从离散数学角度看俄语短语结构

许汉成

(黑龙江大学，哈尔滨 150080)

摘 要： 本文利用数学中的关系和图论理论表示俄语短语结构，尝试初步建立俄语短语结构的数学模型，从而为俄语句法、语义自动分析奠定基础。俄语短语由二个或二个以上词或词形节点构成，构成短语的每个词或词形以及整个短语都是具有复杂特征的对象。从形式语法的角度研究俄语短语结构，是建立俄语单语或者俄汉双语自动处理系统的基础性工作，具有重要的理论和应用价值。

关键词： 自然语言处理；语言模式化；短语结构；俄语句法学

中图分类号： H354.3 **文献标识码：** A

1 引论

众所周知，语言模式化（моделирование языка）是计算语言学的核心问题之一，如果想建立俄语或者俄汉双语自动处理系统，就须首先建立俄语的形式化模型。语言模型主要有两大类：基于规则的和基于统计的。在计算语言学界，前者代表了理性主义学派，后者则代表了经验主义学派（黄昌宁 2002；袁毓林 2001）。我们不想轻率地赞成或者反对任何一种研究角度或方法，具体的研究对象、应用目的以及现实条件决定着我们的选择。本文是作者利用数学中的关系和图论理论建立俄语短语结构模型的一次初步尝试，可以看作是一种理性主义的研究方法。

任何语言的语篇都是由语言单位按照一定规则和层次逐级构成的，从词素到词、从词（词组）到句子、从句子到段落、从段落到语篇。尽管当代语言学的发展水平已经远远超越了句本位的束缚，但是，就自然语言处理而言，句子仍然是最重要的研究对象。句法、语义的形式化始终是计算语言学的核心问题。俄罗斯语言学家曾利用句子模式（структурная схема предложения, модель предложения）表示俄语句子的基本结构（Н.Ю. Шведова 1980：83—190; В.А. Белошапкова 1981：437—454），这些研究为俄语句法—语义形式化作出了贡献。但是，这些研究成果总的来说还属于描写语言学的范畴，离建立自动分析和生成合格的俄语句子的符号系统这一目标还相差很远。

本文试图利用数学中的二元关系和图论理论研究俄语短语结构，建立俄语短语结构的数学模型。俄语短语结构的研究具有重要的理论意义和应用价值，是俄语形式句法—语义形式化的基础。

2 研究短语结构的重要性

如果我们用变量 S 表示任意语句，用 w_i（i=1，2，3，…，n， n∈N，N 是自然数）表示语句 S 里依次出现的词或词形，那么 $S=w_1w_2w_3 \ldots w_{n-1}w_n$。从表面上看，语句 S 是由 n 个词或词形结点组成的一种线性结构，而 n 就是线性结构 S 的长度。但是，语言结构是层次

性的，句子的表面线性关系下面隐藏着层次关系。在语言学研究中，与聚合关系、组合关系一样，层次关系也十分重要。我们应该而且必须明确表示出语言结构的层次关系。英语语法里有句子（sentence）、小句（clause）和短语（phrase）的概念。乔姆斯基的转换生成语法被称为短语结构语法，短语结构语法能够将句子表示成句法树，从而揭示句法结构的层次性（徐烈炯 1988：81—125；冯志伟 1999：244—258）。俄语句法结构也是层次性的，建立俄语的短语结构语法既是可能的，也是必要的。

俄罗斯语言学家经常使用词组这个术语，短语的概念与俄语语言学的词组概念相近，但不完全相同。俄罗斯科学院《俄语语法》认为，“词组是在主从联系（即一致、支配、依附）基础上形成的句法构造。”（信德麟等 1990：488；Н.Ю. Шведова 1980：79）。这意味着俄罗斯科学院这种权威机构不承认并列结构（如 папа и мама, ты или я 等）是词组。尽管我们刚刚开始从形式化的角度研究俄语短语结构，对于短语的概念理解尚且不够深入，不过我们觉得并列结构也应该是短语，因为并列结构在句子或者语句里也是相对独立的单位，是建立句子层次结构所必须考虑的因素。

