蒙古语短语结构树的自动识别

乌兰[1] 关晓炟[2]

[1]内蒙古大学蒙古学学院 呼和浩特 010021

[2]中国科学学院自动化所 北京 100190

摘要：句法分析在自然语言信息处理中处于非常关键的位置。本文在描述蒙古语特点的同时提出蒙古语句子中短语结构分析难点。根据蒙古语自身特点，归纳了短语标注体系，建立了蒙古语短语树库，尝试实现蒙古语句子的自动分析。初次开发的句法分析器的分析准确率达到 62%，自动分析器的测试结果表明该分析器能在较大程度上辨别出短语结构类型，能生成句法树结构，但在短语结构内部关系方面的识别效果还有很大改进空间。最后总结了分析器近期能解决的相关问题。

关键词：蒙古语；短语结构语法；句法树库；自动识别

Automatic recognition of Mongolian phrase structure tree

[1]College of Mongolian Studies Inner Mongolia University Hohhot 010021

[2]Institute of Automation ,Chinese Academy of Sciences Beijing 100190

Abstract: Syntactic Analysis is in a very critical position in natural language processing. This paper, by describing Mongolian characteristics, to puts forwards the difficulty in analyzing phrases structure in Mongolian sentences. According to the characteristics of Mongolian, it induced the phrase annotation scheme, established a Mongolian phrases Treebank and tried to realize automatic analysis of Mongolian sentences. The accurate rate of sentences is 62% by using the Initial analyzer. The result of automatic analyzer shows that it can identify to a large extend the type of phrase structure, and also can generate the structure of syntax Treebank. However, the internal relationship of phrase structure also has much to prove. In the last part of this paper, it sums up the problem which the analyzer can sole at present.

Key Words: Mongolian ; Phrase Structure Grammar; Syntax Treebank; Automatic recognition

1. 引言

20 世纪八十年代开始，蒙古语基于语料库的信息处理研究一直在进行。目前完成了字、词处理阶段的基本任务，步入句子处理阶段。句法分析在自然语言信息处理当中处于非常关键的位置，它能为篇章处理、语义分析提供有效的帮助。句法分析和树库的建设有互相推进的作用，树库作为标准数据评价自动分析器的处理质量，为理论语言学研究提供客观的真实文本标注数据，而自动句法分析为建设大规模树库提供了可能性。

基于短语结构语法来分析句子和建设树库方面，英语、汉语等一些语言的研究有了可喜的成果。几个典型的句法标注语料库有美国的PTB[1]，德国的Tiger[2]，西班牙语有UAM[3]，
还有美国滨州大学的汉语树库CTB[4]和清华大学汉语树库TCT[5]等。蒙古语短语树库标注体系[6]跟清华TCT有相似点，TCT选择了大规模的包含新闻、学术、文学、应用四大体裁的平衡语料文本作为加工对象。[1]它覆盖了汉语“字/词->块 ->句 ->段”等各个层次的句法单元，来形成汉语句子最为详细的句法信息描述。它设计了双标记集的描述体系：一是成分标记集，二是关系标记集。

蒙古语短语结构树可以表示句子较全面的句法信息，包括从词、短语到句子的句法单位。词与词之间的搭配和同现，短语的内部结构和功能分类等，都可以在短语结构中得以体现。
因此我们选择建设短语结构树库来尽可能地反映蒙古语句子组织信息情况，以期尽可能详细地描述蒙古语

收稿日期：2014-5-31 定稿日期：2014-7-22

[1] 本项研究获得中组部万人计划.青年拔尖人才支持计划、教育部项目（13Y740014）、内蒙古高等学校科技英才支持计划、内蒙古自治区蒙古语言文字信息化专项基金的资助。

句子的句法组合信息。

蒙古语短语结构类型和结构内部关系的识别判定是蒙古语句法分析的一项重要内容，也是蒙古语语料库多级加工处理的一个重要环节。蒙古语句法分析研究大多属于基于举例法的句法研究,近期面向信息处理的句法研究也较少见，可以分为有基于短语结构语法分析和基于依存语法分析两种。基于短语结构的句法研究有内蒙古大学所作的对语料库短语标注切分研究[7] [8] [9] [10]；基于依存语法的句法研究有内蒙古大学所作的依存句法分析研究[11]，依存句法与短语结构句法是两种不同体系的研究方法，因此做蒙古语短语结构句法分析器是很有必要。目前这些研究的目标基本统一，通过不同视角对蒙古语句子结构进行分析研究，试图探索蒙古语句子的组合方式和层次结构特征，对句法有一个较清晰的认识，为进一步的计算机处理构绘一定的基础框架。因此本文的蒙古语短语结构树的自动识别研究能为以后构建大规模的蒙古语树库积累经验并将会促进计算机模仿人理解和使用蒙古语的心理过程，为计算机理解蒙古语提供一个行之有效的环节。它还有利于蒙古语句子的结构和性能研究。就应用来说，它在训练基于短语结构的机器翻译、信息检索、信息抽取、问答系统、自动校对等各种应用系统中有着不可缺少的作用和意义。

