Отзыв

О диссертационной работе **Ильи Олеговича Кузнецова**
«Автоматическая разметка семантических ролей в русском языке»
(специальность 10.02.21 — Прикладная и математическая лингвистика)

Более 60 лет назад выдающийся английский лингвист и философ Джон Руперт Фёрс провозгласил тезис, ставший крылатым выражением «You shall know a word by the company it keeps (J. R.Firth, *Papers in Linguistics 1934–1951,* London, Oxford University Press, 1957). По мнению оппонента, не будет преувеличением сказать, что в рецензируемой диссертации этому выражению придается строгий математический смысл.

В самом деле, диссертационная работа И.О.Кузнецова посвящена автоматическому приписыванию семантических ролей актантам предикатных слов в русском тексте, и эта задача не просто успешно решается, но и получает широкое и убедительное теоретическое обоснование.

Диссертационная работа насчитывает 178 страниц и имеет весьма четкую и прозрачную структуру. Основной текст предваряет краткое, но чрезвычайно ёмкое **Введение**, в котором содержится постановка задачи, фиксируется конкретная область исследования, избранная диссертантом (лежащая, по мнению автора, на стыке компьютерной лингвистики и автоматической обработки текстов) и определяется ее место в более широкой области знаний (искусственный интеллект), неформально излагаются основные положения избранной научной области, отмечаются главные вехи в ее развитии, а также характеризуются теоретические принципы и методы, на которых основывается рецензируемая диссертационная работа (это, с одной стороны, теория семантических ролей, восходящая к Ч. Филмору, а с другой, использование корпуса текстов, содержащего ролевую разметку, в системе машинного обучения). Обосновывается актуальность предлагаемой работы, демонстрируется новизна применяемых в ней решений и подходов и отмечается практическая значимость полученных результатов.

Основной текст диссертации состоит из четырех глав, традиционно разбивающихся на разделы (а частично и подразделы) и резюмирующего краткого заключения.

Библиография диссертации насчитывает 102 источника, из которых 17 русскоязычных, а 85 написаны на английском языке.

Первая глава, озаглавленная «Теория семантических ролей и автоматическая разметка актантов», носит в первую очередь теоретический характер. Здесь излагается система семантических ролей Ч.Филмора, построенный на основе этой системы компьютерно-лексикографический ресурс Frame Net, а также концептуально близкая к филморовской система лексических (или тематических) отношений Дж. Грубера. С достаточной степенью детальности описывается разработанный отечественными лингвистами под руководством О.Н.Ляшевской компьютерно-лингвистический ресурс для русского языка FrameBank, в целом основанный на принципах, близких к

филморовским, но отличающимся от них предикатно-специфической ориентацией семантических ролей актантов предикатных слов (прежде всего глагольных). Отмечается также близость этого ресурса к идеям Московской семантической школы, которая в свою очередь восходит к теории «Смысл ⇔ Текст» И.А.Мельчука, А.К.Жолковского и Ю.Д.Апресяна. Именно FrameBank послужил ресурсом, на основе которого строится система приписывания семантических ролей актантам предикатов в русском языке.

Эта же глава содержит тщательно выполненный литературный обзор, в котором характеризуются как работы прошлых лет, так и весьма современные исследования, относящиеся к данной тематике. Львиная доля этих работ приходится на английский язык, для которого задача разметки семантических ролей актантов была поставлена не менее десятилетия назад. Что касается русского языка, то работа И.О.Кузнецова – по существу первый масштабный опыт построения автоматической системы такой разметки.

Основные результаты диссертации приводятся во второй главе, озаглавленной «Система автоматической разметки актантов для русского языка». Данная глава занимает в тексте диссертационной работы центральное место как по объему, так и по содержанию. Она разбивается на пять достаточно крупных и относительно независимых подразделов. В первом из них строго излагается постановка задачи. Указывается, в частности, что система строится на основе предикатно-специфических ролей актантов. Отмечается также, что задача идентификации актантов рассматривается как задача классификации элементов дерева зависимостей, с помощью которых представляется синтаксическая структура русского предложения.

Второй раздел главы 2 задает исходные данные для строящейся системы, а именно, для обучения, тестирования и оптимизации предлагаемого классификатора. Эти исходные данные – не что иное, как коллекция материалов компьютерно-лексикографического ресурса FrameBank. Приводятся конкретные примеры таких материалов, которые обсуждаются со значительной степенью подробности и в исключительно ясных терминах, облегчающих понимание деталей исследования и разработки даже для читателя, не являющегося узким специалистом данной предметной области.

Третий раздел представляет собой структурированное описание системы ролевой разметки актантов, в котором характеризуются все модули, входящие в состав системы. Особый интерес здесь представляет собой завершающий модуль оптимизации, с помощью которого максимально эффективным образом разрешается неоднозначность классификации актантов, если таковая появляется в результате применения предшествующих модулей системы. В этом же разделе характеризуются свойства структуры предложения, для которого производится актантная разметка. Они подразделяются на синтаксические и семантические; к числу первых относится **полный путь** от предиката до его актанта в синтаксической структуре предложения и так называемый короткий путь; падеж актанта или его предложно-падежная форма (образно именуемая диссертантом как финский падеж) и некоторые другие, а к числу

вторых относится лемма, кластер (своего рода парадигматический объект, полученный для леммы предикатов с помощью нескольких часто используемых алгоритмов вроде Chinese whispers и word2vec, а также частеречная принадлежность актанта. Атрибуция последнего свойства как семантического может вызвать сомнения, но диссертант убедительно показывает, что с точки зрения машинного обучения часть речи относится именно к семантическим свойствам.