在研究俄语句法分析的问题过程中，我们切身感觉到了建立俄语短语结构语法的重要性和紧迫性。利用格（падеж, case）、题元（актанты）这些概念研究句法语义是一种通用做法。傅兴尚从事件结构出发利用事格的概念建立了俄语事格语法（傅兴尚 1999）。事格语法反映了俄语句子的情态框架、谓词及其题元结构，是国内俄语形式文法研究方面的一项重大成果。可是这部语法里少了短语结构这个层次，致使句子的分析不够彻底。当然，事格语法还要发展，这是后话，这里只想举这个例子说明短语结构的重要性。看一个实例，令 S=На вечере эти студенты будут читать стихи Пушкина. 图 1 是这个句子的事格语法表示（傅兴尚 2002：199）。

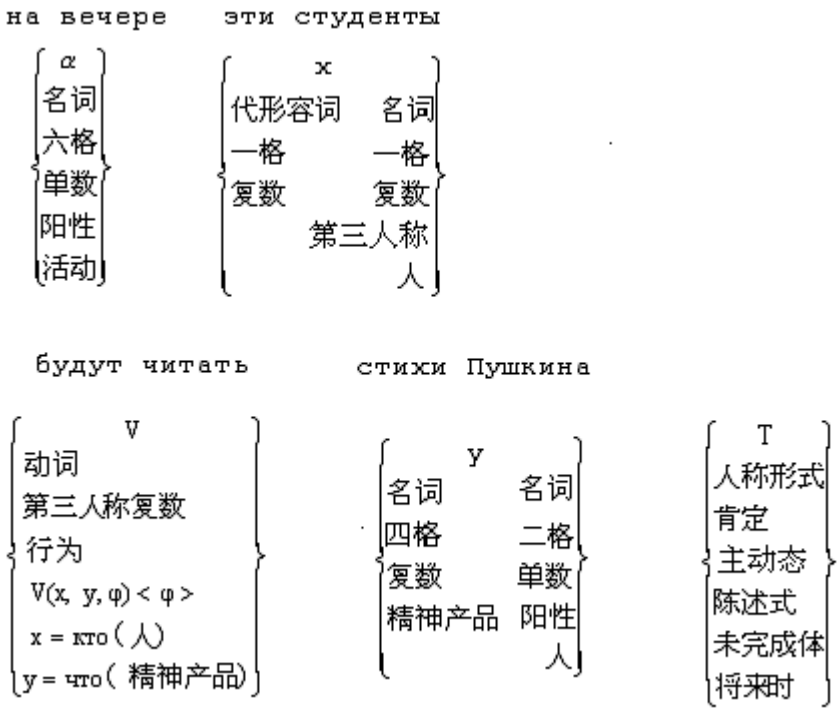

图 1

可以看出，事格语法目前暂时还无法处理类似 на вечере, эти студенты, стихи Пушкина 的短语结构。事格语法必须补充短语规则，才能成为一个完整的语法体系。

再举一个例子。令 S=Вы посмотрели полное описание моделей наших тракторов. 忽略一些细节问题，从事格语法角度看，这个句子的树形图如下所示：

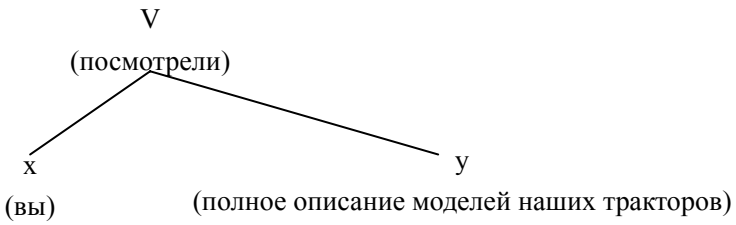

图 2

可以看出，x 由一个词构成，而 y 的结构比较复杂，是一个混合词组。即使我们规定了 описание 的事格框架，成功地分析了 описание моделей 的结构，但是短语结构 полное описание, наши тракторы 没有分析。

我们认为，建立俄语短语结构的形式化描写体系是十分重要的，是全面、彻底地完成句法语义分析的前提。在当代形式语法学界，句子的形式结构一般采用基于短语结构语法（或者改进的短语结构语法）及特征制约机制进行分析，语义问题则利用格语法、谓词逻辑或者内涵逻辑表示，这是一个总的趋势（俞如珍 金顺德，1994；冯志伟，1999）。我们面临着一个紧迫的任务——研究俄语短语结构，建立俄语短语结构的数学模型。俄语的短语结构语法是一个不可逾越的阶段。

