面向动词短语分析的再分类研究

达胡白乙拉

（内蒙古大学 蒙古学学院，内蒙古 呼和浩特 010021）

摘要： 基本词类的再分类研究对动词短语自动分析具有重要作用。根据动词短语分析所面临的具体问题，本文提出了面向动词短语分析的再分类框架，以便改善动词短语的分析效果。

关键字： 动词短语；自动分析；再分类信息

中图分类号码： H085.5　　**文献标识码：** A

蒙古语动词短语分析的主要任务是，利用计算机自动判定动词短语的边界，分析动词短语的内部结构关系。词类信息和形态信息给动词短语的自动分析提供一些必要条件。根据真实语料的统计分析，以并列副动词形式相连接的 VV①组合可以构成动词短语。从非终极符号角度讲，VV 可以与如下四种句法功能结构[1]（Syntactic Functional Structure，简称 SF-结构）相对应：①辅助关系，②状述关系，③宾述关系，④联合关系。这种对应关系归因于颗粒度过大的词类信息。把 VV 实例化（instantiation）后往往只有一种相对应的 SF-结构，也就是说，一般情况下，只能对应于上述四种关系之一。例如，使用 JOGSO/JV BAYI/L_A 来实例化，只能形成辅助关系；假如用 YAGARA/JV ASAGV/L_A 来实例化，就只形成状述关系。BAYI/L_A 是助动词，不能成为中心词。而 ASAGV/L_A 是实义动词，可以成为中心词。在这里，动词的再分类特征导致了截然不同的结构关系形成。因此，为判定 VV 的内部结构关系，有必要再分类 VV 组合中参加的动词。

可见，蒙古语动词短语的自动分析还需要再分类信息的支撑。对基本词类作进一步的子类划分及详细描述各个基本词类的语法属性是消解歧义、分化歧义结构的两个基本手段。[2]

为尽可能详细描述动词短语的类型信息，本课题研究在借鉴传统语法学成果的前提下，采用基于训练语料的再分类（corpus-based subcategorization）描述方法。从训练语料（近 4 万词）里挑选 2751 条动词短语，建立了动词短语库 VPset。根据 VPset 的实例，本文对动词短语的成分进行了再分类研究。

传统语法学，首先把动词分为实义动词和虚义动词，然后把虚义动词分为代动词、概称动词、联系动词和助动词。在蒙古语语料库的词类标注中，代动词和联系动词有专门标记，而助动词和概称动词没有，都用了和实义动词一样的标记。动词短语分析需要区别标记实义动词和助动词。在语法上，助动词有自己的特点，一般不能带宾语或状语，而是辅助别的词附加各种语法意义，或表示其语法变化。一般情况下，判断可否当作助动词可以借助《蒙古语语法信息词典》的动词语法属性库。但是，蒙古语的助动词和实义动词有时同形。在这种情况下，判别助动词往往需要其句法特征和语义特征的支持。蒙古语的 UJE 一词处在并列副动词形式之后往往是助动词，而在名词后往往是实义动词；作为助动词的 UJE 有“尝试”的意思，作为实义动词的 UJE 有“看、读”的意思。例如：BAYILDV/JV UJE（助动词）；NOM UJE（实义动词）。因此，有必要归纳成为助动词的句法条件和语义限制。至于语义特征的形式化，本文暂且不涉及。传统语法学按照助动词所辅助的词类，把助动词分成第一种助动词和第二种助动词。第一种助动词只能辅助动词，而第二种既可辅助动词也可辅助静词。[3]第一种助动词在外形上和实义动词相同，只是在语法功能上有所不同。让计算机判别这类助动词，难点有二：①开放性；②同形歧义分化。至于①，本文限制在 VPset 中出现的 23 条助动词实例上。至于②，本文利用句法条件进行分化歧义的尝试。具体判别过程如下：

