan syllables and its corresponding rule base are established by using Tibetan grammar rules. Its applications are extensively analyzed. Algorithms for automatic spelling of Tibetan syllable are proposed as well. Experiments and case studies validate the rule base of Tibetan.

Key words　Tibetan syllable; rule of Tibetan; rule base; syllable spelling

藏文音节作为构词的成分，有其自身的特征，特别是字母组合上有很多拼写规则。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，只有 28%是可自由选择的[1]，这说明藏文音节中 3/4 的藏文音节是保证依据语法规则来拼写的。在藏文信息处理的应用上，藏文规则在藏文排序中有重要的应用价值。江荻等[2]利用藏文语法提出了构造序、构造级(拼写序)和字符序相结合的排序算法，建立藏文排序模型，为藏文排序在计算机中的实现提供了理论基础。Chilton 等[3]利用藏文规则，对藏文编码国际标准 ISO/IEC 10646 音节进行了排序，通过“collation element”的概念，建立一个“collation element”表，该表通过对藏文规则建立权重分级来解决藏文的排序问题，虽然需要排序的“字符”数量多了许多，但是算法简单并易于实现，在 Mysql、MIMER SQL 和 OpenOffice 2.0 等系统得到了成功应用，但是该方法没有从藏文规则模型的角度来进行讨论。本文依据藏文语法探讨了藏文规则的数学模型，并建立藏文规则库，然后将其应用到藏文音节自动拼写和拼写检查等领域中，说明由规则模型建立的藏文规则方法能够有效解决藏文信息处理研究中的若干基础性问题。

国家自然基金(61165013, 60763010, 61100045, 61262058)和西藏自治区自然科学基金(2010 年 41 号)资助
收稿日期: 2012–06–05; 修回日期: 2012–10–01; 网络出版时间: 2012-10-26 17:54
网络出版地址: http://www.cnki.net/kcms/detail/11.2442.N.20121026.1754.016.html

1 藏文音节结构

藏文音节结构以基字为核心，既有横向拼写又有纵向拼写，前加字、基字、后加字和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，因此具有十分复杂的音节结构。字符在音节中的特定位置可以称为“构造位”。根据藏文的语法，各个构造位上出现的字符的性质与数量均有一定的限制，相互之间形成一种约束关系。

藏文音节中不包括梵音转写藏文，藏文音节的基本结构中构造位共有 7 个，如图 1 所示。

定义 1[4] 构造位上的字符称为构件，根据不同位置分别称为前加字、上加字、基字、下加字、元音、后加字和再后加字，如图 2 所示，我们称之为藏文音节模型–1(简称模型–1)。

每个构造位在藏文音节中的表示为：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、后加字和再后加字来表示在字中的位置。

定义 2[5] 在一个音节中的纵向单位(上下叠加的组合体)叫字丁或叠加字符。例如：“ཀྲ”，“སྒྲ”称为叠加字符。

根据藏文语法，有 5 个前加字，分别是ག, ད, བ, མ和འ; 3 个上加字，分别是ར་མཚ, ལ་མཚ和ས་མཚ; 30 个基字是 30 个藏文字母; 4 个下加字，分别是ཡ་བཏགས, ར་བཏགས, ལ་བཏགས和ཝ་བཏགས; 5 个元音符号中，有 3 个是上元音符号，分别是ཨི, ཨེ, ཨོ, 1 个下元音符号ཨུ，一个隐含元音ཨ; 10 个后加字，分别是ག, ང, ད, ན, བ, མ, འ, ར, ལ, ས; 2 个再后加字，分别是ད, ས。结合模型-1，对每一个构造位

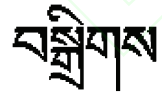

图 1 藏文音节的基本结构
Fig. 1 Structure of Tibetan syllable

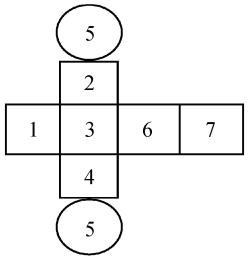

图 2 构造位及模型-1
Fig. 2 Modle-1 and position of construct

的元素集合描述如下。

设基字集合用 B 来表示:

B ={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}。

设前加字集合用 Pr 来表示:

Pr={ག,ད,བ,མ,འ}。

设上加字集合用 U 来表示:

U={ར་མཚ, ལ་མཚ,ས་མཚ}。

设下加字集合用 D 来表示:

D={ྱ,ྲ,ླ,ྭ}。

设后加字集合用 S 来表示:

S={ག,ང,ད,ན,བ,མ,འ,ར,ལ,ས}。

设再后加字集合用 SS 来表示:

SS={ད, ས}。

设元音字符集合用 Tvowel 来表示:

Tvowel={∅, ི, ེ, ོ, ུ}，其中∅表示隐含元音字符。

2 藏文音节规则模型

2.1 藏文音节模型的建立与简化

模型ī 1 是根据藏文的音节结构建立的一个模型，该模型中以基字为核心，在元音和后加字的作用下构成一个音节，在实际写法中除了构造位 3 不能空之外，其余位置均可以为空。当一个基字构成音节时，该音节隐含了元音“མ”和后加字“འ”。模型–1 的笛卡尔积如下式:

$$\mathrm{Pr}\times U\times B\times D\times \mathrm{Tvowel}\times S\times SS=$$
$$\{<p,u,b,d,v,\mathrm{s},\mathrm{ss}>|p\in \mathrm{Pr},\ u\in U,\ b\in B,$$
$$v\in \mathrm{Tvowel},\ s\in S,\ \mathrm{ss}\in \mathrm{SS}\}$$

藏语的语音理论体系中将藏语语音分为元音和辅音。根据藏文的语音特性，对于 30 个辅音字母进行了字性分类，为阳性、中性和阴性 3 种，其中阴性包括准阴性、极阴性和纯阴性 3 种，共计 5 种分类。辅音字母中提取出来的前加字、后加字构件也进行了上述 5 种的分类。根据每个构件的发音特性，字母组合上有很多拼写限制。从上节中的描述可知，模型-1 中的各个构造位的元素是有限的，可以看出经过适当的迭代，藏文全部音节能够被拼写出来，但也产生了很多不符合语法的冗余音节。为了不产生冗余音节，我们通过预组合方式对模型-1 进行简化。

假设 1 根据上加字与基字、下加字与基字、上加字+基字+下加字的固定组合关系，模型-1 中的

构造位 2、3、4 合成为一组。

假设 2 再后加字部分归并到构造位 6 上, 根据后加字与再后加字之间的组合关系, 例如: “གས”,“ནད”的形式归并到后加字集合中。

假设 3 扩展基字的组成成分, 在 30 个辅音字母的基础上, 增加上加字+基字的组合、基字+下加字的组合、上加字+基字+下加字的组合。

假设 4 根据语音强弱关系来决定前加字与扩展基字之间的组合关系。由于上加字、下加字没有语音的强弱特性, 不分阴阳关系, 与基字组合(假设 3 的扩展内容)由基字的语音强弱来决定其阴阳类别。具体组合关系为阳性与阳性、阳性与阴性组合; 阴性与阴性、阴性与中性组合; 中性与阳性、中性与阴性组合; 中性与阳性、中性与阴性组合; 极阴性与中性、极阴性与阴性、极阴性与极阴性组合。

根据假设 1 和 2, 简化的藏文音节模型-2(简称为模型-2)如图 3 所示。模型-2 中, 上加字+基字的组合、基字+下加字的组合、上加字+基字+下加字的组合形式合成到构造位 2 上, 成为基字的一部分, 即模型-1 中的构造位 2、3、4 合成到模型-2 上的构造位 2 上。再后加字归并到后加字集合中, 成为模型-2 的构造位 4, 成为后缀部分。这样, 模型-2 中的构造位 1 是前加位、2 是扩展基字位、3 是元音位、4 是后缀位。

根据假设 3, 设扩展基字集合为 Con, 遵循语法的阴阳关系对扩展基字进行分类, 其元素定义如下:

$$\mathrm{Con}=p\cup \mathrm{ne}\cup \mathrm{pn}\cup \mathrm{qn}\cup \mathrm{vn},$$

