俄语句法结构的模式化描述及操作原理

傅兴尚
黑龙江大学俄语语言文学研究中心 哈尔滨 150080
E-mail: fuxingshang@sina.com

概 要：句法结构的识别是解读句子意义的重要因素，所以，句法分析是自然语言处理的重要环节。本文从俄语作为典型屈折语这一个性特征出发，本着和汉语契合与对接的原则，探讨与俄语自动句法分析相关的几个问题：句法结构的类型；模式化描述的内容和形式；基于这种模型的操作原理。

关键词：俄语；句法分析；模式化；操作原理

On modelization of Russian syntactic structures and processing principles

Fu Xingshang
Center for Russian language and literature studies of Heilongjiang University, Harbin, 150080
E-mail: fuxingshang@sina.com

Abstract: The recognition of syntactic structure is the key to sentence meaning, and thus syntactic analysis composes an indispensable part in NLP. From the viewpoint of Russian as a typical inflective language this paper explores several problems in automatic analysis of Russian sentences in the light of their Chinese counterparts. The problems are about types of syntactic structures, contents and forms of modelization, and processing principles for the model.

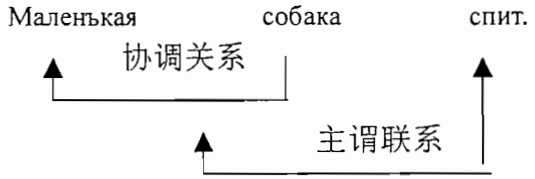
Keywords: Russian, syntactic analysis, modelization, processing principle

我们坚持“自然语言处理要充分顾及语言个性”这一原则。目前，我们正在建造的《面向信息处理的俄语语言知识库》是旨在为俄汉语信息交流服务的语言学保障体系，所以，做到两种语言规则的对接与契合是研究工作中的一个重点。本文研究内容属于俄语句法信息自动化处理问题的一个侧面。从以下几个方面展开讨论：１）论题解析；２）俄语句法结构的类型；３）模式化描述的内容和形式；４）算法。

１．论题解析

语句是最小的交际单位，能够表达相对完整的意义。语句的意义主要决定于两大因素：

一是语句构筑单位的意义；二是这些单位之间的关系，包括线性关系（前后顺序）、层次关系、这种关系的类型和主次地位，可统称句子结构。从不同层面出发，句子结构可分为形态结构、句法结构、语义结构和交际结构。对于自然语言(尤其俄语)处理来说，句法结构的识别是解读句子意义的重要因素。如：Маленькая собака спит.按传统语法分析，其句法结构可表示如下：



其中箭头表示主导词（单位）统辖的从属词（单位）。

自然语言处理有广泛的应用领域——机器翻译、信息自动查询、人机对话、文本自动摘要、 专家系统等，然而，不论那方面的应用领域，实现何种处理目标都离不开句法分析。就实质而言，句法分析的目的就是建立构句单位的层级联系并识别他们的联系类型。其常规表达形式体现为层次分明、主从分明、联系类型明确的句法树。

对于俄语来说，句法分析的入口（вход）是形态自动分析的出口(выход)，即，句中的每个词形已经还原为词位（лексема），并且已经标注出词类、语法意义（性、数、格、体、时、态、人称）等形态（词法）属性，所以，以此为基础建立句法树最重要的是对所有句法结构进行模式化描述，并设计相应的算法。

所谓句法结构(атомная синтаксическая конструкция)指各种构句单位的结构，通常由两个要素构成。例如：形容词＋名词（большая река）就是一个句法结构。语言具有递归性，所以，以有限的句法结构通过递归运用可表示多成分构成的结构，只是需规定运用条件而已。如："形容词＋形容词＋名词"（большая древняя река）可视为两个"形容词＋名词"结构的嵌套：形容词＋（形容词＋名词）。

2．俄语句法结构的类型

综合现有俄罗斯计算语言学与机器翻译研究领域的成果，结合俄语传统语法有关词组章节，我们区分出以下句法结构，每种结构还可能包含有变体形式。

1．1 数词结构，包括 1 ）常规数词结构（表示为 КОЛИЧ，以下同），如: двадцать восемь; одна тысяча пятьсот девяносто шесть.； 2 ）非整数结构（СЛОЖ_ЧИСЛ），如：12,2; 123,555.

1．2 名－数结构(СУЩ-ЧИСЛ)，如: статья 123, пункт 13.

1．3 代词限定语－形容词结构(МОДИФ_ПРИЛ)，如: такой красивый.

1．4 俄式人名结构(ФИО)，如: Иван Глебов; Глебов Иван

1．5 副—形结构(НАР_ПРИЛ) ，如: очень красивый, весьма полезный, особенно хорош.

