多拉 扎西加 • 编著

藏文规范音节频率词典

བོད་ཡིག་ཚེག་ཁྱེམ་བང་མཛོད།

TIBETAN SYLLABLE FREQUENCY DICTIONARY

中国社会科学出版社

多拉 扎西加 • 编著

藏文规范音节频率词典

བོད་ཡིག་ཚེག་ཁྱེམ་བང་མཛོད།

TIBETAN SYLLABLE FREQUENCY DICTIONARY

中国社会科学出版社

图书在版编目(CIP)数据

藏文规范音节频率词典/多拉，扎西加编著. —北京：中国社会科学出版社，2015.1
ISBN 978-7-5161-5663-6

Ⅰ.①藏… Ⅱ.①多…②扎… Ⅲ.①藏语—音节—词典
Ⅳ.①H214.1-61

中国版本图书馆 CIP 数据核字(2015)第 041723 号

出 版 人 赵剑英
选题策划 陈肖静
责任编辑 陈肖静
责任校对 刘 娟
责任印制 戴 宽

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http://www.csspw.cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印 刷 北京君升印刷有限公司
装 订 廊坊市广阳区广增装订厂
版 次 2015 年 1 月第 1 版
印 次 2015 年 1 月第 1 次印刷

开 本 710×1000 1/16
印 张 24.75
插 页 2
字 数 406 千字
定 价 86.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社联系调换
电话：010-84083683

◎ 教育部哲学社会科学研究重大课题攻关项目：藏文《大藏经》十种版本电子资料库建设及其研究（13JZD028）；教育部、国家语委现代藏文规范表意音节频率库及标准研究(MZ115-69)；2012教育部新世纪优秀人才支持计划（NCET-12-0665）
◎ 国家自然科学基金：藏语依存树库的构建（61163043）；基于Ontology的藏文语料库检索关键技术研究（61262053）
◎ 中国语言文学一级学科甘肃省重点学科

前　言

藏文的音节是以字根为中心的语言单位。一个音节中的纵向单位（辅音字母或上下叠加的组合体）叫字丁。如果从字丁组成音节的角度来说，就有单字音节、双字音节、三字音节和四字音节。根据音节的组合规则，究竟有多少规范音节？即藏文理论音节及其数值为几何？任何一个藏文语法书上都找不到答案。但这却是藏文信息处理研究的一个最基础的语言知识问题，也是藏语语言教学中无法避开的基础问题。我们统计并分析了音节频度、累计频度、信息熵以及频级关系等等，其中获得的高频数据基本准确和合理，在此基础上，进一步改进研究方法、增加语料覆盖范围，使其统计结果更加真实和普遍。然而，在我们的研究工作中，当增加语料数量时高频音节分布情况变化不大，但中间频率的音节有一定幅度的变化，低频音节数还会有少量增加。一方面表明当语料数达到一定程度时高频音节种数不再因为语料数量的增加而变化，其分布呈现稳定的较为缓慢的上升态势，另一方面说明语料库还不能完全满足统计的需要，它不能使统计数据达到完全收敛或稳定的状态。因此我们另辟蹊径，在字丁数据结构描述的基础上，以字丁的叠加层数和其字根、上加字、下加字、元音的拉丁转写，以及字根与前加字、上加字、下加字和后加字、又后加字、元音的组合规则描述规范单字、双字、三字和四字音节算法，设计和生成了理论上的单字、双字、三字和四字规范音节，结果发现，藏文理论音节达 18088 个，但这并不表明这所有音节都参与藏文的词汇组织及语言表达。因此，我们必须要知道藏文现实音节情况。

本规范词典着重于藏文表意音节研究，也即在实际语料中出现的音节或人们正在使用的音节，在研究方法上采用依据研究藏文语法理论进行规

则分析、计算生成藏文规范的理论音节，通过语言自省经验和语料库分析验证获得有义项和具有词汇组织功能的藏文规范音节的跨学科研究方法。

藏文音节是指以藏文音节点为界限的音节，梵文音节也以音节点为界限，由于梵文没有严格的音节点来区分音节的规则，因此，在梵文音节中不但包括了单音节梵文，还囊括了部分多音节梵文。在本规范词典中，不管是梵文还是藏文，确定音节的主要依据是音节点。当然，音节点本身未参与字符统计数据。