① 读词条，查询《蒙古语语法信息词典》，判断是否有 TVSALA 属性；

② 如果没有，结束；

③ 如果有，判断句法条件的合格性；

④ 如果不合格，结束；

⑤ 如果合格，判定为助动词。

可见，归纳相关句法条件对这类助动词的判别比较重要。为此，本文对 VPset 中的这类助动词进行统计分析，并提出了相应的判别条件。请看表 5-1：

词条	实义动词②					实义动词前		
	副动词后			形动词后	动词	副动词形式		
	联合	并列	分离	现在将来	干后	联合	并列	分离
AB	+	+	+				+	
OG	+	+						
ALDA	+				+			
GAR	+	+	+					
OD	+							
ONGGERE	+	+	+					
YADA	+	+						
SAGV	+	+						
UJE	+	+						
ABACI	+	+						
YABV	+	+	+					
IRE（IR_E）	+	+	+					
ORO	+	+	+					
EHILE		+				+		
TALBI		+	+					+
CIDA		+						
OCI		+						
BARA		+						
OL		+					+	
DAGVS		+						
ORHI		+						

ABCIRA		+						
SIHA				+				
合计	38	361	43	1	1	1	21	1

表 5-1

为说明上文归纳的句法条件，以动词 AB 为例来了解一下具体判别流程：

① 读词条 AB，查询《蒙古语语法信息词典》，得知有 TVSALA 属性；

② 判断 AB 前面的词是否词类标记为 UIL 的动词；

③ 如果是，判断是否联合副动词，或并列副动词，或分离副动词；

④ 如果是，认定为助动词；

⑤ 如果不是，判断 AB 是否并列副动词形式；

⑥ 如果不是，结束；

⑦ 如果是，判断 AB 后面的词是否词类标记为 UIL 的动词；

⑧ 如果是，认定为助动词；

⑨ 如果不是，结束。

判断是否第一种助动词，以上句法条件也许不够，但我们首先利用比较松的语法限制，然后在大规模真实文本中检验时顺便完善。第二种助动词是词根为 BAI、BOL、A、BO 的动词。这些助动词辅助动词（形动词形式）时不存在歧义问题，而处于静词后就会产生歧义。在静词后，BAI、BOL 可以当作实义动词来使用，表示“存在、占有”、“成熟、再说”等意思。这种情况虽然比较少见，但是其判定涉及到语义、语用因素，仅从语法形式角度很难做出准确判断，本文先不考虑这种区别，只依据词根判断是否属于这种助动词。

限于当前的研究条件，本文把名词再分为如下三类：（1）人名，（2）地名，（3）其它。在蒙古语语料库中，已经对人名、地名进行人工标注加工，因此这样的再分类框架对动词短语分析来说很实用。在人名自动识别方面，在蒙古语文信息处理界已经有了一定的研究[4]，结果表明准确率达到 86%。然而，由于目前还没有真正实现在蒙古语语料库中自动识别标注，动词短语分析还不能直接利用这一研究成果。鉴于以上情况，在本文的再分类框架中仍沿用人名、地名人工标注结果，利用那些人名、地名标记[]和][考察以上三种名词子类和动词组成动词短语的特点，并归纳了相应的分析规则。对名词和动词之间的组合关系描述来说，这样的再分类只能提供一个框架信息，还不足以问题的彻底解决。

根据语气词的意义和用法，传统语法学把语气词再分为疑问语气词、肯定语气词、否定语气词、回忆语气词、推测语气词、转达语气词、责备惊叹语气词、强调语气词、让步语气词、呼唤或感叹语气词等 10 种。从组合角度讲，在蒙古语中，语气词置于中心动词之前和置于中心动词之后两种现象并存。但是，后置语气词与其前面的中心动词往往关系更加密切。例如：BASA MEDE/HU UGEI。在这个例子中 BASA 是叠用语气词,置于相关动词 MEDE/HU 的前面;而 UGEI 是否定语气词,置于 MEDE/HU 的后面。按句法位置，本文把语气词划分为置前语气词和置后语气词。对 VPset 的语气词实例进行统计分类的结果如表 5-2 所示：

