PLAUFL 解放军外国语学院
DISSERTATION SERIES 博士文库

俄语语料库研究

EYU YULIAOKU YANJIU

李绍哲◎著

中国出版集团
世界图书出版公司

PLAUFL 解放军外国语学院
DISSERTATION SERIES 博士文库

俄语语料库研究

EYU YULIAOKU YANJIU

李绍哲◎著

中国出版集团
世界图书出版公司

图书在版编目（CIP）数据

俄语语料库研究 / 李绍哲著. —广州：世界图书出版广东有限公司，2016.4
ISBN 978-7-5192-0804-2

Ⅰ. ①俄… Ⅱ. ①李… Ⅲ. ①俄语－语料库－研究
Ⅳ. ①H35

中国版本图书馆 CIP 数据核字（2016）第 047599 号

俄语语料库研究

策划编辑：刘正武
责任编辑：张东文
出版发行：世界图书出版广东有限公司
（地址：广州市新港西路大江冲 25 号　邮编：510300
网址：http://www.gdst.com.cn　E-mail：pub@gdst.com.cn）
发行电话：020-84451969　84459539
经　　销：各地新华书店
印　　刷：广州市佳盛印刷有限公司
版　　次：2016 年 4 月第 1 版　2016 年 4 月第 1 次印刷
开　　本：787 mm × 1092 mm　1/16
字　　数：334 千
印　　张：19.5
ISBN 978-7-5192-0804-2 / H · 1035
定　　价：42.00 元

前 言

现代俄语语料库的发展具有明显的后发优势。在俄罗斯，俄语语料库的建设还在继续，语料库的应用研究已经开始起步。国内俄语界对俄语语料库及其应用的研究还很罕见，迫切需要系统梳理俄语语料库建设的理论，开展基于俄语语料库的应用研究。

俄语国家语料库是俄语语料库的代表，其语料规模、标注深度、检索系统都属国际先进之列，是最大的俄语学术语料库。本书系统研究俄语国家语料库的建设理论，尝试性进行基于语料库的俄语研究。

俄语国家语料库收录的语料涵盖11—14世纪、15—18世纪、19—21世纪三个阶段，子语料库类型包括平行语料库、方言语料库、诗歌语料库、教学语料库、口语语料库、报纸语料库、句法语料库、重音语料库和多媒体语料库；该语料库的元文本标注符合EAGLES规范，词法标注遵循俄语语言学传统，句法标注以依存语法为基础，语义标注体现了俄罗斯词汇语义学的最新研究成果；检索系统可以实现按词目、词形、语法、语义及其组合的复合检索。

语料库语言学的哲学基础以经验主义为主，研究方法总的来说是自下而上的方法，包括：语料库驱动的方法和基于语料库的方法。但有时也采用自上而下的方法、语料库例证的方法，将定量研究与定性研究相结合。

词汇学是语料库语言学应用的主要方向之一，本书尝试借助语料库方法，对传统词汇学的研究范围和方法进行扩展。在语料库应用部分，对поезд进行历时研究，对名词的动物性和非动物性及其形成原因做了分析，对前置词кроме不同义项在语义、句法、语用方面的差异进行研究，对

только и V_1 (делать / знать / уметь), что V_2 结构在语言不同层面的特点进行了分析，调查了副动词非标准形式在现代俄语中的分布，分析了其演变趋势。

语料库语言学的应用显然不仅限于词汇学领域，俄语中将语料库应用于其他领域的研究尚有待进一步开拓。

由于本人水平有限，而且对俄语语料库进行系统研究、用语料库对俄语语料进行实验性分析在国内尚属首次，许多相关的问题有待于进一步深入研究，许多结论有待于进一步验证，所以本书中难免会有不足之处，甚至疏漏、错误之处，希望读者批评指正。

作者

2015 年 10 月

目 录

绪 论

一、国内语料库研究的现状

语料库语言学是随着语料库的建立而出现的一门相对年轻的交叉学科。学者们在语料库的建立、开发和应用中逐渐形成了一些独特的方法，提出了一些初步的原则，并且对这些方法和原则在理论上进行了探讨和总结，逐步形成了语料库语言学。自20世纪80年代以来语料库语言学取得了迅猛的发展，语料库语言学是以语料库为基础进行语言研究的一种方法，它一方面为各种现有语言学提供一种新的研究手段，另一方面，依据语料库所反映出来的语言事实对现行语言学理论进行批判，提出新的观点或理论。它包括两个方面的内容：①建立语料库的理论和方法；②利用语料库进行语言研究和应用开发的理论和方法。

