现代藏语判定动词句主宾语的自动识别方法*

黄 行　　江 荻

中国社会科学院　民族学与人类学研究所 北京 100081

Email: tzdfc@xinhuanet.com

提　要：本文通过剖析现代藏语判定句主语、宾语以及动词的结构及形式标记，提出识别主语和宾语的方法。其中依据动词的形式和前附修饰成分对宾语与动词的定界有效性可达 99%以上，而采用综合性形式标记对主、宾语的定界可达到 75%以上。文章最后指出，要大幅提高判定句主语的识别率应考虑利用识别宾语和动词时所获取的句法语义等其他信息。

关键词：藏语　判定动词句　主语　宾语　自动识别

Automatic Recognition of the Subject and Object of Linking verbal Sentences in Tibetan

Huang Xing　　Jiang Di

Institute of Ethnology & Anthropology of Chinese Academy of Social Sciences

Abstract: This Paper proposes a method to recognize the subject and object of linking-verbal sentences as well as their predicate verbal structures through analyzing the markers and structures of subjects and objects of sentences. According to our experiment based on the verbal forms and their pre-modifiers the precise rate of recognition is up to 99%. Yet the rate of recognition of subjects is only up to 75% with all those markers to indicate subject chunks. So the further work to improve the rate is to make full of use of the structural and semantic information getting from object chunks.

Keywords: Tibetan　patterns of linking-verb　subject　object　automatic recognition

一 引 论

现代藏语判定句是指以判定动词ཡིན、རེད 以及འཕྲལ （都相当于汉语“是”）[①]为谓语动词的句子。由于人称一致性、与主语的亲疏关系、以及书面语与口语的差别等制约，判定动词在不同情况下可分出好几种不同的形式。主语是第一人称或者与第一人称相关时，动词大多数用ཡིན ；其他情况用རེད 。另外，判定动词的否定形式一般都是在动词前加上否定副词མ“不”，构成མ་ཡིན་，མ་རེད་，མ་འཕྲལ，但是，མ་ཡིན还有一个常用的缩略形式མིན “不是”。

---

*国家自然科学基金资助项目（60173024），中国社会科学院语音学和计算语言学重点实验室资助项目。

① འཕྲལ 是表示谦语的形式，现在拉萨话基本已经不用。但保留了短语性答语འཕྲལ་ཟེར，使用很普遍。

现代藏语基本语序为 SOV，判定动词句也不例外。但与绝大多数其他句式不一样的是，判定动词句的主语和宾语不带明确的词格形式标记，又由于二者具有语序排列上的线性相邻关系，因此，划定二者之间的界限是分析判定句的一个重要环节。

尽管判定句主、宾语之间没有确定的词格形式标记，但相当部分句子仍然存在各种隐性的、或可以判定二者边界的其他形式，其中有些是我们在藏语组块分词方法中设定的组块标记，如名词组块、修饰语组块、非谓动词组块等等［1］，这些组块及其标记的识别方法已在其他专项研究中初步解决［6］、［7］。

本文的目的是初步剖析藏语判定句的基本结构，从中寻找离析主、宾语的依据，为机器自动识别判定句主语和宾语提供分析句法结构上的可能策略。

二　判定句主语的类型及识别条件

一般来说，能够充当判定动词句主语的结构一般都是名词或名词组块，如名词、人称代词、指示代词、疑问代词、数量词等等，动词性短语必须添加名物化标记转化为非谓动词组块才能充当判定句主语。名词组块充当主语还要考虑语义结构类别，如处所名词主语的判定动词可能转义为表示存在的意义，这种句子也可以归入存在句。有些情况下，如对话，主语可以省略。先观察以下例句（s、o、v 分别表示主、宾语和动词）。

(1)དེ་（s/那）བོད་ཀྱི་（藏族的）ལོ་རྒྱུས་（o/历书）རེད་（v/是）།“那是一本藏历。”

(2)ང་（s/我）བོད་རིགས་（o/藏族）ཡིན（v/是）།“我是藏族。”

(3)དེ་རིང་（s/今天）གཟའ་ཟླ་བ་（o/曜日月亮）རེད་（v/是）།“今天是星期一。”

(4)ཀློ་མོའི་གཞས་ཁྲིད་མཁན་（s/女孩领唱者[名物化]）ནེ་མ་（o/妮玛）ཡིན།“女声领唱的是妮玛。”

