现代藏语的句法组块与形式标记*

江　荻

中国社会科学院　民族学与人类学研究所 计算语言学研究室 北京 100081

Email: jiangdi@public3.bta.net.cn

摘　要：本文定义和描述了现代藏语句法组块的基本类型以及相关的形式标记，并在此基础上提出藏语自动分词的组块方法。而实施组块分词方法的措施包括按照一定顺序原则识别组块的形式标记，通过各类标记函数集、辅助词表，以及从组块中抽取的句法信息确定组块的边界，然后对组块进行分词和词性标注。进一步的设想是对组块进行归并，使其与藏语句法成分形成一致关系，达到消除嵌套组块和利于句法理解的目的。

关键词：现代藏语　句法组块　形式标记　组块分词策略

On syntactic chunks and formal markers of Tibetan

Jiang Di

Department of Computational Linguistics

Institute of Ethnology & Anthropology of Chinese Academy of Social Sciences

Abstract: The present paper describes the basic types of syntactic chunks and their formal markers in modern Tibetan. And a scheme of automatic word-segmentation based on chunks has been provided according to the features of Tibetan syntactic structures. In order to carry out the scheme some primary work must be done as pre-processing, setting up small tables of formal markers of each chunk, the verbal paradigm, special tables of differentiating cases and homographs, etc.. Then with all these information we can find the left and right bounders of each chunk, and go on segmenting words with a dictionary and tagging within chunks.

Keywords: Tibetan　syntactic chunks　formal markers　segmentation based-on chunks

一　句法标记与句法结构

现代藏语句法结构是较典型的中心词短语结构，突出的特征表现在以名词为中心词的名词短语和以动词为中心词的动词短语，其次还有形容词为中心词的短语和少量句法位置游离性较大的成分。提出这种观点至少可以获得以下几种依据的支持：

一是从藏语句法成分来看，句子的基本语序和成分是：主语＋宾语＋谓语，即 SOV，其

* 本项研究获国家自然科学基金资助，项目批准号 60173024。

中，主语、宾语基本由名词短语充当，谓语由动词和形容词短语充当。名词一般不能做谓语，动词和形容词短语一般也不能充当主、宾语和名词的前修饰语，除非添加名物化标记转换为非谓动词结构。这两项规则在藏语句法规则中基本是严格的。二是修饰语与中心词的句法位置关系和语义关系，名词中心词的修饰语可以前置也可以后置，凡后置的都是无标记形式，包括代词、指示词、数词、形容词，以及非谓动词结构；而前置的一般都带属格标记，包括名词、代词、指示词、形容词和非谓动词结构。动词中心词的修饰语也分为两类，时体后缀、人称／意愿后缀，以及语气词分布在动词之后，副词以及带词格标记等状语性短语分布在动词之前，形容词或名词短语做补语也放在动词之前。这些修饰语与中心词之间呈现为指示或限制的语义关系。第三方面，短语与短语之间的联系不单纯是语序关系，还有更明显的形式标记，突出表现为名词的词格，它出现在短语的末端，总是包容了名词中心词的后修饰语词串，具有指示短语的形式标记和句法功能作用，即“S 词格 O 词格 V”。

根据以上分析，藏语基本句法结构可以用“名词短语（NP）＋名词短语（NP）＋动词短语（VP）来描述。但由于词格等标记所标示的词串与上述短语并不完全一一相应，如内嵌于名词短语中的属格结构、非谓动词结构、以及句法成分位置的可能变动（如 OSV）等等，同时考虑到充当主、宾语的成分还可以是独用的代词、指示词、数量词或者非谓动词结构，为避免与句法分析上的功能性短语相混，我们可以把各种句法短语以及带形式标记的结构或词串统一称为组块。[1]

上文讨论引出的另一个概念是句法结构形式标记。藏语是一种具有较丰富形式标记的语言，形式标记自然地把藏语句子分割为若干组块，清楚地描述各种可能的组块及其形式标记，就有可能预设我们的自动分词方案。按照我们的设想，藏语自动分词的基本路线是将句子切分为较小的组块片段，然后在组块内部对有限的词语进行词的切分和词性标注。因此，如何识别句法形式标记和确定组块边界是藏语自动分词的重要步骤。[2]