语料库语言学两方面内容的研究还在不断推进，但从事后者研究的学者比前者要多，原因在于：一方面，目前学界已经形成了一套较为成熟的建立语料库的理论和方法；另一方面，建立一个有影响的语料库需要大量的人力、物力和财力，不是某个学者单独所能完成的任务。因此，国际上语料库语言学研究的热点集中在基于语料库的语言学研究和语料库的开发应用上，近年来国际英语语料库语言学年会提交的论文所涉及的内容可以证明这一点。

当前国际上语料库语言学的研究热点集中在以下几个方面：

- 语料库在语言研究中的作用；
- 词汇、语法和语义的探索；

- 语篇和语用研究；
- 语言的变迁与发展；
- 跨语言研究；
- 语料库软件开发。

国内语料库语言学的研究则是两方面内容并重：一方面，多个语料库的建设项目正在进行；另一方面，对现有的国际、国内语料库的开发应用研究同步推进。从中国知网（CNKI）的学术趋势[①]来看，从1997年到2015年关注语料库研究的中国学者总体呈上升趋势，见图0.1。在中国知网学术期刊和硕、博士论文文献数据库中以“语料库”为关键词进行检索，在1994年至2015年间粗略检索到4375篇论文或学位论文。论文所涉及的研究课题分布情况见图0.2。

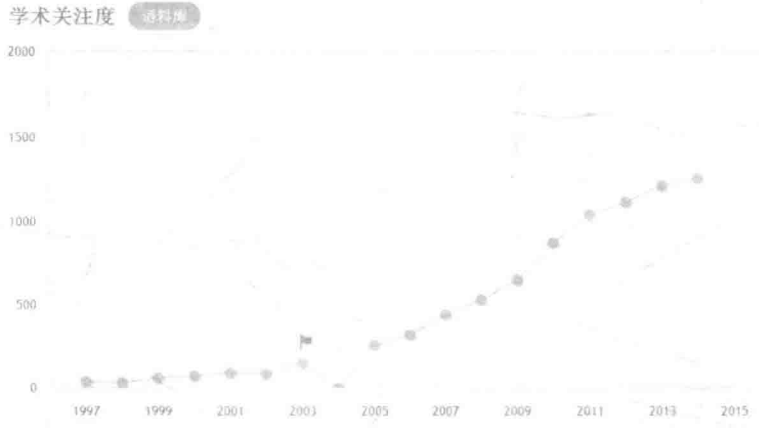

图0.1：中国学者1997—2015年对语料库的学术关注度（数据来源：CNKI）

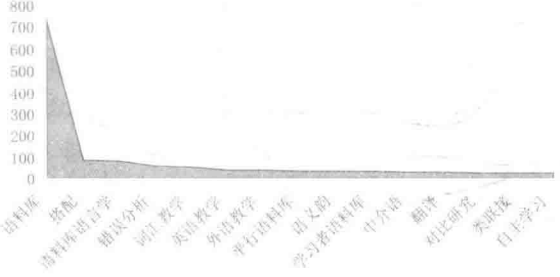

图0.2：1994—2015年语料库相关研究课题分布（数据来源：CNKI）

① 学术趋势的网址 http://trend.cnki.net

从文献调查来看，语料库语言学在国内发展的历史只有20多年，学者们对语料库本身的研究占绝大部分。王臻在中国知网利用中国期刊全文数据库进行搜索，将检索项设定为“题名”，检索词设定为“语料库”，在1994—2006年的“核心期刊”范围内进行检索，共查询到169篇文章，除去其中关于会讯的文章8篇，共有161篇文章纳入考察范围。按照文章的研究内容进行了简单的分类统计，其结果直观地展现了我国国内对语料库研究的大概情况，见表0.1和表0.2。①

表0.1：1994—2006年国内语料库研究情况统计

	非英语	英语（135）							合计
		理论型（18）		应用型（117）					
		引介综述	前瞻性	教学研究	语料库研究	语言研究	翻译	词典编纂	
篇数	26	14	4	52	26	18	11	10	161
比例（%）	16.15	8.70	2.48	32.30	16.15	11.18	6.83	6.21	100

表0.2：不同语种语料库研究情况统计

	英语	汉语	日语	法语	蒙语	俄语	维语	粤语	合计
篇数	135	17	2	2	2	1	1	1	161
比例（%）	83.85	10.56	1.24	1.24	1.24	0.62	0.62	0.62	100

从表0.1和表0.2来看，国内语料库研究主要集中在英语学界，从初期的理论引介开始发展到语料库建设研究和具体的实践应用性研究，进而发展到预测语料库研究的未来走向。可见，近30年的不断探索已使英语语料库语言学找到自己的合理定位。转视非英语学界，已有不少语种的专家开始致力于语料库的研究和开发，相比之下，俄语界的语料库研究较为滞后，还鲜见研究语料库语言学两方面内容的文章。

综上所述，国内俄语界对语料库和语料库语言学的研究尚属薄弱环节，

① 王臻. 俄语语料库语言学研究现状与瞻望 [J]. 中国俄语教学，2007（2）：44—47.

