蒙古语句子成分前置结构的计算机处理研究

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： 蒙古语句子成分前置结构的计算机处理研究是蒙古语自动句法分析的重要组成部分。人工标注句法结构的语料文件 HANABE（共 735 个句子）中有 167 个句子结构，由于句子成分前移而构成表面上与分析句子结构的最终判断规则 S=>NP VP 有出入的格局。笔者认为，深层次上，这些前置结构与短语结构语法基本公式 S=>NP VP 有等价性。因此，本文采用归纳新规则的方法，扩充了最终检验能否成功分析整个句子的判断规则 S=>NP VP。

关键词： 蒙古语；句法结构；句子成分前置；计算机处理

中图分类号码： H085.5　　**文献标示码：** A

蒙古语句子成分前置现象的计算机处理研究属于语言学基础应用研究，是自动句法分析（syntactic parsing）研究的不可分割的一部分。系统研究蒙古语句子成分前置现象，并采取适当的计算机处理策略，对自动句法分析、语料库多级加工以及诸如机器翻译、自动校对等自然语言处理系统的研制起到重要作用。

一、　问题的提出

众所周知，短语结构语法的基本公式 S=>NP VP 的含义是一个句子由名词性短语和动词性短语两个部分组成。因此，我们在语料库加工的重要环节—短语标注研究中，以 S=>NP VP 为判断句法分析成功与否的最终原则。在能够自动界定短语的前提下，如果计算机把一个句子最终成功分成 NP VP 两个部分，我们认为该句子的短语结构分析成功，否则失败。然而，人工标注蒙古语句子的短语结构时，我们发现有些句法成分移位造成表面上与 S=>NP VP 有出入的格局。而且，这种句法结构不论种类还是数目都不少。因此，这种句法结构的计算机处理策略自然而然地被列入现代蒙古语语料库加工的重要研究议题当中。

二、　研究基础和方法

（一）浅加工的语料库

蒙古语句子成分前置现象的计算机处理研究是在浅加工的现代蒙古语语料库的基础上进行的探索性研究。蒙古语语料库的词性标注与统计是华沙宝教授在国家自然科学基金项目的支持下主持完成的重要研究成果。其采用的词语分类体系是 16 分类体系，标记说明如下：JIN（名词）、UIL（动词）、 TEN（形容词）、TON（代词）、TON（数词）、HMN（量词）、CON（时位词）、DAY（副词）、DAU（后置词）、HOL（连接词）、SVL（语气词）、HND（情态词）、DUR（摹拟词）、AYA（感叹词）、TOU（代动词）、HOU（连接动词）。内蒙古大学蒙古语文研究所较早制定的该分类体系，随着蒙古语文信息处理研究的深入开展，受到来自蒙古语文信息处理各领域的挑战。这一点促使早日制定反映蒙古语文本身特点的词语详细分类体系及其标记规范。虽然内蒙古大学蒙古学学院语言研究所已经初步提出了更完善的词语分类体系及其标记集，但目前尚未应用于现代蒙古语文词性标注语料库中。因此本文暂且采用 16 分类体系及其标记集。另外，该语料库中浅加工的内容还有切分蒙古语词语变形形式（词语变形形式前加中线—、斜线/表示），标注人名、地名（人名前加方括号 [] 表示；地名前加][表示）和一些复合形式（构成复合形式的各成分之间用=、&等来连接）等。

（二）试用短语标记集及其说明

蒙古语中，前置句子成分既可以是单词，也可以是短语。因此，在蒙古语句子成分前置结构的形式句法规则中采用短语标记是理所当然。然而，至今蒙古语中还没有通用的、确实反映蒙古语本身特点的短语分类体系及其标记规范。因此，蒙古语语料库中人工标注短语时，根据标注实践和参照蒙古语语法信息词典中拟采用的短语分类体系及其标记集， 我们提出了一个试用短语标记集。其简要说明如下： NP（名词短语）、VP（动词短语）、AP（形容词短语）、RP（代词短语）、MP（数词短语）、MQP（数量词短语）、OP（方位词短语）、GP（后置词短语）、DP（副词短语）、HD（引语）。

