基于 Mealy 机的藏文字构件分解

才让卓玛[1,2],李永明[1],才智杰[2]
(1.陕西师范大学计算机科学学院,陕西西安 710062; 2.青海师范大学计算机学院,青海西宁 810008)

摘 要: 藏文字构件分解是藏文信息处理的基础,具有重要的理论价值和广阔的应用前景.针对藏文字构件的复杂性与多样性,文章通过分析现代藏文字的构字规则和结构特点,研究了藏文字构件的分解过程,利用 Mealy 机的输出字符与移动一一对应的特性描述了藏文字构件的行为语义,给出了对于任意字符串能否被 Mealy 机分解的判定定理及基于 Mealy 机的藏文字构件分解算法,并设计实现了基于 Mealy 机的藏文字构件分解系统,验证了算法的可行性.

关键词: 藏文信息处理; Mealy 机; 构件; 构字分解
中图分类号: TP391 **文献标识码:** A **文章编号:** 0372-2112 (2015)05-0935-05
电子学报 URL: http://www.ejournal.org.cn **DOI:** 10.3969/j.issn.0372-2112.2015.05.016

Components Decomposition of Tibetan Words Based on Mealy Machines

CAI Rang-zhuoma[1,2], LI Yong-ming[1], CAI Zhi-jie[2]
(1. *College of Computer Science, Shaanxi Normal University, Xi'an, Shaanxi* 710062, *China*;
2. *College of Computer Science, Qinghai Normal University, Xining, Qinghai* 810008, *China*)

Abstract: Components decomposition of Tibetan words is the basic work of Tibetan information processing, it provides significant theoretical value and has wide application perspective. Since the complexity and variety of components of Tibetan words, this paper studies the process of components decomposition of Tibetan words by analyzing the grammars and the structure of Tibetan words, gives a component decomposition algorithm and judgment theorems about decomposition based on the one-to-one relationship between each output character and its state transition of Mealy machine, then verify the validity of the algorithm by a component decomposition system based on Mealy machine.

Key words: Tibetan information processing; Mealy automata; components; component decomposition

1 引言

藏文信息处理中,凡与字相关的研究都不仅需要分析字结构,而且需要确定构件的位置特征.因此,藏文字构件分解是藏文信息处理的基础.藏文字是以基字、前加字、上加字、下加字、后加字、重后加字及元音等1至7个藏文字符为构件的二维拼音文字,其中前加字、后加字和重后加字与基字横向拼写,上加字、下加字和元音与基字纵向拼写[1].前加字、基字、上加字、下加字、后加字、重后加字和元音是构成藏文字的最小单位构件,简称构件;上加字、下加字及元音与基字纵向排列而成的字符组合称组合构件.藏文字构件的复杂性与多样性是丰富而严谨的现代藏文文法的产物,同时也是藏文信息处理的重点和难点.长期以来,对藏文字构件的研究一直被业内人士所关注,文献[2,3]以藏语口语材料中3926个常用字为研究对象,统计了藏文字长和构词频度、声、韵母结构方式及组合构件频度;文献[4]对《中华大藏经·丹珠尔》中藏文字频度和构件进行了统计;文献[5]从字符、构件、音节和词汇的角度统计了词典中的词汇;文献[6]对特定藏语语料做了字频、音节频度的统计;文献[7]对19380个藏文字的字长、结构方式、构件的频度及组合构件进行了统计;文献[8]基于常用词典对藏文字及构件频度进行了统计.上述研究为藏文字构件分解提供了重要的参考数据,但还存在如下缺憾:(1)研究范围局限于特定词典或语料中,不具有普遍性;(2)研究成果为小语料库上以手工与计算机辅助方式结合得到的构件频度,没有解决构件自动分解问题,不具有广泛的应用性.

收稿日期:2013-09-18;修回日期:2014-10-24;责任编辑:蓝红杰
基金项目:国家自然科学基金(No.61262051,No.11271237,No.61163018);国家社科基金(No.13BYY141,No.14BYY1322);教育部"春晖计划"合作科研项目(No.Z2012093);"长江学者和创新团队发展计划"创新团队资助项目(No.IRT1068)

自动机理论在文字识别、词法分析及人工智能等领域一直被广泛应用[9~12],尤其是带输出的自动机—Mealy 机[13,14]. 由于 Mealy 机处理串时输出字符与移动一一对应的特性有利于描述藏文字构件的分解过程,本文通过将任意现代藏文字作为 Mealy 机的输入,构件分解及位置特征值作为 Mealy 机的输出,研究基于 Mealy 机的藏文字构件分解算法,旨在解决藏文字构件自动分解问题,使其具更广泛的参考意义和应用价值.

