俄语基本名词性构句块模式研究

叶其松

（黑龙江大学，黑龙江 哈尔滨 150080）

摘 要: 构句块的自动化分析起着承接俄语自动形态分析和句法—语义分析的作用，是俄语自然语言处理的重要模块。本文从基本名词性构句块的内部结构和功能属性入手，结合俄语作为屈折语的特点，探索短语一级语言单位模式化及模式多层级、形式化描写的途径，旨在为最终实现名词性构句块的自动化分析和识别奠定基础，并为俄语自动句法—语义分析提供语言学保障。

关键词: 自然语言处理；基本名词性构句块；模式化

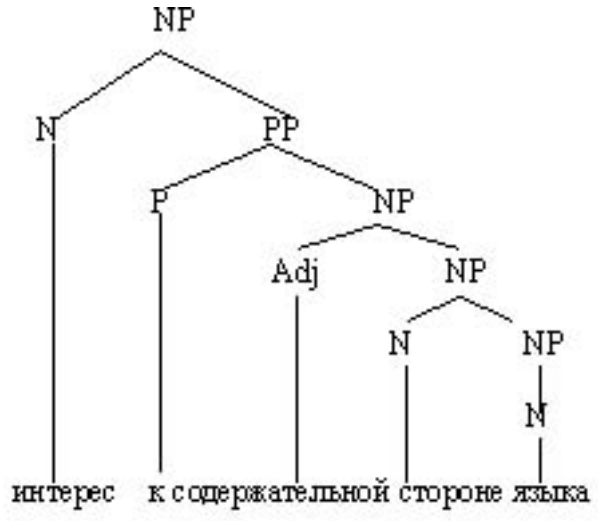
中图分类号: H085 **文献标识码:** A

1 引言

众所周知，语句是最小的交际单位，对其结构的理解，语言学家的观点大相径庭。汉语语法从句子成分出发，区分出主语、谓语、宾语、状语等成分。其中，主语、谓语和宾语是主要的句子成分，一个句子的结构可以用主语+谓语+宾语表示。乔姆斯基的转换生成语法认为，名词短语（noun phrase，NP）和动词短语（verb phrase，VP）是构成句子的两大基本部件，任何一个句子（sentence，S）都可以表示为 S→NP+VP。格语法的创立者、美国语言学家 Fillmore 把句子分成情态（modality）和命题（proposition）两部分，可以用公式 S→M+P 表示（杨成凯 1986：37）。Г. А. Золотова 则认为句法素（синтаксема）是句子的直接构筑单位。她将“句法素”视为俄语中“最小的、不可分割的语义－句法单位”。范畴语义特征、形态特征和句法功能被认为是区分句法素的重要特征。在俄语句子中，句法素体现以下 3 种基本句法功能：1）作为独立的单位使用；2）作为句子的组成部分使用；3）作为词组（或词的组合）的组成部分使用。根据句法素在句子中所起的句法功能数量的多寡将其分为自由型句法素（свободная синтаксема）、限制型句法素（обусловленная синтаксема）和连接型句法素（связанная синтаксема）。在限制型句法素和连接型句法素中又可以区分出各种具体的“位”（позиция）。不仅是句子，像超句子统一体，乃至语篇等更大的语言单位都由句法素组合而成。（Г. А. Золотова 1988：4–5）俄语事格语法是从自然语言处理角度提出的可操作性强、高度形式化的俄语定性化描写体系。事格语法认为，客观世界由事件组成，事件映射到语言中，体现为一个个句子，句子的抽象模式可以表示为 V(x，y，z) < a >，动词 V 是事件中的“代表”， x，y，z 和 a 则是事件的参与者。在交际过程中，根据各项参数取值的不同，可以生成变化无穷的句子。（傅兴尚 1999：45）。

我们认为，句子是由有限的构句块（синтаксический блок）组成的。构句块是指句子中某一片段（отрезок），常以该段的第一个词和最后一个词为分割边界。每个构句块都包含主导词，由主导词继承整个构句块的语法属性。句子中的基本构句块包括动词性构句块、名词性构句块、副动词构句块和形动词构句块（傅兴尚 2004：41）。本文拟研究俄语名词性构

句块（以下简称 NP 构句块）。先看下例：В последние 50 лет в лингвистике возрос интерес к содержательной стороне языка。该句包含 3 个独立的 NP 构句块：последние 50 лет，лингвистике，интерес к содержательной стороне языка，主导词分别是 лет，лингвистике，интерес。但是 3 个构句块内部结构是不相同的：在构句块 последние 50 лет 中，последние 和 50 是 лет 的修饰语；лингвистике 中主导词本身就构成一个构句块，интерес к содержательной стороне языка 的结构比较复杂：интерес 是整个构句块的主导词，其中“嵌套”另外一个 NP 构句块 содержательная сторона языка。用树形图表示构句块 интерес к содержательной стороне языка 的结构如下：



