蒙古语助动词标注与分析*

达胡白乙拉　萨仁图雅
内蒙古大学蒙古学学院蒙古语文研 呼和浩特 010021
E-mail: dabhvrbayar@163.com　bsrty@163.com

摘　要：蒙古语助动词的判别涉及蒙古语词法、句法和语义问题，是蒙古语语法学较难的研究课题之一。作者对助动词在真实文本中的分布特征进行分析，描述与助动词共现词语的形态变化、词类等特征，归纳蒙古语23 条常用助动词的判别规则。在此基础上，研制助动词标注软件，对现代蒙古语语料库进行标注，分析标注错误例子，改进了助动词标注软件。再测试表明，改进后的规则可以准确判别100 万词级现代蒙古语语料库的绝大多数助动词。

关键词：蒙古语；助动词；标注；分析

Tagging and Analyzing Mongolian Auxiliary Verbs

Dabhurbayar　Sarantuyag_a
School of Mongolian Studies, Inner Mongolia University, Hohhot 010021
E-mail: dabhvrbayar@163.com　bsrty@163.com

Abstract: It is one of the arduous tasks in Mongolian linguistics to distinguish Mongolian auxiliary verbs because it has close relations to many aspects of the Mongolian language such as morphology, syntax and lexical meaning. The authors analyzed the distributions of Mongolian auxiliary verbs in real text, described the morphological and POS features of their co-occurred words and concluded the rules of distinguishing the 23 auxiliary verbs which are frequently used in Mongolian language. In terms of these rules, we developed a tagger and tagged the auxiliary verbs in Mongolian corpus. The test of the tagged results shows that most of the auxiliary verbs in the corpus can be tagged correctly in light of the improved rules.

Keywords: the Mongolian language, auxiliary verbs, tagging, analyses

1　前言

自然语言处理中，常常利用对基本词类的再分类信息来分化歧义，提高分析精度。蒙古语助动词的标注与分析是蒙古语动词再分类信息的获取任务之一，也是蒙古语语料库加工的重要研究课题。

蒙古语语法学，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。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，第二种既可以辅助动词也可以辅助静词。第一种助动词在外形上与实义动词相同，在语法功能上不同。因此，第一种助动词的标注与分析，实质上是一种动词功能歧义的分化任务。

第二种助动词是词根为 BAI、B0L、A、BO 的动词。这类助动词可以辅助动词，也可以辅助静词。出现在动词后边的时候，只有辅助中心动词的功能；出现在静词后边的时候，辅助静词，构成谓语结束式。现代蒙古语中也有少数词根为 BAI、B0L 的实义动词，可以出现在静词后边，

*本项研究得到国家社会科学基金项目（07CYY024）、内蒙古自治区哲学社会科学规划项目（06B057）和国家自然科学基金项目（60763003）的支持。

表示“存在、占有”、“成熟、再说”等意思。因此，在静词后边出现的词根为 BAI、B0L、A、BO 的动词在句法语义上有歧义。这种情况比较少见，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。

在语法上，助动词有自己的特点，一般不能带宾语或状语，而是辅助其他词语，表示各种语法意义，构成各种语法变化形式。

目前，在蒙古语语料库标注中，不分第一种助动词和第二种助动词，标记使用了 VZ。从标注结果看，只对第二种助动词进行了标注，而没有对第一种助动词进行区分标注。对于第二种助动词和实义动词的同形歧义问题，也有待进一步探索和解决。因此，本文的研究目标就成为在蒙古语语料库中区分实义动词和助动词，对第一种助动词进行标记。

2 判别规则

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，在真实语料中，判别第一种助动词，解决与实义动词同形歧义问题，往往需要其语法特征、语义和语用特征的支持。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）；N0M UJE（实义动词）。因此，有必要归纳助动词判别的各种特征。

怎样归纳这些特征？对助动词识别来说，语法特征的作用大呢，语义特征的作用大呢？还是语用特征的作用大呢？至于这些问题,我们先归纳形态、词类、上下文等方面的特征，进行对真实语料的标注与分析，探索这些特征的实际作用。