（三）形式规则所用其它标记及其含义

形式规则所用其它标记及其含义说明如下：（1）用 | 表示逻辑“或”关系；=>表示可以改写的意思;(2)词类标记后的 [] 里填写了词语变形形式等语法信息或非句号标点符号;短语标记后的 [] 里填写了构成本短语的最后一个词的变形形式等语法信息或非句号标点符号；短语标记后的花括弧{}里标明了该短语的形式结构；（3）在形式规则中没有标记句号；（4）S’=>NP VP， S’ =>NP AP 是贯通下面各形式规则中的全局性规则；（5）实例中，在花括弧{}后标记了句法范畴；这里的花括弧{}是用来标明短语之间的分隔，与形式规则中的花括弧不一样。

（四）研究方法

蒙古语句子成分前置结构的计算机处理研究，以人工标注为基础，以自动标注为手段，从而达到成功分析句子前置结构为目标。在人工标注的过程中所制定的一切标注规范及所采取的方法，都为自动标注提供有力的语言学依据，从而保证自动句法分析的正确性。

研究蒙古语句子成分前置结构的具体步骤如下：训练集的创建->人工标注句子结构->统计前置句子结构->形式化规则集的创建->自动标注前置句子结构->实验结果分析->自动标注前置句子结构。其中，从人工标注向自动标注转变是整个标注工作的关键环节。因为在这个转变过程中，归纳出有关前置成分的结构规则和相关句法特征，并为自动标注提供根本依据。自动标注的准确率在很大程度上会依赖这些句法语义形式规则。下文论述的形式句法规则的归纳研究就是上述研究方法的局部体现。

三、 计算机处理用形式规则集

（一）训练集的创建与句子结构的人工标注

现代蒙古语句子成分前置现象的计算机处理研究属于语言信息处理基础研究，它的使用目标不是针对某个专门应用系统。在不同的语域中，语法特征的分布特点不会一样。因此，现代蒙古语句子成分前置现象的计算机处理研究，对语料的选取方面应该严格要求它的一般性。鉴于以上考虑，我们选用现代蒙古语词性标注语料库文件 HANABE(HABVR-VN NARAN BEGEJING-ECE)，为人工归纳句子成分前置结构的形式规则的基础语料。

在此基础上，我们人工分析了 HANABE 的所有句子，并对句子中的短语结构和功能进行了标注，从而建成了面向自动句法分析研究用训练集。在此基础上，我们挑选出含有前置成分的句子结构库，以供蒙古语句子前置结构的计算机处理研究。人工标注示例：

```
{
   {
      {
         ENE TON:02H83* 8
         HU SVL:02H83* 8
      }RP
      {
```

```
        SIGVRGA/N-V JIN:02H83* 8
        EDUR-UD-TU JIN:02H83* 8
      }NP
   }NP
   , :02H83* 8
   {
      {
        []GANGGAM_A-YIN JIN:02H83* 8
        DVGVYILANG-VN JIN:02H83* 8
      }NP
      HOMOS JIN:02H83* 8
   }NP
   , :02H83* 8
   {
      {
        BOHO TON:3:02H83* 8
        HUCU-BER-IYEN JIN:02H83* 8
      }NP
      TEMECE/L_E UIL:02H83* 9
   }VP
. :02H83* 9
}S
```

（二）状语前置句子结构的统计

我们人工标注了现代蒙古语词性标注语料库文件 HANABE（共 735 句），其中有 167 个句子的结构是由于句子成分前置而造成表面上与 S=>NP VP 有出入的格局。目前，在我们标注的语料库中，状语提前的有 43 个句子（一个状语前置的句子有 41 例，两个状语子前置的句子有 2 例），宾语提前的有 7 个句子，谓语提前的有 114 个句子。另外，也有前置两个不同句子成分的语言现象。至今，在我们人工标注的语料中有 3 个句子的结构属于这种类型。就其原因，主要是由于语用需要，比如为了强调，使本属于谓语部分 VP 的某些状语成分提前。下文对句子成分前置现象按前置成分的数目和内部结构的特点，在不同层次上进行了分类，并在此基础上给出了相应的形式规则，以供计算机处理。

（三）句子成分前置结构的形式规则集

根据前置成分的句法功能，首先分为谓语前置、宾语前置、状语前置、状语和宾语同时前置、

状语和谓语同时前置等五种类型。然后，按前置成分的数目、内部结构特点，对以上五种类型进行详尽分类，并给出相应的形式子规则。比如，根据前置状语的数目，首先分为一个状语前置的句子结构和两个状语前置的句子结构。然后按内部结构特点，对这两个大类进行详细划分，并给出相应的形式子规则。由于篇幅的限制没有对每一条子规则都给出实例，而仅给出了一个带有粗略结构标注（没有标明具体词语的词类和短语内部结构的详细关系）的典型例子。

