

国家社会科学基金青年项目
教育部留学人员科技活动择优资助项目
中国博士后科学基金
黑龙江省博士后基金

资助



语
言
学
论
丛

俄汉 机器翻译系统中的 语义识别

以第五格名词为例

Семантическая интерпретация
падежных форм в системах
русско-китайского
машинного перевода:

На примере творительного падежа

李 爽 著



北京大学出版社
PEKING UNIVERSITY PRESS

国家社会科学基金青年项目
教育部留学人员科技活动择优资助项目 资助
中国博士后科学基金
黑龙江省博士后基金



语
言
学
论
丛

俄汉 机器翻译系统中的 语义识别

以第五格名词为例

Семантическая интерпретация
падежных форм в системах
русско-китайского
машинного перевода:

На примере творительного падежа

孙 爽 著



北京大学出版社
PEKING UNIVERSITY PRESS

图书在版编目(CIP)数据

俄汉机器翻译系统中的语义识别:以第五格名词为例/孙爽著.
—北京:北京大学出版社,2012.7

(语言学论丛)

ISBN 978-7-301-20838-0

I. ①俄… II. ①孙… III. ①俄语—机器翻译—研究

IV. ①H355.9

中国版本图书馆 CIP 数据核字(2012)第 132861 号

书 名: 俄汉机器翻译系统中的语义识别:以第五格名词为例

著作责任者: 孙 爽 著

组 稿 编 辑: 白 渤

责 任 编 辑: 李 哲

标 准 书 号: ISBN 978-7-301-20838-0/H · 3086

出 版 发 行: 北京大学出版社

地 址: 北京市海淀区成府路 205 号 100871

网 址: <http://www.pup.cn>

电 子 信 箱: zbing@pup.pku.edu.cn

电 话: 邮购部 62752015 发行部 62750672 编辑部 62759634

出版部 62754962

印 刷 者: 北京彩云龙印刷有限公司

经 销 者: 新华书店

880 毫米×1230 毫米 A5 6.5 印张 140 千字

2012 年 7 月第 1 版 2012 年 7 月第 1 次印刷

定 价: 29.00 元

未经许可,不得以任何方式复制或抄袭本书之部分或全部内容。

版权所有 侵权必究 举报电话

电子邮箱

Предисловие 1

Исследования Сунь Шуан посвящены чрезвычайно важной и востребованной в современный период, разработанной теме создания лингвистических представлений для систем машинного перевода с флективных языков на изолирующие. В центре исследования работы находится русско-китайская языковая пара, однако предлагаемые решения применимы для более широкого класса языков.

Данная работа опирается на очень подробный и глубокий анализ значительного числа произведений научной литературы, в которых представлены исследования по теории падежа и предлагаются модели описания механизма падежей и установления передаваемых ими значений, а также анализ достижений и результатов работы существующих практических систем русско-китайского машинного перевода. В работе достигнута поставленная цель исследования— предложены способы решения сложной проблемы перевода русского творительного падежа с учетом того, что одна и та же словоформа в творительном падеже может иметь целый ряд различных значений, каждое из которых должно переводиться на китайский язык отличным от других способом. Для успешного решения этой проблемы автор обращается к глубинно-семантическим аспектам представления языковой реальности. Для достижения цели исследования автором решены поставленные задачи: выявлен максимально полный инвентарь семантических ролей творительного падежа в русском языке; определены те средства, которые являются наиболее частотными и стилистически нейтральными эквивалентами соответствующих значений творительного падежа в китайском языке; на основе анализа представительной выборки

аутентичных примеров из русских текстов разных функциональных стилей достоверно определены те семантические признаки лексического значения маркированной творительным падежом словоформы и те формальные и семантические признаки ее локального контекста, которые обуславливают значение показателя творительного падежа в его конкретном употреблении; найдены способы репрезентации, необходимой для целей перевода синтаксической и семантической информации в словарных статьях русско-китайского комбинаторного словаря; построены правила распознавания семантических ролей, выражаемых творительным падежом в русском языке; разработаны правила перевода, обеспечивающие выражение этих семантических ролей средствами китайского языка.