2 藏文字的构字规则与结构分析

根据藏文文法[1],元音字母不能单独成字,辅音字母可单独成字,也可与元音字母拼合成字;前加字、上加字、下加字必须添在对应基字之前(上或下),元音可与所有基字接合. 现代藏文构字规则如表 1 所示.

表1 现代藏文字构字规则

前加字添加规则		上加字添加规则		下加字添加规则		重后加字添加规则	
ག	ཅ་ཀ་ཅ་ཀྱ་ད་ནཁ་ཡ་ཤ་ས	ར	ཀ་ཁ་ང་ཇ་ཉ་ཏ་ད་ན་བ་མ་ཙ་ཛ	ཡ	ཀ་ཁ་ག་ང་ཌ་ད་ན་བ་མ	ད	ན་ར་ལ
ད	ཀ་ང་ག་ང་བ་མ	ལ	ཀ་ག་ང་ཅ་ཇ་ཀ་ད་ན་བ་ཧ	ར	ཀ་ཁ་ག་ཀ་ཐ་ད་ན་བ་མ་ཤ་ཧ	ས	ག་ང་བ་མ
བ	ཀ་ཅ་ཀ་ཅ་ག་ང་ཇ་ཉ་ཀ་ཏ་ད་ན་ཁ་ཉ་ཤ་ས	ས	ཀ་ག་ང་ཉ་ཀ་ད་ན་ང་བ་མ་ཙ	ལ	ཀ་ཁ་བ་ཟ་ར་ས		
མ	ཁ་ཆ་ཐ་ཚ་ག་ཇ་ད་ཏ་ང་ན་ཉ་ཤ			ཤ	ཀ་ཁ་ག་ཉ་ད་ཚ་ཞ་ཟ་ར་ལ་ཤ་ཧ		
འ	ག་ཇ་ད་ན་ཏ་ཁ་ཆ་ཐ་ཚ	后加字可添接在任何辅音字母后					

藏文字中基字和组合构件之一必须存在且仅能存在一个,其余构件可缺省. 从单位构件角度看,一个藏文字由 1 至 7 个构件组成[15];若把组合构件看作一个整体,则一个藏文字由 1 至 4 个构件组成,即藏文字结构有两种形式:(1)[前加字][+ 上加字]基字[+ 下加字][+ 元音][+ 后加字][+ 重后加字];(2)[前加字]组合构件[+ 后加字][+ 重后加字]. 如藏文字“ བསྒྲུབས ”在第一种形式下由前加字、上加字、基字、下加字、元音、后加字和重后加字等 7 构件组成,在第二种形式下该字由前加字、组合构件、后加字和重后加字等 4 构件组成. 本文将仅含基字或组合构件的藏文字称单构件藏文字(简称单构件),其余类型根据构件个数称为二构件、三构件及四构件藏文字(简称多构件),分别用 L_1、L_2、L_3和 L_4表示,并用 F、H、R、U、V、L、L'表示前加字、上加字、基字、下加字、元音、后加字、重后加字等构件的位置特征,藏文字结构分类及示例见表 2. 由藏文字结构可确定各构件的位置属性,如 L_2型藏文字“ ཚོས ”便可确定其各个构件的位置属性;反之,除少量的“辅音字母 + 辅音字母 + 辅音字母”型藏文字,如“ བཤགས་མགས་བངས་དངས ”;(根据文法它们既属于 RLL'型,也属于 FRL 型,因此本文称之为不确定类型藏文字)”等外,其余藏文字结构也可唯一确定. 对于不确定类型藏文字,本文根据频度统计结果进行了分类.