(5)ས་སྐྱའི་ཀྱིག་མཐའ་ལ་ས་སྐྱ་རེས་ཡས་（s／萨迦称作萨迦派[名物化]）ག་རེ་（o／什么）ཡིན་ན།“萨迦派的萨迦是什么意思？”

(6)ལྷ་སའི་སྐྱིད་ཆུ་ནི་（s／拉萨吉曲河）ཡར་ཀླུང་གཙང་པའི་ཡན་ལག་ཆུ་བོ་ཞིག་（o／雅鲁藏布江支流）རེད་།“拉萨吉曲河是雅鲁藏布江的支流。”

(7)དྲུང་ཚང་གཅིག་ནི་(s／一户人家)ཆུ་འཁོར་གཉེར་པ་(o／管水磨人)རེད(v)་།གཅིག་ནི་(s／一)ལྕགས་དྲུང་མཁན་(o／打铁人)རེད(v)་།“一户是水磨经管人，一户是打铁的。”

(8)རྗེ་མོ་གླང་ན་ཡིན་ན་（s/珠穆朗玛峰）འཇམ་གླིང་ཐོག་གི་རི་བོ་མཐོ་ཤོས་（o/世界上最高的山）རེད་།“珠穆朗玛峰嘛，是世界最高峰”

(9)ས་ཁ་འདི（s/这地图）་ཁྱོད་རང་གི་（o/你的）རེད་པས་།“这幅地图是你的吗？”

(10)རི་ལྟ་ཆེ་བ་དེ་（s／大颗粒）ས་མཐོ་ལི་ས་ཆུ་ལ་འཚོད་པ་ལི་ནད་བཅོས་སྨན་ཡིན་（o／治疗高原水土不服[非谓短语]）རེད་།“大颗粒的是治高山反应症的。”

在这些例句中，我们分别看到充当主语的词类可以是指示代词(1)，人称代词(2)，时间名词(3)，一般名词(6)、(8)，数词(7)，非谓动词短语(4)、(5)。尽管这些词类或者组块并没有附加显明的词格标志，但从藏语语序共性可以知道，指示词（含不定指示词）、代词、数量词这类词作为修饰语普遍后置于名词或名词组块，它们都属于封闭类词类，具有主语煞尾标记作

用(9)、(10)。而出现在句首的指示词、代词、数量词属于独用性质，其自身就蕴含了标记性质。至于复数后缀，也置于名词、代词等名词组块之后，属于比较明确的形式标记。

真正可以作为判定句主语的形式标记有几类，一类是添加语气词ནི་、ད（较少）、ཡིན་ན་等具有典型指示话题的标记形式，它们出现在主语之后，恰好把主语与宾语分隔开来(6)、(7)、(8)，[1]其次是与此功能类似的范围副词ཡང་（书面形式为ཡང）“也”(11)；第三类是非谓动词组块，它们具有明显的名物化形式标记，也是比较普遍的现象(14)；最后还有数量上略少一些的现象，如带ལ་标记的主语，这是因为判定动词兼有表示“存在、领有”的意义(12)、(13)）。

(11)ཟེ་དགོས་ཡང་ཀྱི་རྒྱུ་མཚན་ཡིའེ་དེ་རེད་།“道理就在这儿。”

(12)དེའེ་སྒྲོམ་དེའི་སྣང་ལ་（s/那个书架上）བོད་ཡིག་གི་དེབ་（o/藏文书）ཀུང་ཀུང་རེད་།“那个书架上面全部是藏文书。”

(13)ས་གདན་ལ་（s/地毯）སྟོང་ཟེ་ཅིག་སྟོང་ལྔ་བཅུ་（o/一千五百元）རེད་།“地毯要一千五百元。”

(14)ཁྱེད་རང་གིས་གསུངས་པ་（s/您说的[非谓短语]）དཀའ་རང་（o/这样）རེད་།“您说的正是这样。”

通过从 10 万字口语语料中抽取的 154 个判定句统计后（参见表 1 的数据），发现各种不同的主语标记情况很值得重视，如语气词标记（ནི་/ཡིན་ན）、指示词标记（འདི་/དེ）、范围副词（ཡང་ ）以及非谓动词组块的名物化标记达到全部语料的 63%，加上ལ་标记、代词和数词等非典型标记则比率达到 77%以上。

表 1　判定句主语的标记类型统计数

标记类型	ནི་/ཡིན་ན	འདི་/དེ	ཡང་	名物化标记	ལ་	代词	数词	主语省略	名词
数量	37	42	5	14	10	6	4	14	22
比率(%)	24.0	27.9	3.3	9.1	6.5	3.9	2.6	8.4	14.3