Композиционно работа состоит из введения, трех глав, заключения, списка использованной литературы и Интернет-ресурсов. Во введении обосновывается актуальность темы, указываются объект и предмет исследования, цели и задачи, научная новизна, теоретическая и практическая значимость работы, положения, выносимые на защиту, описываются материал и методы исследования; в заключении суммируются результаты проведенного исследования.

В первой главе представлена теоретическая база исследования: рассмотрены различные теории падежа и существующие подходы к описанию падежных значений, а также проблема перевода падежных значений на китайский язык. В первом разделе автор прослеживает историю развития представлений о падеже и эволюцию от чисто синтаксических представлений к таким, которые учитывают семантические свойства падежа. Второй раздел посвящен истории развития и современному состоянию систем автоматического перевода, а также в нем содержится анализ работы существующих систем русско-китайского машинного перевода с точки зрения перевода падежей.

Во второй главе рассматривается общая структура системы русско-китайского автоматического перевода в аспекте перевода падежных

показателей. В частности, указывается, что предполагаемая в диссертации система русско-китайского машинного перевода, в рамках которой может быть реализован предлагаемый автором способ обработки падежей, должна представлять собой двуязычную систему машинного перевода третьего поколения, дается подробное описание блоков и словаря системы.

Третья глава посвящена семантико-синтаксическому анализу различных конструкций с творительным падежом и представлению его результатов в виде формализованных правил распознавания семантических ролей именных групп с вершиной, оформленной творительным падежом. Для каждой выделенной семантической роли предлагаются правила ее выражения средствами китайского языка.

Проведенное исследование наглядно показывает, что обеспечение адекватной передачи значений русских падежей лексическими или синтаксическими средствами китайского языка возможно при условии разработки словаря, в котором представлена информация для распознавания семантических ролей именных групп, выражаемых при помощи падежей.

Работа Сунь Шуан—очень серьезное и глубокое исследование. Очень точно поставлены цели и задачи, результаты исследования убедительно доказаны на очень значительном фактическом материале. Теоретическая значимость работы Сунь Шуан заключается в том, что разработан и применен подход на основе глубинно-семантических представлений, в котором отражено подтвержденное результатами исследований положение о зависимости между лексическим значением существительного и типичными для его падежных форм значениями; уточнены представления о свойственных творительному падежу в русском языке семантических функциях и получены полные и точные сведения о характеристиках контекстов, в которых эти функции реализуются. На основе выявленных способов выражения установленных значений словоформ в творительном падеже построена система соответствий между языковыми механизмами выражения этих значений в русском и

китайском языках, которая может служить методологической основой для создания моделей перевода с флективных языков на изолирующие. Автор убедительно показывает, что для адекватного описания соответствий между языками, которые существенно отличаются по своему строю, необходимо обращение к глубинным семантическим характеристикам языковых объектов.

Результаты данного исследования Сунь Шуан несомненно имеют очень высокую практическую ценность и могут быть использованы для создания усовершенствованных систем машинного перевода с русского языка на китайский язык. Исследование Сунь Шуан имеет большую практическую значимость для развития современных информационных технологий, а также в актуальной и динамично развивающейся сфере межкультурных коммуникаций. Предложенные автором методы являются очень полезным источником проектных решений при разработке систем машинного перевода третьего и четвертого поколений и других типов информационных систем, в которых предусмотрены компоненты обработки естественного языка. Кроме того, полученные результаты могут быть использованы при создании курсов и учебных пособий по русскому языку для тех обучающихся, чьим родным языком является китайский, и для русскоязычных студентов, изучающих китайский язык.

