藏语文-音自动规则转换及其实现

Rules for the auto-transformation of Tibetan text to IPA

李永宏

Li Yonghong

摘要：为满足语言学、音韵学和工程语音学的需要，该文根据现代藏文与 3 大方言语音之间的对应规律和藏文正字法，提出了从文字上对藏文声母和韵母拆分的“字丁分解法”，实现了藏文到各方言国际音标的自动转换。并对算法和实现过程进行了详细的阐述，建立了藏语 13 个方言点的方音数据库。方音数据库的建立为藏语方言研究和语言教学提供了科学、方便的工具，为藏语标准音的制定、推广及应用提供原始的语音材料，也能作为藏语语音识别和语音合成的标音基础。

Abstract：Modern Tibetan orthography and the sound rules relating between written text and Tibetan dialects were analyzed to develop rules for the transformation of Tibetan text to IPA for linguistics, phonology and engineering phonetics analyses. The transformation separates the Tibetan single syllable words into initial and final parts and then uses an automatic IPA transformation from the Tibetan text to the speech sounds of the dialects. A corpus of thirteen Tibetan dialects has been automatically developed. The transformation system provides a convenient tool for Tibetan dialect study an language teaching, for establishing original speech materials, for improving Standard Tibetan tongue an for engineering phonetics study.

关键词：藏文信息处理 藏语方言 国际音标 藏语文-音转换

Key words： Tibetan information processing; Tibetan dialect; International Phonetic Alphabet; Tibetan text to IPA transformation

0 引言

藏语属于汉藏语系藏缅语族藏语支，主要分布在中国西藏自治区和四川、云南、青海、甘肃等 5 省区，使用人数达 500 多万，有古老的拼音文字及浩瀚的文献典籍[1]。 在世界范围内，尼泊尔、不丹、巴基斯坦、印度等国家的境内也有一部分地区使用藏语。

在藏文初创时，以藏语口语为基础，在语音上，严格按照一字一音的原则，准确标记；在语法和词汇上，以口语为规范书写。随着语言的发展，藏文和口语失去严格对应，字母的标音功能减弱，不同地区的藏语朝着各自不同的方向发展变化，形成了各具特色的方言土语。

国内学术界将藏语主要分卫藏、安多和康 3 大方言。它们之间的差别主要表现在语音方面，如有无声调、有无清浊声母的对立、辅音韵尾的多寡等，其次是词汇和语法的差别。语音感知和科学研究的结果表明，卫藏方言与安多方言在“两端”，康方言介于中间。各方言的发展速度也不尽一致，卫藏方言最快，次为康区方言，安多方言发展最慢。安多话是藏语中保留古面貌较多的藏语方言，有很多特殊的语言现象，例如：音调不具有区别词义的功能，音节只有习惯调；有较多的复辅音等。虽然 3 大方言在口语上有较大差距，但与文字各有严格的对应规律，因此，各方言都可以拼读文字[2]。

由于藏语没有类似汉语普通话的口语标准音，所以各方言之间交际有一定的困难，这成为困扰一代又一代藏学专家的难题。虽然国内提出藏语标准语的方案很多，但都难以完全达到一致，而藏语各方言内部的语音特点和方言之间的语音差异的研究为科学、合理的制定藏语标准语提供了依据。此外，本文提出的文-音转换方案不仅能够为音韵学的研究提供方便的工具，也能作为藏语语音识别和语音合成的标音基础。

本文重点对藏语单音节词从文字到不同方言语音的转换进行了详细阐述，并在 windows 系统上进行了程序实现，建立了藏语方音数据库。藏语采用基于大字符集编码系统，国际音标采用基于 Unicode 的编码系统，藏语方言语音主要采用北大孔江平教授实地调查的藏语方言数据库（泽库和红原除外），方言点包括：拉萨、日喀则、德格、巴塘、泽库、夏河、同仁、循化、

化隆、红原、天峻、道孚、阿柔，共 13 个点。藏语词表来源于《藏汉大字典》、《安多口语字典》、《拉萨口语字典》、《格西曲扎藏文辞典》、《新编藏文字典》、《藏文同音字典》、《藏语文课本（小学 12 册、初中 6 册、高中 6 册）》6 部藏文字典和 24 册藏语文课本，共 13 万余条词汇。