让计算机判别第一种助动词，难点有二：①开放性；②同形歧义分化。第一种助动词有多少个,哪些动词可以成为这类助动词？至于这些问题,蒙古语语法学至今还没有一个穷尽的定论。作者从现代蒙古语语法著作和现代蒙古语语料库的部分语料中分析选定 23 条助动词，根据其出现的上下文和本身的特征归纳了助动词判别规则，然后利用这些规则，对现代蒙古语语料库进行自动标注和人工分析。蒙古语助动词判别规则如下表所示：

词条	实义动词					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	+	+	+					
0D	+							
ONGGERE	+	+	+					
YADA	+	+						
SAGV	+	+						
UJE	+	+						
ABACI	+	+						
YABV	+	+	+					
IRE（IR_E）	+	+	+					
0R0	+	+	+					

EHILE		+				+		
TALBI		+	+					+
CIDA		+						
0CI		+						
BARA		+						
0L		+					+	
DAGVS		+						
0RHI		+						
ABCIRA		+						
SIHA（SIH_A）				+				
合计	38	361	43	1	1	1	21	1

下面以动词 AB 为例说明上表归纳的语法特征与规则：

① 判断 AB 前面的词是否实义动词；

② 如果是，判断是否联合副动词，或并列副动词，或分离副动词；

③ 如果是，认定助动词；

④ 如果不是，判断 AB 是否并列副动词形式；

⑤ 如果不是，结束；

⑥ 如果是，判断 AB 后面的词是否实义动词；

⑦ 如果是，认定助动词；

⑧ 如果不是，结束。

这样的判别过程会涉及实义动词的判定问题。对于这个问题，我们采取利用词根排除方法。蒙古语中，词根不是“BAYI/（BAI）或 B0L/（B0L）”的动词都属于实义动词。在词根为“BAYI/（BAI）或 B0L/（B0L）”的动词中也有一部分是实义动词，但是这些实义动词不会与其后出现的第一类助动词组合。因此，具体判别的过程中，我们可以把实义动词定义为词根不是“BAYI/（BAI）或 B0L/（B0L）”的动词。

3 标注软件

根据上文归纳的语法特征和规则，用 VC 语言编写了一个标注程序。程序可以对词性标注的现代蒙古语语料进行第一种助动词的区分标注。标注格式如下：

… …HAR_A Ac TANGGIS-VN Ne1 SIGVRG_A Ne2][ANGgLI Nt2][FRANcI-YIN Nt2 HOLGE Ne1 0NGG0CATV Ac ANGGI-YI Ne1 DAGARI/GSAN Ve1 YABVDAL Ne1 CAG Ne1 AGVR-VN Ne2 MEDEGE/N-U Ne1 BVI Ve2 B0L/0/GSAN-DV Vz **AB/CV Ve1/Vz1** HELE/BEL Ve1 GENEDTE-YIN Dc HEREG Ne1 B0L/0/N_A Vz .（黑海风暴袭击英、法轮船的事件，对于已有天气预报的时代来说，是一件突然事故。）… …

… …J0CID-I Ne2 page210 YAGV/N-V Ra EMUN_E Oa SU'-TEI Ne2 CAI Ne2 SIR_A Ac T0S0-BAR Ne2 DAYILA/GAD Ve1 , W, DARAG_A Oa NI Sf ARIHI Ne2 MIH_A-YIN Ne2 JUIL-I Ne1 GARGA/JV Ve1 **IRE/N_E Ve2/Vz1** .（首先以奶茶、黄油招待客人，然后才以酒肉之类的食品招待。）… …

上面的例子中，Vz1 是第一种助动词的标记。

在标注结果的获取上，可以句子形式提取，并按照助动词的不同进行分类和统计标注例子的数量。对 100 万词级现代蒙古语语料库的标注结果分类统计情况如下表所示：

序号	助动词	标注次数	助动词	标注次数
1	AB|AB/	517	0R0|0R0/	94
2	OG|OG/	399	EHILE|EHILE/	105
3	ALDA|ALDA/	3	TALBI|TALBI/	45
4	GAR|GAR/	425	CIDA|CIDA/	562
5	0D|0D/	10	0CI|0CI/	25
6	ONGGERE|ONGGERE/	39	BARA|BARA/	15
7	YADA|YADA/	50	0L|0L/	225
8	SAGV|SAGV/	178	DAGVS|DAGVS/	16
9	UJE|UJE/	278	0RHI|0RHI/	17
10	ABACI|ABACI/	14	ABCIRA|ABCIRA/	33
11	YABV|YABV/	360	SIH_A|SIHA/	14
12	IR_E|IRE/	449		
合计				3873