所谓藏语句法形式标记也就是句法功能虚词，包括词格标记、名物化标记、时态助词、复数后缀、敬语语素、语气词、作修饰语的指示词和数量词、以及名词、动词、形容词、副词等词类的构词词缀。这些标记进入话语后附着于不同的结构或词串，规定或限定了这些结构或词串的句法性质，并能揭示组块之间的句法关系。

从这个角度思考，我们可能获得丰富的句法形式标记，充分利用这些形式标记，并通过带标记结构之间或与无标记结构之间的相关句法关系就能细化对句子结构的分析。例如，非谓动词结构可能充当主语、宾语，也可能只是主语或宾语成分中的一部分，甚至是修饰语中的一部分，通过名物化标记把较长的词串细分能够避免匹配上的组合型歧义，还可能提供分析其它组块乃至全句所需的信息。再如，带属格标记的名词前置修饰语不直接充当基本句法成分，但能为我们确定名词右边界与属格之间的组块。所以，按照组块及其标记阐述句法结构，能够为机器处理文本提供明确的识别形式标志。

本文的目的是初步分析构成现代藏语组块的各类标记，以及这些标记的性质和类别、以及所组成的句法组块的功能，然后提出依据组块自动分词的技术路线和处理方法。

---

[1] 这个定义可能导致组块之间的嵌套，我们的处理办法是分两步走，先处理最小组块，在块内分词与标注，然后进行组块的功能性归并，详见［7］。总的目标是充分利用藏语形式标记获得最优化的识别和处理方法。

[2] 限于篇幅，本文不能细述各类标记的具体情况，亦未能举例。请参考文献［7］。

二　句法标记分类与组块的类别描述

藏语的句法标记包括：词格标记；名物化标记；动词语尾；指代词；构词词缀。其中，词格标记最复杂，有施格、受格、位格、与格、对象格、从格、比较格等等。名物化标记指添加在动词或动词短语后使之构成非谓动词结构，能够充当主语、宾语等成分，具有名词性功能。动词语尾数量众多，主要包括时体后缀、人称／意愿后缀、语气词、助动词、趋向动词。这一类句法标记是谓语的构成部分，一般情况下，单个动词构成谓语的情况不多，大多数句子都要在动词之后添加时、体助词或者语气词来构成谓语。指代词，包括人称代词、疑问代词、指示代词、不定指示词、连词、复数后缀、敬语语素。这一类只指作为名词修饰语的情况。构词词缀。藏语双音节或多音节的名词、动词、形容词、副词等词类在构词上分别有自身的形式特点。尽管这些形式特点不足以完全将其与它类词区分开来，但在局部范围内作为其他词法和句法的辅助识别手段还是有一定的形式价值的。

依据以上五类句法形式标记来确定句法组块并不能覆盖全部的句法结构，藏语中还存在一些零标记现象。因此，还应该考虑藏语的普遍语序和句法成分结构位置所提供的信息。

根据现代藏语结构，总的语序位置是：SOV，而其扩展的句法语序是：主语＋（间接宾语）＋（直接宾语）＋（结果补语）＋（状语）＋动词＋（状态补语）。这些句法成分带不带形式标记或者带何种形式标记，取决于动词的类型、组块与组块之间的句法关系。

以下我们根据藏语中较为普遍的现象定义了 7 种类型的句法组块，其依据是综合考虑了词类、标记以及它们的句法分布特点，而它们的句法功能也是由这些要素决定的。

1 名词组块　名词或名词性短语可以分布在多种句法位置，充任各种句法成分，但根据其所带词格标记和句法位置，大致具有三种句法功能。

充当主、宾语功能组块，如带གིས（gis）类施格标记名词组块充当句子施事主语，带ལ／ར（la／ra）类词格标记组块充当“领有、获得、需要”等动词的主语。带ལ（la）类对象格标记组块充当心理动词指向的对象宾语，带ལ（la）类受益格标记组块充当三价动词的间接宾语，在变化动词句子里，带དུ（du）类标记组块充当成事宾语，以及充当某些动词的补语。还有一些充任主语或宾语的名词组块不带标记（即通格形式或零形式标记），这些名词组块一般需要利用其他组块中抽取的信息以及组块定界后的剩余特征确定其边界位置。