根据藏语内在特点和程序处理中的实际需要，本文在已有的术语基础上[3]，约定了一些新的术语，例如“基字丁”，“带音基字丁”，“不带音基字丁”等。

1 藏语音节的基本概念

藏文是在梵文天成体的基础上发展而成的一种拼音文字，共有 30 个辅音字母，4 个元音符号，其中/a/为零位。藏文有一套严格而完整的字母组合排列规则，它的字符流是两维呈现，自左向右横向书写，传统藏文文法根据字母在音节中的结构位置，将字母分为“基字”、“上加字”、“下加字”、“前加字”、“后加字”和“再后加字”，基字为整个藏字的核心，30 个辅音字母都可以做基字，元音不能做基字，其中鈲[i]，鈴[e]，鈉[o]，加在基字丁上面，鉚[u] 加在基字丁下面，不带元音符号的默认为是元音[a]，30 个辅音字母中有 5 个可做前加字（“ག་ད་བ་མ་འ”），3 个上加字（“ས་, ་ར་, ་ལ”），4 个下加字（“ཡ་, ་ར་, ་ལ་, ་ཝ”），10 个后加字（“ག་ང་ད་ན་བ་མ་འ་ར་ལ་ས”），再后加字是后加字中的“ད”（今不用）和“ས”，当再后加字出现时，后加字和再后加字的组合在现代藏文中只有四种形式（“གས་ངས་བས་མས”），否则不符合藏文的正字法[4]。

藏文大字符集编码方案在计算机中是以上下叠加的字母为一个整体进行编码的，称之为字丁，例如“བསྒ”就是一个带有上加字（ས）和下加字(ར)和元音(ེ)的字丁。藏文音节包含的字丁数，称之为位长，一个藏语音节位长最小值为 1，最大值为 4，例如“བསྒྲབས”位长为 4。包含有基字的字丁，称之为“基字丁”，带元音的基字丁称为“带音基字丁”，不带元音的基字丁称为“不带音基字丁”，实际藏文中所有的基字丁都是带音基字丁(没有元音符号表示带元音[a])，不带音基字丁是在字丁分解中，为了研究方便而引入的。

2 国际音标及其拉丁转写的实现

为了能够得到较为全面的藏语词汇，经过长达一年的词典录入和校正，共收录了 6 部藏文字典和 24 册藏语文课本，总词汇达 13 万余条，查重后得到 9 万余条藏文词汇，其中单音节（不包括梵音藏文和借词）有 0.5 万余条，占到总词汇量的 5.6%；双音节词汇有 4.3 万余条，占到总词汇的一半；三音节词汇大约有 2 万余条，四音节词汇大约有 1.6 万余条，四音节以上词汇大约有 0.5 万余条，如图 1 所示。

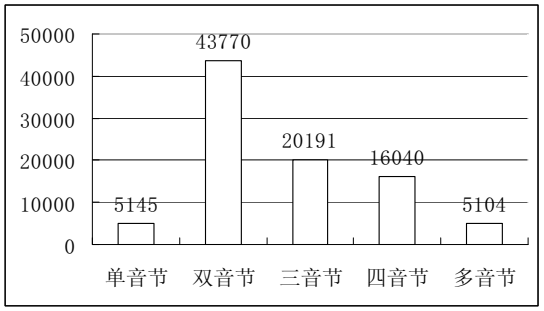

图 1 音节数量分布图

无论是词汇还是篇章，文字到国际音标转换的基础是藏文单音节的转换，转换过程如图 2 所示。核心思想为：对拟要转换的单音节藏文进行声韵母的分离，对分离的声母和韵母分别进行音标转换，然后合并生成国际音标，对有声调的方言还需要加上声调。程序中需要已经处理好的字丁分解表、声母音标转换表、韵母音标转换表的支持。