2 基本 NP 构句块的模式化及其次范畴化

上图表明，NP 构句块具有层级性。句子中最“底层”的、合乎句法规则的、以名词为主导词的构句单位称之为基本 NP 构句块。虽然句子中的 NP 构句块千变万化，但是基本 NP 构句块的模式是有限的。借助相应的规则对基本 NP 构句块加以组配，可以生成各种类型的 NP 构句块。

根据俄语名词的组配性能，可区别以下 8 类基本 NP 构句块模式：1）Adj N_1—形容词+名词一格（如 утренний час，стальная воля，отличный студент）；2）Adv N_1—副词+名词一格（如 прогулка ночью，поворот налево，совсем дурак）；3）Pron N_1—代词+名词一格（如 весь народ，мой словарь）；4）Num N_f—数词+名词（如 два стола）；5）N_1 V—名词一格+动词不定式（如 возможность учиться，приказ наступать，мастер плавать）；6）N_1 P N_f—名词一格+前置词+名词（лобовь к народу，робость перед народами，наблюдение за морем）；7）N_1 N_f—名词一格+名词（如 ожидание автобуса，владение языком）；8）ОДНОР СУЩ—同质名词短语（如 кофе или чай，сын и дочь）。

我们知道，计算机通常按照条件执行相关指令，算法设计要求对语言单位进行定性化描写。上述 8 类基本 NP 构句块模式是高度抽象化的、概括性很强的构句单位。为了便于操作，必须实现模式的次范畴化。次范畴化指的是根据某些鉴别特征（形态特征、语法意义等）将语言单位细化（实现语法或语义层面具体化）的过程。在基于规则的自然语言处理系统中，把握次范畴化的适宜度至关重要。标准过于宽泛，容易导致语言单位承载的各项信息不足，给设计算法带来困难或引起歧义；划分过细便于对语言规则的操作，但这会占用很大的内存，影响运算速度。因此，次范畴化应该以面向操作任务为原则，以实现语言单位的可计算性为目标。具体涉及到 NP 构句块模式的次范畴化，可选取以下鉴别特征：1）N_f 或 P 的形态特

征；2）修饰语的词汇－语法类别；3）修饰语与主导词间的语义关系；4）联系用语的不同。

基本 NP 构句块模式的次范畴化，需要强调以下两点：1）某一类基本 NP 构句块模式的次范畴化，往往只选取部分鉴别特征。构句块模式 Adj N_1 的次范畴化，我们选取的鉴别特征包括修饰语的词汇－语法类别和修饰语与主导词间的语义关系两项，最终区分出 Adj N_1（属性，красный стол）、 Adj N_1（材料，каменный домик）、Adj N_1（事物所有者，отцовская шляпа）等 17 类带有语义标注的 NP 构句块模式。其中，表示数量意义的构句块模式中的修饰语一般为 сотый，тысячный，некоторый（如 сотные строители）等具有数量意义的形容词。联系用语的不同只用于模式 ОДНОР СУЩ 的次范畴化。2）次范畴化具有层级性。各层级次范畴化所选取的鉴别特征也不尽相同。如模式 N_1 P N_f 的次范畴化由 3 个层级组成：第一层级的次范畴化围绕 N_f 的形态特征展开，可以得到 N_1 P N_2，N_1 P N_3 等模式。选取 P 的形态特征进行第二层级的次范畴化后，可以得到 N_1 от N_2， N_1 к N_3， N_1 перед N_5 等比较具体的模式。第三层级次范畴化选取的鉴别特征是构句块内部的语义关系，最终我们得到诸如 N_1 от N_2（客体，освобождение от пустяков），N_1 от N_2（空间，тропика от леса），N_1 от N_2（时间，письмо от двацатого мая）等带有语义标注及相关语言信息的基本 NP 构句块模式。

次范畴化后基本 NP 构句块模式表现为一个层级系统。位于顶部的是诸如 Adj N_1 等基本构句块模式，数量少，概括性强；基本模式下面是第一级次范畴化后形成的模式，随后是第二级次范畴后形成的模式，越底层的模式数量越多，越具体。