4 结果分析

在面向人的传统语法学中，助动词的判别研究实属难题之一。因为这类助动词不仅与整个句法结构的分析相关，而且还和词汇意义和上下文密切相关。既然要判别，那就要有个判别的标准。根据蒙古语语法学著作的相关论断和真实语料中的助动词特点，作者制定如下标准，用来人工判别计算机标注结果的正确性。人工鉴别标准如下：

A. 句法标准：至少与一个动词连着出现，只作为中心词的辅助成分。

B. 意义标准：说明中心动词的语法意义，例如趋向、受益、情态、动词的体范畴。

C. 形式标准：看能否省略，或看省略以后词汇意义有什么变化没有。

例子：

（1）… … H0RIYA/N-V-BAN Ne1 DARVG_A-DV Ne1 0CI/JV Ve2 CILOGE Ne2 **AB/V/GAD** Ve1 , W, NOHOD-TEI-BEN Ne2 SAL/HV Ve2 Y0S0 Ne1 HI/JU Ve1 ,… …（向团长请假，向同志们告别， … …）

（2）… … CINGG_A Ac HUCUTEI Ac ARG_A Ne1 HEMJIY_E Ne2 **AB/CV** Ve1 H0RIGLA/N Ve1 J0GS0GA/GSAN Ve1 BAYI/N_A Ve2 .（… … 采取强有力的措施进行制止。）

第一个例子中，“AB/V/GAD”没有和动词连着出现，也不能省略；第二个例子中，“AB/CV”虽然和动词连着出现了，但句法结构上它没有辅助中心动词的作用，而是自己担任了名词短语“CINGG_A Ac HUCUTEI Ac ARG_A Ne1 HEMJIY_E Ne2”的支配成分，也不能省略。因此，根据上述三个标准，以上两个例子中的词根为“AB/”的动词都不能成为助动词。

为了系统说明具体测试分析的过程，作者以 AB 一词为例，展开讨论在计算机标注结果中出现的错误类型、出错原因以及相应的解决办法。关于 AB 一词，程序标注了 517 条例子。根据上述三个标准人工判别后发现其中有 34 例为错误标注结果。错误类型和解决办法的探讨如下：

1）把复合词的一部分误标为助动词

例子：

… … HALTAR Ac N0HAI Ne1 … … HAMAR Ne1 H0NGSIYAR-IYAR-IYAN Ne1 SINGSI/JU Ve1 UJE/GED Ve1/Vz1 **AB/V/N** Ve1/Vz1 D00R_A-BAN Oa IDE/GSEN Ve1 UGEI

Ve2 , W, … …（… … 花狗 … … 用鼻子和嘴嗅闻知后，没有马上吃掉 … …）

… … VHVSHI/N Ve2 HODEL/JU Ve2 H0YAR M TAtAR-I Ne1 **AB/CV** Ve1/Vz1 HAYA/HV Ve1 GE/TEL_E Vx , W, … …（… … 冲上去想干掉两个鞑靼人 … …）

上述两个例子中，程序把复合词“AB/V/N Ve1/Vz1 D00R_A-BAN Oa”（马上、立即）和“AB/CV Ve1/Vz1 HAYA/HV Ve1”（干掉）分开理解，从而误标其中的“AB/V/N”和“AB/CV”为助动词。这种错误归因于蒙古语语料库中的复合词标注还不够完善。这样的错误共有 24 例，占 AB 一词标注错误总数的 70.59%。通过蒙古语复合词词库的更新和扩充，可以减少此类错误的出现。

2）把作为中心词的动词误标为助动词

蒙古语的助动词与其邻近的动词构成辅助结构关系。也就是说，在辅助关系中才有可能存在助动词。从这个角度讲，蒙古语动词相关的句法结构关系的准确判断有时候可以成为助动词判别的重要基础。换句话说，助动词的识别有时候以句子结构关系的判别为前提。程序标注结果中，出现了如下 2 种误标情况：