其中, p 为阳性、ne 为中性、pn 为纯阴性、qn 为准阴性、vn 为极阴性, 分别取值如下。

p={ཀ,ཅ,ཏ,ད,ཙ,ཀ,ཁ,ཤ,ཀྱ,ཀྲ,གྲ,ཏྲ,གྱ,སྐ,སྟ,སྤ,ཀ,ཤ,'ཤ,ཁ,ཐ,ཏ,ཙ,སྟ}。

ne={ཁ,ཆ,ཐ,ཚ,ཚ,ཁྲ,ཁྱ,ཐྱ}。

pn={ག,ང,ད,ན,ཉ,ཞ,ཟ,འ,ཡ,ར,ལ,ས,ཀྱ,ཏྲ,གྲ,ཤྲ,ཀྲ,ཏྲ,ཤྲ,ཁྲ,ཁྱ,ཐྱ,ཉ,ཏྲ,ཏྱ,ཐྱ,ཁྲ,ཁྱ,ཐྱ}。

qn={ར,ལ,ཤ,ཚ,ཟ}。

vn={ང,ཨ,ན,ཤ,ཉ,ཐ,ཏྲ,ཤྲ,ཏྱ,ཐྱ,ཚ,ཁྱ,ཐྱ}。

设归并后的后缀字符集合为 Tpostfix

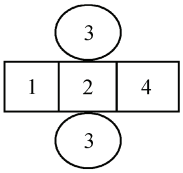

图 3 藏文音节模型-2

Fig. 3 Model-2 of Tibetan syllable

Tpostfix={ག,ང,ད,ན,བ,མ,ཨ,ར,ལ,ས,གས,ངས,བས,མས,ནད,རད,ལད}

模型-2 的笛卡尔积为

$\mathrm{Pr}\times \mathrm{Con}\times \mathrm{Tvowel}\times \mathrm{Tpostfix}=$

$\{<p,b,v,s>|p\in \mathrm{Pr},\ b\in \mathrm{Con},$

$v\in \mathrm{Tvowel},\ s\in \mathrm{Tpostfix}\}$。

根据模型–2, 前加字与 Con 中元素组合仍能产生不符合语法的冗余音节, 以下将利用假设 4 的前加字与基字之间的阴阳关系来进一步简化模型–2。

根据藏语的语音理论, 对前加字集合进行阴阳关系分类, 设前加字集合 *Pr* 为

$\mathrm{Pr}=\mathrm{pp}\cup \mathrm{pne}\cup \mathrm{ppn}\cup \mathrm{pvn},$

其中, pp 为阳性, pne 为中性, ppn 为纯阴性,pvn 为极阴性, 分别为 pp={བ}, pne={ག,ད}, ppn={ཨ}, pvn={ས}。

已知前加字集合与基字字符集合的笛卡尔积为:

$\mathrm{Pr}\times \mathrm{Con}=\{<x,y>|x\in \mathrm{Pr},\ y\in \mathrm{Con}\}$。

根据前加字与基字的组合关系, 存在如下几个前后关系。

关系 1 为阳性与阳性、阳性与阴性组合, 则 $R_1\subseteq \mathrm{Pr}\times \mathrm{Con}$, 且

R_1={< བ , ཀ >,< བ , ཅ > ,< བ , ཏ >,< བ , ཙ > ,< བ , ཀ >,< བ , ཁ > ,< བ , ཤ >,< བ , ཀྱ >,< བ , ཀྲ >,< བ , གྲ >,< བ , ཏྲ >,< བ , གྱ >,< བ , སྐ >,< བ , སྟ >,< བ , སྤ >,< བ , ཀ>,< བ , ཁྲ >,< བ , ཐ >,< བ , ཉ > , < བ , ཏྲ >,< བ , ཤྲ >,< བ , ཁྲ >,< བ , ཁྱ >,< བ , ཐྱ >,< བ , ཉ >,< བ , ཏྲ >,< བ , ཏྱ >,< བ , ཐྱ >,< བ , ཁྲ >,< བ , ཁྱ >,< བ , ཐྱ >,< བ , ཉ >,< བ , ཏྲ >,< བ , ཏྱ >,< བ , ཐྱ >,< ད , ཁ >,< ད , ཉ>,< ད , ཐ >,< ད , ཏྲ >,< ད , ཏྱ >,< ད , ཐྱ >,< ད , ཁྲ >,< ད , ཁྱ >,< ད , ཐྱ >,< ད , ཉ>,< ད , ཏྲ >,< ད , ཏྱ >,< ད , ཐྱ >}