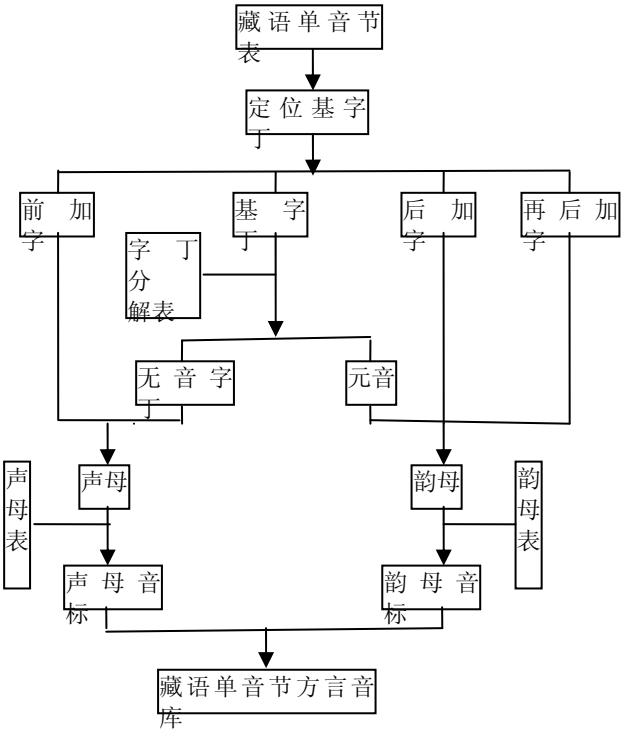

图 2　单音节音标转换框图

2.1　基字丁的定位

根据藏语的拼写方式，元音是叠加在基字丁上的，要完成藏语的声韵母分离，首先要正确的找到基字丁，这也是整个转换过程中最关键的一步。

虽然藏文有一套严格而有规律的字母组合规则，但是在一些特殊的组合形式中，并不能依靠规则直接判断，存在着歧异。例如：前加字+（'གས','ངས','བས','མས'），首字丁和中字丁都符合做基字的条件，但显然只有一种情况是正确的。因此本文按照音节位长的不同，考虑藏语单音节词的所有的字母组合类型，分别进行基字丁的判断。

类型	表现形式	基字丁	例词
单字丁	单个字丁	1	ཀ
双字丁	前加字+非后加字&非“ལེ་ཀྱ་ཏེ”	2	གཡུ
	前加字+后加字	1	གབ
	非前加字+后加字	1	དེད
	首字丁+“ལེ་ཀྱ་ཏེ”	1	ཀེའེ
三字丁	前加字+非（'གས','ངས','བས','མས'）	2	བཀྲོན
	前加字+（'གས','ངས','བས','མས'）	2	འབགས
		1	གངས
	非前加字+（'གས','ངས','བས','མས'）	1	ཡངས
	前加字+中字丁+“ལེ་ཀྱ་ཏེ”	2	གཁའེ
四字丁	前加字+基字丁+（'གས','ངས','བས','མས'）	2	གཡབ གས

表 1　藏语整字基字丁的确定

注：1 表示基字丁为第 1 个字丁；2 表示基字丁为第 2 个字丁。

2.2　基字丁分解

在藏文音节中找到基字丁后，利用藏文字丁拆分表，把基字丁包含的辅音成分和元音成分分开。基字丁分解表包括 550 多个现代藏文字丁，其中不带元音符号的字丁，本身带有[a]音，因此把字丁分解为不带音基字丁（无音字丁）和 5 个元音，ཨ[a]、ཨི[i] 、ཨུ[u] 、ཨེ[e] 、ཨོ[o]，如表 2 所示。

表 2　藏文字丁拆分表

字丁	无音字丁	元音	字丁	无音字丁	元音
ཀ	ཀ	ི	ཀེ	ཀ	ེ
ཀོ	ཀ	ོ	ཀཻ	ཀ	ཻ
ཀུ	ཀ	ུ	ཀཽ	ཀ	ཽ
ཀྲ	ཀ	ྲ	ཀྲཱ	ཀ	ཱ
ཀླ	ཀ	ླ	ཀླཱ	ཀ	ཱ

2.3 音节拆分

把藏文单音节基字丁中包含的辅音成分和元音分开，基字丁前面如果有字丁就是前加字，后面有字丁就是后加字和再后加字，这便得到单音节拆分表，前加字、后加字、再后加字都可以为空，如表 3 所示。

藏文	前加字	基字丁			后加字	再后加字
		带音	不带音	元音		
བསྒྲུབ	བ	སྒྲ	ས	ུ	བ	
གཡོག	ག	ཡ	ཡ	ོ	ག	ག
བདེབས	བ	དེ	ད	ེ	བ	ས