（1）把辅助关系中的中心词误标为助动词

例子：

… … GAGCA Ac []JVNDA Nt1 LA Sh JORIGUDLE/GED Ve1 **AB/V/GSAN** Ve1/Vz1 UGEI Ve2 .（… …只有准达（人名）一人拒绝，没有要。）

… … TAtAR Ne1 OHID Ne1 INIYELDU/N Ve2 J0GS0/GAD Ve2 **AB/HV** Ve1/Vz1 UGEI Ve2 BAYI/HV-DV&NI Ve2 TERE Rj T0GLAGAM-IYAN Ne1 0RHI/JV Ve1 , W, … …（… … 鞑靼姑娘们站着逗笑,但不接纳，他扔掉玩具，… …）

上述两个例子中，程序把辅助关系“AB/V/GSAN Ve1/Vz1 UGEI Ve2”和“AB/HV Ve1/Vz1 UGEI Ve2”中的中心动词“AB/V/GSAN”和“AB/HV Ve1/Vz1”误标为助动词。其原因是，程序只看前面一个词语的变形形式，而且还不能识别短语边界。这样的错误共有 2 例，占 AB 一词标注错误总数的 5.88%。目前，通过判断当前词语两边的语法特征，可以纠正这类错误的一部分。

（2）把并列关系中的中心词误标为助动词

按照连接成分的性质，可以分为如下 2 种情况：

1）把词语并列关系中的动词误标为助动词

例子：

… … HELEGEI Ac B0L/BACV Vz , W, OG/CU Ve1 **AB/V/BAL** Ve1/Vz1 COM Rx B0LCIHA/HV Vz YVM Sb , W, … …（… … 虽然结巴，但在给要的事情上还是可以的 … …）

… … ARIHI/N Ne2 HVNDAG_A Ne1 ERGULCE/JU Ve1 , W, OG/CU Ve1 **AB/HV-BAN** Ve1/Vz1 T0GTA/L_A Ve2 .（… … 举起酒杯，说定给要的关系。）

上述两个例子中，程序把词语并列关系“OG/CU Ve1 AB/V/BAL Ve1/Vz1”和“OG/CU Ve1 AB/HV-BAN Ve1/Vz1”中的动词“AB/V/BAL ”和“AB/HV-BAN”误标为助动词。其原因是，程序还不能识别短语结构关系。这样的错误共有 2 例，占 AB 一词标注错误总数的 5.88%。目前，还没有有效的办法可以纠正这类错误。

2）把短语并列关系中的动词误标为助动词

例子：

… … GAR Ne1 DAMJIGVL/V/N Ve1 SARIGALA/HV Ve1 $ATV Ne1 $ATV-BAR Ne1 UN_E Ne2 OSHE/HU Ve1 YABVDAL-I Ne1 ERES Ac ARG_A Ne1 HEMJIY_E Ne2 **AB/CV** Ve1/Vz1 H0RIGLA/HV Ve1 HEREGTEI Ac .（… …对于接连倒卖环环增加价位的事情，采取严格的措施，必须制止。）

… … DALAI Ne1 D0T0R_A-ACA Oa NIGE/N M DVSVL Ne1 VSV Ne2 **AB/CV** Ve1/Vz1 T0M0RAGVL/HV Ve1 SIL-IYER Ne1 AJIGLA/BAL Ve1 … …（… … 从海里要一滴

水，用扩大境观察...）

上述两个例子中，程序把短语并列关系“ERES Ac ARG_A Ne1 HEMJIY_E Ne2 AB/CV Ve1/Vz1 | H0RIGLA/HV Ve1 HEREGTEI”和“DALAI Ne1 D0T0R_A-ACA Oa NIGE/N M DVSVL Ne1 VSV Ne2 AB/CVVe1/Vz1 | T0M0RAGVL/HV Ve1 SIL-IYER Ne1 AJIGLA/BAL Ve1”中的动词“AB/CV Ve1/Vz1”误标为助动词。其原因是，程序还不能识别短语结构关系。这样的错误共有 6 例，占 AB 一词标注错误总数的 17.65%。目前，还没有特别有效的办法完全克服这类错误。