表2 藏文字结构分类及示例

结构类型	示例		
L_1 型藏文字	ག	རྐྱ	བསྒྲུབ
L_2 型藏文字	གང	བསོམ	སྒྲུབས
L_3 型藏文字	གངས	བསོམས	སྒྲུབས་
L_4 型藏文字	འབས་	བསྒྲུབས	བསོམས

3 基于 Mealy 机的藏文字构件分解

3.1 构件分解模型

藏文字构件分解包括藏文文本读入、识字并过滤非藏文字符、计算构件数、识别紧缩字、构件分解、构件位置特征分析和藏文字结构特征分析等,其分解模型如图 1 所示.

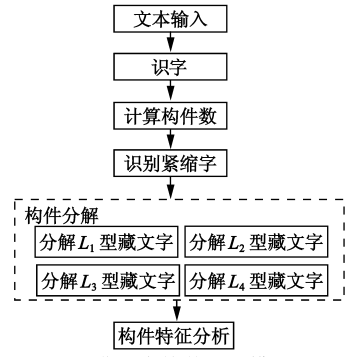

图1 藏文字构件分解模型

3.2 构件分解 Mealy 机

定义 1 (单构件分解 Mealy 机) $M_1=(Q,\Sigma,\Delta,\delta,q_c,\lambda)$,其中,$Q=\{q_c,q_h,q_r,q_v,q_u\}$是状态的非空集合,$q_c$是初始状态,$q_h(q_r,q_v,q_u)$分别表示 M_1读过一个上加字(或基字、或元音、或下加字)后到达的状态. 输入字母表 $\Sigma=\Sigma_H\cup\Sigma_R\cup\Sigma_U\cup\Sigma v$,$\Delta=\{H,R,U,V\}$是输出字母表. $\delta: Q\times\Sigma\to Q$. 对任意$(q,x)\in Q\times\Sigma$,$\delta(q,x)=p$ 表示 M_1在状态 q 读入字符,将状态变成 p,将读头指向下一个字符. 用 $q\xrightarrow{x/y}q'$表示 M_1的一个转移,其中 $q,q'\in Q$,$x\in\Sigma$,$y\in\Delta$,标记 x/y 表示$\delta(q,x)=q'$,$\lambda(q,x)=y$,如图 2 所示.

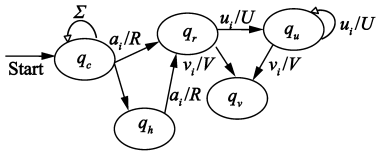

图2　分解L_1型藏文字Mealy机

定义 2　**（多构件分解 Mealy 机）**　设 Q 是状态的非空集合，$\Sigma=\Sigma_F\cup\Sigma_H\cup\Sigma_R\cup\Sigma_U\cup\Sigma_V\cup\Sigma_L\cup\Sigma_{L'}$，$\Delta_1=\Delta\cup\{n\}$是输出字母表，其中 $\Delta=\{F,H,R,U,V,L,L'\}$，$\lambda$ 是从 Q 到 Δ_1的映射，如果对任意$(q,x)\in Q\times\Sigma$，δ 满足以下 2 个条件，则称 $M_m=(Q,\Sigma,\Delta_1,\delta,q_0,\lambda)$是分解多构件藏文字 Mealy 机.

(1) $|x|=1,\delta(q,x)=p,\lambda(q,x)=y,y\in\Delta$;

(2) $|x|>1,\delta(q,x)=q_c,\lambda(q,x)=n\in\Delta_1$.

$q\xrightarrow{x/y}q'$表示 M_m的一个转移，其中 $q,q'\in Q,x\in\Sigma,y\in\Delta_1$. 当 $x=C,y=n$ 时用$q\xrightarrow{C/n}q'$表示当前状况接受了组合构件，其中 $C\in\Sigma$，且$|C|>1,n\in\Delta_1-\Delta$. 图 3 ~5 分别表示 L_2型、L_3型及 L_4型多构件分解 Mealy 机.