表 3　藏语单音节拆分表

2.4 藏文声韵母组合

藏语和汉语类似，也是单字单音的声韵母结构。对预转换的藏文查找藏语单音节拆分表，就很容易从文字上进行声韵母分离了，结果如表 4 所示。

藏文声母=前加字+不带音基字丁

藏文韵母=元音+后加字+再后加字

现代藏语从文字上来看，声母共有 213 个，韵母有 77 个[1]，理论上声韵母组合能产生 16401 个单音节藏文，而实际存在的藏字大约占理想组合的 1/3。

序号	藏文	汉意	藏文声母	藏文韵母
1	བསྒྲུབ	力量	བསྒྲ	ུབ
2	གཡོག་ས	靠	གཡ	ོག་ས
3	བདེབས	盖子	བད	ེབས

表 4　藏文声韵母分离表

2.5 声调的处理

由分开的藏文声韵母，分别查找声母音标对照表和韵母音标对照表，便能得到藏语声母和韵母各自对应的国际音标，然后直接组合就能得到整个音节对应的国际音标。但是，对有声调的藏语方言来讲，其音节的实际调值与声母和韵母都有直接的关系，需要在声母音标对照表和韵母音标对照表中进行标识。

以藏语拉萨话为例，声调有六个，用五度标调法记音，应记为 55，114，52，13，51 和 132。但实际上 51 调和 132 调有“ʔ”韵尾，因此藏语拉萨话的声调也可以处理成 4 个，以宽式记音可标作：55，14，53（52 和 51）和 12（12 和 132）。拉萨话的声调有长短之分，长韵母为长调 55 和 14，短韵母为短调 53 和 12；也有高低之分，清声母音节为高调 55 和 53，浊声母音节为低调 14 和 12[5] [6] [8]，如表 5 所示。

	声母		韵母	
	清	浊	短韵	长韵
形式	古藏文清声母	古藏文浊声母	vp、v、$\tilde{v}^{ʔ}$、$v^{ʔ}$、$vm^{ʔ}/vŋ^{ʔ}$	vr、vm/vŋ、v:、$\tilde{v}$:、vv:
调号	1	2	3	4
调型	高调	低调	短调	长调
调值	55,53	14,12	53,12	55,14

表 5　藏语拉萨话声调表（一）

注：调号为方便理解和程序处理而引入。

以上分析可知，藏语拉萨话单音节声韵母组合调号共有 4 种类型，其对应的实际调值，如表

6 所示。

韵 调	清声 短韵	清声 长韵	浊声 短韵	浊声 长韵
调号	13	14	23	24
调值	53˥	55˥	12˩	14˩

表 6　藏语拉萨话声调表（二）

注：藏语有些方言词汇调值不稳定用“00”来表示。

藏语其它方言的声调相对要复杂的多，黄布凡在《藏语方言声调的发生和分化条件》(1994)中提到，藏语有声调方言都经过有自然声调阶段，自然声调起源于声母和韵尾的附带特征。音位声调起源于声母和韵尾的演变导致自身辨义功能的减弱和转移。各方言声调分化并非都是清高浊低，而是条件各异，自成系统[7]。这就为声调的自动标注带来很大的困难，还需要对每一个方言点的声调做系统的分析，才能够更科学、合理的表怔藏语在不同方言中的实际的声调。

2.6　音标转换

根据得到的藏文声母和韵母，查找声韵母音标对照表（参考文献[1]），得到各自对应的音标和调号。

声母音标对照表包含现代藏文的 213 个声母，字段有声母对应的拉丁转写，包括拉萨、日喀则、德格、巴塘、泽库、夏河、同仁、循化、化隆、红原、天峻、道孚、阿柔，共 13 个点的国际音标，字段包括结构藏文声母、拉丁转写、每个方言点的音标和调号，其中安多方言内的点没有声调字段。韵母音标对照表包含 77 个藏文韵母，字段同声母音标对照表。