2. 蒙古语描述特点

蒙古语是黏着性语言，词是一般可以分解为词根（词干）和词缀两个部分，有的词根可以单独使用，词干上加接构词词缀可以派生新词，在派生词上再接续构词词缀或构形词缀还可以构成新词或增添语法意义，蒙古语中较长的多音节词一般都是几个构词词缀和构形词缀依次相加的结果[12]。它富有形态变化。静词类有格、领属、数范畴的形态变化。动词类有式、体、态等范畴变化和连接形、兼役形变化。句子中词与词的句法关系是通过这些形态变化来表达的。在《信息技术-信息处理用蒙古文词语标记集》[13]（GB/T 26235-2010）（下面简称《国标》）把蒙古语的构形附加成分分为数范畴、格范畴、领属范畴、形容词级范畴、数词变化形式、祈使式、陈述式、副动词、形动词、名动词、态范畴、体范畴、附属等 13 大类。在句子中这些构形附加成分的出现如例子：

[]LeNIN/Nt1（列宁） IRE/Ve2+JEI/Fs11（来过了）（列宁来过了）。

ABV/Ne1（父亲）JOBSIYERE/Ve1+BE/Fs14（同意了）（父亲同意了）。这些句子里的动作是通过动词陈述式的表示过去时的词缀“JEI，BE”来表示这个行为已经完成了。

蒙古语形态变化丰富，例如:“ABV/Ne1 JOBSIYERE/Ve1+N_E/Fs21”“ABV/Ne1 JOBSIYERE/Ve1+HU/Ft12”“ABV/Ne1 JOBSIYERE/Ve1+JU/Fn1” “ABV/Ne1 JOBSIYERE/Ve1+GSEN/Ft11”(爸爸同意)，以上四个短语是通过不同的构形附加成分来表达“爸爸同意”这个行为。是在动词词根“JOBSIYERE”上加不同的词缀“N_E” “HU”“JU”“GSEN”来表达“爸爸同意”这个行为，在不同的语境里分别使用，但基本语义不会发生变化。在短语结构分析当中无论它的动词有多少变化，它就是体述关系的短语。

在蒙古语的构形附加成分中“格”表示名词和其他词的关系以及它在短语和句子中的功能。蒙古语的“格”通过在静词之后接续某种词缀来表示[14]。比如，在蒙古语句子中有时名词和名词会发生关系，例如:“MONGGOL=HELEN/NT-U/Fc11 HICIYEL/Ne1(蒙古语课程)”,这个短语中的两个名词是所属关系，因此在两个名词之间加入蒙古语的“属格”，即“U/Fc11”，相当于汉语的“的”；除此之外，名词与动词也可能会发生关系，例如:“VSV/Ne2-BAR/Fc51 VHIYA/Ve1+GSAN/Ft11（用水洗)”,此时就要在名词后面加“工具格”，即“BAR/Fc51”。在《国标》里把蒙古语的“格”分为主格、属格、与格、宾格、从格、工具格、共同格、联合格、定格等九种。蒙古语的“格”在句子中可以充当主语、定语、宾语和状语等句子成分。在蒙古语短语结构句法树库里，短语结构内部关系的宾述关系、体述关系、状述关系、定体关系的内容跟“格”有很大的关联。这样一来，“格”对蒙古语短语结构句法树库的影响是可想而知的。

两种格之间的歧义问题是自动分析器的一个困难点。比如，短语结构分析句子的时候蒙古语的间接宾语与状语有的时候很难区分。静词的工具格有的时候构成间接宾语，有的时候构成状语。在个别情况下，同样一个形式有时可以表示宾语也可以表示状语。例如:“MORI/Ne1-BAR/Fc51 YABV/Ve2+N_A/Fs21（骑马走)” 中是宾语，“SVRGAGVLI/Ne1-BAR/Fc51 TOGORI/Ve1+Y_A/Fb11（校园里逛)”是状语。两个短语都是“名词-工具格 动词”形式，但是句子中做的成分却不一样。还有些传统语法学论著明确指出成分句的主语可以以宾格形式存在。这意味着宾述关系和体述关系之间一定会产生同形歧义问题。[15]蒙古语主格是零