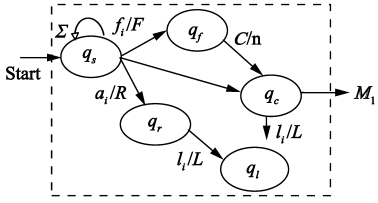

图3　分解L_2型藏文字Mealy机

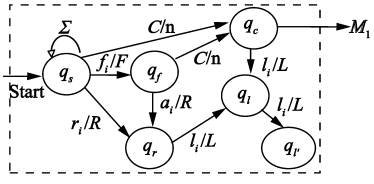

图4　分解L_3型藏文字的Mealy机

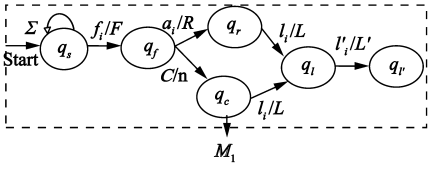

图5　分解L_4型藏文字的Mealy机

定义 3　**（执行片断和运行）**　Mealy 机 M 的执行片断可以定义为状态、输入与输出的一个有穷序列，记作 $w=q_0\xrightarrow{x_1/y_1}q_1\xrightarrow{x_2/y_2}q_2\cdots q_{n-1}\xrightarrow{x_n/y_n}q_n$，其中，$q_0$为开始状态. 如果 $q_i(0<=i<=n)$是终止状态，则称 w 为 M 的一个运行.

定义 4　**（迹、全迹、字、标注、语言）**　设 $w=q_0\xrightarrow{x_1/y_1}q_1\xrightarrow{x_2/y_2}q_2\cdots q_{n-1}\xrightarrow{x_n/y_n}q_n$是 Mealy 机 M 的一个执行片断，由 w 的输入与输出组成的序列称为 M 的一个迹，记作 $trance(w)=x_1/y_1,x_2/y_2,\cdots,x_n/y_n$. 如果 w 为 M 的一个运行，则 $trance(w)=x_1/y_1,x_2/y_2,\cdots,x_n/y_n$称为 M 一个全迹，其中输入序列 $x_1,x_2,\cdots,x_n$称作一个字，输出序列 $y_1,y_2,\cdots,y_n$称作这个字的标注.

例 1　分解“ ”时 $q_0\xrightarrow{h_3/H}q_h$和 $q_0\xrightarrow{h_3/H}q_h\xrightarrow{a_3/R}q_r\xrightarrow{u_2/U}q_u\xrightarrow{x4/V}q_v$分别为 M_1的一个执行片断和运行，h_3/H和 $h_3/H,a_3/R,u_2/U,v_4/V$ 分别为迹和全迹，序列(h_3,a_3,u_2,v_4)表示该字，(H,R,U,V)为该字标注，依次为上加字、基字、下加字和元音.

3.3　构件分解算法

算法 1　**单构件(L_1型)藏文字分解算法**

输入：(1)x 是L_1型藏文字，$x\in\Sigma$;

(2)上加字标记集合 $flag_1=\{h_1,h_2,h_3\}$;

(3)基字标记 $flag_2=\{a_1,a_2,\cdots,a_{30}\}$;

输出：序对$(x,y)\in\Sigma\times\Delta,\Delta=\{H,R,U,V\}$

主要步骤：

Do {For each $x\in\Sigma$}

step1：构造不含上加字 H 的执行片断集

if $(flag_1=\phi)$ && $(flag_2!=\phi)$}

$S=\{q_s\xrightarrow{x1/R}q_r\ \|\ q_s\xrightarrow{x1/R}q_r\xrightarrow{x2/U}q_u\ \|\ q_s\xrightarrow{x1/R}q_r\xrightarrow{x2/V}q_v\ \|\ q_s\xrightarrow{x1/R}q_r\xrightarrow{x2/U}q_v\xrightarrow{x3/V}q_u\ \|\ q_s\xrightarrow{x1/R}q_r\xrightarrow{x2/U}q_v\xrightarrow{x3/U}q_u\}$

step2：不含上加字 H 的迹 $trance(w)=\{x_1/R\}\ \|\ \{x_1/R,x_2/V\}\ \|\ \{x_1/R,x_2/U\}\ \|\ \{x_1/R,x_2/U,x_3/V\}\ \|\ \{x_1/R,x_2/U,x_3/U\}$

step3：output(x,y)}

step4：构造含上加字 H 的执行片断集

if $(flag_1!=\phi)$ && $(flag_2!=\phi)$}

$S=\{q_s\xrightarrow{x1/H}q_h\xrightarrow{x2/R}q_r\ \|\ q\xrightarrow{x1/H}q_h\xrightarrow{x2/R}q_r\xrightarrow{x3/U}q_u\ \|\ q_s\xrightarrow{x1/H}q_h\xrightarrow{x2/R}q_r\xrightarrow{x3/V}q_v\ \|\ q_s\xrightarrow{x1/H}q_h\xrightarrow{x2/R}q_r\xrightarrow{x3/U}q_u\xrightarrow{x4/V}q_v\}$

step5：含上加字 H 时的迹 $trance(w)=\{x_1/H,x_2/R\}\ \|\ \{x_1/H,x_2/R,x_3/U\}\ \|\ \{x_1/H,x_2/R,x_3/V\}\ \|\ \{x_1/H,x_2/R,x_3/U,x_4/V\}$

step6：output(x,y)