Четвертый раздел второй главы посвящен детальному описанию модуля глобальной оптимизации результатов работы системы по классификации элементов текста как актантов предикатов. Этот модуль выполнен на основе применения средств целочисленного программирования.

Наконец, пятый раздел характеризует некоторые технические особенности реализации системы: фиксируются форматы представления данных, задаются состав и параметры работы промежуточных модулей (морфологических, синтаксических и т.д.), описываются библиотеки программ, с помощью которых реализуется программный комплекс.

Чуть менее объемной, но отнюдь не менее важной содержательно является третья глава диссертации. Она озаглавлена «Экспериментальная оценка и результаты» и посвящена методам оценки качества созданной системы автоматической разметки семантических ролей актантов предикатных слов, а главное, практическому применению этих методов в процессе машинного обучения. Здесь формулируются критерии оценки, фиксируются применяемые метрики, как основные (такие как полнота, точность и F-мера), так и созданные автором специально для разработанной им системы. Глава 3 разбивается на четыре раздела, в которых подробно описываются критерии оценки качества, процедура такой оценки, а отдельно приводятся ее результаты. Оценка производится по целому ряду параметров и наборов применяемых свойств. Выделяется несколько лучших алгоритмов, конфигураций и вариантов работы системы, учитывается зависимость между ними, выделяются наиболее значимые и наименее значимые свойства. Показано, что наилучшие результаты дает комбинация синтаксических и семантических свойств, в то время как учет только одних семантических свойств приводит к ухудшению параметров работы системы.

Заключительная, четвертая глава диссертации подводит основные итоги диссертации и намечаются возможные пути продолжения работы, при котором можно ожидать приращения результатов. Среди этих путей диссертант видит, во-первых, применение альтернативных методов машинного обучения (в частности, интерпретируемого алгоритма машинного обучения, в котором вместо применяемого в работе метода опорных векторов использовались бы деревья принятия решений), во вторых, в применения ряда методов обучения без учителя, а, в третьих, в усовершенствовании и расширении используемых лингвистических ресурсов.

Следует подчеркнуть, что диссертация И.О.Кузнецова выполнена весьма и весьма тщательно, написано очень хорошим, грамотным, богатым и доходчивым языком. Основные положения диссертации излагаются последовательно, логично, за ходом изложения легко следить даже начинающему специалисту. У оппонента нет

сомнения, что данная диссертация будет использована не только как источник новой научной информации, но и как весьма качественное учебное пособие в области машинного обучения применительно к компьютерно-лингвистическим системам различного назначения.

Стоит отметить также, что ряд формулировок диссертации представляется очень удачным в общенаучном и даже философском плане. Одну такую формулировку, приведенную в самом начале текста, на стр. 5, хотелось бы отметить особо: «Такой подход помещает автоматическую обработку языка в контекст более общей задачи обработки сигналов (signal processing), что не обязательно является адекватным подходом к анализу феномена с богатой внутренней структурой, каким является естественный язык.».

В Заключении на стр. 165 диссертант справедливо отмечает, что в последние годы российская компьютерная лингвистика по спектру решаемых задач приближается к западным стандартам, имея в виду, что пока эти стандарты выше российских. Не вызывает никакого сомнения, что рецензируемая диссертация делает значительный шаг на этом пути.

Актуальность и научная новизна диссертационной работы И.О.Кузнецова несомненны: эта работа принципиально решает проблему ролевой разметки семантических актантов предикатных слов для русского языка, которая поставлена и выполнена впервые.

Замечания оппонента к работе немногочисленны и носят непринципиальный характер: некоторые из них, добавим, носят скорее статус не замечаний, а вопросов.

1. К сожалению, в диссертации очень мало внимания уделяется рассмотрению конкретных результатов разбора построенной системой конкретного языкового материала, как удачных, так и ошибочных. Это, увы, является общим свойством как конференционных и журнальных статей, так и диссертационных работ, выполняемых в русле компьютерной лингвистики. Между тем именно таких примеров ждет от таких работ специалист, интересы которого выходят за рамки такой лингвистики – они, бесспорно, служат интересам взаимного влияния теоретического языкознания и прикладных компьютерно-лингвистических задач.

2. Оппонент не вполне согласен с предлагаемым диссертантом разделением «нетрадиционной» лингвистики на компьютерную лингвистику и автоматическую обработку текста (диссертант использует вместо последнего термина выражение «автоматическая обработка языка, которая представляется мне не слишком удачной калькой с английского). По мнению оппонента, автоматическая обработка текста – это просто самое распространенное приложение компьютерной лингвистики.