带有གི（gi）类属格标记的组块都是名词的前修饰语。充任状语的名词组块带有各种形式的标记，如ལ（la），ར（ra），ནས（nas），ལས（las）等，分别表示动作发生的时间、处所、方式，或者表示比较、材料、工具等意义。

从以上名词或名词短语带不同标记所具有的句法功能来看，尽管它们的句法位置和功能存在很大的变动性，但名词所构成的组块自身却是确定的，即：NP＋（形式标记）。

2 形容词组块　藏语形容词自身具有一定的格式，包括最主要的带后缀པོ（po）的形式，以及其他重叠形式和比较级／最高级形式。但这些构成形式自身存在一些变体，且部分还与其他句法上构词或构形的形式同形，因此他们只能算作隐性标记，还需要通过其他句法现象加以制约才可能加以利用。

在句法位置上，形容词作为名词的后修饰语是其主要句法位置。一般来说，在这个位置上，大多数形容词都可以根据其自身形式加以识别，只有少量的还需要采用别的方法与其他

词类区分。不过，形容词作修饰语的这种情况结合句法结构来看，可以归入后修饰语组块。

形容词作状语和补语一般都位于动词之前，并且带ལ（la）类形式标记（拉萨口语用“形容词＋ཡས（byas）”构成副词作状语），因此，对这个位置上的形容词组块的识别比较容易把握。另外，形容词作宾语也处在动词之前，但一定是判定动词句，且形容词不带其他标记，不会与其它功能混淆。形容词自身做谓语往往需要带一定的（谓词）语尾形式，结合其自身的隐性标记和有限的语尾形式，判定这类作谓语的形容词组块也是可行的。

值得说明的是，形容词做状语和谓语都存在一定的形式标记，以功能的角度看，似乎可以分别归入修饰语组块和谓语动词组块。但形容词作补语和宾语又难以归入名词组块，为此，我们考虑定义形容词组块还是会有利于下一步的句法分析。

3 非谓动词组块　现代藏语的动词或动词短语充当主语、宾语、修饰语等成分时必须添加句法上的形式标志从而转化成名词性成分，这样的形式化标记称为名物化标记。由名物化标记构成的词串不再具有谓语的功能，可以称为非谓动词组块。

非谓动词组块的标记来源于各种虚化的名词，有些虚化程度高，不再具有构成其他词类的功能，有些尚处在虚化之中，还可与其他语素构词。从意义上观察，各种名物化标记表达不同的组块意义，如“ཐུ”，原义为“材料、事物”等意义，附在动词现在时形式后表示要做的事情；“ཡས”，原义为“上面，上方”，附在动词现在时形式后表示动作本身、动作的对象、动作用具等等意思。名物化标记的含义对动词形式也有一定制约，如，“ཀ / ཏ（pa / ba）”，原为名词后缀，附在动词过去时形式后面构成非谓动词组块，表示与动作有关的事物或东西。

非谓动词组块最主要的句法功能是充当句子的主语、宾语、和修饰语，充当修饰语而前置时一般要带属格标记。

4 谓语动词组块　藏语的谓语动词是句子的核心，其句法位置是句子的末端。单独以动词煞尾的句子不多，一般总是带有一些其它成分，其扩展格式是：{（状语）＋动词＋（助动词情态和趋向）＋（后缀时体_人称 / 意愿）＋（语气词）}。

谓语动词组块是句子中包含信息最丰富的部分。一般来说，动词后缀包含了表达人称一致（自称 / 他称）、时（现在 / 过去等）、体（已然 / 未然，将行 / 即行等）、意愿（自主 / 不自主）、熟知亲见（亲见结果 / 泛泛推知）、动作趋向（趋向说话人或离开说话人）等等语法意义，而语气词则包括了陈述、疑问、反诘、推测、商榷、强调等等含义。