在现实音节考察中，即便不断增加语料量，但音节种数的增量仍然非常缓慢，停留在5000多个音节，因此，我们采取提高词种数的办法专门建立相应语料库，随着词种数的大幅增加，音节种数也明显提高。

通过建立四种不同的语料库进行考察，最后我们得出的结论是，藏文规范音节总数为9111个，其中纯藏文规范音节8263个，梵文转写音节848个，这正是目前藏文的音节总的使用情况。

我们认为，通过语料库考察藏文音节，可以有两种不同的结论，即，动态频率与静态频率。在3000万字符通用语料库中考察到的5000多个音节既属于动态频率，而在36万多词种中统计出的音节则是静态频率。后者的优点在于每一个词只出现一次的情况下统计到的音节频率，反映了藏文音节在整个语言的词汇系统中所扮演的角色和主次序列，不仅音节种数大大增加，且在整个语言系统中具有一定的稳定性。因此，将词种语料库的统计作为规范音节的分布数据及使用度依据，其可信度和可靠性有了更好的保证。

目　　录

第一章　藏文音节频率词典

第一节　主题内容

关于藏文音节的数量，古代藏族学者也曾讨论过，并提出了一些基本的音节逻辑组合方式，但都没有一个确切的数据。近代以来，也有很多高僧大德对这方面进行过研究，其中最有代表性的当数色多活佛，他在《色氏语法》里较为详尽地描述了音节组合规则，堪称全面而深入。到了现代，藏文研究多侧重于语法和格以及音律研究，对于音节涉猎甚少，然而，仍有部分学者对此较为重视，其中已故原西北民族学院教授才丹夏茸大师是很具代表性的人物，他在《藏文文法详解》一书中，详细测算并得出可写可读的藏文音节有17532个。提出一种基于逻辑推理算法，即，先找出206个基本字，然后乘以4个元音和五个前加字、33个上加字、43个下加字以及9个后加字（其中a'后加字一般省略），虽然这个方法使该项研究较之以前有了一个数字结果，但生成能力太强，音节量过大，把很多废字死字也顺带囊括进来，藏文音节出现了“鱼龙混杂”现象。同时，还存在以下几个方面的问题：（1）以9个后加字为限，未能包括在有前加字的三字音节中有10个后加字的情况；（2）尚未添加新造字hpha及部分元音和后加字的组合；（3）有些字根不与元音组合的也包括在其中；（4）没有考虑翻译译音专写的部分字；（5）没有囊括古藏文文献中的元音反写字；等等。虽然前人做了很多工作，但都属于藏文可读写音节研究，并未区分表意和非表意音节。

第二节　藏文音节确定原则

藏文音节是指以藏文音节点为界限的音节，梵文音节也以音节点为界限，由于梵文没有严格的音节点来区分音节的规则，因此，在梵文音节中不但包括了单音节梵文，还囊括了部分多音节梵文。在本规范词典中，不管是梵文还是藏文，确定音节的主要依据是音节点。

本规范词典着重于藏文表意音节研究，也即在实际语料中出现的音节或人们正在使用的音节，在研究方法上采用依据研究藏文语法理论进行规则分析、计算生成藏文规范的理论音节，通过语言自省经验和语料库分析验证获得有义项和具有词汇组织功能的藏文规范音节的跨学科研究方法。

第三节　适用范围

本规范词典适用于藏文信息字典的建设、藏文基础词汇的研究和通用词表的建设、藏语教材建设、藏语分级词汇研究以及藏文智能输入法的开发、藏文正字软件的开发、藏文信息检索系统的开发等方面。