藏文单音节音标直接由声韵母对应的音标叠加得到，音节的实际调值由对应的声韵母的调号进行组合，然后查找表 6 就得到对应的 5 度调值。

拉丁转写=声母拉丁转写+韵母拉丁转写

音节音标=声母音标+韵母音标+声调

经过以上步骤，最后得到藏语单音节方言音库，如表 7 所示。

如果需要转换的是多音节藏语词汇或语篇，会带来更为复杂的问题。首先根据音节点，把多音节切分为单音节，对每个单音节分别转换成各自的国际音标。此时，连续变调成为我们考虑的首要问题，拉萨话的连续变调规则研究的比较多，规律也相对清楚，康方言的连续变调比较复杂，很多调值变化并不稳定，需要进一步的做深入调查和分析；安多方言牧区话和农区话差别比较大，传统说法认为安多方言没有区别意义的声调系统，但安多方言习惯调的内部规律，还在进一步的研究中。

2.7　结论

我们可以认为古藏文就是古代藏文的实际音值，字音转换是根据藏语不同方言的现代拼读法与古代藏文的对应关系建立的现代藏语方音数据库。从转换的结果来看，首先，安多方言由于其音系相对复杂，声韵母数量较多，其音节数量在 1600 左右，同音词概率为 5145/1600=3.2，有些同音词的习惯调值并不完全相同，还需要做进一步的研究，卫藏和康方言音节数量大致都在 1200 左右，但因其都有声调，所以同音词的概率要远远小于安多藏语，而且声调在不同的方言内部，分布也有很大的差异。其次，个别词汇的实际音值和利用规则转换的国际音标之间并不是完全相同，也就是说某一方言内部的同一声（韵）母在不同的词汇中其发音并不相同，其原因主要有以下几点：

（1）由于其它语种或藏语其它方言的影响，造成部分词汇发音并不符合大的系统性规律，属于语言接触的范畴。

（2）语音的演变并没有系统性的完成，有些内部词汇还保留其演变前的发音。

（3）由于藏语方言发展不平衡，藏语方言内部的部分借词，其发音并不遵从本语言点的演变规律和拼读规则。

藏语方言内部的复杂性也证明了语言发展是有内部层级性的，这就给语言工作者提出了更高的要求，必须根据研究的目的对藏语进行多角度，多层次的综合分析。

3　结束语

本文通过对藏语 3 大方言 13 个点的语音数据采集和音系整理，6 本藏文字典和 24 册藏语文课本的统计分析，结合藏文在计算机内部的处理机制，提出了“字丁分解”的藏文-音转换方法，建立了 500 多个藏语字丁的分解表，213 个

藏文声母音标对照表，77 个藏文韵母音标对照表，完成了 5145 个藏语单音节文字到语音的自动转换，建立了藏语方言语音数据库。

本文的研究成果对藏语语言学研究、语言文字教学和藏文信息处理领域的许多方面，具有重要的学术价值和广泛的应用价值。

（1）利用计算机能够识别的符号表达文字和语音信息，便于计算机进行存储、传递和数据处理，有效地保护民族语言文化。

（2）为藏语方言研究、语言教学提供了科学、方便宜用的工具。

（3）在文字和语音的标注上，提供一致的平台和符号，便于研究人员的阅读和相互交流。

（4）为语音的韵律特征分析和包括识别、合成的工程领域的研究，提供字音转换功能。

（5）为藏语标准音的制定、推广及应用研究提供原始语音材料。

一方面为了能够更好的完善藏语方音数据库，还需要进行藏语方言的调查，陆续的加入更多的方言点；另一方面，对调查的方言语料进行声学分析，建成藏语方言语音声学参数数据库，研究藏语方言的发展和内部差异，更好的为语言学，语音学研究，甚至工程语音学服务。

参考文献

[1]格桑居冕，格桑央京．藏语方言概论[M]．北京：民族出版社，2002

[2]金鹏．藏语简志[M]．北京：人民出版社，1983

[3]于洪志．计算机专用藏文文字术语探讨[J].，《术语标准化与信息技术》，1997，3：26-27

[4]周季文．藏文拼音教材[M] 北京：民族出版社，1982

[5]胡坦．藏语(拉萨话)的声调研究[J]．民族语文，1980，3：22：36

[6]谭克让，孔江平．藏语拉萨话元音、韵母的长短及其与声调的关系[J]．民族语文，1991，2：12：21

[7]黄布凡．藏语方言声调的发生和分化条件[J]．民族语文，1994，3：1：9

[8]孔江平．藏语（拉萨话）声调感知研究[J]．民族语文，1995，3：56-64