关系 2 R_2 为阴性与阴性、阴性与中性组合, 则 $R_2\subseteq \mathrm{Pr}\times \mathrm{Con}$, 且

R_2={< ཨ, ཀ > , < ཨ, ཅ > , < ཨ , ཏ > ,< ཨ , ཙ >,< ཨ , ཀྱ >,< ཨ , ཀྲ > , < ཨ, ད > ,< ཨ , ཀ >,< ཨ , ཁ > ,< ཨ, ཤ > ,< ཨ , ཀྱ >,< ཨ , ཀྲ >,< ཨ , ཁ >,< ཨ , ཚ >,< ཨ, ཟ >,< ཨ, ཡ >,< ཨ, ཚ >,< ཨ, ཟ >,< ཨ, ཡ >,< ཨ, ཚ >,< ཨ, ཟ >,< ཨ, ཡ >,< ད , ཀ > , < ད , ཁ > }

关系 3 R_3 为中性与阳性、中性与阴性组合, 则 $R_3\subseteq \mathrm{Pr}\times \mathrm{Con}$, 且

R_3={< ག, ཅ> , < ག, ཏ> , < ག , ཙ> ,< ག , ད>,< ག , ཁ>,< ག, ཤ>,< ག, ཡ> ,< ག , ཀ>,< ག , ཤ>,< ག , ཀྱ> , < ག, ཁ> }

关系 4 R_4 为中性与阳性、中性与阴性组合, 则 $R_4\subseteq \mathrm{Pr}\times \mathrm{Con}$, 且

R_4={< ད, ཀ> , < ད, ཅ> , < ད , ཙ> , < ད, ཀྱ> , < ད, ཁྱ> , < ད , ཐྱ>, < ད , ཀ>, < ད , ཅ>,< ད, ཙ>,< ད, ཀྱ> ,< ད , ཁྱ>,< ད, ཐྱ>,< ད , ཀ>,< ད, ཁ>,< ད ,

བྲ>,<ད,ང>,<ད,མ>,<ད,ཀྱ>,<ད,བྲ>}

关系 5 R_5为极阴性与中性，极阴性与阴性，极阴性与极阴性组合，则$R_5 \subseteq \mathrm{Pr} \times \mathrm{Con}$，且

R_5={<མ,ཁ>,<མ,ཆ>,<མ,ཐ>,<མ,ཚ>,<མ,ཁྱ>,<མ,ཁྲ>,<མ,ག>,<མ,ཇ>,<མ,ད>,<མ,ཛ>,<མ,གྱ>,<མ,གྲ>,<མ,ང>,<མ,ཉ>,<མ,ན>}。

根据关系 1~5，在模型-2 的基础上进一步简化，形成藏文音节模型-3(简称为模型-3)，即模型-2 中的构造位 1 和 2 预组合成模型-3 的构造位 1，如图 4 所示。在模型-3 中，构造位 1 是前缀部分，2 是元音部分，3 是后缀部分，笛卡尔积为

$\mathrm{Trule} \times \mathrm{Tvowel} \times \mathrm{Tpostfix} =$

$\{<b,v,s> | b \in \mathrm{Trule},\ v \in \mathrm{Tvowel},\ s \in \mathrm{Tpostfix}\}$,

其中，Trule 为规则字符集，是模型-3 中构造位 1 中的元素，包括 Con 中的元素和关系 1~5 中形成的元素，其余两个集合与模型-2 的集合一致。根据藏文语法，基字、元音和后加字组合才能构成藏文音节，所以在 Trule、Tvowel、Tpostfix3 个集合中按顺序各取一个元素能够组合成藏文音节。