3 关系、图论和俄语结构短语模型

短语结构数学模型是一个反映俄语短语结构的抽象的、形式化的数学结构。任何语言结构都可以看成是对象，对象都具有一定的特征，特征由属性和值偶对组成，根据描写需要，对象的复杂特征集里的属性—值偶对可以增删（许汉成 傅兴尚，2003）。这就是说，任何语言结构，包括短语，都可以看成是一个二元组 D=（E，Q），D 表示语言结构，E 表示构成语言结构的词、词形及其组合的集合，Q 是语言结构特性的复杂特征集（冯志伟 1996：272—316；刘颖 2002：66—69）。

按照结构，《俄语语法》将词组分为简单词组（простое словосочетание）、复杂词组（сложное словосочетание）和混合词组（комбинированное словосочетание）3 类（Н.Ю. Шведова, 1980：79—82；信德麟等，1990：488—490）。简单词组是建立在单一的强联系或弱联系（如 новый дом，бояться грозы，звонок из редакции，идти пешком）、双重强联系（如 отдать книгу ученику，вбить гвоздь в стену）或者强弱联系（如 открыть дверь гостю，отдать часы починить）的基础上的句法结构。复杂词组则建立在两种或两种以上联系的基础上，但这些联系是由一个主导词规定的，并以不同的方式结合（如 неожиданный приход начальника，лежать на диване с книгой，осмотр врачом больного）。混合词组也建立在两种以上联系的基础上，但这些联系是由不同的主导词规定的（如 увлеченно читать интересную книгу，иметь ровный край намотки，дать рекомендацию по приобретению оборудования）。显然，复杂词组和混合词组都是在简单词组基础上按照句法规则组合而成的。因此，我们可以首先从建立简单词组的数学模型开始，然后用递归的方法按照一定规则生成复杂词组和混合词组。自然语言异常复杂，为了简化问题，本文不研究复杂词组和混合词组的生成问题，同时也不研究固定词组和成语的问题。

《俄语语法》提到了词组元数问题，认为简单词组可以分为二元的与三元的，并且认为 перевести книгу с английского на русский 可算作四元（Н.Ю. Шведова, 1980：81）。显然，《俄语语法》的作者不把前置词算作短语的元。这也有一定道理。前置词表达一个实词对另一个实词在词组或句子中的关系，从而表达这些词所称谓的事物与动作、状态、特征之间的关系。然而，不把前置词看成词组的元数是从语义角度考虑问题的，短语结构首先是一个句

法结构（形式）问题。从句法角度看，前置词的语法功能非常重要，起着中心词的作用，支配着它后面名词的形式。各种英语语法都把“前置词+名词”作为一个短语结构。所以，我们姑且把前置词也算作短语结构的元数。这样，由二个词或词形构成的短语称为二元短语，如 наши тракторы, модели тракторов 等。对于二元短语，集合 E 的基|E|=2。如果一个词组既是简单词组，同时又是二个短语，它就是二元简单短语。可以看出，подарить дочке куклу 是一个三元简单短语，модели наших тракторов 是一个三元短语，但是 модель наших тракторов 不是简单短语，其中 наши тракторы 是一个二元简单短语，可以看成是一个整体，这个整体又限定 модели。至于 звонок из дома 这样的词组，它是一个三元简单短语，按照短语语法可以表示为：

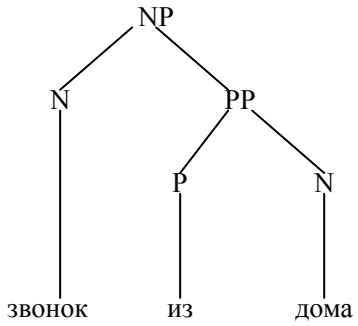

图 3

在说明了二元简单短语的概念之后，就可以用数学上的二元关系概念建立俄语二元简单短语的数学模型了。二元关系是离散数学的重要概念。假定 A，B 是两个集合，R 是笛卡尔乘积 A×B 的一个子集，则 R 就是 A 到 B 的一个关系。关系实际上是一个集合，由所有有序对(a，b)(a∈A，b∈B，(a, b)∈R)构成(李大友 1999：14—18)。假如 A={хороший, плохой}，B={человек, книга, читать}，那么 A×B 上的修饰关系为 R={(хороший человек)，(хорошая книга)，(плохой человек)，(плохая книга)}，如表 1 所示：