Е. Б. Козеренко

2011.08 Москва

Предисловие 2

Исследование Сунь Шуан посвящено проблеме семантической интерпретации падежных форм русского языка для целей их машинного перевода (МП) на китайский язык. В силу сложности и объемности этой темы, она рассматривается на примере одного падежа — творительного (ТВОР), представляющего особую трудность для автоматического семантического анализа в связи с многообразием его синтаксических и семантических функций.

В работе Сунь Шуан проявила похвальную основательность. Во-первых, установив, что значение падежа можно отождествить с семантической ролью актанта или сирконстанта, обозначаемого именной группой (ИГ), в ситуации, обозначаемой предикатом элементарной предикативной конструкции, Сунь Шуан на основе изученной литературы составила максимально полный унифицированный инвентарь семантических ролей, выражаемых при помощи творительного падежа в русском языке. Во-вторых, Сунь Шуан приступает к решению поставленной задачи только после критического рассмотрения основных подходов и методов, применяемых в современных системах машинного перевода, которое приводит её к выводу о том, что для адекватной интерпретации падежных граммем в системе необходим семантический компонент, включающий словарь с разнообразной семантической информацией, правила приписывания русским именным группам их семантических ролей и правила перевода этих ролей на китайский язык. Для формальной записи семантической информации автор берет метаязык «Онтологической семантики» С.Ниренбурга и В. Раскина [Nirenburg, Raskin 2004] — прикладной семантической теории,

ориентированной на задачи автоматического семантического анализа текстов на ЕЯ. Этот шаг можно только приветствовать, поскольку использование универсального семантического языка в правилах приписывания падежным формам семантических ролей позволит этим правилам естественно встраиваться в общую систему семантического анализа. Результаты, полученные Сунь Шуан в ходе проведенного ей семантических ролей, имеют несомненную теоретическую ценность.

Работа вносит существенный вклад в формальное семантическое описание русского языка творительного падежа в виде правил приписывания семантической ролей. Полученные результаты имеют большую значимость для разработки лингвистического обеспечения систем русско-китайского МП, а также в преподавании русско-китайского языка как и иностранного, в особенности в китайской аудитории.

Представленная работа свидетельствует о больших способностях и трудолюбии Сунь Шуан, также высокую квалификацию как специалист и в области теории языка, в области русского языка, и в такой актуальной области прикладной лингвистики, как машинный перевод.



И.М.Кобозева

2011.07 Москва

作者自序

在语言学与众多学科发生广泛交叉的今天，自然语言处理得到了飞速的发展，然而俄汉机器翻译系统的性能同其他语言的处理结果相比显然还有相当一段距离。中国数学家和语言学家周海中曾指出：要提高机器翻译的质量，首先要解决的是语言问题而不是程序问题。单靠若干程序来做机译系统，肯定是无法提高机译质量的。《俄汉机器翻译系统中的语义识别》一书正是从语言学角度出发，顺应发展趋势，旨在解决俄汉机器翻译系统的性能问题。

该研究独辟蹊径，以尼林布尔克和拉斯金的本体论语义学为基础，以俄罗斯国家语料库的真实语料分析为例，在机译过程中引入语义特征信息，以俄语中不带前置词第五格为例建立语义识别规则及其俄汉翻译规则。本书在内容结构上主要包括三部分。第一部分主要研究了格结构所确定的情景参与者的语义角色描述手段；中国和俄罗斯机器翻译的历史和现状；从俄语格意义的角度分析俄汉在线翻译。第二部分提出了建立俄汉机器翻译系统的设想，在该框架内提供对相应格的处理方法，详细研究了俄汉组合词典的词条编写。第三部分为该研究的核心部分，建立了俄语中不带前置词第五格结构的语义识别规则和俄译汉规则。该研究对创建俄汉机器翻译系统、建立俄语语言信息库、俄汉平行语义词典以及俄语教学都具有重要的理论价值和实践应用前景。