2.2 藏文音节规则库

定义 3 字符序列 xyz($x \in \mathrm{Trule}$, $y \in \mathrm{Tvowel}$, $z \in \mathrm{Tpostfix}$)，称字符序列 xyz 为藏文音节。

由定义 3 可知，符合语法的、并由点(该点称为音节点)分的藏文字符串为藏文音节。

定义 4 $\forall x \in \mathrm{Trule}$，且 x 满足 $x \in R_i$ ($R_i \subseteq \mathrm{Pr} \times \mathrm{Con}$, i=1,…5)，则 x 称为规则字符。

由定义 4 可知，根据 Pr 与 Con 的 R_1 至 R_5 组合关系构成的字符串组合，但不构成一个音节的为规则字符。例如“དཀ”为规则字符，它不能构成一个音节。

在规则库的构建中，Trule 与 15 个“ཡ་བཏགས་ཅན”规则字符构成主要规则表的内容，如表 1 所示。其中 15 个“ཡ་བཏགས་ཅན”规则字符是指表 1 中的“ཡ་བཏགས་ཅན།*”所在行对应的字符。Tvowel 集合、Tpostfix 集合、4 个常用符号、10 个藏文数字作为辅助规则表的内容，其中 Tvowel 集合是模型-3 中构造位 2 的元素，Tpostfix 集合是模型-3 中构造位 3 的元素，辅助规则表如表 2 所示。可以看出规则表(表 1)是模型-3 中藏文音节的前缀部分，规则表 2 是模型-3 中的元音集合、后缀字符集合和其他一些字符。

规则表(表 1)中，第一组为 30 个藏文字母；第二组为 3 个上加字与基字组合的 33 字符；第三组为 4 个基字与下加字组合的 42 个字符；第 4 组为 5 个前加字、基字和其他叠加字符组合的 105 个字符；第 5 组为上加字、基字和下加字组合的 14 个叠加字符。辅助规则表列出了 4 个显示元音字符和 1 个隐含元音；4 个常用符号；10 个藏文数字符号；10 个后加字；7 预组合的再后加字符。需要说明的是两个规则表可根据应用需求的不同而增加和减少。

2.3 歧义规则处理

在表 1 中与前加字ག组合的{གད, གན, གས}3 个规则字符、与前加字ད组合的{དག, དང, དབ, དམ}4 个规则字符、与前加字བ组合的{བག, བད, བས}3 个规则字符和与前加字མ组合的{མག, མད, མང, མན}4 个规则字符共计 14 个存在歧义现象。这些规则字符在藏文中不仅能够单独构成一个音节，而且跟其他“后加字”连接也能构成另外的一个藏文音节。

在具体应用中针对这 14 个规则需要另加判断条件，即“判断这 14 个规则后面是否有音节点，若有，则为一个音节而非规则，否则为规则”。

3 规则方法的应用研究

藏文音节规则模型(TSRM)是根据模型-3 以建构和解构方式建立的一种方法，简称为规则方法，本文中以建构方式进行自动拼写藏文音节。以解构方式进行藏文音节的拼写检查算法，可参见文献[6]。

3.1 自动拼写藏文音节方法

在实际书写的现代藏文音节中，包含 5 种类型的音节：1)Trule 中大部分元素构成的音节，在实际的语法中含有隐含元音和“འ”黏着后加字；2)Trule 集合与 Tvowel 集合元素组合构成的音节，含有隐含的“འ”黏着后加字；3)Trule、Tvowel 与 Tpostfix 集合元素组合构成的音节；4)特殊的藏文音节，例如“སྤྱིའུ”，“ནྭ”，“གཏོའ”，“ཞྭའང”形式的特殊藏文音节；5)常见的梵音转写藏文音节，例如：“པདྨེ”形式的藏文音节。