形式，特别是在体述关系、定体关系里出现的频率较高，因此只能依靠“格”来辨别是不够的，还需要词性、语义等信息。这样一来，这些歧义对句法分析器分析短语内部关系带来很多困扰。辨别上文提到的两个短语句子成分的时候我们依靠大脑的语言知识和理解能力，但是计算机处理方式是形式化，类似于上文提到的短语就很难辨别出来。

蒙古语的语序比较灵活，但是中心词的位置基本上是固定的（除了特殊句型以外），处在后部分，蒙古语的句子结构是主宾谓（SOV）形式。这对分析器产生短语结构类型有了基本的理论依据。如把《名词-动词》 《动词-动词》 《副词-动词》 《摹拟词-动词》 等以动词为中心词的短语称之为动词短语。 例如：“YEHE/Ac-BER/Fc41 HI/Ve1+JU/Fn1（大量做）”是动词短语,因为“HI/Ve1+JU/Fn1（做）”是中心词，处在短语的后部分，它的词性是动词。

虚词对句子分析中也占有自己的位置，比如用后置词、时位词、连接词等来连接两个词或者短语。有些内部关系通常通过一些虚词后它的特征会很明显，我们可以通过这些虚词来确定短语内部关系，比如它含有‘BOGED/MORTEGEN/BA/BOLVN/,/、’等词或符号的时候是联合关系。

3. 蒙古语短语树库简介

3.1 蒙古语短语树库的词语类标记

产生短语结构树的时候第一步工作是进行固定短语标注工作，我们使用固定短语标注系统和结合人工校对，用“=”号连接。标注语料实例如下：“EHE ORON”要用“=”符号连接起来，是“EHE=ORON”（“祖国”的意思）形式。有些固定短语还会被漏掉，因此需要人工校对和加以修改。实例：“HOMON TOROLHITEN”->“HOMON=TOROLHITEN”（人类）。

在此基础上，我们要词法标注。2010 年内蒙古大学与中国科学院合作并研制了基于统计的词法分析器-Mglex 分析器。它能标注出蒙古语词干（词根）词性信息和构形附加成分的相关信息，准确率达 97.7%。单个词上标注的格式为《词根（词干）/词性标记+词缀/词类标记》 。如：单个词根 ABV/Ne1，词根+连写词缀 AHI/Ve2+GVL/Fe11+BA/Fs14；

《_》 ：蒙古语中分开写的元音，如‘OGERECILE/Ve1+N_E/Fs21’中‘_E’是分开写的元音，与‘N’一起才看成是一个音节。

《+》 ：在连写词缀（附加成分）前面标注， ‘YABV/Ve2+N_A/Fs21’中‘N_A’前面有加号，是在说明它与前面‘YABV’的连写词缀。

《-》 ：静词类格范畴，领属范畴，复数范畴的分写词缀前面标注此符号。如：‘ABV/Ne1-YIN/Fc11’中，‘YIN’是前一个词‘ABV’的分写词缀。

《=》：用这个符号连接的词有固定短语，也有专有名词。如：YASV=CINAR/Yn(质量)；DVMDADV=VLVS/NT（中国）

《[]》 :人名前面用这个符号。如：[]WeN=JIYA=BVV/Nt1（温家宝）

《][》 :地名前面用这个符号。如：][TAYIWAN/Nt2（台湾）

Mglex 分析器目前还没有对固定短语词法标注的功能。所以对固定短语词性进行了人工标注，参考了德.青格乐图等人研制的《现代蒙古语固定短语与发信息词典》标注形式是用“=”号连接的词后面有个斜线再写词性。蒙古语固定短语分为复合词（Y）、习用语（X）、成语（K）、固定词（J）、名词术语（NT）等五大类，再把复合词分为名词性复合词（Yn）、形容词性复合词(Ya)、代词性复合词(Yr)、时位词性复合词(Yo)、动词性复合词(Yv)、副词性复合词(Yd)等六种；习用语分为名词性习用语(Xn)、形容词性习用语(Xa)、动词性习用语（Xv）等三种；成语分为名词性成语（Kn）和动词性成语（Kv）。实例如下：“HODEGE=TOSHON/Yn-V/Fc12”（农村的），这里‘Yn’是表示名词性复合词 。

3.2 蒙古语短语结构树库短语标记集

蒙古语短语结构树库的标记集是参考了蒙古语传统语法学中关于词组类型和词组内部关系的分类及命名方法。如，在蒙古语里中心词处在词组的最后部分，即中心词的词性就是词组的词性。[16]词组内部关系分为体述关系、定体关系、宾述关系、状述关系、联合关系和辅助关系等。[17]蒙古语传统语法上大部分著作上认为词组是实词与实词组合的，我们认为词组是短语的一部分，短语可以是实词与实词，虚词与虚