3. На стр. 20 указывается, что «каждая лексема имеет набор синтаксических валентностей, определяющих, какие слова или категории данная лексема может иметь в качестве вершины и зависимых» Если говорить о задаче классификации актантов в том виде, в каком она поставлена в диссертации, то разумно говорить лишь о ситуации, когда актант синтаксически зависит от предиката. Если уж расширять состав

синтаксических валентностей, то, на взгляд оппонента, следовало бы хотя бы коротко рассмотреть противопоставление активных, пассивных и разрывных валентностей, предлагаемое И.М.Богуславским.

4. На стр. 17 в примере Pat came [to the library]$_{\text{Источник}}$ [from the cafeteria]$_{\text{Цель}}$ перепутаны пометы «источник» и «цель».

5. На стр. 21 неверно воспроизведена модель управления *курить* из ТКС И.А.Мельчука и А.К.Жолковского – *курить из трубки*, а не **курить из табака* :

1=X [кто вдыхает]	2=Y [из какого устройства]	3=Z [какое вещество тлеет
сущ. им.п.	сущ. в.п.	--
сущ. им.п.	сущ. в.п.	из сущ. р.п

6. На стр. 45 неточное воспроизводится представление структуры FrameNet:

[Abby]*Buyer* bought [a car]$_{Goods}$ [from Robin]*Seller* for [$5000]$_{Money}$.

В оригинале предлог *for* входит в состав валентности «money».

7. На стр. 54 на рис. 3 – отмечается, что во FrameBank при предикате *купить* указывается, что его пациенс – это $S_{вин}$ и $S_{парт}$. Между тем следовало бы указать, что это не обязательно партитивный, а родительный падеж. Синтаксически идентифицировать партитивное значение родительного падежа очень трудно. В НКРЯ такие конструкции классифицируются именно как примеры с родительным падежом.

8 В двух местах, по-видимому, по техническим причинам даются неверные библиографические ссылки. Так, на стр. 33-34 сказано, что «авторы опираются на теоретический аппарат теории связывания [Mel'čuk, 1988]». Между тем автор теории связывания – Н.Хомский, а вовсе не И.А.Мельчук. На стр. 54 говорится: «FrameBank представляет собой корпусно-лексикографический ресурс, описывающий лексические конструкции русского языка с помощью специальным образом размеченных предложений из Национального корпуса русского языка [Апресян, Богуславский, Иомдин, 2005].» Указанная ссылка касается не FrameBank, а СинТагРуса.

9. На стр. 148 приводится не очень ясное рассуждение о кореферентности в примерах типа ***Иван*** *хочет купить автомобиль, чтобы* ***ездить*** *на нём в деревню* и . ***Иван*** *был бы рад купить автомобиль, чтобы* ***ездить*** *на нём в деревню.* Как кажется, здесь речь идет о субъекте предиката *ездить,* которым является слово *Иван.* Однако никакой кореферентности здесь нет, можно лишь говорить об опущенном актанте

Разумеется, эти мелкие замечания никоим образом не влияют на общую чрезвычайно высокую оценку работы И.О.Кузнецова.

Добавим в завершении отзыва, что статьи автора на тему диссертации, а также автореферат диссертации, написанный исключительно ёмко и четко, вполне адекватно отражают основные положения работы.

Из сказанного можно заключить, что диссертационная работа И.О.Кузнецова полностью соответствует пунктам 9 и 10 Положения о присуждении ученых степеней и является научно-квалификационной работой, в которой содержится решение задачи, имеющей существенное значение для прикладной и математической лингвистики, а ее автор полностью заслуживает присуждения искомой ученой степени кандидата филологических наук.

Л.Л.Иомдин,
и.о. зав. лабораторией компьютерной лингвистики ИППИ РАН,
кандидат филологических наук,
ведущий научный сотрудник.

Адрес: Институт проблем передачи информации им А.А.Харкевича РАН,
Москва, 103051, Б.Каретный пер. 19, стр. 1
тел (499) 699-4927, E-mail: iomdin@iitp.ru

К теме диссертации относятся следующие статьи оппонента:

1. Iomdin L. Automatic Text Processing and Deeply Annotated Text Corpora of Russian: Interaction and Mutual Impact // Jazykovedné štúdie XXXI. Bratislava: Jazykovedný ústav Ľ. Štúra Slovenskej akadémie vied, 2014. C. 136-146. ISBN 978-80-224-1391.

2. Iomdin L., Petrochenkov V, Sizov V. Tsinman L. ETAP parser: state of the art // Computational Linguistics and Intellectual Technologies. International Conference (Dialog'2012). Moscow: RGGU Publishers, 2012. Issue 11(18). P. 830-843. ISSN 2221-7932.

3. Л.Л.Иомдин, Б.Л.Иомдин. Валентности русских предикатных существительных и микросинтаксические конструкции // Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции «Диалог» (2014). М.: Изд-во РГГУ, 2014. Вып. 13 (20). С. 219-231.

Л.Л.Иомдин

18 апреля 2016 года.

Подпись
Иомдина Л.Л.
УДОСТОВЕРЯЮ
Зав. канцелярией