3 基本 NP 构句块模式的信息标注

信息标注重在为次范畴化后的基本 NP 构句块模式提供必要信息。在基于规则的处理系统中，信息标注具有重要意义。基本 NP 构句块模式的信息标注，在理论上为 NP 的研究提供了新的内容，拓宽了研究者的视野；在操作中可为建造语言知识库和实现 NP 构句块的自动处理提供信息源。确定一个合理、统一、开放的信息标注集是实现信息标注的必要前提。从实现 NP 构句块自动处理的角度出发，我们确定一个包括构句块的形式化表达、语义关系、主导词、例词、汉化语序等 5 项内容组成的信息标注集，无法归入标注集，但对于 NP 构句块的自动处理具有重要价值的信息，可体现在备注中。

构句块的形式化表达是用形式化语言表征基本 NP 构句块模式化的结果。在数学中，用公式 $a^2+b^2=c^2$ 表示直角三角形两个直角边与第三条边之间的长度关系。在自然语言处理中，借用元语言表达手段表示语言单位的内部结构，便于对其理解和计算。

语义问题是目前机器翻译中的重点和难点。实现语言单位的“句法－语义一体化描写”是自然语言处理中不可避免的趋势，句法分析侧重对语言单位结构的理解，语义分析着重阐释语言单位的意义。在基本 NP 构句块模式化过程中，本文力求明确模式内部的语义关系。

构句块的语法属性集中体现在主导词上。Adj N_1，Adv N_1，Pron N_1 等模式中只有一个 N_1，N_1 即为构句块的主导词。N_1 P N_f 和 N_1 N_f 模式中的主导词一般为 N_1，ОДНОР СУЩ 模式中存在两个（或两个以上）的主导词。确定模式的主导词，便于实现构句块的规约（即 NP→N）并减少 NP 构句块中的“节”点。

基于规则建模的主要依据是研究者的语言知识。基于规则建造的模式分析效率较高，但是主观性较大，往往需要经过真实文本的验证。本文为每个基本 NP 构句块模式配备相应例词，供读者检验。汉化语序是模式相应的汉语对等翻译形式。本文根据不同上下文为 NP 构句块提供不同汉化语序的方案，增强了译文的合理性和准确性。

备注中的内容主要体现为：1）基本 NP 构句块模式对其组成要素的要求。构句块并不是主导词与修饰语的任意组合，对各组成要素的形态特征、语义类别进行规定是 NP 构句块

自动处理的重要辅助模块；2）NP 构句块模式的变体形式。在自然语言处理中，不同的变体形式往往包含在一个模式中，一方面可以节省模式占用的内存空间，另一方面有利于模式的查询，提高运算速度。实现对 8 类基本 NP 构句块模式的次范畴化和信息标注后，所有的模式按顺序排列。现在我们从中选出部分模式，来展示 NP 构句块模式化的过程。

N_1 P N_f：以 N_f 的形态特征进行第一层级的次范畴化。
N_1 P N_2：选取 P 的形态特征进行第二层级的次范畴化。
N_1 от N_2：根据模式内部的语义关系可进行第三层级的次范畴化。

至此，模式的次范畴化已经完成，可以对第三层级次范畴化后得到的基本 NP 构句块模式进行信息标注。

模式表达式： N_1 от N_2；模式内部的语义关系：行为－客体意义；模式主导词：N_1；例词：отличие от нас 等；模式汉化语序：对 N_2 的 N_1/ N_1 N_2；备注：该模式中的 N_1 可以为：1）защита，охрана，освобождение，гарантия 等表示摆脱威胁、困境等意义的动名词。例如，защита от непогод，освобождение от пустяков 等；2）отказ，отличие，отречение，отвлечение，отнятие 等部分带有前缀 от-的动名词，如 отречение от трона，отказ от просьбы 等。

模式表达式：N_1 от N_2；模式内部的语义关系：事物－空间意义；模式主导词：N_1；例词：тропинка от леса 等；模式汉化语序：N_2 旁的 N_1。

模式表达式：N_1 от N_2；模式内部的语义关系：事物－时间意义；模式主导词：N_1；例词：телеграмма от пятницы，письмо от двацатого мая 等；模式汉化语序：N_2 的 N_1

…

N_1против N_2

…

N_1 P N_3

…

4 基本 NP 构句块模式的操作原理

自然语言处理是一个包含词法分析、句法分析、语义分析等若干处理模块的复杂过程。各处理模块是相对独立的，即完成相应的操作任务，又紧密相连，集中体现为前一个处理模块的结果对后一个模块的分析产生直接影响。构句块的分析作为句法分析（或句法－语义分析）的预处理模块，该处理模块应建立在成熟的词法分析技术基础上，词法分析的出口即为构句块分析的入口。以 Я читаю книгу 为例，经过词法分析，对句中的词形进行形态还原并赋予相应的词法信息，结果如下：