除此之外，动词本身还包含了更重要的信息。一是动词的语义类别属性，如关系动词，性状动词，动作动词，变化动词，趋向动词，心理动词，评议动词，相互动词，述说动词等，不同语义结构的动词将导致动词呈现出不同的句法结构，如及物与不及物，自主与不自主，瞬时还是延时，论元数量不同（配价数）等等。二是动词自身的形式交替，现在时形式（原形）、未来时形式、过去时形式、命令式形式、自动词形式、使动词形式等。

谓语动词所蕴含的信息是藏语自然语言理解或者自动分词的重要知识来源，提取其中的知识有利于我们总结其中的句法规则并进行形式化操作，而利用这些信息可以大幅提高分词的准确性和句法理解的精细程度（文献[3]）。

5 前修饰语组块　凡处在名词性中心词之前或者动词性中心词之前修饰或限定中心词的成分都构成前修饰语组块。修饰名词性中心词的组块一般都带有ཀྱི（gi）类属格标记，主要包括名词、代词、指示词、非谓动词组块、以及少量形容词（带ཀ / ཏ（pa/ba）等后缀）。

修饰动词中心词的成分主要包括副词、形容词、以及带各种词格标记表示方式、来源、

处所、工具、目的等等的名词组块。其中，副词最普遍，种类很多，有些带标记，有些不带标记。如程度副词、方式副词、时间副词、范围副词、否定副词等等，这些副词一般都处在动词之前作状语，构成副词性前修饰语组块（个别副词，特别是否定副词的位置较特殊[3]）。

6 后修饰语组块　凡处在名词性中心词之后或者动词性中心词之后修饰中心词的成分构成后修饰语组块。充当名词后修饰语的主要有形容词、指示代词、疑问代词、数量词。这几类词的基本顺序是：{名词＋（形容词）＋（单纯疑问代词）＋（量词）＋（数词）＋（指示代词／不定指示词）}。从形式上看，形容词基本都是带后缀ཏོ/ཏོ（po/bo）的性质形容词；指示代词作后修饰语，自身用通格，其后不加其他标记的话则它自身可看作类似英语定冠词这样的标记；不定指示词作后修饰语，不能单用，可以看作不定冠词，也是句法标记。

位于动词性中心词之后的修饰语很少，例如，形容词修饰形容词置于作为中心词的形容词之后。个别程度副词放在动词或形容词前与放在动词或形容词之后意义不同，如ཆེས（ches）“非常，太”在形容词词根前表示程度加强，在后面则表示程度过分。

7 从句组块　利用各种连词和关联短语构成复句。由于复句类别甚多，同一类连词可用于不同类复句，以及连词与其他词类部分同形，情况复杂，我们暂时不讨论这类现象。

三　组块的边界预测及分词过程

按照组块分词的方法，具体过程大致可以分出四个步骤（图1）。其中第一个步骤主要针对藏语各种虚词标记的变体形式，采用续连规则集对其进行根词归一化的处理。由于藏语形式标记变体基本都遵循传统文法所描述的续连规则，我们只须为每一类选取一个代表形式，并构建函数条件（因变量）便可建立相关的形式变量集。有关讨论请参见文献[5]。

确定组块边界是本项研究的重点。根据上文的讨论，组块右边界一般可以通过形式标志的识别来确定[4]、[5]、[7]，并确定其位置参数。对于其左边界，有两种情况需要甄别。一，左向无相邻标记，则一直向左上溯至最靠近的标记，该标记右侧的第一个位置可初步确定为该组块的左边界标志；二，左向相邻位置存在一个标记，这类情况大多是“名物化标记＋属格标记”，或者“名物化标记＋对象格标记”一类，可以分别标志其位置参数，然后继续向左上溯左边界标