表 1

	человек	книга	читать
хороший	√	√	
плохой	√	√	

俄语简单短语结构具有以下重要属性：联系方式（一致关系、支配联系和依附联系）、语义关系（客体关系、限定关系、疏状关系、补足关系等）、短语类型（名词性短语、动词性短语、形容词性短语、副词性短语等）。本文的限定关系是指 новый дом, наши тракторы 之间的这种说明、限制名词性词语的关系；而疏状关系是指 ночевать в лесу 这种修饰、说明动词性词语的关系，这也与《俄语语法》在术语使用上略有不同。我们把这些属性放入复杂特征集，用这个特征集来描写短语结构。

我们用赋权图 G（V，E，f，g）来表示俄语的词或词形对象以及它们所构成短语的特征。其中 V 是图的顶点的集合，表示词或词形结点，E 是图的边，由 V 中的元素两两连接而成，代表节点之间的关系（如果词形之间没有关系，则没有连线），f 和 g 分别表示节点和边的特征（属性和值矩阵）。这样，短语 хорошая книга 可以用下图表示：

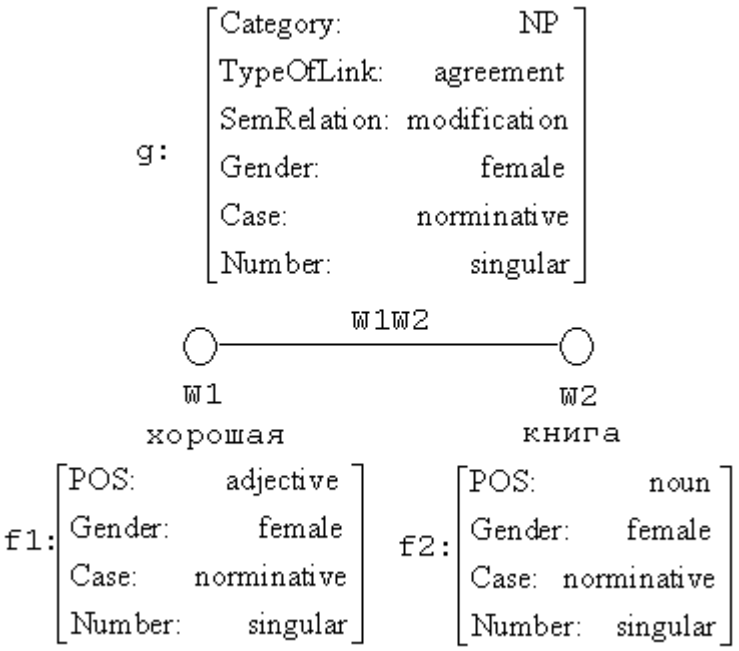

图 4

这里，POS 是 part of speech 的缩写，即词类，Gender, Case, Number, Category 分别表示性、格、数、短语类型、短语联系性质、短语元素语义关系。w_1、w_2 分别代表 хорошая 和 книга 这两个词形，显然，w_1、$w_2 \in V$，边 $w_1w_2 \in E$。f_1、f_2 分别表示节点 w_1、w_2 的特征。g 则表示短语结构的特征。Category, TypeLink, SemRelation 表示表示短语类型，NP, agreement 和 modification 分别表示名词短语、一致、修饰或限定关系。可见，利用二元关系理论（必要时可以扩展到多元关系）可以描述俄语短语结构的数学模型的形式语义特征，如果规定词构成短语的规则，那么上述模型还可以说明由词生成短语的过程，这一思想显示了一种建立俄语短语结构模型的诱人前景。