本书是在作者原博士学位论文的基础上修改加工而成。在本稿付梓之际，我要衷心地感谢所有关心、支持和帮助过我的老师、亲人和朋友。

首先,我要特别感谢莫斯科国立大学的两位恩师,И.М. Кобозева教授和Ю.Н. Марчук教授。两位导师在我攻读硕士和博士学位的五年半时间里,给予了细心指导,倾注了大量的心血,耳提面命,将我引上了学术道路,并激励我锲而不舍,把握前沿,勇攀高峰。恩师的谆谆教导让我刻骨铭心,恩师的为人、为学、为师之道带给我深深的震撼,是我永远学习的榜样。在成书过程中,俄罗斯信息科学院的Е.Б.Козеренко、俄罗斯国立人文大学的М.Б. Рукодельникова、莫斯科国立大学的М.В. Всеволодова、俄罗斯科学院信息传输问题研究所的Л.Л. Иомдин等教授都提出了很多宝贵意见,在此一并表示感谢。这些专家们深厚的知识底蕴、严谨的治学态度深深地影响着我,激励着我,使本书最终得以完成。

同时,我要把这本书,献给我的父母和所有亲人。感谢家人多年来一如既往的支持和鼓励,让我心无旁骛地投入到学习和研究中,并给我不断增加探索的勇气和动力。

最后,我要真诚地感谢东北林业大学有关领导和老师对专著出版给予的关怀、帮助和支持。

本书的出版得到了国家社会科学基金青年项目(项目编号:11CYY063)、教育部留学人员科技活动择优资助项目(项目编号:41311401)、中国博士后科学基金(资助编号:2011M500700)和黑龙江省博士后基金的资助,同时得到了北京大学出版社的大力支持、协助。作者一并致以最诚挚的谢意!

该研究目前仍处于理论研究的层面,尚未经俄汉机器翻译系统进行检验。囿于学识有限,书中难免有纰漏和不当之处,恳请各位专家和学者们批评指正,不吝赐教。

孙 爽

2011年12月于哈尔滨

ОГЛАВЛЕНИЕ

Введение	1
Глава 1. Проблема перевода морфологического падежа в системах автоматического перевода с флективного языка на изолирующий	10
1.1. Падеж как средство выражения семантической роли актанта ситуации, обозначаемого именной группой	10
1.1.1. <i>Понятие о падеже в теории языка: морфологический падеж в русском языке и выражаемые им семантические роли</i>	10
1.1.2. <i>Средства выражения синтаксических и семантических ролей в китайском языке</i>	21
1.1.3. <i>Семантика творительного падежа в русском языке</i>	31
Выводы	46
1.2. Общая стратегия разработки систем машинного перевода	47
1.2.1. <i>История и современное состояние российского и китайского машинного перевода</i>	47
1.2.2. <i>Существующие подходы к созданию систем машинного перевода: современное состояние российского и китайского машинного перевода</i>	55

1.2.3. Анализ образцов онлайн-перевода с русского языка на китайский в аспекте перевода русских падежных показателей	63
Выводы	72
Глава 2. Общая структура системы машинного перевода с русского языка на китайский в аспекте перевода падежных показателей	73
2.1. Общие сведения о предлагаемой архитектуре русско-китайского машинного перевода	73
2.2. Русско-китайский словарь системы и содержащаяся в нем синтаксическая и семантическая информация	81
2.2.1. Зона части речи (POS)	83
2.2.2. Зона переводного эквивалента (TRANS)	83
2.2.3. Зона синтаксических признаков (SYNT)	84
2.2.4. Зона семантических признаков (SEM)	84
2.2.5. Зона моделей управления (МУ)	86
Глава 3. Определение семантических ролей творительного падежа и правила их перевода с русского языка на китайский	90
3.1. Семантические роли синтаксисом творительного падежа, записываемые в моделях управления в словаре	90
Выводы	102
3.2. Перевод обусловленных синтаксисом творительного падежа	103
3.2.1. Синтаксема творительного падежа (ПВОР) в составе пассивной конструкции	106