词，实词与虚词之间都可以组合，即短语包含词组。

表 1 蒙古语短语结构类型标记

短语结构类型	标记代码	短语结构类型	标记代码
名词短语	NP	方位词短语	OP
形容词短语	AP	时位词短语	TP
代词短语	RP	动词短语	VP
数词短语	MP	副词短语	DP
量词短语	QP	后置词短语	GP
语气词短语	SP	情态词短语	HP

表 2 蒙古语短语结构内部关系标记

短语结构内部关系	标记代码	短语结构内部关系	标记代码
宾述关系	t	体述关系	u
状述关系	b	辅助关系	s
联合关系	h	定体关系	d
复指关系	j	总括关系	x

3.3 蒙古语短语结构分析标注规范

在同一层面上采用二分的形式。顺序为：从大到小，从左到右，一步一步分析。每部分采用对称的大括弧，在闭弧后紧跟相应的短语标记。

例 如 ： ORCIL/Ne2 AJV=AHVI/Yn-YI/Fc31 YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21 ./Wp1

第 一 步 分 为 《 ORCIL/Ne2 AJV=AHVI/Yn-YI/Fc31 YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21》和《./Wp1》两部分。

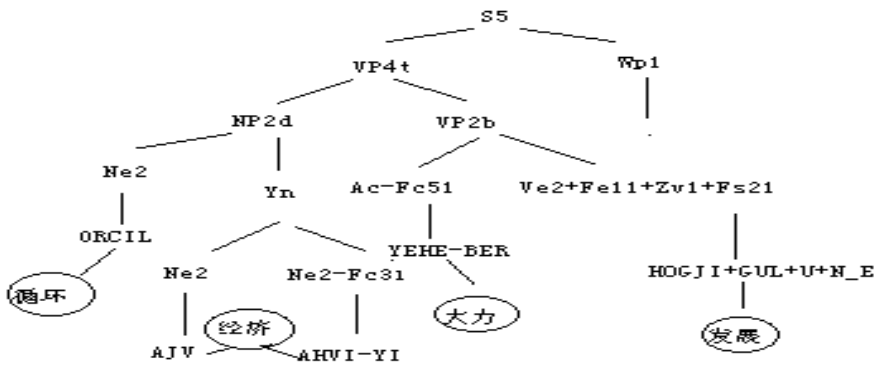
第 二 步 把 《 ORCIL/Ne2 AJV=AHVI/Yn-YI/Fc31 YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21 》 部 分 分 为 《 ORCIL/Ne2 AJV=AHVI/Ne2-YI/Fc31 》 和《YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21》两部分来分析。

第三步把《ORCIL/Ne2 AJV=AHVI/Yn-YI/Fc31》部分分为《ORCIL/Ne2》和《AJV=AHVI/Yn-YI/Fc31》两部分来分析。

第四步把《YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21》部分分为《YEHE/Ac-BER/Fc51》和《HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21》两部分来分析。

分 析 出 来 的 句 子 ：{{{ORCIL/Ne2 AJV=AHVI/Yn-YI/Fc31}Np2d {YEHE/Ac-BER/Fc51 HOGJI/Ve2+GUL/Fe11+U/Zv1+N_E/Fs21}VP2b}VP4t ./Wp1}S5。如图 1

图 1 短语树形结构图



“大力发展循环经济”短语树形结构如图 1。S5 意为词数有 5 个的句子；VP4t 意为动词短语、词数为 4、宾述关系；Wp1 意为标点符号；NP2d 意为名词性短语、词数为 2、定体关系；VP2b 意为动词性短语、词数为 2、状述关系；Ne2 是不可数名词，对应的词是“ORCIL”；Yn 是名词性复合词，由 Ne2（对应的词是 AJV)和 Ne2-Fc31(对应的词是 AHVI-YI,AHVI 是词,YI 是它的附加成分)组成;Ac-Fc51 对应的词是 YEHE-BER，这里 YEHE 是词，BER 是附加成分；Ve2+Fe11+Zv1+Fs21 是指 Ve2 为词根加了三个词缀 Fe11、Zv1、 Fs21 的一个动词。以此标记。

3.4 蒙古语短语树库语料分布特点

在有 20 万词级的标注词类的蒙古语短语树库语料上进行短语结构分析。树库语料有 20201 条句子，句子词数最少的有 2 个词，最多的有 76 个词。语料选取于《100 万词级现代蒙古语语料库》和一些政府文件材料。对训练集 19201 条句子进行 12 种短语结构类型和 8 种短语结构内部关系的统计如下：