1．6 同质形容词组结构(ОДНОР_ПРИЛ)，如： хороший, плохой и злой; первой и

единственной.

1 . 7 同质副词组结构(ОДНОР_НАР) ，如： плохо и хорошо

1 . 8 同质动词不定式组结构(ОДНОР_ИНФ) ，如： пить или курить

1 . 9 俄式日期结构(ДАТА)，如： август 1968 года.， 1 сентября 1939 года .

1 .10 形/副比较极结构(СРАВН-СТЕПЕНЬ) ，如： гораздо сильнее; значительно больше и умнее.

1 .11 副—动结构(НАРЕЧ_ГЛАГОЛ) ，如： злостно нарушать ; тяжело жить.

1 .12 形—名结构(ПРИЛ-СУЩ).，如：длинная унылая дорога; единственному настоящему другу.

1 .13 数量副词—名词结构(НАР-ЧИСЛ-СУЩ)，如：много очень простых ребят; мало красивых женщин.

1 .14 择选(элективная)结构(ЭЛЕКТ_ИГ)，如：одной из аудиторий; второй из нас

1 .15 数—名结构(ЧИСЛ-СУЩ)，如： сорок восемь попугаев; улица двадцати шести бакинских комиссаров.

1 .16 名一名所属结构(ГЕНИТ_ИГ)，如： в стране непуганых идиотов; рука Москвы.

1 .17 简单比较级结构(ОТСРАВН)，如：краше тебя, краше твоего дома

1 .18 前置词—名词结构(ПГ)，如： на холм; в краю степей; в большом просторном доме

1 .19 同质名词组结构(ОДНОР_ИГ)，如：Это было сказано руководителям （отдела и всего проекта.）

1 .20 否定词—动词结构(ОТР_ФОРМА)，如：не любить.

1 .21 动词一直接补语结构(ПРЯМ_ДОП)，如：рубить дрова; есть кашу; не любить маму.

1 .22 动词—动词不定式结构(ПЕР_ГЛАГ_ИНФ)，如：пойти выпить; позвать гулять.

1 .23 名—形后置限定结构(ПРИЛ_ПОСТПОС)，如：Зрелище это производило （ впечатление необычное, пугающее и очень неприятное）. Представьте себе （жизнь скучную, одинокую）, когда существование ваше никого постороннего не может интересовать.

1 .24 名一后置形容词独立语结构(СУЩ_ОБС_ПРИЛ)，如： ...мальчикам, большому и маленькому,.... ； ...брат и сестра, совсем больные,...

1 .25 被多个相同的连接词(повторяющиеся союзы)或间断性连接词（разрывные союзы）分割的同等成分结构(P_C_*)

这种结构主要包括 P_C_ОДНОР_ПРИЧ, P_C_ОДНОР_СУЩ, P_C_ОДНОР_МС, P_C_ОДНОР_ИНФ, P_C_ОДНОР_ДЕЕПР 等类型，如：то днем, то вечером，... не только вчера, но и сегодня.., ... хотя и очень больной, но довольно сильный... , ...если не писать, так читать... , ...как лежащие под столом книги, так и спрятанные в шкаф папки... и т.п.

所谓间断性连接词指 не...,а\но; не только...,но (и)...: как..., так и...; хотя...,зато...; пусть...,но...; хотя (и)..., но\а\да...; не то что\ чтобы..., но\а...; если (не)...,

то\так…

1.26 名词—形动词结构(ПРИЧ_СУЩ)，如： дом, построенный на холме.

1.27 名词—定语从句结构(ПРИДАТ_ОПР)，如： дом, кторый построили на холме. постепенно разрушался.

1.28 副词—谓语副词结构(НАР_ПРЕДИК)，如： очень интересно.

1.29 分析型比较级结构(АНАТ_СРАВН) ，如：более сильный, менее привлекателен

1.30 动词—尾缀结构（Г_ка)，如：пойдем-ка, давайте-ка, давай-ка

1.31 主谓结构(подлежа_сказуе) ，如：Кино мне нравится.

3．模式化描述的内容和形式

程序的基本单元体现为“条件—动作”的偶对。在这里，“动作”主要指合成，即把作为终端字符串的句中词形按照句法结构特征逐层合成为非终端字符串，直至归并为起始符S。所以，对句法规则从以下几个方面进行描写：1）句法分布结构，体现词类、形态属性以及线性序列等方面信息；2）合成的条件或附加条件；3）合成结果的表达；4）合成后的主导词和语法属性；5）汉化语序；6）实例。受篇幅所限，不能对第二部分列举的所有结构进行描述，在此只节选其中两个。

3．1 数词结构

3．1．1 常规数词结构（КОЛИЧ）

1）句法分布结构：俄文数词串（包括序数词）。

2）合成的条件或附加条件：连续排列。

3）合成结果的表达：КОЛИЧ。

4）合成后的主导词和语法属性：最后一个词，继承最后一个词的语法属性。

5）汉化语序：逐词翻译，如果最后一个词为序数词，前加“第”。

6）实例：сорок второй; двадцать два

3．1．2 非整数结构（СЛОЖ_ЧИСЛ）

1）句法分布结构：数字＋逗号＋数字

2）合成的条件或附加条件：无

3）合成结果的表达：СЛОЖ_ЧИСЛ

4）合成后的主导词和语法属性：逗号前的数字

5）汉化语序：替换

6）实例：12,2; 123,555.