第四节　术语和定义

一　藏文音节

由藏文字丁构成的最小的语言基本单位。

二　藏文字丁

一个音节中的纵向单位（辅音字母或上下叠加的组合体）叫藏文字丁。

三　藏文理论音节

按照语法理论能够读写的音节叫做理论音节。通过生成统计，共有18088个。

四　藏文规范音节

符合音节生成的规律并且具有表意或词汇组合作用的音节叫做藏文规

范音节。

五　藏文字符层数

表示包括元音的叠加层数，简记为：　字符层数或 int iNumber

六　频次

指调查对象在调查语料中出现的次数。

七　频率

指的是某一调查对象的频次与整个语料所含调查对象总频次的比值。计算公式：其公式为：Fi=Ni/N*100%。其中 Fi 为调查对象 i 的频率，Ni 为调查对象 i 的出现次数，N 为语料中调查对象出现的总次数。它能说明调查对象的使用度。

八　累计频率

每一个对象与其前面对象的频率之和。在全部语料中所占的比重。一个对象的总出现次数在语料调查范围内所占的比重。覆盖率有着同义作用。

九　信息熵

对信息源 X 的各符号的自信息量取统计平均，可得每个符号的平均自信息量为：

$$H(X)=-\sum_{i=1}^{m}p(a_i)\log_2 p(a_i)$$

这个平均自信息量 $H(X)$ 称为信息源 X 的熵(entropy)，单位为 bit/符号。

第二章　藏文音节组合规则

第一节　藏文规范单字音节

单字规范音节有以下4类。特别说明：除字根外的其它如前加字、后加字、次后加字、上加字、下加字等用字母表示时不加元音“a”。

一　单层单字音节

藏文30个辅音字母。

二　双层单字音节

（一）30个辅音字母与4个元音的组合；

（二）“字根＋下加字”的结构形式：

1. 字根为ka、kha、ga、pa、pha、ba、ma，下加字为y；

2. 字根为ka、kha、ga、ta、tha、da、na、pa、pha、ba、ma、sha、sa、ha，下加字为r；

3. 字根为ka、ga、ba、ra、sa、za，下加字为l；

4. 字根为ka、kha、ga、ca、nya、ta、da、tsha、zha、za、ra、la、sha、sa、ha，下加字为w。

（三）“字根＋上加字”的结构形式：

1. 字根为ka、ga、nga、ja、nya、ta、da、na、ba、ma、tsa、dza，上加字为r；

2. 字根为ka、ga、nga、ca、ja、ta、da、pa、ba、ha，上加字为l；

3. 字根为ka、ga、nga、nya、ta、da、na、pa、ba、ma、tsa，上加

字为 s。

三　三层单字音节

（一）“字根＋上加字＋下加字”的结构形式：

1. 字根为 ka、ga、ma，下加字为 y，上加字为 r；

2. 字根为 ka、ga、pa、ba、ma，下加字为 y，上加字为 s；

3. 字根为 ka、ga、pa、ba、ma，下加字为 r，上加字为 s；

4. 字根为 ts，上加字为 r，下加字为 w。

（二）“字根＋上加字＋元音”的结构形式

元音为 i：

（ i ）字根为 nya、ta、da、ma、tsa、dza、ja，上加字为 r；

（ ii ）字根为 ca、ja、ta、da，上加字为 l；

（iii）字根为 ga、nya、ta、da、ma，上加字为 s。

元音为 u：

（ i ）字根为 ka、ga、nga、ja、ta、da、na、ma、tsa、dza，上加字为 r ；

（ ii ）字根为 ka、ta、da、ba、ha，上加字为 l；

（iii）字根为 ka、ga、nga、nya、ta、da、na、pa、ba、ma，上加字为 s。

元音为 e：

（ i ）字根为 ka、ga、ja、nya、ta、ma、tsa、dza，上加字为 r；

（ ii ）字根为 ca、ta、da、ha，上加字为 l；

（iii）字根为 ka、ga、nya、ta、da、na、pa、ba、ma，上加字为 s。

元音为 o：

（ i ）字根为 ka、ga、nga、ja、nya、ta、da、na、ba、ma、tsa、dza，上加字为 r；

（ ii ）字根为 ka、ca、ja、ta、da、ha，上加字为 l；

（iii）字根为 ka、ga、nga、nya、ta、da、na、pa、ba、ma、tsa，上加字为 s。

（3）“字根＋下加字＋元音”的结构形式

元音为 i：

（ i ）字根为 ka、kha、ga、pha、ba、ma，下加字为 y；