语气词	置后 语气词	置前 语气词	举例
UGEI	+		TANI/HV UGEI
YVM	+		(*HOMOJIL*) OLGO/HV YVM
YVMSAN	+		BODO/GSAN YVMSAN
CV(CU)	+		ASAGV/HV CV (*UGEI-BER*)
LA(LE)	+		ILEREGUL/HU LE (*TEDUI*)
BOL	+		HODELMURILE/HU BOL
CINI	+		INGGI/GSEN CINI
NI	+		NIS/HU NI
BVI	+		JUIRLE/HU BVI
VV(UU)	+		OGOCI/N_A VV
SIV(SIU)	+		HELE/GSEN SIU
DA(DE)	+		DVVGAR/V/N_A DA
BISI	+		TORO/GSEN BISI
BILE	+		ERGU/GSEN BILE
BIJE	+		UJE/GSEN BIJE
AJI	+		HAMIYAR/HV AJI
A	+		(*MEHELE/JU BAYI/G_A BISI*) BAYI/HV A
ULU		+	ULU MEDE/HU
BUU		+	BUU IDEGUL/U/GEREI
BITEGEI		+	BITEGEI JOBA
ESE		+	ESE TOGDA/GSAN (*MAyIG-TAI*)
BASA		+	BASA YARILCA/BA

表 5-2

蒙古语传统语法学对情态词的分类也按照其意义来进行。本文对情态词的再分类和语气词一样，也是按照句法位置分为置前情态词和置后情态词。在蒙古语中，置后情态词和它前面的词语关系更为紧密。例如：JABAL TOGDA/GSAN UJELTE OBOR BAYI/HV HEREGTEI。 其中，JABAL 是置前情态词，而 HEREGTEI 是置后情态词。VPset 中出现的置前情态词有 LABTAI, YERU--DEGEN；置后情态词有 HEREGTEI, HEREG=UGEI, CIHVLATAI, YOSOTAI, BOLOLTAI。从数量看，构成动词短语的情态词中，置后情态词占 92.86%。

蒙古语动词短语自动分析可能还需要更详细的再分类信息。例如，把代词分为人称代词、反身

代词、指示代词等。通过在真实文本中的自动分析测试，再决定还应该再分类哪些词类和怎么分类。在本文仅以蒙古语基本动词短语的自动分析为例，提出了面向动词短语分析的再分类框架。完整而详实的、真正起作用的再分类体系的研制往往和动词短语的结构特点有密切关系，需要进一步研究。

注释

①本文引用了“面向信息处理的蒙古语文标记集”。

②在处理的时候，先看作非 BAI（BAYI/）、BOL 词根动词。

参考文献

[1]冯志伟.自然语言的计算机处理[M].上海:上海外语教育出版社，1996. 185.

[2]俞士汶.关于汉语短语结构体系及其描述方法的说明[M].北京:北京大学计算语言学研究所，1994，5.

[3]内蒙古大学中国语言文学蒙语教研室.现代蒙古语[M].呼和浩特:内蒙古人民出版社，1964. P.760

[4]那顺乌日图等.蒙古文人名自动识别研究[A].孙茂松,陈群秀主编.语言计算与基于内容的文本处理[C].北京:清华大学出版社，2003，7. .97-102.

[5]确精扎布.关于蒙古语词类[J].内蒙古大学学报，1962，(1).

[6]白音门德.关于时间词[D].中国蒙古语文学会第七次年会，1999，7.

[7]侯敏,孙建军.汉语机译中汉语分析的策略问题[J].语言文字应用，1996，(3).

[8]清格尔泰.蒙古语语法[M].呼和浩特:内蒙古人民出版社，1991.

[9]James Allen. Natural Language Understanding[M]. The Benjamin/Cummings Publishing Company, Inc. 1995.

A study of Subcategorization for verb phrase Analysis

Dabhurbayar

(Mongolian Language Institute, College of Mongolian studies,Inner Mongolia University,Huhhot，010021 China）

Abstract: Subcategorization for Mongolian basic word class takes and important role in automatic analyses of Mongolian verb phrases. In this paper, a subcategorization frame for verb phrase analysis is proposed according to the problems in research of Mongolian verb phrase analysis.

Keywords: verb phrase; automatic analysis; subcategorization

收稿日期: 2005-11-20;

基金项目: 国家自然科学基金项目（60263001）; 国家社会科学基金项目（02BYY036）

作者简介: 达胡白乙拉（1977—），男，蒙古族，内蒙古科右中旗人。内蒙古大学蒙古学学院蒙语所博士，主要研究方向为计算语言学与蒙古语文信息处理。