}while$(\Sigma=\phi)$

定理 1　对于给定的字符串 w 是否被 M_1识别是可判定的.

证明　将串 W 作为 M_1的输入串，从 M_1的初始状态开始，逐步模拟 M_1的运行过程. 如果在状态 q 下读到 W 中符号 a 时，$\delta(q,a)$没有定义就表示 M_1不能接受串 W；否则，当 M 读完串 W 时，都可以到达它的某一个确定状态，即 W 被 M_1接受.

算法 2　**多构件藏文字分解算法**

输入：$x=x_1x_2$或 $x=x_1x_2x_3$或 $x=x_1x_2x_3x_4$;//x 是 L_2, L_3 L_4型藏文字之一

输出：序对$(x,y)\in\Sigma\times\Delta,\Delta=\{F,H,R.U,V,L,L'\}$

主要步骤：

Do { For each $x\in\Sigma$)}

step1:构造不含组合构件的执行片断

$q_s \xrightarrow{x_1/R} q_r \xrightarrow{x_2/L} q_l \parallel q_s \xrightarrow{x_1/F} q_f \xrightarrow{x_2/R} q_r \xrightarrow{x_3/L} q_l$ or $q_s \xrightarrow{x_1/R} q_r \xrightarrow{x_2/L} q_l \xrightarrow{x_3/L'} q_l$' $\parallel q_s \xrightarrow{x_1/F} q_f \xrightarrow{x_2/R} q_r \xrightarrow{x_3/L} q_l \xrightarrow{x_4/L'} q_l$';

step2:生成相应的迹 $trance(w) = \{x_1/R, x_2/L\} \parallel \{x_1/F, x_2/R, x_3/L$ or $x_1/R, x_2/L, x_3/L'\} \parallel \{x_1/F, x_2/R, x_3/L, x_4/L'\}$;

step3:输出文字与标注 output(x, y);

step4:构造包含组合构件的执行片断 $q_s \xrightarrow{x_1/F} q_f \xrightarrow{x_2/n} q_c$ or $q_s \xrightarrow{x_1/n} q_c \xrightarrow{x_2/L} q_l \parallel q_s \xrightarrow{x_1/n} q_c \xrightarrow{x_2/L} q_l \xrightarrow{x_3/L'} q_l$', or $q_s \xrightarrow{x_1/F} q_f \xrightarrow{x_2/n} q_c \xrightarrow{x_3/L} q_l \parallel q_s \xrightarrow{x_1/F} q_f \xrightarrow{x_2/n} q_c \xrightarrow{x_3/L} q_l \xrightarrow{x_4/L'} q_l$';

step5:调用 M_1分解组合构件 $q_c \xrightarrow{call} M_1$;

step6:相应的迹 $trance(w) = \{x_1/n, x_2/L$ or $x_1/F, x_2/n\} \parallel \{x_1/n, x_2/L, x_3/L'$ or $x_1/F, x_2/n, x_3/L\} \parallel \{x_1/F, x_2/n, x_3/L, x_4/L'\}$

step7:输出文字与标注 output(x, y);}while$(\Sigma = \phi)$

定理 2 对于给定的字符串 w 是否被 M_m识别是可判定的.

证明 (略).

例 2 藏文字“ཨོན”分解过程.

step1:$x =$ “ཨོན”是 L_2型藏文字($x = x_1 x_2$且满足条件 $length(x_1) > 1$),执行片断为 $q_s \xrightarrow{x_1/n} q_c \xrightarrow{x_2/L} q_l$;

step2:生成迹 $trance(w) = x_1/n, x_2/L_4$后调用 M_1分解组合构件;

step3:输出藏文字与其各个构件.