（ ii ）字根为 ka、kha、ga、tha、ta、da、pa、pha、ba、ma、sa、ha，下加字为 r；

（iii）字根为 ga、ra，下加字为 l。

元音为 u：

（ i ）字根为 ka、kha、ga、pha、ba、ma，下加字为 y；

（ ii ）字根为 ka、kha、ga、da、pha、ba、sa、ha，加字为 r；

（iii）字根为 ka、ga、ba、za、ra、sa，加字为 l。

元音为 e：

（ i ）字根为 ka、kha、ga、ba、pha、ma，加字为 y；

（ ii ）字根为 ka、kha、ga、pha、ta、da、ba，加字为 r；

（iii）字根为 ka、ga、sa，加字为 l。

元音为 o：

（ i ）字根为 ka、kha、ga、pa、pha、ba、ma，加字为 y；

（ ii ）字根为 ka、kha、ga、ta、da、pa、pha、ba、sa，加字为 r；

（iii）字根为 ka、ga、ba、za、ra、sa，加字为 l。

（4）字根为 ga，下加字为 r、再下加 w。

四　三层单字音节

（一）“字根＋上加字＋下加字＋元音”的结构形式：

元音为 i：

（ i ）字根为 ma，上加字为 r，下加字为 y；

（ ii ）字根为 ka、ga、pa、ba、ma，上加字为 s，下加字为 y；

（iii）字根为 ka、ga、pa、ba、ma，上加字为 s，下加字为 r。

元音为 u：

（ i ）字根为 g，上加字为 r，下加字为 y；

（ ii ）字根为 ka、ga、pa、ma，上加字为 s，下加字为 y；

（iii）字根为 ka、ga、pa、ba、na，上加字为 s，下加字为 r。

元音为 e：

（ⅰ）字根为 ka，上加字为 r，下加字为 y；

（ⅱ）字根为 ka、ga，上加字为 s，下加字为 y；

（ⅲ）字根为 ga、na、pa、ba、ma，上加字为 s，下加字为 r。

元音为 o：

（ⅰ）字根为 ka、ga，上加字为 r，下加字为 y；

（ⅱ）字根为 ka、ga、pa、ba、ma，上加字为 s，下加字为 y；

（ⅲ）字根为 ka、ga、pa、ma、na，上加字为 s，下加字为 r。

第二节　藏文规范两字音节

规范双字音节可分为以下 4 类。

一　单层平行两字音节

两个字符层数都为 1，第一个是 30 个辅音字母之一，第二个是除 v 以外的 9 个后加字之一；特别：la 后不能加 la（注：后续文本中为便于理解，后加字字母后不加元音 a）。

二　前加字与双层字符

首字符为 5 个前加字之一，第二个字符为 2 层，其中：

（一）前加字为 b：

1. 字根为 ta，上加字为 l；
2. 字根为 ka，下加字为 r；
3. 字根为 ra，下加字为 l；
4. 字根为 da，上加字为 r。

（二）前加字为 v：

1. 字根为 pha，下加字为 y；
2. 字根为 da、pha，下加字为 r。

（三）前加字为 g：

第二个字是“字根+元音”且字根为 ca、ta、da、sa、zha、za、ya、sha；

（四）前加字为 d：

字根为 ga，下加字为 r。第二个字是“字根+元音”且字根为 ga

（五）前加字为 m：

第二个字是“字根＋元音”且字根为 kha、ga、cha、ja、nya、tha、da、na、tsha、dza。

三　前加字与三、四层字符

首字符为 5 个前加字之一，第二个字符层数大于等于 3，其中：

（一）前加字为 b：

1. 第二个字是“字根+上加或下加＋元音”且字根为 ja、dza、ka、ga、nga、ta、da、sa、za；特例：前加字 b 与字根 da 结合时不带下加字。

2. 第二个字是“字根+上加+下加＋元音”且字根为 ka、ga。

（二）前加字为 d，第二个字是“字根+下加＋元音”，字根为 ka、ga、pa、ba、ma。

（三）前加字为 m，第二个字是“字根+下加＋元音”，字根为 kha、ga。特例：前加字 m 与字根 da 的搭配不带下加字 l。

（四）前加字为 v，第二个字是“字根+下加＋元音”，字根为 kha、ga、pha、ba、da 且不带下加字 l。

四　首字丁两层及以上

第一个字符层数大于等于 2，下加字不是 w，第二个字符为除 v 以外的 9 个后加字之一；当字根为 ha、后加字为 ng 时字根可带下加字 w；字根为 ga，下加字为 r、再下加 w 时，可以带粘着虚词 s 和 r。