利用这种短语结构数学模型分析俄语短语结构，对其进行分类，建立类似乔姆斯基扩充式短语结构语法，确定每类短语结构的分析和生成算法，这一切显然已经不能在本文完成了。

4 俄语主要短语类型

由并列连接词（и, или, не только..., но и...）连接的短语是并列短语。组成并列短语的各个词形之间没有主次之分，因而并列短语在句子里的句法作用基本相同。主从短语则不同，主从短语由两个或两个以上词或词形组成，其中有一个占有支配地位，决定着从属词的语法形式和整个短语的性质，占主导地位的词被称为中心词（head），占从属地位的词被称为从属词（adjunct）。俄语短语的主从关系主要有 3 种：一致关系，支配关系和依附关系。为了句子语义分析的需要，可以将主从关系短语收缩或展开。

详尽论述俄语短语结构已经超出了本文的任务。但是，为了有一个直观的印象，我们还是根据 A.B. Сокирко（2000）的论文，将俄语自动处理过程中常见的短语结构列举出来，如下表所示：

表 2

类型	简称	释例
数词短语	КОЛИ	двадцать восемь
数字序列，中间有标点符号	КОЛИ	12, 2
词表规定的个别名词+数量标识符	СУЩ_ЧИСЛ	статья 123
姓—名—父称短语	ФИО	Петров Петр Владимирович
表示比较等级的词（如 очень）+形容词或形动词	НАР_ПРИЛ	очень красивый
同等形容词	ОДНОР_ПРИЛ	первый и единственный
同等副词	ОДНОР_НАР	плохо и хорошо
同等不定式	ОДНОР_ИНФ	стоять или лежать
同等形容词比较级	ОДНОР_ПРИЛ	красивее и моложе
日期短语	ДАТА	август 1968 года, 12 июня 99 г.
时间短语	СЛОЖ_ПГ	с первого августа по двадцатое сентября
形容词或副词分析形式比较级	СРАВН-СТЕПЕНЬ	гораздо сильнее
副词+动词	НАРЕЧ_ГЛАГОЛ	злостно нарушать
一个或若干个形容词+名词	ПРИЛ_СУЩ	длинная унылая дорога
副词性数词+名词短语（复数二格）	НАР_ЧИСЛ_СУЩ	много очень простых ребят
数词+名词短语	ЧИСЛ+СУЩ	сорок восемь попугаев
名词+名词/名词短语（二格）	ГЕНИТ_ИГ	модель трактора
前置词短语	ПГ	на холме
同等名词短语	ОДНО_ИГ	мама и папа
否定词+动词	ОТР_ДОП	не любить
网址和电子邮件地址	ЭЛ_АДРЕС	www.list.ru
动词+不定式	ГЛАГ_ИНФ	пойти выпить

上面表格显示了俄语短语结构的复杂性，即便如此，还是不能保证这张表已经列出了全部俄语短语结构。由于个人语法观点不同，分类标准不同，短语结构的类型也会变化。无论怎样，将俄语句子的线性结构映射到一个层次性短语结构是一个复杂的过程。但是，这始终是一个十分诱人的题目，具有重要的理论和应用价值。

5 主从短语的收缩与展开

我们定义两个运算符相反的收缩（>>）与展开（<<），其操作对象是主从短语。收缩是指将主从短语的从属词暂时隐藏起来，从图形上看仿佛“折叠”起来一样，而展开则是指将隐藏的短语结构的从属成分重新显现出来。由于主从短语的主导词的词性常常决定着整个短语的句法功能，可以先用“>>”运算，将俄语句子里的所有主从关系词组“折叠”起来，反复使用这个运算，直到只剩下中心词，这样我们就可以得到一个核心句子结构。如短语 полное описание моделей наших тракторов