本文所讨论的拼写藏文音节方法只考虑了前 3 种的藏文音节，后 2 种属于特殊现象的音节的，情况比较复杂未列入本文考虑范围。

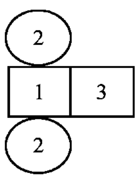

图 4 藏文音节模型-3

Fig. 4 Model-3 of Tibetan syllable

表 1 藏文规则表

Table 1 Table of Tibetan syllable rules

གསལ་ཡིད་ཀྱི་ཡོ	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
ཀ་མཉོ་ཅན།	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
ཁ་མཉོ་ཅན།	ཀྱ	ཁྱ	གྱ		ཅྱ	ཆྱ	ཇྱ		ཏྱ	ཐྱ	དྱ		པྱ	ཕྱ	བྱ	མྱ	ཙྱ	ཚྱ	ཛྱ											
ག་མཉོ་ཅན།	ཀྲ	ཁྲ	གྲ						ཏྲ	ཐྲ	དྲ		པྲ	ཕྲ	བྲ	མྲ													
ཡ་བཏགས་ཅན།	ཀྱ	ཁྱ	གྱ	ང	ཅྱ	ཆྱ	ཇྱ	ཉ	ཏྱ	ཐྱ	དྱ	ན	པྱ	ཕྱ	བྱ	མྱ	ཙྱ	ཚྱ	ཛྱ	ཝ	ཞ	ཟ	འ	ཡ	ར	ལ	ཤ	ས	ཧ	ཨ
ར་བཏགས་ཅན།	ཀྲ	ཁྲ	གྲ						ཏྲ	ཐྲ	དྲ		པྲ	ཕྲ	བྲ	མྲ													
ལ་བཏགས་ཅན།	ཀླ	ཁླ	གླ										པླ	ཕླ	བླ	མླ													
ཤ་བཏགས་ཅན།*	ཀྵ	ཁྵ	གྵ	ངྵ				ཉྵ																					
ཁ་མཚོན་འཇུག	ཁཀ	ཁག	ཁང	ཁཅ	ཁཆ	ཁཇ	ཁཉ	ཁཏ	ཁཐ	ཁད	ཁན	ཁཔ	ཁཕ	ཁབ	ཁམ	ཁཙ	ཁཚ	ཁཛ	ཁཝ	ཁཞ	ཁཟ	ཁའ	ཁཡ	ཁར	ཁལ	ཁཤ	ཁས	ཁཧ	ཁཨ
ད་མཚོན་འཇུག	དཀ	དག	དང	དཅ	དཆ	དཇ	དཉ	དཏ	དཐ	དད	དན	དཔ	དཕ	དབ	དམ	དཙ	དཚ	དཛ	དཝ	དཞ	དཟ	དའ	དཡ	དར	དལ	དཤ	དས	དཧ	དཨ
ག་མཚོན་འཇུག	གཀ	གག	གང	གཅ	གཆ	གཇ	གཉ	གཏ	གཐ	གད	གན	གཔ	གཕ	གབ	གམ	གཙ	གཚ	གཛ	གཝ	གཞ	གཟ	གའ	གཡ	གར	གལ	གཤ	གས	གཧ	གཨ
བཀྱེས་པ་ཅན།	ཀི	ཁི	གི	ངི	ཅི	ཆི	ཇི	ཉི	ཏི	ཐི	དི	ནི	པི	ཕི	བི	མི	ཙི	ཚི	ཛི	ཝི	ཞི	ཟི	འི	ཡི	རི	ལི	ཤི	སི	ཧི	ཨི

表 2 辅助规则表

Table 2　Table of secondary rules

元音字符	常用符号	数字符号	后加字	组合的后缀字符
ཨ	་	༠	ག་	གས་
ཨི	।	༡	ང་	ངས་
ཨེ	༄	༢	ད་	བས
ཨོ	༅	༣	ན་	མས
ཨུ	༼	༤	བ་	ནད
	༽	༥	མ་	རད་
		༦	འ་	ལད་
		༧	ར་	
		༨	ལ་	
		༩	ས	

自动拼写藏文音节系统以模型-3为依据，取 Trule 中 209 个元素，Tvowel 中 5 元音个元素，Tpostfix 后缀字集合中 17 个元素作为自动拼写藏文音节对象。

根据藏文语法，模型-3 中的任何“基字”与后缀字之间在元音的作用下可以进行任意的组合，这里的“基字”可以包含 *Trule* 集中的任何元素。模型-3 的藏文语法如下所示：