1.谓语前置的句子结构及其形式规则

1.1 VP{VP{HD HOU[/JU]} UIL[/GED]} S'{NP VP}=>S

例子：{{{《CASV OCOHEN CU HODEL/U/GE UGEI-BER %UU? UGEI! BISI》GE/JU}VP HELE/GED}VP {{BAG-VN DARVG_A}NP {TOLOGAI-BAN SEGSUR/BE}VP}S'.}S

1.2 VP{HD HOU[/JU|…]} S'{NP VP}=>S

1.3 VP{NP[…] VP[…]} S'{NP VP}=>S

2.宾语前置的句子结构及其形式规则

2.1 RP[…] S'{NP VP}=>S

2.2 NP[-I|…] S'{NP VP}=>S

例子：{{NIS/HU ONGGOCA/N-V DALDA ORO/GSAN TERE JUG-I}NP {{HOMOS}NP {NAM SIRTE/JU JOGSO/N_A}VP}S'.}S

2.3 JIN[…] S'{NP VP}=>S

2.4 VP[…] S'{NP VP}=>S

3.状语前置的句子结构及其形式规则

3.1 一个状语前置的句子结构及其形式规则

3.1.1 JIN S'{NP AP}=>S

3.1.2 GP S'{NP VP}=>S

3.1.3 NP[…|-TV,|…] S'{NP VP}=>S

例子：{{TAL_A NVTVG-TV}NP, {{IRUGEL DAGVLAL-VN HABVR}NP {IREL_E}VP}S'.}S

3.1.4 VP[…] S'{NP VP}=>S

3.1.5 OP S'{NP VP}=>S

3.1.6 JIN S'{NP VP}=>S

3.1.7 TON S'{NP VP}=>S

3.2 两个状语前置的句子结构及其形式规则

3.2.1 OP[,] VP S'{NP VP}=>S

例子：{{[]GANGGAM_A-YIN DVGVYILANG-VN GER-UD-UN HOGORONDO}OP, {ENDE=TENDE UGEI}VP {{MADAI-TAI BORDOG_A}NP {$OBOYILCA/JV BAYI/L_A}VP}S'.}S

3.2.2 CON[…] SVL S'{NP VP}=>S

4.状语、宾语同时前置的句子结构及其形式规则

4.1 OP{NP CON} NP{NP SVL} S'{NP VP}=>S

4.2 OP{UIL[/GSEN-ECE] CON[,]} NP{VP[/GSAN] NP[/N-ECE]} S'{NP VP}=>S

例子：{{CILOGELEGDE/GSEN-ECE HOYISI}OP, {ARAD-VN JSAG-VN ORDO/N-ACA MALCID-TV TVSALA/JV TALBI/GSAN JIGELELGE-YIN MONGGO/N-ECE}NP {{[]ARSLAN}NP {DORBE/N JAGV=DALAN=NAYIMAN=TUME/N TOGORIG JIGELE/N AB/CV, TABIN=HOYAR HONI HVDALDV/N AB/CAI}VP}S'.}S

5.谓语、状语同时前置的句子结构及其形式规则

VP{HD HOU[/JU]} NP[-TAI] S'{NP VP}=>S

例子：{{《ABV-YIN HAGVCIN NI HODEL/U/GED IRE/JEI》GE/JU}VP {NELIYED SANDVRVGSAN MAyIG-TAI}NP {{SIRUNCECEG}NP {HARIGVLBA}VP}S'.}S

四、等价性

按照管辖约束论中的语迹理论，句子成分移位时在原来的句法位置上留下语迹。通过一定的转换规则，我们可以把这些前置成分还原到相应的语迹位置上，当然这样的还原不会改变原句的基本意思。因此，以上那些形式规则在理论上与 S=>NP VP 等价。