表 3 各结构类型出现频次比例

类型	VP	AP	NP	OP	HP	QP	MP	DP	SP	TP	GP	RP
总量	67993	6017	54235	2270	19	494	1269	154	256	552	923	2361

从表格 3 我们可以看出蒙古语短语树库语料各结构类型中，动词短语（VP）出现的频次最高，占全部结构类型的 49.8%，其次是名词短语（NP）和形容词短语各占 39.7%和 4.4%。出现频次最低的是情态词短语（HP），占 0.014%。

表 4 各关系类型在语料库中出现频次

关系标记	t	b	d	u	h	x	s	j
总量	17493	21469	41757	14555	10874	3177	26648	539

从表格 4 我们可以看出蒙古语短语树库语料各关系类型中，定体关系（d）出现的频次最高，占全部关系类型的 30.6%。复指关系（j）出现的频次最低，占全部关系类型 0.4%的比例。

表 5 各关系类型在不同结构类型中出现频次比例

组合	AP	DP	GP	HP	MP	NP	OP	QP	SP	RP	TP	VP
d	1549	21	40	0	670	37704	706	386	11	182	296	178
s	1647	77	25	19	265	9552	192	31	47	1632	101	13044
x	15	20	845	0	106	454	1254	47	124	167	78	70
b	698	8	2	0	10	415	32	6	1	83	8	20202
t	323	4	2	0	8	361	25	0	4	17	1	16745
u	849	6	10	0	48	1542	14	6	1	152	9	11915
h	913	16	1	0	157	3731	48	20	67	64	58	5795
j	10	2	0	0	4	449	0	0	0	63	1	10

从表格 5 上我们能看出蒙古语短语树库分布特点 ,横看全部定体关系（d）里名词短语（NP）的定体关系占 90.32%，并且在名词短语里与其他各内部关系相比，定体关系占 69.5%。因此定体关系主要出现在名词短语里。在全部状述关系（b）和宾述关系(t)里动词短语（VP）中的状述关系和宾述关系各占 94.09%和 95.27%。因此，状述关系和宾述关系主要是在动词短语里出现。体述关系在动词短语里出现的频次最高，占 81.8%，名词短语里占的比例 10.5%,因此体述关系很大一部分是出现在动词短语里。情态词短语在短语类型中占的比例是最少并且内部关系只出现了辅助关系（s）。再看联合关系（h），名词短语、动词短语、形容词短语中出现的频次都比较高，各占 34.3%、53.3%、8.4%比例。复指关系（j）在名词短语中出现的频次最高，占 83.3%，代词短语中占 11.6%，这样我们就能知道复指关系大部分情况下是在名词短语和代词短语中出现。从表格整体上来看，名词短语、动词短语、代词短语、形容词短语等类型出现的频率高的同时，它们的各内部关系出现的频率也高，因此在蒙古语短语树库中实词性的短语占的比例高。

4.蒙古语自动分析器开发

本节介绍蒙古语自动分析器的分析方法。分析器采用“移近-归约”[18]的确定性方法，它是将分析过程

看成是一步步作用于输入句子之上的分析动作的序列。分析的输入为已经分词并带有词性标注的句子，分析过程主要的数据结构为一个栈(S)和一个队列(Q)，输入的<词,词性>对按顺序存储于队列中，栈中存放分析过程中每一步产生的部分句法树，对于每一个分析步骤，其状态由当前栈和队列中的内容表示。本文采用SVM分类器对当前的状态做出动作决策。

其分析动作主要是建立词和词之间的关系。动作模式分为两类。第一为“移进(shift)”动作，代表从队列中取出第一个元素并将其压入栈顶。第二类“规约(reduce)”动作，代表连续出栈两次，将栈顶的两个元素合并为一个新节点，两个元素分别作为新节点的左右孩子，按照规约产生新节点的标记类型，对规约进行分类。由于分析动作只有“移近(Shift)”和“规约(Reduce)”两种类别，可训练出关于分析动作的分类器。在分析过程中，分类器可用来预测分析动作。

特征主要是围绕两个焦点节点选取，焦点节点是指，在当前状态下栈中的第一个和第二个节点，其可能为叶子节点，也可能为分析过程中产生的句法子树。每当采用一个分析动作时，就会得到一个新的状态。在训练阶段，特征及其对应的分析动作组成训练数据；在分析阶段，由分类器在获得的特征的基础上做出分析动作决策。当队列为空，且栈中全部节点规约到一个根节点下时，分析过程结束。

设 S 为栈，Q 为队列，i，j 为节点序号，k 为后缀序号，则所选特征如下表 6 所示：

表 6 蒙语分析器特征模板

特征类别	特征	说明
词根特征	SiW	Si 词根
	SiP	Si 词根类别
	QjW	Qj 词根
	QjP	Qj 词根类别
后缀特征	SikW	Si 的第 k 个后缀
	QjkW	Qj 的第 k 个后缀
	SikP	Si 的第 k 个后缀类别
	QjkP	Qj 的第 k 个后缀类别