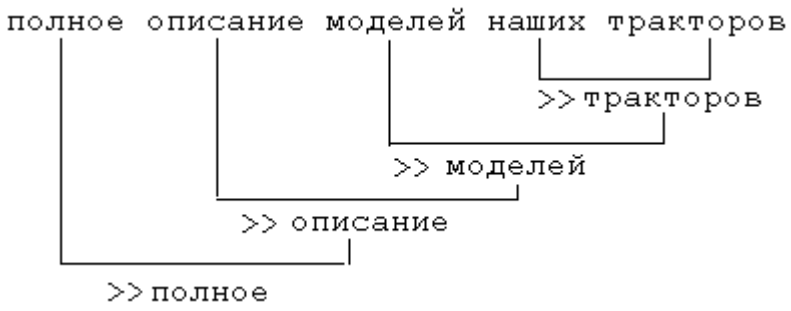

图 5

经过收缩运算就变成了一个由名词“описание”代表的名词性结构，从而大大简化了句法结构。假如有一个句子 S = Он прочитал полное описание моделей наших тракторов，经过收缩运算，句子 S 变成了一个非常简单的结构 S=Он прочитал ОПИСАНИЕ，这个简单结构按照《俄语语法》可以表示成 N_1V_f，也可以用事格语法表示成 Прочитал(он, ОПИСАНИЕ)，当然这里“ОПИСАНИЕ”代表整个 NP 短语，NP = полное описание моделей наших тракторов. 扩展（“<<”）则是收缩的相反，恢复短语原来成分和结构。

6 结束语

短语结构是一个十分重要的问题，是建立俄语各种形式语法的基础。本文的俄语短语结构数学模型既描写了短语成分及其特征，同时也描述了整个短语结构本身的特征。充分描写短语结构的特征可以为句法—语义分析提供必要的信息。从《俄语词汇信息的形式表达》到《俄语短语结构的数学模型》，我们似乎看到了建立俄语形式语法的依稀模糊的路径。这条路径便是充分考虑各层次语言结构对象以及对象之间关系的特征（即属性—值偶对），利用数学、逻辑工具建立、完善俄语形式语法。

俄语形式语法以及俄汉语的对比研究不仅是俄语教学、俄汉语翻译这些传统语言应用领域的需要，更是解决信息时代中俄交流所面临的新问题（如机器翻译、自动摘要、信息检索、数据挖掘等）的需要。我们认为，我国的俄语研究既要吸收俄罗斯语言学的丰富营养，同时也可以借鉴英美形式语法理论，跟踪国际语言学研究前沿，补充、更新国内俄语语言研究的方法。

参考文献

[1] Белошапкова В. А. 1981 Современный русский язык [M]. М.

[2] Сокирко А. В. 2000 Семантические словари в автоматической обработки текста (по материалам системы Диалинг), канд. дис. [D]. http://www.aot.ru

[3] Шведова Н. Ю. 1980 (ред.) Русская грамматика [M]. I-II, М.

[4] 冯志伟 1999 现代语言学流派 [M]，西安：陕西人民出版社。

[5] 冯志伟 1996 自然语言的计算机处理 [M]，上海：上海外语教育出版社。

[6] 傅兴尚 1999 现代俄语事格语法 [M]，北京：军事谊文出版社。

[7] 傅兴尚 2002 范畴·规则·操作 [M]，哈尔滨：黑龙江人民出版社。

[8] 黄昌宁 2002 统计语言模型能做什么？[J]，语言文字应用，第 1 期。

[9] 李大友 1997 离散数学 [M]，北京：清华大学出版社。

[10] 刘　颖 2002 计算语言学 [M]，北京：清华大学出版社。

[11] 信德麟 张会森 华劭 1990 俄语语法 [M]，北京：外语教学与研究出版社。

[12]徐烈炯 1988 生成语法理论[M]，上海：上海外语教育出版社。
[13]许汉成 傅兴尚 2003 俄语词汇知识的形式表达[J]，外语学刊，第 1 期。
[14]俞如珍 金顺德 1994 当代西方语法理论[M]，上海：上海外语教育出版社。
[15]袁毓林 2001 计算语言学的理论方法和研究取向[J]，中国社会科学，第 4 期。

A Mathematical Model of Russian Phrase Structures

XU Han-cheng

(Heilongjiang University, Harbin 150080，China)

Abstract: The author takes the first step to establish a mathematical model of Russian phrase structures utilizing the mathematical tools of relation and graph theories. The model is intended to contribute to the automatic syntactic and semantic analysis of Russian language after further development and refining. Russian phrases are represented as mathematical structures from two or more nodes of words or word forms. The words or word forms and the phrase itself can all be characterized with a complex features set. The study of Russian phrase structures from formal approaches is considered fundamental to and indispensable for monolingual or multilingual NLP systems in which Russian language is involved.

Keywords: NLP; language modeling; phrase structures; Russian syntax

收稿日期： 2004-07-28

作者简介： 许汉成（1965- ），男，陕西汉中人，解放军国际关系学院副教授。主要研究方向：俄语语言学，计算语言学。

[责任编辑：孙淑芳]