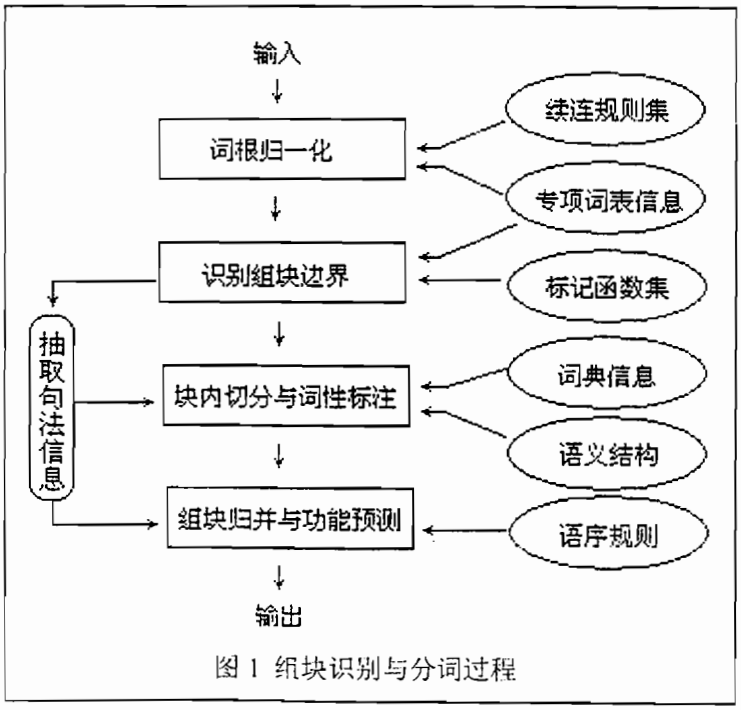

图1 组块识别与分词过程

记。应该指出，左向相邻标记不包括指示词、数量词、复数后缀这类标记。

为了顺利识别组块的边界，我们建立了两类辅助词表（参考文献[6]，其中词表和辅助词表的论述很值得借鉴，藏语的处理方法实际与汉语相当接近）。一类是标记函数集，主要内容包括各类标记形式以及标记的功能描述和类别，其作用是鉴别文本中的标记及组块。另一类是小型辅助词表，如动词词形表[3]，动词后缀表，带ར/ས（ra/sa）韵尾词形表[4]，同形异类词形表。这些表的作用主要是辅助识别右边界标记，排除与标记同形的语素。

藏语组块识别的一个重要原则是顺序性，组块识别的先后取决于各组块所依赖知识的类型和识别策略。从藏语特征来看，藏语句子宜采用逆扫描方法，因为各种句法标记都处在组块的后面（右端）。一般来说，各组块的识别顺序大致是：谓语动词组块，非黏着型词格名词组块，非谓动词组块，话题语气词组块，从句组块，黏着型词格前置修饰语组块，黏着型词格名词组块，形容词组块，后修饰语组块。在识别标记及确定组块边界的同时，组块分词中还有一项重要任务，即抽取组块内的句法信息，特别是谓语组块所蕴含的信息。这些信息对于全句句法结构的判断和支持其它组块识别都能发挥作用[3]。

确定了组块的左右边界以后，可以在组块内部对有限的词语进行词典匹配和分词，同时还给匹配成功的词语标注词性。对未登录词以及人名、地名等难以匹配的现象，允许容错处理。如果词典中包含了语义结构信息，则可进一步利用组块抽取信息进行分析。

按照形式标记识别组块并非句法分析和自动分词的目标。一旦完成组块识别和块内词语切分、标注，可以考虑趋向功能性句法短语的分析。由于属格组块或前修饰语组块、后修饰语组块、部分非谓动词组块，以及形容词组块与句法成分并不完全对应，而且形成组块之间的复杂嵌套格局，不利于句法结构的线性分析[2]，为此有必要结合藏语句法结构位置或语序规则消除这些组块，使它们汇入更大的组块之中（或可以称为句法成分短语）。而名词组块、充当主、宾语的非谓动词组块等也可以根据它们的句法位置，并利用已经抽取的句法信息给它们赋予句法成分的标记。这样的分析可能为进一步的篇章分析，或者语义理解奠定基础[1]。下面两个分词例句中，{} 表示相应于句法成分的功能性短语，() 表示具体组块，[] 表示组块与标记管辖的关系。