总结上面的讨论，藏语判定动词句主语的识别相当部分可以通过各种形式标记及其组块加以解决，但对于非典型标记的句子以及无标记句子的主语识别则还需要结合宾语的情况以及其他句法信息综合解决，所以下面我们讨论判定动词句的宾语情况。

三　动词的形式与宾语的类型

藏语动词居于句尾，宾语的位置在它前面，因此，讨论宾语特征之前，应先了解动词的各类附加形式。判定动词句的动词形式一般都是简单形式，即ཡིན，རེད，ཨཁོ་及其否定形式。但为了表达某些其他语法意义或语气，判定动词还可以附加一些动词后缀或者句末语气词。

判定动词之后出现最多的成分是疑问语气词，如拉萨口语中多用ཡིན་བས་、རེད་བས（一般问句）、ཡིན་ལ（特殊问句），早期书面语的疑问语气词是ནམ་、དམ་等。另外，书面语中陈述语气词ནོ་等也可添加在判定动词之后，如“དེ་ནི་གསེར་གྱི་རི་ཡིན་ནོ་།那是金子的山”(15)。表示推测语气的判定动词句用“ཡིན”加上语尾“གྱི་རེད”、“ཟ་རེད”、“ཟ་ཡོད”等，例如，

[1] ཡིན་ན原本是连词，在拉萨口语里经常用作表示语气或语气转折的虚词，因此这里看作主语标记。

ཁོང་（s/他）རེ་ཨོ་འབྲི་མཁན་（o/画画者）ཡིན་གྱི་རེད་（可能是）།“他可能是画家”⑯；再如，ཚ་ལྷི།་རང་（s）ཨེ་ཅིན་ནས་（o）ཡིན་པ་རེད་།“哎呀，你原来是北京来的”⑰。其他还有少量特定的后缀用法，如“ཀྱག”加在“རེད”后表示突然发现某种原来不知道的情况，如ང་གཉིས་ཀྱི་ཁྱིམ་ཚང་（s）སྒྲུ་འབྲེལ་（o）རེད་ཀྱག།“咱俩的宿舍是隔壁呀”⑱。

上文已经指出，判定动词句的宾语没有特定的词格形式，因此还需要排除它与动词之间作为修饰动词的前置状语成分。除了否定副词置于动词之前外，少量表示范围或频次的副词也出现在动词之前。不过，我们统计的句子中只有两例出现过前置状语（不计否定副词），而且是同一个词“ཀུན 全都（是）”。例如，ད་ལྟ་གསུངས་པ་དེ་ཚོ་（s）སྐད་ཀྱགས་ཡང་བའི་རེ་（o）ཀུན་རེད་པ་།“刚才讲的全是著名的山吧”⑲。下面另外举出两个例句。

⑳ཟང་ཁེན་གྱི་ལྟད་མོ་（s）ཞབས་བྱེ་（o）ཀུང་ཀུང་རེད་།“明天的节目尽是跳舞。”

㉑དེབ་གནས་གྱི་ནང་ལ་ཡོད་བའི་དེབ་（s）བོད་ཡིག་（o）ཤ་སྟག་རེད་།“老师家里的书全部是藏文。”

下面我们来讨论宾语。充当判定动词句宾语的成分包括名词或各类名词组块、代词、数量词等等，动词性短语充当判定句的宾语同样需要添加名物化标记转化为非谓动词组块。除此之外，能够充当判定动词宾语的还有形容词，或者带属格的短语结构。各类句式如下。

㉒བཟོ་རིག་པ་ནི་（s）བཟོ་སྒྲུན་ལག་ཚལ་གྱི་ཚན་རིག་ཅིག་（o）རེད་།“工巧明是一门技术工艺的科学。”

㉓སྐྲུལ་གཏོད་གཏང་མཁན་（s/创建者）སུ་（o/谁）རེད།“创建人是谁。”

㉔ཞེད་ཁར་རྒྱག་ཡས་ལྷད་གཅོ་བོ་（s）ག་རེ་（o）རེད་།“主要施什么肥料。”

㉕ང་ཚོ་ཚང་མ་（s）བོད་ཡིག་བསྦབ་མཁན་（o）ཡིན་།“我们全部是学藏文的人。”