与英语界和汉语界相比在该领域的研究差距较大，究其原因，我们认为主要有以下几点：

（1）国内某个语种语料库研究的状况取决于语言对象国家的语料库发展水平，也就是说，国内俄语语料库研究的状况在很大程度上受制于俄罗斯国内语料库的发展水平。俄罗斯的俄语语料库基本是在2000—2004年建成，有的语料库还处在建设之中，因此，俄罗斯的语料库建设理论和实践刚刚起步，这种状况制约了国内俄语界对俄语语料库的相关研究。

（2）语料库语言学是语言学和计算机科学交叉形成的一门学科，进行语料库语言学的研究要求具备语言学、计算机和数学方面的专业知识。我国俄语界学者的语言学理论功底扎实，这已是世界俄语界的共识，但既具备语言学知识又比较精通计算机和网络技术的学者不多，能将数学方法应用于语言学研究的就更少，这些因素客观上影响了我国俄语界对语料库语言学的研究。

（3）基于语料库的应用研究必须系统地了解和掌握俄语语料库的结构、功能和使用，然而目前国内缺乏对俄语语料库的细致梳理和系统的研究，还不了解俄语语料库的文本构成、标注方法、具有的功能和使用方法。虽然我国俄语界有不少先知先觉的学者在研究中已经使用到了语料库，但有的应用还称不上语料库的研究方法。因此，对俄语语料库的陌生阻碍了基于语料库的应用研究。

基于以上分析，对俄语语料库建设和应用两个方面进行系统研究不仅是必要的，而且是十分迫切的。

二、本书的理论意义和实践价值

随着俄语语料库，特别是俄语国家语料库建设的日臻完善，国内俄语界通过借鉴俄罗斯已有研究成果建立自己的俄语语料库，并开展俄语语料库应用研究的条件基本成熟，但前提是要对俄语语料库及其建设思想有系统的了解。因此，本书将系统梳理俄语语料库建设的理论和思想，探讨基于俄语语料库的语言学研究方法，尝试对俄语中的某些语言现象做实证性研究。

本书的理论意义在于：通过对比不同语种语料库选取语料的原则，将语

料选取引向科学化；通过分析文本分类原则和标注方法将语料库建设引向规范化；系统研究俄语国家语料库建设理论和思想，对于国内建立相关俄语语料库具有重要的理论意义和借鉴价值；对基于语料库研究方法的分析，可以拓展俄语语言学研究的思路。

本书的实践价值在于：国外俄语语料库建设中开发的工具和软件可以用于国内俄语语料库建设、俄语搜索引擎开发和相关俄语语言工程中；使用“自下而上”的方法，通过俄语语料库发现以前无法观察到的语言现象，从而提出俄语语言研究的新课题；使用“自上而下”的方法，验证学者提出的语言理论假设。

此外，不同语种在语料库建设方面所面临的问题各不相同，如对汉语来说自动分词技术和词性标注是难点，对俄语来说形态还原、自动标注、同音异义现象的过滤等是难题，但所有语言的语料库在语义标注和语义消歧方面都存在不足，认清这些问题，可以为后续研究明确方向。

基于语料库进行语言学研究为研究人员提供了一种新的思维角度，克服语言研究者靠“直觉”和“内省”判断所产生的主观性和片面性，语料库的研究方法在词典编纂、语言定量分析、语言教学、词汇研究、句法研究、语义研究和作品风格分析等领域已显现出了强大的生命力。以俄语为语料进行基于语料库的应用研究在国内尚不多见，在这一领域开展研究非常必要。

三、本书的目的和研究方法

语言研究的方法总是与语言研究的目的密切相关的，无论是描写的研究方法还是阐释的研究方法，它们之间并不存在非此即彼的排斥关系。从研究目的来看，有的研究为语言描写和语言教学服务，有的研究为语言工程提供基础支持，还有的研究专门通过语言来探索人类心智以及语言与社会的关系。“语料库语言学将数据收集与理论论述有机地结合在一起，使我们对语言的理解发生了质的变化。”[①] “在过去几十年里，语料库以及语料库研究对

① Halliday M. A. K. Corpus studies and probabilistic grammar [C]// English Corpus Linguistics: Studies in Honor of Jan Svartvik. London: Longman, 1991: 30–43.