རྗེས་འཇུག་ཡི་གེ་བཅུ་པོ་ཡི·······.མིང་གཞིའི་ཡི་གེ་ཀུན་ལ་འཇུག。

通过如上所述的语法和模型-3 可以建立自动拼写藏文音节方法，如图 5 所示。

对于上述 3 种不同的藏文音节，前两种音节由系统的 Trule 与 Tvowel 集合的组合完成自动拼写，第 3 种音节由 Trule、Tvowel 与 Tpostfix 集合的组合完成自动拼写。

根据自动拼写藏文音节系统，拼写算法通过算法 1 和算法 2 来实现自动拼写过程。算法 1 是从 Trule 集合中取一个元素，与 Tvowel 集合中元音组合拼写出藏文音节；算法 2 是从 Trule 集合中取一个元素，在 Tvowel 的作用下与 Tpostfix 集合中的每一个元素结合拼写出藏文音节。具体算法设计如下。

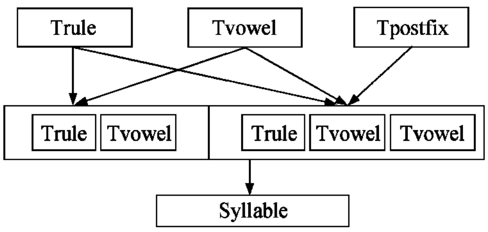

图 5 自动拼写藏文音节方法

Fig. 5　Automatic spelling method of Tibetan syllable

算法 1 规则集 Trule 与 Tvowel 元音字符集进行组合，构成藏文音节。

```
输入:
Trule: 规则字符集合
Tvowel: 元音字符集合
输出: 藏文音节
begin
创建输出对象 outfile;
创建规则集的字符串类型 latter;
创建元音集的字符串类型 vowel;
创建藏文音节的字符串类型 syllable;
for  each latter∈Trule do
        for  each vowel∈Tvowel  do
            syllable=latter+vowel;
             outfile. writetofile(syllable);
        end
    end
end
```

算法 1 包括主要步骤如下：1) 创建输出对象 outfile; 2) 创建规则元素变量 latter; 3) 创建元音元素变量 vowel; 4) latter 与 vowel 组合构成藏文音节。

算法 2 规则集 Trule 与 Tvowel 元音字符集、Tpostfix 集合(不含空集)进行组合，构成藏文音节。

```
输入:
Trule: 规则字符集合
Tvowel: 元音字符集合
TPostfix: 后缀字符集合
输出: 藏文音节
begin
创建输出对象 outfile
创建规则集的字符串类型 latter
创建元音集的字符串类型 vowel;
创建后缀集的字符串类型 postfix
创建藏文音节的字符串类型 syllable
for each latter∈Trule do
      for  each vowel∈Tvowel  do
            for each postfix∈TPostfix  do
                syllable=latter+vowel+poatfix;
                 outfile.writetofile(syllable);
            end
      end
end
end
```

表 3　存在歧义的音节表

Table 3　Table of ambiguous syllable

音节表														
不符合语法的 45 个规则					黏着字འ的样例					前加字作为基字的音节				
གཅ	གསྐ	བད	འཇ	མཆ	ཀའ	ཁྱའ				གགས	བལ	འགས		
གཉ	དཀ	བཞ	འཐ	མཇ	གིའ	ཁྱིའ				དགས	བར	འངས		
གཏ	དགེ	བཀ	འད	མཉ	གུའ	ཁྱུའ				དངས	བན	འབས		
གད	དད	བཙ	འཚ	མཐ	གེའ	ཁྱེའ				དབས	ཕགས	འམས		
གན	དབ	བཟ	འབ	མད	གོའ	ཁྱོའ				དམས	ཕངས	འཉད		
གཙ	དབ	བཤ	འཆ	མན	ཀའ	ཆྱའ				དཉད	མབས	འཚད		
གཞ	དམ	བས	འཇ	མཚ	ཀིའ	ཆྱིའ				དངད	མམས	འལད		
གཟ	བག	འག	མཁ	མཇ	ཀུའ	ཆྱུའ				དལད	ཚཏད			
གཡ	བཅ	འཁ	མག		ཀེའ	ཆྱེའ				བགས	མངད			
གའ	བཀ	འཆ	མང		ཀོའ	ཆྱོའ				བངས	མལད			