从理论角度讲，这种错误也可以出现在从句并列关系中。纠正这类错误，需要以句子中所有短语或从句之间界限和句法关系的正确识别为前提。因此，目前无法完全解决这类错误。而且，随着语料量和种类的增加，这样的错误增加与否，也是不确定的。根据上述讨论，我们利用复合词标注的更新与当前词语双边词语的语法特征的相关性参考的办法改进了标注软件。作为结果，对助动词 AB 的标注错误由上述 34 例减少到 8 例，标注准确率达到 98.37%。

通过人工测试和分析，改进了助动词的标注软件，对 100 万词语语料库标注结果进行了自动标注。标注结果如下：

序号	助动词	标注次数	标注错误	出错率
1	AB|AB/	491	8	1.63%
2	OG|OG/	391	5	1.28%
3	ALDA|ALDA/	1	0	0%
4	GAR|GAR/	411	25	6.08%
5	0D|0D/	6	2	33.33%
6	ONGGERE|ONGGERE/	32	3	9.38%
7	YADA|YADA/	44	0	0%
8	SAGV|SAGV/	128	66	51.56%
9	UJE|UJE/	162	27	16.67%
10	ABACI|ABACI/	11	2	18.18%
11	YABV|YABV/	331	19	5.74%
12	IR_E|IRE/	365	84	23.01%
13	0R0|0R0/	68	32	47.06%
14	EHILE|EHILE/	102	2	1.96%
15	TALBI|TALBI/	44	4	9.09%
16	CIDA|CIDA/	498	16	3.21%
17	0CI|0CI/	18	2	11.11%
18	BARA|BARA/	11	2	18.18%
19	0L|0L/	188	12	6.38%
20	DAGVS|DAGVS/	13	0	0%
21	0RHI|0RHI/	17	0	0%
22	ABCIRA|ABCIRA/	33	7	21.21%
23	SIH_A|SIHA/	14	0	0%
合计		3379	318	9.41%

5 结语

蒙古语助动词的判别涉及蒙古语词法、句法和语义问题，是蒙古语语法学中较难的研究课题之一。作者对助动词在真实文本中的分布特征进行分析，描述与助动词共现词语的形态变化、词类等特征，归纳蒙古语23条常用助动词的判别规则。在此基础上，研制一个基于规则的助动词标注软件，对现代蒙古语语料库中进行标注，测试标注结果，分析标注错误的原因，改进了助动词标注软件。在蒙古语助动词判别规则中着重集成助动词本身的形态特征、前后一个词语的形态变化和词类特征。测试结果表明，这样的规则可以准确判别100万词级现代蒙古语语料库中的绝大部分助动词。蒙古语助动词的判别有时候以整个句子或短语的结构分析为先决条件。这样的情况下，目前还不能以自动方式准确判别助动词。

参 考 文 献

[1] Nicholas Poppe. Grammar of Written Mongolian. Otto Harrassowitz, Wiesbaden, Germany. 1954.

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

[3] Borjigin Schenbaatar. The Chakhar Dialect of Mongol. The Finno-ugrian Society, Helsinki. 2003.

[4] 冯志伟. 机器翻译研究. 北京:中国对外翻译出版公司，2004.

[5] 詹卫东,. 汉语短语结构定界歧义类型分析及分布统计. 中文信息学报，1999，3.

[6] 周强. 汉语短语的自动划分和标注. 中文信息学报，1997，1.

[7] 内蒙古大学蒙古语文研究所. 现代蒙古语. 呼和浩特:内蒙古大学出版社，2005.

[8] 清格尔泰. 现代蒙古语语法. 呼和浩特:内蒙古人民出版社，1999.

[9] 华沙宝. 蒙古语词类标注系统. 计算语言学文集. 北京:清华大学出版社，1999.

[10] 那顺乌日图. 蒙古语语法信息词典框架设计. 内蒙古大学博士学位论文，2000.

[11] 达胡白乙拉. 蒙古语基本动词短语自动识别研究. 内蒙古大学博士学位论文，2005.

[12] 内蒙古大学蒙古语文研究所. 面向信息处理的蒙古语词语标记集. 技术文件，2004-2008.