本文中 i 取值为{1，2，3}，j 取值为 1，k 取值为{1，2}。

5. 实验结果分析

5.1 测试集

树库语料中训练集为 19201 条句子，测试集为 1000 条句子。图 2 是训练集句子长度折线图，句子长度为 2 个词到 69 个词的分布图。最高点在 e8（8 个词），接着句子越长出现的频次越低。图 3 测试集句子长度折线图，句子长度为 6 个词到 76 个词的分布图。最高点在 e8，接着句子越长出现的频次越低。对比两张图，它们句长特点很相似，所以测试集是适合进行实验的语料。

图 2 训练集句子长度折线图

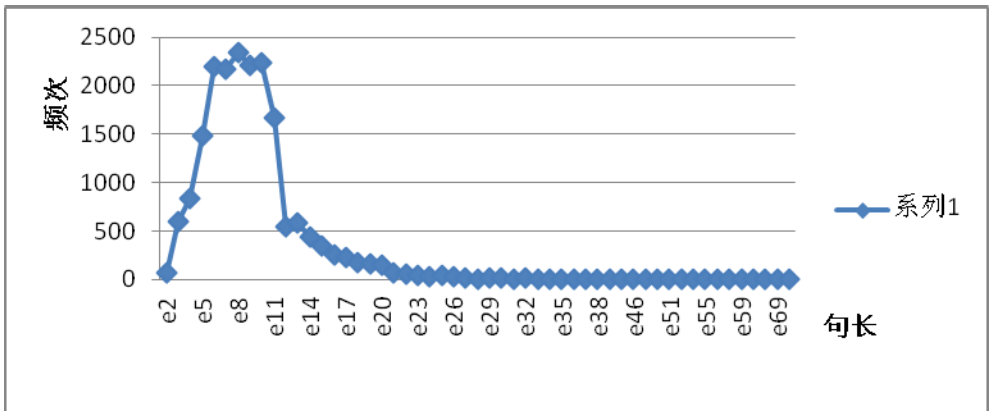
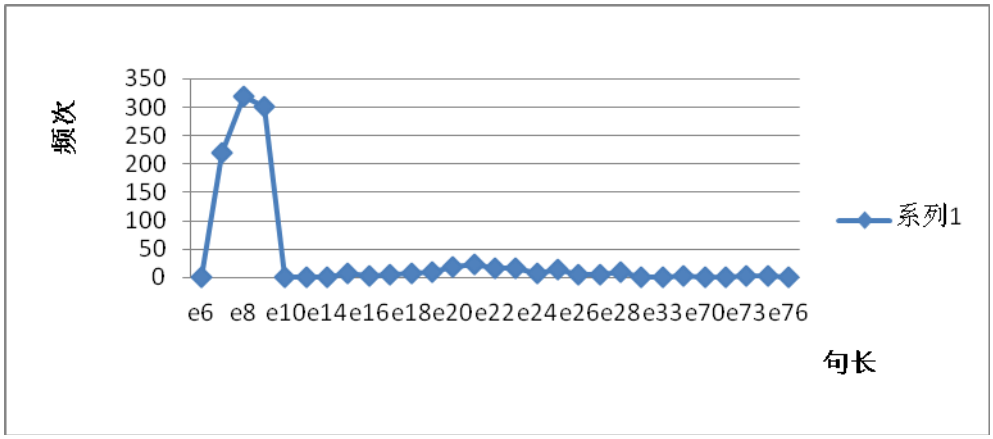


图 3 测试集句子长度折线图



自动分析测试集 1000 条句子，并统计了结构类型出现频次。表格 7 中结构类型出现最多的是动词短语（VP），占 52.55%，其次是名词短语和形容词短语，各占 41.52%和 2.5%。语气词短语（SP）和量词短语（QP）出现频次最少，各占 0.033%。测试集里没有出现情态词短语和副词短语。对表格 3 和表格 7 进行对比，我们能看出它们在短语主要结构类型上形成相似的分布特点。

表 7 自动分析测试集的各结构类型出现频次比例

类型	AP	GP	MP	NP	OP	QP	RP	TP	SP	VP
总量	227	121	99	3727	49	3	5	25	3	4717

5.2 评测指标设计

分析器性能的评价采用常规的评价指标，及准确率（P）、召回率（F),和 F 值(F)，假设分析结果中正确的短语个数为 A，测试集中标准的短语数量为 B，分析结果中的短语数量为 C，则 P=A/C, R=A/B, F=2PR/(P+R)。其中，一个短语分析正确的判断依据为：当且仅当短语的成分标记及边界划分均正确。

