

语篇索引技术在藏语文本处理中的应用*

江 荻

中国社会科学院民族所计算语言学室

摘要 本文阐述了语篇索引技术在语料库处理中的基本功能，并在我们自行开发的 CATC 系统中实现了对藏语现在时动词及其语尾标记的分析，提出藏语动词按其内在时间结构分类的观点。

关键词 语篇索引 藏语动词 语尾标记

The Application of Concordance Technology to Tibetan Corpus

Jiang Di

Department of Computational Linguistics,

Institute of Nationality Studies, Chinese Academy of Social Sciences

Abstract: This paper introduces the primary functions of concordance technology, and analysis the relations between Tibetan verbs and their ending markers in our CATC software system. As a result, we propose a view that Tibetan verbs can be classified according to their inherent temporal features.

Keywords: Concordance Technology Tibetan verbs Ending Markers

一、语篇索引技术概要

语篇索引是一种非常有效的语言分析技术。语篇索引早期仅用于词典编纂，我国前辈学者曾巧妙地音义兼译为“勘考灯”（concordance）¹。近些年来，随着计算机语料库

* 本研究获中华社科基金和中国社会科学院重点课题基金资助。

¹ 承蒙胡坦教授相告，特此致谢。

结点重合无持续意义的时点结构，称为分时时状 (|++++|) 如 so sovi sa gnas khag la vkhrab kyi yog red “各自在当地演出”，意为“某甲在甲地演出”，“某乙在乙地演出”。如：

lo gcig dus bzhi tshang mar gangs babs babs par sdod kyi yog red. 一年四季都下雪。(分时)
 nga tsho nyin re las nyin re yag ru yag ru vgro gi yog red. 我们一天一天好起来。
 deb vdir sgor mo gnyis gnas kyi yod pa red. 这本书值 2 元钱。(关系)

本项实验充分利用语篇索引 CATC 功能对藏语动词句法展开全新的研究，首次提出按动词内在时间结构分类的观点，并给予了论证。另外，在全部提取的表示现在时的动词句中，例外的情况极少，而且主要集中在关系类动词上。示例见表 4。

表 4：动词类别与语尾标记的对应关系示例

kho ngavi nang la kha lag za zam pa des ldem byed ngavi grogs po de lha sar las ga byed nga nam rgyun gnam grur bsdad byas yong nga tshos mes rgyal la dgav zhen byed	gi yod. gi yod. gi yod. gi yod. kyi yod.	他正在我家吃饭。 那桥在颤摇。 我那个朋友在拉萨工作。 我一般坐飞机回来。 我们热爱祖国。
rmugs pa vthib bzhas, nga tshos gangs ri mthong khas sa ngas bshad pavi skad cha khyod rang da dran rlung tshub gchig lang blo bzang pha yul dran char pa tog tsa bab kha sa chu tshod brgyad par vdir lha mo vkhrab	gi mi vdug. gi vdug gas? gi vdug. gis(gi vdug). kyi vdug. kyis(gi vdug)	太暗了，我们看不见雪山。 昨天我的话现在还记得吗？ 刮起一场暴风雨。 洛桑想家（乡）。 正在下小雨。 昨天八点，这儿正在演戏。
lung pa vdivi ming la ga re zer byas tsang sa cha de gyad la bod ljongs gyi vbru mdzod zer sa gdan de tsho phyi rgyal la mang po btsong da dung sa cha kha shas la lo gcig la lo tog thengs gnyis bsdu	gyi yod red? gyi yog red. gi yog red. gi yog red.	这个地方叫什么？ 那些地方称为西藏的粮仓。 那些地毯大量销往国外。 还有的地方一年收两季呐。

在现代藏语动词研究中，学者们一般采用内省的方法进行语法及结构的研究。认为藏语动词作谓语涉及时、体、态、人称等概念，并归纳出动词的自主性、意志性、可观察性、使动/自动、及物/不及物、亲知/非亲知、新知/引述、ergativity 等范畴。然而，关于这些范畴的形式和功能的描述尚不精细，例外很多。特别是，人们一直无法开展句法研究的量化分析，缺乏验证过程。本项研究从真实文本中提取出带现在时动词语尾的句子 91 例，分析出的例外情况仅 7 例，只占 7.7%。因此，可认为动词语尾是不同动词类别的形式标记。

表 5：动词类别与语尾标记的对应关系

	gi yod		gi vdug			gi yog red	
动词类别	动态		动态	心理	变化	关系	动态
谓语时状	过程	事件	过程（他称）	静态	状态	匀质	分时
时间结构	-----	~~~~	-----	-----	-----	-----	++++

实际上，我们也可以用动词做关键词来进行语篇索引，这样，语尾标记将表现出某

种与动词一致的趋向性，形成某种语法上的搭配关系。例如动词 vjug（下面以过去式 bcug 形式为例）接在另一个动词后表示使动时通常要在二者之间加上 ru/du，形成某种“动词+ru+bcug”格式，有“使做...，让做”的意思。但不加 ru 的现象也是存在的。另外，对所统计的句子做右索引发现语尾标记有 pa red 而没有 pa yin，这又意味着该类句式与语法的人称范畴和动词的自主/不自主范畴有关系。

表 6: 从动词索引上下文搭配关系

	gar vkhrab	bcug	gnang pa de dang dmangs khrod kyi
kho ra -vi skra la rgyug shad rgyag ru		bcug	nas rgyu shad brgyab
	ud yin pa shes	bcug	nas tshang mas kho la ngo rgol byas
vi sdod sar sa dong rim pa dgu vdru ru		bcug	pa dang
byas rta rgyug vgran bsdur byed ru ma		bcug	pa red .
khams gling sde zer sa der vkhrung ru		bcug	pa red . lha sras de
sprevu de rtsed mo sna tshogs rtse ru		bcug	pa red . snang sa
	cha de vi bdag po byed	bcug	pa red . ge sar gyis bdud rje btul
gzhan dag gi nang la ngal gso rgyag ru		bcug	. bu dang pha ma