从计算语言学方法的角度看，这只关系到分析层次的增加与减少，不会引起语义上的变化。因为我们可以把诸如 S=>NP S'规则改写成 S=>NP NP VP 形式。而前一个 NP 是该句子谓语的一个组成部分，易位后与 VP 可以构成担任谓语成分的新 VP。这样也能还原成 NP VP=>S 的形式。很明显，前一个是二叉树，后一个是三叉树。计算语言学理论与方法的有关论著已经证明二叉树和多叉树的等价性，并在此基础上实现了多叉多标记树模型（Multiple-branched and Multiple-labeled Tree Model，简称“MMT 模型”）。根据 MMT 模型，1981 年成功进行了汉-法/英/日/俄/德多语言 MT 试验。这是在自然语言处理领域（NLP domain）充分证明了这种等价的准确性。

由于等价性，以上归纳的各形式句法规则与 S=>NP VP 同样可以检验蒙古语句子结构的合格性。从短语标注策略角度看，这意味着最终检验能否成功分析整个句子结构的判断原则不仅仅是一条了。因此，我们不必使句法结构符合普遍语法规则 S=>NP VP 而苦苦地建立一些转换规则。这将一方面大大增加自动句法分析的灵活度，另一方面会减少计算复杂度。

当然，随着训练集的扩充，蒙古语句子成分前置结构，不论数目还是种类都会增加。因此，为了提炼出更可靠的形式句法规则，我们打算进一步扩充我们的训练集到 20 万词，同时会增加合理的语料种类。从以上归纳的形式规则，不难看出彻底研究句子成分前置结构的计算机处理是以蒙古语各短语类型的计算机处理研究为基础。目前，在这方面致力的研究主要有内蒙古大学蒙古语文研究所华沙宝教授承担的国家自然科学和社会科学基金项目“蒙古语语料库的短语标注研究”、青格乐图教授主持完成的国家社会科学基金项目“蒙古语固定短语研究”、巴达玛敖德斯尔副教授的博士学位论《面向机器翻译的汉蒙短语转换规则研究》和笔者的硕士学位论文《面向信息处理的现代蒙古语名词短语结构规则研究》等。然而这些研究表明蒙古语自动句法分析研究刚刚开始，需要解决的问题还很多，需要我们不断地探索研究。

参考文献

[1]James Allen.Natural Language Understanding[M]. Redwood city in America:The Benjamin/Cummings Publishing Company, 1995.

[2]Mark Baltin, Chris Collins.The Handbook of Contemporary Syntactic Theory[M].Beijing:Foreign Language Teaching and Research Press, Blackwell Publishing Ltd, 2001.

[3]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.

[4]俞士汶．语法知识在语言信息处理研究中的作用[J]．语言文字应用，1997，(4)：81—87.

[5]徐烈炯．生成语法理论[M]．上海：上海外语教育出版社，1988.

[6]汉语句子的句法树标注规范(V2.0)[M]．智能技术与系统国家重点实验室编写， 清华大学计算机科学与技术系。

[7]确精扎布．有关蒙古语词组的几个问题[A]．探索与硕果[C]．呼和浩特：内蒙古大学出版社，2002.156—168.

[8]清格尔泰．关于句法结构分析[A]．探索与硕果[C]．呼和浩特：内蒙古大学出版社，2002.169—178.

[9]华沙宝．蒙古语短语标注策略[J]．中央民族大学学报(哲学社会科学版)，2003，(5):90—100.

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

[11]巴达玛敖德斯尔．面向机器翻译的汉蒙短语转换规则研究[M]．内蒙古大学博士学位论文，2003.

[12]达胡白乙拉．面向信息处理的现代蒙古语名词短语结构规则研究[M]．内蒙古大学硕士学位论文，2002.

A Study of Computer Processing of Syntactic Structure with Advanced Ingredients

Dabhurbayar

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

Abstract: Computer processing of syntactic structure with advanced ingredients is the important part of automatic syntactic parsing of Mongolian language. In the lingual material made up of 735 sentences which are labeled with phrase structure tag by human, there are 167 sentences which include advanced syntactic ingredients. The author thinks that the 167 sentences are the same with the other sentences in deep structure. Therefore, in this article some new formal syntactic rules which describe those 167 sentences' structures are concluded for successful automatic syntactic parsing.

Key word: Mongolian language；syntactic structure；advanced ingredient;computer processing

收稿日期： 2004-04-16;

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博士研究生。2003年赴北京广播学院播音主持艺术学院应用语言学系研修半年，主要研究方向为计算语言学与蒙古语文信息处理。