目前分析器的效果：

B: 9802

C: 9942

A: 6175

precision: 0.621102

recall: 0.629973

f-measure: 0.625506

5.3 实验结果分析

表 8 自动分析测试集的各关系类型在不同结构类型中出现频次比例

组合	AP	GP	MP	NP	OP	QP	SP	RP	TP	VP
d	67		94	3201	4	3	0	0	0	0
s	81		3	264	0	0	2	5	8	283
x	0	80	0	15	42	0	0	0	17	5
b	8		0	0	0	0	0	0	0	852
t	0		0	0	0	0	0	0	0	888
u	43	41	0	60	2	0	0	0	0	1150
h	28	0	2	183	1	0	1	0	0	539

表 9 原语料测试集的各关系类型在不同结构类型中出现频次比例

组合	AP	DP	GP	MP	NP	OP	QP	RP	TP	VP	SP
d	94	1	0	95	3371	0	12	1	0	10	0
s	67	0	0	5	254	0	0	4	0	377	3

x	0	0	0	0	26	0	0	0	0	7	0
b	6	0	0	0		0	0	0	0	1245	0
t	0	0	0	0	0	0	0	0	0	1470	0
u	20	0	0	0	20	0	0	0	0	243	0
h	30	1	0	3	406	0	1	0	0	430	0
j	0	0	0	0	1	0	0	0	0	0	0
xx	0	0	70	0	0	57	0	0	17	0	0

表格 9 里出现的“xx”不是内部关系标记，而是在分析规范里规定的 GP、TP、OP 三个类型的内部关系不标注情况，对原语料进行统计的时候我们就把这三种类型的内部关系暂时用 “xx”代替统计出来了。

对测试集 1000 条句子进行自动分析，表格 8 是对自动分析测试集的各关系类型在不同结构类型中出现的频次统计。表格 8 和表格 9 进行对比，动词短语和名词短语在短语总数所占比例较高，情态词短语、语气词短语所占比例最低等数据统计情况上我们得出分析器能较好的产生短语树结构。在识别内部关系方面名词短语的定体关系、辅助关系等方面分析器有较好的效果。识别定体关系达 94%，辅助关系达 91.7%。总括关系出现错误最多的是在后置词短语里，在人工标注的时候后置词短语不标注内部关系，但在分析器里凡是结构类型都有标注内部关系，因此表格 8 中后置词短语里出现了 80 次的总括关系和 41 次的体述关系。这种情况对分析器的正确分析内部关系有一定的影响。识别联合关系也是比较差的，特别是在静词性短语里。状述关系和宾述关系涉及到歧义问题，分析器分析错误出现较多。体述关系多以主格形式出现，主格没有具体的形式格符号，因此体述关系的识别也是有较大的困难。

5.4 错误实例分析

在面向人的传统语法中，短语内部关系的辨别也是有一定的难度，尤其在歧义部分。测试分析器的测试集是 1000 条句子，句子词数最少的有 6 个词，最多的有 76 个词。句子平均长度为 10.777。从测试结果上来看，词数越多的句子自动分析出现的错误越多。分析器标注形式是括号相对应方式，分析出来的标注形式如下：

[VP-s [VP-u [NP-d [NP-d Ed-UN/Fc11=JASAG/Yn-VN/Fc11 HORW_A=TOHIRAGVLVL/Yn] EJEMDE/Ve1+L/L-I/Fc31] [VP-u [NP-d VLAM/Dx [NP-d NIGE/Mu ALHVM/Ne1]] CINGGADH_A/Ve1+BA/F4]] ./Wp1]（更进一步加强财政宏观政策。）

[VP-s [VP-u [NP-d DALAI=TANGGIS/Yn-VN/Fc11 EHI=BAYALIG/Yn-I/Fc31] [VP-u [NP-d JUI/Ne2 JOHISTAI/Ax] [VP-h[NP-s[VP-h NEGEGE/Ve1+N/Fn3 ASIGLA/Ve1+HV/Ft12] BA/Cj] HAMAGALA/Ve1+N_A/F1]]]./Wp1]（合理开发和保护海洋资源。）