三、语篇索引技术的多维性

语篇索引技术涉及语言研究的许多方面。以根词化 (lemmatization) 技术来说，由于藏语具有较为丰富的黏着型词缀，如表示施动、受动、领属、工具等的部分名词格词缀黏着在根词上，因此，做文本分析时就必须建立词表的匹配功能，使之凸显出来进行统计和分析。如，rgyal po “国王（原形）”，rgyal povi “国王的（属格）”，rgyal pos “被国王（施动格）”，rgyal por “向国王（业格表趋向）”等都是同一根词的语法形式。

词表 (wordlist) 技术是通过比较两个词表来描绘文本的特征。其中一个为参照词表，来源于大规模语料统计后逐步定型的熟词表，另一个是研究者从文本中提取的供研究的词表。研究词表中获取的认识可以经过大词表验证。如果同时还使用关键词(keywords)技术，则可能提供文本体裁、内容、词频顺序、搭配模式等诸多方面的信息。

其他重要的技术还有语料注解 (annotation) 和词类标注 (word-class tagging)，句法分析 (syntactic parsing) 以及文本预处理等。

语篇索引技术在统计方面也有特色。如全文统计可提供字符数、词数、词次 (tokens) 数、词类(types)数、平均词长、1 字符词数、2 字符词数、……、句数、平均句长等等。

近年，语篇索引技术所涵盖的范围越来越广，而且在软件转化方面也朝着集成化方向发展。80 年代使用较广泛的有 OCP(Oxford Concordance Program), WordCruncher, TACT 等，这些软件的早期版本功能有限，后期的版本则逐步增加了如预索引功能和搭配检索功能等等。90 年代中期以来，新一代语篇索引软件具备了处理大规模语料的能力和已标注语料的处理能力，而且集成了多功能统计包。如 LEXA 系统包含了 60 个左右适合词汇、语法、主题等语言分析的相关程序。还有一些语篇索引系统软件是从大型语料库项目中产生的，如“国际英语语料库工程”(International Corpus of English)就产生了 ICECUP(Corpus Utility Program)软件系统。“英国国家语料库”(British National Corpus)

开发出具有注解功能的 SARA 系统。另外，这些新一代软件大多都能处理世界各类字符型文字的语言。最近，笔者见到牛津大学出版社的 WordSmith 系统就很有特色，既能处理生语料也能处理标注语料，其“或/与/非”功能能较好地检索非连续序列语言现象。处理的语言达到 30 余种，但无法处理汉字，藏文，日文，阿拉伯文这类文字的文本。

四、结语

从藏语实践来看，语篇索引技术不仅可行，而且相当有效。从目前经验来看，人们认识到的显性语法规则应采用大规模真实语料验证，在反复摸索中逐步认识语言中的其他隐性规则。而语篇索引技术将在这个认识过程中发挥重要作用。

参考文献

- 1、Graeme Kennedy, 1998, *An Introduction to Corpus Linguistics*, Longman.
- 2、Ma Zhiyi, Zhan Xuegang, Yao Tianshun, *Temporal Analysis for Text Understanding Based on situations*, PICCIP1998, P280-287.
- 3、Roger Gasider & Paul Rayson, 1997, *High-level annotation tools. in Corpus Annotation*. Longman.
- 4、Wolfgang Klein, 1994, *Time in Language*. P120-141. Routledge. London.
- 5、陈平, 1988, 《论现代汉语时间系统的三元结构》, 《现代语言学研究》P142-180. 重庆出版社.
- 6、郭锐, 1997, 《过程和非过程--汉语谓词性成分的两种外在时间类型》, 《中国语文》, P162-175.
- 7、江荻, 1999, 《藏语拉萨话现在时的标记及功能》, 《民族语文》1999 第 5 期.
- 8、土丹旺布, 索多, 罗秉芬编著, 1995, 《拉萨口语会话手册》, 中央民族大学出版社.
- 9、周季文, 谢后芳《藏文文法》, 中央民族大学藏学系油印教材, 1994.

附：依据 CATC 分析的部分藏语动词分类

动态动词（过程，事件，分时） za（吃）； vtshong（卖）； vbri（写）； byed（做）； bsrab（教）； mchod（喝，敬语）； bo lo bkyon（打球）； gzhas gtong（唱歌）； sgrung deb lta（看小说）； vchang（吠叫）； las ga byed（工作）； lha mo vkhrab（演戏）； langts（起床）； rogs ram byed（帮助）； ldem byed（摇晃）； tha mag vthen（抽烟）； brje（兑换） dgav zhen byed（热爱）； tshong rgyag（做生意）

变化动词（状态） don（出现）； gnyid sad（醒）； thob（获得）； rnyed（找到）； gnyid khug（睡着）； ltod gyeng（放松）； thon skyed byed（产生）； pham nyes yong（失败）； brjed（忘）； slebs（到达）； gog（脱落）； chad（断）； mthong（看见）； go（听见）

心理动词（静态） ha go（知道）； ngo shes（认识）； dgav（喜欢）； dran（记起，想起）； shes（会，懂，明白）； zhed（怕）； brjed vgro（忘记）； bsam（想）； mgo na（头疼）； ngo tsha（害羞）； ltogs（饿）； na（病）

关系动词（匀质） yin/red（是）； zer（称名，姓[什么]）； mthun（符合）； vdra po byed（相似）； vdra ba（好象）； mtshungs/mnyam（等于）； mtshon（像）； gnas（值[钱]）