3.4 构件分解统计结果

为了进一步验证算法的有效性,研发了基于 Mealy 机的藏文字构件分解系统,并对 2129234 字进行了构件分解与统计.实验证明本系统能够正确分解符合现代藏文拼写规则的文字(本算法没有考虑梵文和由于音译等产生的不符合现代藏文拼写规则的藏文字).统计显示在 2129234 字的藏语语料中,基字分布基本均匀(最高出现频率为 11.5%,最低出现为频率 0%),具体分布统计情况见图 6.元音共出现 675151 次,4 个元音出现频率依次为 31.8%、21.6%、16.4%和 30.2%,分布基本均匀;上加字共出现 148324 次,3 个上加字出现频率依次为 40.3%、8.9%和 50.8%,分布不太均匀;下加字共出现 236526 次,3 个下加字出现频率依次为 69.4%、25.5%、4.5%、0.6%,分布极不均匀;前加字共出现 249330 次,分布基本均匀;后加字共出现 766420 次,分布不太均匀.2 个重后加字的出现频率悬殊较大,分别为 0%与 100%.前加字、上加字、元音、下加字、后加字及重后加字分布统计情况见图 7.

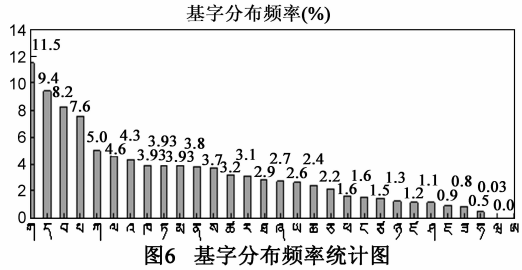

图6 基字分布频率统计图

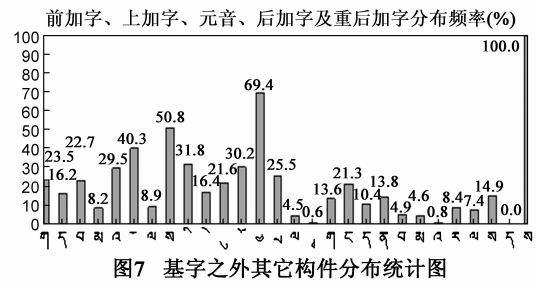

图7 基字之外其它构件分布统计图

4 结束语

针对藏文字构件的复杂性与多样性,文章通过分析现代藏文字结构与构字规则,研究了基于 Mealy 机的藏文字构件分解算法,开发了基于 Mealy 机的藏文字构件分解与统计软件.实验证明本系统能够正确分解符合现代藏文拼写规则的藏文字(本算法没有考虑梵文和音译文本中不符合现代藏文拼写规则的藏文字.藏文字构件分解算法在输入法研究、字属性分析、搜索引擎、机器翻译、网络信息安全、语音合成、语料库建设以及语义分析等方面具有重要的理论价值和广阔的应用前景.基于本研究成果,今后将进一步扩大语料库规模,研究外来语和音译文本中不符合现代藏文拼写规则的藏文字属性,进一步完善构件分解系统.

参考文献

[1] 格桑居冕.实用藏文文法[M].成都:四川民族出版社,1987.
Gesang Jumian. Practical Tibetan Grammar[M]. Chengdu: Sichuan Nationalities Publishing House, 1987.(in Chinese)

[2] 江荻,董颖红.藏文信息处理属性统计研究[J].中文信息学报,1994,9(2):37-44.
Jiang di, Dong Ying-hong. Research on property of Tibetan characters as information processing[J]. Journal of Chinese Information Processing, 1994, 9(2): 37-44.(in Chinese)

[3] 江狄,董颖红.藏字叠加结构线性处理统计分析[J].中文信息,1994,(4):44-46.
Jiang di, Dong Ying-hong. Linear statistics analysis of Tibetan character overlay structure[J]. Chinese Information Processing, 1994,(4):44-46.(in Chinese)

[4] 扎西次仁.《中华大藏经·丹珠尔》藏文对勘本字频统计分析[J].中国藏学,1997(2):122-133.
Xiciren Zha. The frequency statistics for Gangyur of Chinese of Tripitaka[J]. China Tibetology, 1997(2): 122-133.(in Chinese)