1 “我很喜欢看西藏歌舞”。

{[ང]ས} ་ {[(བོད་ཀྱི་) (སྒེའུ་གར་ལྟ་) ཡས] ་ལ} ་ {(དགའ་པོ་ཞེ་བྱས) ་པོད}

我 施格 西藏 属格 歌舞 看 名物化 对象格 喜欢 很 动词后缀

2 “所谓俚语就是少数人才懂得话”。

{[(ཚིག་རྙིང་ཅེར་)ཡས་]དེ་ར[[[(གཅོ་བོ་ལྷི་ཀུང་ནས་)ཀྱིས་][(ཀུ་གོ་)པ་]འི་] ་(སྐད་ཆ་)དེ་རེད་}

俚语 称为 名化 那 主要 人 少数 最 施格 知道 名化 属格 话 那 是

四 结语

采用组块方法处理藏语文本的用意很明显，即尽量简化分词的复杂度并提高句法自动分析的精确性。而客观上，藏语句法事实本身（如句法标记）也为组块分词方法预设了形式识别的优越性。其一，组块在结构上所存在的形式标记具有较强的系统性和普遍性，绝大多数

的句子都是由各类组块串联而成，而且形式标记易于识别，这就可能减少分词对词典的依赖，节省了分词过程的开销。其二，预先识别形式标记有利于消除同形词语和跨组的歧义现象。其三，对句法组块的识别作为一种预分词手段，能够相当程度上为全面的分词提供所需的句法信息，这些信息在后续组块处理和自动分词中能够发挥重要作用。

我们划分藏语组块的经验性基础并不是单纯的功能性短语结构类型，如主谓结构、述宾结构、述补结构、联合结构、偏正结构等等。但在本质上，本文的指导思想仍然是功能性的。定义名词组块、非谓动词组块表面上看不涉及句法功能，但这正是组块多功能性的体现，一旦确定组块的句法位置，其功能也就随之确定。藏语中还有一些组合，如数量短语、重叠结构，我们并不列入组块分类，这是因为我们同时还要考虑形式标记和语序作用。当数量短语直接充当句法成分（即非修饰成分），那么标记原则同样会把它作为名词组块来对待，而它作为修饰成分则又归入修饰语组块，所以不必另列。

从功能角度观察以上组块的分类，可以看出完全单一的结构与功能对应的情况基本没有。即使我们在表中列出各类组块的标记，仍然会出现多种对当关系，因为标记本身也是交叉对应的。所以划分标记类型和定义组块只是藏语句法分析的一个必要手段，为了达到藏语文本句法分析和自动分词的目的，还需要借助语序分析、组块信息提取以及句法成分与组块对当分析等等描述方法。

在组块分词策略的基础上，我们已经展开了多项关于藏语文本的具体研究，积累了相关的经验，也获取了大量数据。最近，我们结合组块分词设想，进一步对不同句式展开分词研究，并初步取得一些成效，验证了组块分词的可行性。这里限于篇幅，不能展开论述。就整个藏语分词研究来说，现有的各项研究还处在起步阶段，需要对藏语语法作更细致的分析，并借鉴英语、汉语对文本识别和分词的经验。请同行专家批评指导。

主要参考文献

[1]苑春法，陈刚，黄昌宁："基于词性和语义知识的汉语句法规则学习"，《中文信息学报》2001 年第 3 期。

[2]张昱琪，周强："汉语基本短语的自动识别"，《中文信息学报》，2002 年第 6 期。

[3]江荻："现代藏语谓语动词的识别与信息提取". Paper for 20th ICCPOL'2003。

[4]Jiang Di. Long Congjun. The Markers of non- finite VP of Tibetan and its automatic recognizing strategies. Paper for 20th ICCPOL'2003.

[5]Kang Caijun. Jiang Di. The Methods of Lemmatization of Bound Case forms in Modern Tibetan. In *Collectanea of Phonetics and Computational Linguistics* (forthcoming).

[6]孙茂松："汉语自动分词研究的若干最新进展"，《辉煌二十年－中国中文信息学会二十周年学术会议》，清华大学出版社，2001 年。

[7]江荻："现代藏语组块分词的方法和过程概述"，《民族语文》，2003 年第 4 期。

[8]Herbert Bruce Hannah. 1912. A grammar of the Tibetan language. Motilal Banarsidass Publishers Private limited. Delhi 1996.