я（代词，单数，第一格，я）
читаю（动词，单数，第一人称，主动态，未完成体，行为，читать）
книгу（名词，单数，第四格，语言作品，книга）

进入构句块分析阶段，首先考虑各种类别基本 NP 构句块的组配顺序，同时完成对构句块的规约（NP→N），直至实现整个 NP 构句块的分析。这里列举几条组配规则，以展示其分析步骤。

规则 1：8 类基本 NP 构句块模式的组配顺序为：A（Adv N_1，Adj N_1，Pron N_1，Num N_f）→B（N_1 P N_f，N_1 N_f，N_1 V）→C（ОДНОР СУЩ）。用此规则分析 красная шапка сестры и джинсы братаNP 构句块，规约结果可表示为｛〔（красная шапка） сестры〕и （джинсы

брата)}。

规则 2：当修饰语本身带有接格关系时，优先进行规约。对于 NP 构句块 знакомый мне человек 的规约顺序为〔(знакомый мне) человек〕。

规则 3：对于上面 A 组内的模式，经常出现一个主导词前存在多个修饰语的情况，这时优先规约距离主导词近的修饰语。例如 NP 构句块 пять красных шапок 的组配顺序为〔пять (красных шапок)〕。当修饰语前出现若干同质修饰语时，优先规约同质修饰语。对 красивая французская машинаNP 构句块进行规约，结果如下：〔(красивая французская) машина〕。

规则 4：当一个构句块中“嵌套”另一个构句块时，优先规约被嵌套的结构。构句块 интерес к содержательной стороне языка 的规约顺序如下：{интерес к 〔(содержательной стороне) языка〕}。

对俄语句子中的 NP 构句块进行自动化处理时，除引入组配规则外，还需解决 NP 构句块边界测定、后置定语的分析和识别、歧义消除等一系列技术问题，限于篇幅，不再赘述。

5 结束语

句法－语义分析的一个热点是注重句子的局部分析，内容涉及基本名词短语（Base NP）的确定、短语边界的划定、语块（或组块）分析等（赵铁军等 2000：157－175）。虽然上述研究的出发点和方法各不相同，但宗旨大体一致，即为后来的句法－语义分析作准备，提高句法分析的质量。就某种意义而言，构句块分析也属于一种局部分析的方法。由此可见，构句块分析具有广阔的研究价值和应用前景。

本文以 NP 构句块的内部结构为基础，以构句块的模式化和形式化为表达手段，力求对俄语基本 NP 构句块进行定性化描写，为实现 NP 构句块的自动分析和识别提供信息源，为短语分析以及其他构句块的分析积累经验。

参考文献

[1]Золотова Г. А. 1988 Синтаксический словарь[Z], M.

[2]傅兴尚 1999 现代俄语事格语法[M]，北京：军事谊文出版社。

[3]傅兴尚 2001 基于事格文法的俄语词汇知识库[M]，哈尔滨：黑龙江人民出版社。

[4]傅兴尚 2004 俄语句法结构的模式化描述及操作原理[A]//语言计算与基于内容的文本处理（2004 年第七届计算语言学联合学术会议）[C]，北京：清华大学出版社。

[5]傅兴尚 2004 俄语句法信息的自动化处理（基本构句块及其识别算法）[J]，解放军外国语学院学报，第 1 期。

[6]杨成凯 1986 Fillmore 的格语法理论（上）[J]，国外语言学，第 1 期。

[7]赵铁军等 2000 机器翻译原理[M]，哈尔滨：哈尔滨工业大学出版社。

Research on Modeling BaseNP of Russian

YE Qi-song

(Heilongjiang University, Harbin 150080, China)

Abstract: Automated analysis of BaseNP is regarded as an important module, connecting the automated

morphological analysis with the syntax-semantic analysis of Russian. Based on structural and functional attributes of BaseNP, this paper investigates the modeling phrases in a sentence and their multi-stage formalized description of Russian as a typical inflectional language. We hope that this paper can ultimately lay a foundation for automated analysis and recognition of BaseNP as well as provide a linguistic model for the syntax-semantic analysis of Russian.

Key words: NLP; BaseNP; Modeling

收稿日期： 2004-10-30

作者简介： 叶其松（1980-），男，安徽霍山人，黑龙江大学俄语学院教师。主要研究方向：计算语言学。

[责任编辑：孙淑芳]