㉖ཐག་（s）ཉེ་པོ་（o）རེད།“距离很近。”

㉗རྒྱུ་ཆ་དེ་ཚོ་（s）དངོས་གནས་ཀྱ་ཆེ་པོ་ཞེ་དགོས་（o）རེད།“这些资源真的很重要。”

㉘ང་སྐོར་（s）བུ་བ་བདུན་པའི་ནང་ལ་（o）མ་རེད་པས།“望果节不是在七月份吗？”

㉙ཨ་གའི་དྲག་ལོག་གསར་པ་དེ་（s）ངའི་（o）རེད་།“那边的那件新衣服是我的。”

㉚ཀུ་མོ་རང་（s）ལྷད་པ་ག་ནས་（o）ཡིན་།“姑娘你是哪个地方的？”

㉛ཁོང་ནི་（s）རང་རྒྱལ་མི་དམངས་ཁྲད་ཀྱི་སྐྱེན་ཀྱགས་ཡོད་བའི་སྨན་པ་ཞེག་（o）ཡིན་།“他是我国历史上民间一个有名的医生”。

这些句子中包括了宾语的最主要类型，带指示代词或不定指示词的名词或名词组块㉒、㉛，疑问代词㉓、㉔，非谓动词组块㉕，形容词㉖、㉗，la 类位格名词组块㉘，属格名词组块㉙，从格名词组块㉚。显然，其中独用或者作后修饰语的代词、指示词、数量词以及名物化标记等形式都可能作为宾语与动词划界的形式标志，也可以作为判断宾语类型的参考性依据。为此，我们调查了相关的现象，获得了表 2 所列的数据。

表 2　判定句宾语的标记类型统计数

标记类型	不定指示词 ཅིག	疑问代词	指示代词	形容词	名物化标记	从格 ནས	位格 ལ	无标记名词组块
数量	32	15	13	19	10	2	3	60
比率(%)	20.8	9.7	8.4	12.3	6.5	1.3	2.0	39.0

从统计数据来看，不定指示词ཅིག（“某个，某些”）及其变体形式（ཞིག，ཤིག），无疑是名词性宾语的重要标志，占全部数据比率的1/5强，这与判定动词句式自身有关，宾语往往表示无定性事物(30)。名物化标记出现在判定动词前也是典型的形式化标志，它所构成的非谓动词组块具有名词性质。带从格和位格的宾语数量不多(28)、(30)，但也可以从形式标记加以判断。疑问代词和指示代词无论是自身作宾语还是作名词宾语的后修饰语，也能够清楚地与动词部分离析。形容词处在动词前与动词部分划界不难，但决定它的宾语性质还要判断它是否名词的后修饰语。以上统计数据中没有属格作宾语的情况(29)，它应该也是可识别的标记。最后，我们要指出，充当宾语的无标记名词组块中，有两个是专有地名，7 个是复杂的数量名组块，还有3个带形容词后修饰语，其它名词构成类型复杂，无法一一细分。

小结以上宾语与动词的讨论，虽然我们给出了充当宾语词串的组块特征，但实际上，单纯依靠动词类型和形式特征已基本可以解决判定句宾语与动词谓语分界问题。在154个例句中，正确识别率达99%以上，只有一例令我们调整了局部分块方法。即句子(32)，其中表示尊敬的语素ལགས 与书面语的判定动词同形，这类问题只需限定判定动词不能相连出现或动词 ལགས 不能出现在ཡིན、རེད之前便可解决。

(32) ཁྱོད་རང་（s/您）ཨ་མེ་རི་ཀ་ནས་པེབས་པའི་[从美国来（非谓动词）的（属格）]རྒྱལ་ཁབ་ཡུའེ་ཏོན་ཀུན་ལགས་（o/[约翰逊＋敬语素]）ཡིན་པས་།“您是从美国来的约翰逊先生吗？”

(33)བསྡོམས་ནས་（s/总计）ཇོར་མོ་བཞི་བརྒྱ་དགུ་བཅུ་（o）ཏག་ཏག་རེད་།“总共四百九十元。”

还有一类属于积累经验的问题，即动词前的修饰语包含哪些具体词语。因为形容词宾语和修饰名词中心词的形容词都位于动词之前，怎样区分它们与修饰动词的副词需要逐个甄别。例(33)的“ཏག་ཏག”有两个意思，“恰好，正好”和“纯粹、单独”，一般认为是形容词，实际分为两个词性较好。如果句法上作为副词理解，应该收入判断动词前修饰语的词表。