第三节　藏文规范三字音节

规范三字音节有以下 3 类。

一　单层平行三字音节

（一）第一个字是 20 个纯基字，第二个字是 g、ng、b、m 之一且又后加字是 s；

（二）第一个字为 5 个前加字之一，第三个字为 10 个后加字之一：

1. 前加字为 b，字根为：ka、ga、ca、ta、da、tsa、zha、za、sha、sa；

2. 前加字为 v，字根为 kha、ga、cha、ja、tha、da、pha、ba、tsha、dza；

3. 前加字为 g，字根为 ca、nya、ta、da、na、tsa、zha、za、ya、sha、sa；

4. 前加字为 d，字根为 ka、ga、nga、pa、ba、ma；

5. 前加字为 m，字根为 kha、ga、nga、cha、ja、nya、tha、da、na、tsha。

二　首字符层数大于一

首字符层数大于一，下加字不是 w，第二个是 g、ng、b、m 之一，第三个是 s。特殊：当第一个字符是 dwa，后加字是 g、ng，又后加字为 s。

三　首字符为前加字

第二个字符层数大于 1，第三个字为 9 个后加字之一：

（一）前加字为 b：

1. 第二个字是单字字根 ka、ga、ca、ta、da、tsa、zha、za、sha、sa 与元音的组合。

2. 第二个字是上加字 r 与字根 ka、ga、nga、ja、nya、ta、da、na、tsa、dza 的组合；上加字 l 与字根 ta、da 的组合；上加字 s 与字根 ka、ga、nga、nya、ta、da、na、tsa 的组合。

3. 第二个字是字根 ka、ga 与下加字 y 的组合；字根 ka、ga、sa 与下加字 r 的组合；字根 ra、sa 与下加字 l 的组合。

4. 第二个字是上加字 r 与字根 ka、ga 及下加字 y 的组合；上加字 s 与字根 ka、ga 及下加字 y 的组合；上加字 s 与字根 ka、ga 及下加字 r 的组合。

（二）前加字为 v：

1. 第二个字是单字字根 kha、ga、cha、ja、tha、da、pha、ba、tsha、dza 与元音的组合。

2. 第二个字是字根 kha、ga、pha、ba 与下加字 y 的组合；字根 kha、ga、da、pha ba 与下加字 r 的组合。

（三）前加字为 g：

第二个字是字根为 ca、nya、ta、da、na、tsa、zha、za、ya、sha、sa 与元音的组合。

（四）前加字为 d：

1. 第二个字是单字字根 ka、ga、nga、pa、ba、ma 与元音的组合；

2. 第二个字是字根 ka、ga、pa 与下加字 y 的组合，或者字根 ka、ga、pa、 ba 与下加字 r 的组合。

（五）前加字为 ma：

1. 第二个字是单字字根 kha、ga、nga、cha、ja、nya、tha、da、na、tsha 与元音的组合；

2. 第二个字是字根 kha、ga 与下加字 y 或 r 的组合。

第四节 藏文规范四字音节

第一个字是 5 个前加字之一，后加字为 g、ng、b、m 而又后加字为 s：

一 前加字为 ga

第二个字为单字字根 ca、nya、ta、da、na、tsa、zha、za、ya、sha、sa 及这些字根与元音的组合。

二 若前加字 da

（一）第二个字为单字字根 ka、ga、nga、pa、ba、ma 及这些字根与元音的组合；

（二）第二个字为字根 ka、ga、pa、ba、ma 与下加字 y 的组合；或字根 ka、ga、pa、ba 与下加字 r 的组合。

三 前加字为 ba

（一）第二个字为单字字根 ka、ga、ca、ta、da、tsa、zha、za、sha、sa；

（二）单字字根 ka、ga、ca、ta、da、tsa、zha、sha、sa 与元音的