[5] 卢亚军,马少平,张敏,罗广.基于大型藏文语料库的藏文

字符、部件、音节、词汇频度与通用度统计及其应用研究[J].西北民族大学学报,2003,24(48):32 – 42.
Lu Ya-jun, Ma Shao-ping, Zhang Ming, Luo Guang. Researches of Calculations of Tibetan characters, pieces, syllables, vocabulary and universal frequency and its applications[J]. Journal of Northwest Minorities University, 2003, 24(48): 32 – 42. (in Chinese)

[6] 王维兰,陈万军.藏文字、音节频度及其信息熵[J].术语标准化与信息技术,2004,2:27 – 31.
WangWei-lan, Chen Wan-jun, The frequency and information entropy of Tibetan character and syllable[J]. Terminology Standardization & Information Technology, 2004, 2: 27 – 31. (in Chinese)

[7] 高定国,龚育昌.现代藏字全集的属性统计研究[J].中文信息学报,2005,19(1):71 – 75.
Gao Ding-guo, Gong Yu-chang. A statistically study on the qualities of all modern Tibetan character set[J]. Journal of Chinese Information Processing, 2005, 19(1): 71 – 75. (in Chinese)

[8] 艾金勇,李永宏,于洪志.藏文字形结构计量统计分析[J].计算机应用,2009,29(7):2029 – 2031.
Ai Jin-yong, Li Yong-hong, Yu Hong-zhi. Statistical analysis on Tibetan shaped structure[J]. Journal of Computer Application, 2009, 29(7): 2029 – 2031. (in Chinese)

[9] 张大方,张洁坤,黄昆.一种基于智能有限自动机的正则表达式匹配算法[J].电子学报,2012,40(8):1617 – 1622.
Zhang Da-fang, Zhang Jie-kun, Huang Kun. A regular expression matching algorithm with smart finite automaton[J]. Acta Electronica Sinica, 2009, 29(7): 2029 – 2003. (in Chinese)

[10] 赵力,邹采荣,吴镇扬.基于 3 维空间 Viterbi 算法的汉语连续语音识别[J].电子学报,2000,28(7):84 – 87.
Zhao Li, Zou Cai-rong, Wu Zhen-yang. Recognition of Chinese continuous speech based on 3-Dimension Viterbi search [J]. Acta Electronica Sinica, 2000, 28(7): 84 – 87. (in Chinese)

[11] 蒋宗礼,姜守旭.形式语言与自动机理论[M].北京:清华大学出版社,2007.
Jiang Zong-li, Jiang Shou-xu. Formal Languages and Automata Theory[M]. Beijing: Tsinghai University Press, 2007. (in Chinese)

[12] John E. Hopcroft, Rajeev Motwani, Jeffrey D. Ullman. Introduction to Automata Theory, Language, and Computation [M]. Beijing: China Machine Press, 2009.7.

[13] Wuu Yang. Mealy Machines are a better model of lexical analyzers[J]. Computer Languages, Systems & Structures. 2002, 28 (3): 273 – 288.

[14] IllyaI. Reznykov, Vitaliy I. Sushchansky. On the 3-state Mealy automata over an m-symbol alphabet of Growth order[nlogn/2logm][J]. Journal of Algebra 2006, 304: 712 – 754.

[15] 江荻,康才俊.书面藏语排序的数学模型及算法[J].计算机学报,2004,27(4):527 – 529.
JiangDi, Kang Cai-Jun. The sorting mathematical model and algorithm of written Tibetan language[J]. Chinese Journal of Computers, 2004, 27(4): 527 – 529. (in Chinese)

作者简介

才让卓玛　女,藏族,1970 年出生,青海海西人,陕西师范大学博士生,青海师范大学教授,主要研究领域为藏语自然语言处理、人-机语音交互.
E-mail: cr-zhuoma@163.com

李永明　男,教授,1966 出生,陕西省大荔县人,陕西师范大学博士生导师,研究方向:拓扑学、智能系统分析、可计算与复杂性理论.
E-mail: liyongm@snnu.edu.cn

才智杰　男,藏族,1970 年出生,青海乐都人,青海师范大学教授、硕士生导师,研究方向:藏文信息处理、藏语自然语言处理.
E-mail: Czjqhsd@163.com