自动分析“更进一步加强财政宏观政策”，在内部关系标注上出现错误。在整个句子中前半部分[VP-u [NP-d [NP-d Ed-UN/Fc11=JASAG/Yn-VN/Fc11 HORW_A=TOHIRAGVLVL/Yn] EJEMDE/Ve1+L/L-I/Fc31] 和后半部分[VP-u [NP-d VLAM/Dx [NP-d NIGE/Mu ALHVM/Ne1]] CINGGADH_A/Ve1+BA/F4]]是宾述关系（t），而不是体述关系(u).因为“EJEMDE/Ve1+L/L-I/Fc31”有宾格“I/Fc31”，这是在说前半部分和后半部分是直接的宾述关系。 在[VP-u [NP-d VLAM/Dx [NP-d NIGE/Mu ALHVM/Ne1]] CINGGADH_A/Ve1+BA/F4]] 后半部分里，“CINGGADH_A/Ve1+BA/F4”（加强）是中心词，前面的“VLAM/Dx NIGE/Mu ALHVM/Ne1（更进一步）”是修饰加强的程度。它们之间的关系应该是状述关系（b），而自动分析的句子里出现的是体述关系（u）。

自动分析“合理开发和保护海洋资源”，在结构类型和内部关系标注上都出现了错误。前半部分[VP-u [NP-d DALAI=TANGGIS/Yn-VN/Fc11 EHI=BAYALIG/Yn-I/Fc31] 和后半部分[VP-u [NP-d JUI/Ne2 JOHISTAI/Ax] [VP-h[NP-s[VP-h NEGEGE/Ve1+N/Fn3 ASIGLA/Ve1+HV/Ft12] BA/Cj] HAMAGALA/Ve1+N_A/F1]]] 有宾格“I/Fc31”,这是在说前半部分和后半部分是直接的宾述关系。在 [VP-u [NP-d JUI/Ne2 JOHISTAI/Ax]部分里，结构类型分析错误，“JUI/Ne2 JOHISTAI/Ax”应该是形容词性短语

（AP），而不是名词性短语。在[NP-s[VP-h NEGEGE/Ve1+N/Fn3 ASIGLA/Ve1+HV/Ft12] BA/Cj]部分里，“BA”是辅助前面的动词性短语“NEGEGE/Ve1+N/Fn3 ASIGLA/Ve1+HV/Ft12”，所以它的结构类型也是动词性短语，而不是名词性短语。

除了识别短语结构上出现一些错误以外识别内部关系方面的错误比较多。比如，分析动词短语的状述关系和宾述关系的能力各达到 68.43%和 60.4%；体述关系在语料里出现了 283 次，自动分析器分析出来的句子的体述关系有 1296 次.显然分析器对短语结构内部关系的识别方面需要很大的空间去研究和改进。

6.结论

蒙古语短语树库自动分析对蒙古语的句子处理层面上的重点之一。从人工标注和自动分析情况看标记集所包含的短语结构类型和内部关系类型是合理的，该标记集标注的树库包含了丰富的句法信息。在此基础上研制的蒙古语自动句法分析器在一定程度上解决了短语结构人工分析的问题。分析器能准确的产生树库结构，这对蒙古语的句法分析方面也是个进步。在短语内部关系方面的处理还不是较好的效果，因此在接下来的研究中分析错误句子的同时总结出错误点，为内部关系的进一步研究提供更多的理论依据。

参考文献：

[1][2][3][4]王跃龙，姬东鸿.《汉语树库综述》.当代语言学.2009(1).P47-55；
[5]周强.《汉语树库标注体系》.中文信息学报.2004（4）.P2-P7；
[6]达胡白乙拉.《现代蒙古语句法结构树库的建设》.内蒙古大学学报.2011（6）.P18-30；
[7]华沙宝.《蒙古语短语标注策略》.中央民族大学学报.2003（5）哲学社会科学版.P98-100；
[8]达胡白乙拉.《面向信息处理的蒙古语名词短语结构研究》.内蒙古大学硕士学位论文.2002；
[9] 吉仁花.《面向信息处理的蒙古语形容词短语结构规则研究》. 内蒙古大学硕士学位论文.2004；
[10]德.青格乐图.《现代蒙古语固定短语语法信息词典详解》 .呼和浩特：内蒙古教育出版社，2005；
[11]斯.老格劳.《现代蒙古语依存句法自动分析研究》.内蒙古大学博士学位论文.2011；
[12]德力格尔玛，高莲花，其木格.《蒙古语与汉语句法结构对比研究》 .北京：民族出版社，2013 年；
[13]中国电子标准化研究所、内蒙古大学等.《信息技术-信息处理用蒙古文词语标记集》 （GB/T 26235-2010）；
[14]包满亮.《蒙古语构形词缀研究》.中央民族大学博士学位论文.2007；
[15]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；
[16][17]清格尔泰.《现代蒙古语语法》(修订版）.呼和浩特：内蒙古人民出版社，1999 年；
[18]马骥，朱慕华，肖桐，朱靖波.《面向移进归约句法分析器的单模型系统整合算法》.中文信息学报.2012(3)。