语言研究以及语言应用研究进行了一场革命。”[①] 因此，本书的目的是通过对比不同语种语料库建设的实践，系统研究建设俄语语料库的理论，特别是俄语国家语料库的建设思想，通过实证研究来实践俄语语料库在语言研究中的应用。

在语料库的建设方面本书采用对比分析与描写归纳相结合的方法；在语料库的应用方面既采用“自下而上”的方法，以发现并分析以往因条件所限而未能注意的语言现象，又采用“自上而下”的方法，分析语言现象，验证提出的语言理论假设。两种研究中都将定量分析与定性分析、“内省法”与“例证法”相结合。

四、本书的结构

本书正文分为5章，各章标题和主要内容如下：

第一章 世界背景下的俄语语料库

现代语料库发轫于20世纪50年代，勃兴于80年代。本章将俄语语料库放在主要的英语语料库、汉语语料库以及东欧、日本的语料库背景之下，通过对比各语料库在文本构成、规模、语料标注方面的特点和成就，分析某些语料库在语料库语言学发展中的地位和作用，划分语料库发展的三个阶段，弄清俄语语料库在世界语料库背景中的定位，为下一章的研究提供素材。

第二章 语料库语言学理论

界定语料库和语料库语言学的基本概念，分析其经验主义的哲学基础；从不同角度探讨语料库语言学的学科定位；总结并归纳语料库语言学的研究方法，包括：语料库数据驱动的方法和基于语料库的方法，但有时也采用语料库例证的方法；分析语料库建设中影响语料库质量和应用价值的因素以及不同层次语料标注的原则和理论；最后，对语料库的管理和应用进行简要的阐述。

第三章 俄语国家语料库研究

① Hunston S. Corpora in Applied Linguistics [M]. Cambridge: Cambridge University Press, 2002: 1.

重点研究俄语国家语料库的结构和文本组成，分析元文本标注的方法及其遵循的规范，研究该语料库在词法、句法、语义等语言层面进行标注的理论与实践，探讨与主语料库不同、独具特色的句法标注语料库、平行文本语料库、口语语料库和方言语料库的建库理论与方法，为国内建设同类语料库以及开展基于俄语国家语料库的各项语言研究奠定基础。

第四章 俄语国家语料库的使用

本章通过图文结合、Q&A 的方式说明如何使用俄语国家语料库，主要涉及如何访问俄语国家语料库、如何构建个人子语料库、如何在语料库中进行检索、诗歌语料库的特点和平行语料库的特点。

第五章 基于语料库的实证研究

本章依据俄语国家语料库的数据资源，尝试进行语言学实证研究，限于笔者的知识结构，研究主要限于词汇领域。词汇学是语料库语言学应用的主要方向之一，借助语料库方法，传统词汇学的研究范围和方法得到了扩展：词汇及词义的历时演变研究、词汇的用法以及语法范畴的分布、词汇的支配模式和搭配的特点等。

第一章　世界背景下的俄语语料库

现代语料库的发展大致可以划分为三个阶段，20 世纪 50 年代至 80 年代为第一个阶段，80 年代至 90 年代为第二个阶段，90 年代以后为第三个阶段，不同阶段建立的语料库分别称为第一代、第二代、第三代语料库。这种划分方法对英语语料库是比较合适的。

1959 年，英国语言学家 R. Quirk、G. Leech 和 S. Greenbaum 组织发起了“英语用法调查”（the Survey of English Usage，简称 SEU）项目，目的是全面描述英语语法。他们合作建立了百万词次的英国英语口语和书面语语料库，后来在 80 年代初将纸质语料电子化，形成计算机版本的语料库。这是从传统语料库过渡到计算机语料库的重要标志。而将大量语料进行电子化处理，并用计算机进行标注加工、存储以及检索、取样和统计分析等则是近 30 年来的事。1961 年，以 N. Francis 和 H. Kucera 为首的一批语言学家和计算机专家会集在美国的布朗大学，合作建成了世界上最早的机读语料库 Brown University Standard Corpus of Present-Day American English Corpus，即布朗语料库（Brown Corpus）。[①] 1975 年，Svartvik 与他隆德大学的同事把 SEU 语料库中的口语部分转变为计算机可读的形式，最后在 80 年代建立了 London-Lund Corpus of Spoken English，即伦敦—隆德英语口语语料库（LLC）。“这 3 大经典语料库为现代语料库语言学奠定了坚实的基础，因为它们所采用的不是通过臆想或杜撰的语言，也不是仅仅从某些著作抄录下来的若干例句，而是有目的有系统地收集的大量在现实生活中使用的书面语和口头语，并使

① W3-Corpora project. The Brown Corpus [EB/OL]. (1998-02-05) [2015-04-18]. http://www.essex.ac.uk/linguistics/external/clmt/w3c/corpus_ling/content/corpora/list/private/brown/brown.html.