3．1．3 混合结构

如果是诸如«20 тысяч»的混合字符序列，则采取如下解决方案：

令ЦК—表示数字字符串，Xn –属于集合 {тысяча, миллион, миллиард}， n – 表示当前词形的语法意义（им——格, вн—四格， рд—二格，ед—单数，мн—复数），МЧ–属于小数词集合 {два, три, четыре}, БЧ – 表示除去小数词集合和“один”的任何大数词，则：

a) ЦК + Хрд (3 тысяч)：СЛОЖ_ЧИСЛ 获得 им/вн, мн

б) МЧим + Хрд,ед (две тысячи)：СЛОЖ_ЧИСЛ 获得 им/вн,мн

в) БЧим + Хрд,мн (двадцать тысяч)： СЛОЖ_ЧИСЛ 获得 им/вн,мн

г)“один” + из + Хрд (одной из тысячи)：同 ПРИЛ_СУЩ

3．2 形一名结构(ПРИЛ-СУЩ).

3．2．1 常规结构

1) 句法分布结构：长尾形容词（或等价物）＋名词（或等价物）

2) 合成的条件或附加条件：两者在性、数、格上构成协调关系。

3) 合成结果的表达：ПРИЛ-СУЩ

4) 合成后的主导词和语法属性：名词（或等价物）及其语法属性

5) 汉化语序：形容词（或等价物）＋的＋名词（或等价物）

6) 实例：очень большая река

3．2．2 形容词独立语一人称代词结构

1）句法分布结构：逗号或句首标记＋X＋逗号+人称代词或专有名词，X属于集合{单一形容词\形动词，形容词\形动词等价物，同质形容词组}

2）合成的条件或附加条件：两者在性、数、格上构成协调关系。

3）合成结果的表达：ПРИЛ-СУЩ

4）合成后的主导词和语法属性：人称代词或专有名词及其属性

5）汉化语序：X＋的+人称代词或专有名词

6）实例：Вернувшись поздно,（ усталый и очень недовольный, он） мгновенно уснул.

3．2．3 数范畴不一致的形一名结构（A）

1）句法分布结构：同质形容词组或形容词（或等价物）＋名词

2）合成的条件或附加条件：各个形容词为单数形式，名词为复数形式，二者只在格范畴上保持一致。

3）合成结果的表达：ПРИЛ-СУЩ

4）合成后的主导词和语法属性：名词，继承名词的语法属性

汉化语序：同质形容词组或形容词（或等价物）＋的＋名词

6）实例：с красной и синей бабами

3．2．4 数范畴不一致的形一名结构（B）

1）句法分布结构：形容词（或等价物）＋同质名词组

2）合成的条件或附加条件：形容词（或等价物）为复数形式，同质名词组中的第一个名词为单数形式，且与前面形容词在格范畴上保持一致。

3）合成结果的表达：ПРИЛ-СУЩ

4）合成后的主导词和语法属性：同质名词组，继承形容词（或等价物）的语法属性。

5）汉化语序：形容词（或等价物）＋的＋同质名词组

6）实例：усталым дяде и тете

4．操作原理

为了行文和理解方便，我们引入句法块的概念。句法块（Синтаксическая группа）是指句子中的某一片段（отрезок），常以该段的第一个词和最后一个词为分割边界，其中标有主导块（词或另一个句法块）。句法块有以下表现形态：1）构句词形；2）由词形与词形合成的句法块，如：ПРИЛ-СУЩ；3）词形与句法块合成的句法块；4）句法块和句法块合成的句法块。句法块不能间断，句块之间只能是一个包含在另一个之中，不能相互交叉。

所谓操作原理就是要展现利用上列结构模式化规则实现句法自动分析的动态步骤和算法。我们对各种结构的描写是一套归并操作模式，也就是句法规则，依此把构句单位（词形或句法块）归并为新的句块。处理过程中，需调用所有的规则以便把当前输入词形（按从左到右的顺序）与右边相邻的词形合成新的单位，如果合成成功，则用新的句块再与相临的词形合成，否则，以右边的词形为输入单位再继续合成。当然所有的句法规则应该按一定的顺序排列。这个顺序与合成句法块的层级顺序是一致的，例如，应归并副－形结构，再合成形－名结构，这样才能正确识别“очень красивый город”：ПРИЛ-СУЩ(НАР_ПРИЛ(очень красивый) город)。