3.2.2. Синтаксема творительного падежа в безличной конструкции	119
3.2.3. Синтаксема творительного падежа в составе сравнительной конструкции	124
3.2.4. Синтаксема творительного падежа в конструкциях со значением «оценочной квалификации»	129
3.2.5. Синтаксема творительного падежа в составе параметрической определительной конструкции	134
3.2.6. Синтаксема творительного падежа в составе ориентационной конструкции	136
3.3. Перевод свободных синтаксисом творительного падежа	142
3.3.1. Распознавание семантических ролей Инструмента и Объекта сравнения	145
3.3.2. Распознавание семантических ролей Маршрута и темпоральных	162
Выводы	175
Заключение	176
Правила распознавания семантических ролей и правила перевода на китайский	179
Литература	184
Словари	192
Internet-ресурсы	193
Список сокращений	194

Введение

Известно, что возникновение прикладной лингвистики имеет длинную историю, однако, становление ее как автономной научной дисциплины относится к относительно недавнему прошлому. По мере расширения информатизации современного общества возрастает значение прикладной лингвистики. Бурное развитие прикладной лингвистики связано с интеграцией компьютера и традиционной лингвистики, изучающей структуру языка, законы его развития и способы использования. Прикладная лингвистика оптимизирует функции языка, формирующие и поддерживающие традиции чтения и понимания или обеспечивающие контакты между разными народами. При этом соответствующие проблемы все больше входят в круг интересов как прикладной, так и теоретической лингвистики.

Как писал Б.Ю. Городецкий, «проблемы взаимосвязи науки и практики сегодня преломляются для лингвистики через призму требований глобальной компьютеризации. Язык носит всепроникающий характер. Любая проблема информатики, искусственного интеллекта, экспертных систем и—шире—рационализации человеческой деятельности имеет языковой аспект». [Городецкий Б.Ю. 1989]

По сложившейся еще в советский период терминологической традиции частью прикладной лингвистики является компьютерная лингвистика, к которой относят практически все, что связано с использованием компьютеров в языкознании. К области компьютерной лингвистики относится и машинный перевод (МП) или автоматический перевод (АП).

Машинный перевод понимается как выполняемое на компьютере действие по преобразованию текста на одном естественном языке в эквивалентный по содержанию текст на другом языке, а также результат такого действия¹. Научно-исследовательские и опытно-конструкторские работы по автоматическому (машинному) переводу ведутся во многих странах мира уже много лет. Параллельно с созданием систем МП разрабатывались и разрабатываются теории, которые можно было бы положить в основу создания таких систем. Интересно отметить, что такие теории, вследствие сложности задачи МП, как правило, охватывают не только в собственном смысле перевод

1 Цитируется по сайту «Кругосвет»: <http://www.krugosvet.ru>

с одного естественного языка на другой, но и многие смежные и актуальные, часто глобальные теоретические вопросы человеческого владения и пользования языком. [Марчук 2006: 244]

Развитие персональных компьютеров, а также распространение Internet и внедрение Internet обусловили существенное повышение спроса на автоматический перевод. Задача преодоления языкового барьера становится все более актуальной. Автоматический перевод нужен людям больше, чем когда-либо ранее. Однако создание систем автоматического перевода представляет собой трудную задачу.

МП (АП) есть комплексная теоретическая и практическая задача, в которой отражаются и концентрируются многие проблемы компьютерной лингвистики—от слова до способов анализа содержания и синтеза словоформы, предложения и целого текста.

Актуальность исследования.

Актуальность исследования определяется необходимостью совершенствования систем автоматического перевода с русского языка на китайский в отношении адекватности передачи значения русских падежей. Обращение к исследованию семантических ролей и возможности их использования в системе лингвистического обеспечения русско-китайского автоматического перевода обуславливается тем, что, с точки зрения грамматической теории, главные функции