算法 2 包括主要步骤如下: 1) 创建藏文后缀元素变量 postfix; 2) latter、vowel 与 postfix 构成藏文音节,其余释义与算法 1 一致。

经过实验测试, 算法 1 产生 1045 个音节, 算法 2 能够产生 17765 个音节, 共计 18810 个藏文音节。需要说明如下。

1) 由算法 1 中的前加字+基字+隐含元音构成的 45 个音节确实存在不符合语法的情况, 如表 3 所示。

2) 由算法 2 产生的藏文音节中与黏着字“འ”组合的1045个音节中除去45个不符合语法的音节, 共计 1000 个音节与现代藏文音节写法不一致, 但语法上没有错误, 样例如表 3 所示。

3) 当 5 个前加字作为基字, 后跟后加字和再后加字时, 产生的音节在基字定位上存在歧义, 如表 3 所示。字体加粗部分表示前加字作为基字的部分, 但计算机难以识别哪个是基字, 如图 6 和 7 所示, 其中图 6 以“ས”为基字, 图 7 以“ང”为基字, 表示基字定位上存在歧义。

4) 常有的“ཀྱྭའ”等形式的藏文音节没有包含在自动拼写音节方法中, 特别是许多生僻的音节还需要语言学专家进一步论证。

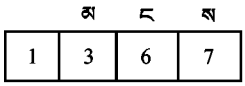

图 6 基字+后加字+再后加字

Fig. 6　Root+suffix+post suffix

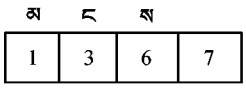

图 7 前加字+基字+后加字

Fig. 7　Prefix+root+suffix

根据算法 1 和 2 自动拼写出来的藏文音节中, 去掉 45 个规则, 再增加 15 个“ཨ་བཏགས་ཅན”规则字符, 共计 18780 个音节, 这与参文[7]中提到的两个数字是接近的, 一个数字是萨迦索南孜木提到的 18745 个藏文音节, 另外一个是才旦夏茸提到的将近 18000 个的藏文音节。

4　结论与展望

本文简单回顾了藏文信息处理的历史过程, 概述了针对藏文音节开展的研究工作, 分析了藏文音节模型在藏文信息处理过程中的应用情况, 并讨论了藏文音节的结构、组成成分、拼写规则。接着根据藏文音节的拼写规则建立藏文音节数学模型和规则库, 并提出了规则方法的自动拼写藏文音节算法。通过编程自动拼写出 18810 藏文音节。下一步的工作将是利用藏文音节的规则方法, 提高该方法在实际应用中的有效性, 并继续开展规则方法在藏文词语校对、短语校对、藏文分词等领域的应用研究。

致谢　研究工作得到西南交通大学“四川省云计算重点实验室”老师和同学的帮助, 也得到了西藏大学藏文信息技术研究中心(藏文信息技术国家地方联合工程研究中心)老师的帮助, 表示衷心感谢。

参考文献

[1] 江荻. 书面藏语的熵值及相关问题 // 中文信息处理国际会议论集. 北京: 清华大学出版社, 1998

[2] 江荻, 康才畯. 书面藏语排序的数学模型及算法. 计算机学报, 2004, 27(4): 524 - 529

[3] Chilton R R. Sorting unicode Tibetan using a multi-weight collation algorithm[EB/OL]. [2012–08–10]. https://collab.itc.virginia.edu/access/wiki/site/26a34146-33a6-48ce-001e-f16ce7908a6a/sorting%20tibetan.html

[4] 扎西次仁.《中华大藏经•丹珠尔》藏文对勘本字频统计分析. 中国藏学, 1997, 2: 122–133

[5] 王维兰, 陈万军. 藏文字丁、音节频度及其信息熵. 术语标准化与信息技术, 2004, 2: 27–31

[6] 珠杰. TSRM 的藏文拼写检查算法. 软件学报, 审稿中

[7] 才旦夏茸. 才旦夏茸全集. 北京: 民族出版社, 2007: 44–45