有些句法结构歧义忽略不计，即归并结果是约定性的。如：对 древние стены города 的处理结果是 генит_иг(прил_сущ(древние, стены), города)（генит_иг 表示受第二格名词限定的名词短语，прил_сущ 表示被形容词限定的名词短语），形容词与名词的合成先于名词与名词的合成，这是硬性规定。

根据上述规则和操作原理，可合成层次分明的句法块，最后归结为主语块和谓语块的合成。这一步骤比较复杂，在此简要展示。

定义：句法名词是指 1）名词及其等价物；2）主体性代词；3）诸如 каждый, один,другой,тот,который 等代形容词。

性、数一致规则：两个句块 1）数范畴一致；2）如果都是单数，性范畴也一致。

人称、数一致规则：两个句块 1）数范畴一致；2）人称范畴一致。

令 X 是潜在主语的语法属性，Y 是潜在谓语的语法属性。则，建立主谓结构关系，当且仅当：

1) 具有过去时或者短尾的属性，且

（a）X 具有第一人称第二人称的标记，X，Y 都是复数形式，或者 Y 是阳性或阴性；

ты вышел； я вышел； ты была； мы приежали

（b）X 与 Y 在性、数范畴上协调。

он вышел； поезд ушел； девочка красива； девочки красивы； мальчик красив

2) 动词有现在时或将来时语法标记，且满足以下两个条件之一：

（a） X 是第一人称和第二人称，且与 Y 在人称和数上一致；

я выйду； ты выйдешь

（b） X 和 Y 具有共同的数范畴。

они выйдут

3) 动词有命令式的标记，X 具有第二人称的语法属性并且在数范畴与 Y 一致。

Садись ты

4) X 为 НАР_ЧИСЛ_СУЩ, ЧИСЛ_СУЩ 等名词等价物时，Y 作为潜在谓语应该满足以下条件之一。

（a）动词具有单数、中性、过去时的标记；

（b）动词具有复数、过去时的标记；

（c）动词具有现在时或者将来时单数第三人称的标记；

（d）动词具有现在时或者将来时复数第三人称。

5. 结论

句法分析过程不仅合成句法块，最后通过主谓结构合成句子，而且还在同时不断消除形态歧义。至于句法歧义的消除尚需后句法分析以及调用语义词典进行语义分析后才能最终解决。关于通过语义分析消除句法歧义问题后有专论，此不赘言。另外，我们目前只设计了数词结构和形一名结构的识别模块，并在黑龙江大学计算语言学研究所自行建设的 3M 俄语语料（主要来自俄语网页）上进行了测试。结果显示，数词结构识别精确率为 86%，召回率为 82%；形一名结构识别精确率为 74%，召回率为 73%。出错原因主要有三个方面：规则覆盖不全面，规则选择顺序尚需调整，以及部分语料噪音。总体看来，这样的结果对句法分析来说还不够实用，因为单个句法块的错误往往造成整个句子结构识别的错误，所以下一步工作除了逐步实现其它识别模块外，还要继续完善规则集，并在真实语料处理的基础上引入规则的统计性分布信息。

参 考 文 献

[1] Герд А.С. Структурная и прикладная лингвистика ,С.-Петербургский университет 1998.

[2] Герд А.С. Прикладное языкознание ,С.-Петербургский университет,1996.

[3] Григорьев Н. В. Восходящий алгоритм построения дерева зависимостей для системы ЭТАП3//Труды Международного семинара Диалог' 1999 по компьютерной лингвистике и ее придлжениям, том 2 1999.

[4] Золотова Г. А. Синтаксический словарь М., 1988.

[5] Невзолова О. А. Алгоритм сегиентации предложений на простые составляющие//Труды Международного семинара Диалог' 2000по компьютерной лингвистике и ее придлжениям, том 2 2000

[6] Марчук Ю.Н. Проблемы машинного перевода М.,1983.

[7] Марчук Ю.Н. Основы компьютерной лингвистикиы М.,2000.

[8] Мельчук И. А.Опыт теории лингвистических моделей «смысл текст» Школа М.,1999.

[9] 傅兴尚：《现代俄语事格语法》， 军事谊文出版社， 1999 年。

[10] 傅承德：《自然语言理解的方法与策略》， 河南人民出版社， 2000 年。

[11] 赵铁军等：《机器翻译原理》，哈尔滨工业大学出版社，2000 年。