四　实验结果及讨论

本项试验是将藏语句子的识别分为两部分。第一部分识别是否判定句以及确定谓语动词与宾语的分界节点（参见第三节和文献［8］），第二部分判定主语。

实验的结果是，不能识别主语的句子32句，其中无标记名词主语句19句，主语省略的13句。而识别错误的7句，其中无标记名词3句，省略主语的1句，其他还有4例属于其他句子。也就是说，无标记名词句和省略主语的判定句完全不能识别。

在识别过程中，我们根据第二节的统计数据建立了一些关于主语的识别规则。（1）凡非句首且最右侧出现的语气词ནི་及ཡིན་ན་都判定为主语界标（ནི་也会偶尔出现在宾语成分之后，本数据中没有）；（2）凡不带属格、施格、位格标记的人称代词（含后附自指代词རང་）、指示代词及其复数形式（名物化标记后出现指示代词也归属词类），以及非谓动词组块均可判定为主语，但与动词相邻的འདི་འདྲས（“这样，这些”）属于宾语（习惯用法）；（3）数词（基数词和序数词）可独立作主、宾语，或位于作主、宾语的名词中心词之后，其自身可以看作是主语的标记；（4）以上三项都不能够与谓语动词相邻；（5）la标记的情况更复杂，需要向后继续搜索是否存在以上三类标记，但搜索节点止于至少动词前两个节点，当其

它标记出现并满足相关条件后，放弃 la 作为主语标志的尝试，反之则可以作为主语标志。

为实现以上条件，我们还建立了相关的函数与小型词表作为判断依据。如语气词词表、人称代词、疑问代词与指示代词词表、格标记词表、名物化标记表，以及基数词和序数词词表。为了判断动词与宾语分界点，也利用了课题的动词词表和动词语尾（词缀）表，并建立了可能出现在判定动词前的前置副词词表。

在误识的非名词主语和主语省略句子中，有些属于给定的规则不充分，如ཁ་གདན་[坐毯]ཚ་གཅིག་（s/一对）ལ་རྒྱ་ལོ་བཀྲ་བཅུ་（o/八百元）རེད་།“坐毯呐八百元一对。(34)”这是因为没有考虑数词之后仍然可能存在其他标记（此例应为ལ 标记主语）。还有一些情况很复杂，出现了多种符合规则的标记，如果从最左向起始来满足规则经常会出错。例如，ཁྱེད་རྣམས་ཚོ་（你们）ང་ཚོ་（我们）འབྲོག་པའི་ས་ལྟེ་ལ་（牧民的地区）སེབས་པས་（来[名物化]）དཀོན་པོ་（珍贵）རེད་།“你们到我们牧区来很不容易”(35)，这是一个非谓动词组块作主语的句子，可是判断为代词短语作主语了。主语省略句识别错误也是因为长词串宾语中存在一些以上用来识别主语的标记。主语省略的判定动词句是一个值得进一步讨论的问题。

总起来说，寻找主语标记的时候要尽可能向后溯，规则更细致，但是这样做难免不与宾语部分发生纠纷。所以，我们还需要从宾语结构中寻找关联要素。例如，当与动词相邻的宾语标记为ཅིག、ལ、ནས、名物化标记以及形容词后缀པོ 时，至少考虑其左向一个节点为宾语或宾语组成部分，如果与动词左向相邻的是指示代词或者疑问代词，则至少当前节点为宾语，以此来判断主语后溯的长度。当然，更细致、更具普遍性的规则需要从大规模语料库中逐步建立。最后应该指出，我们在其他藏语句式中一般采用逆向扫描方式识别组块，因判定动词句的特殊性而改用主宾语的顺向识别定界、宾语与动词的逆向识别定界，这个策略似乎还需进一步修正。总之，要大幅提高藏语判定句主、宾语的识别率，一个可能的方向是依赖从宾语获取的句法语义结构关系来解决无标记主语的判断问题以及主宾语识别的实现顺序问题。

参考文献

[1]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。

[2]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。

[3]吴云芳，段慧明，俞士汶：“‘是’字句主语和宾语的自动界定”，《中文信息学报》，2002 年第 2 期。

[4]周季文，谢后芳：《藏文阅读入门》，云南民族出版社，1998 年。

[5] Jiang Di: Recognition and Information Abstraction of Finite Verbs in Modern Tibetan, Paper for 20th ICCPOL'2003.

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

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