УТВЕРЖДАЮ

Первый проректор
учреждения образования
«Гродненский государственный
университет имени Янки Купалы»

_____А. Е. Каревский
«___» _____2023 г.

ОТЗЫВ ОППОНИРУЮЩЕЙ ОРГАНИЗАЦИИ
учреждения образования
«Гродненский государственный университет имени Янки Купалы»
на диссертацию **Голяк Юлии Дмитриевны**
«Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)»,
представленной на соискание ученой степени кандидата филологических наук
по специальности 10.02.21 – прикладная и математическая лингвистика

Соответствие содержания диссертации заявленной специальности и отрасли науки

Диссертация Голяк Юлии Дмитриевны «Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)», выполненная на кафедре прикладной лингвистики филологического факультета Белорусского государственного университета, соответствует специальности 10.02.21 – прикладная и математическая лингвистика и отрасли науки «филология». Предметом исследования является автоматическое дополнение русскоязычных пользовательских запросов в корпоративных системах информационного поиска, что соответствует подобластям 2.7 и 2.15 «лингвистическое обеспечение автоматизированных информационных систем» области исследования «Моделирование процессов восприятия, хранения, преобразования и передачи информации на естественных языках», предусмотренной паспортом специальности 10.02.21 – прикладная и математическая лингвистика, утвержденным приказом Высшей аттестационной комиссии Республики Беларусь от 13.11.2017 № 260.

Научный вклад соискателя в решение научной задачи с оценкой его значимости

Актуальность исследования, выполненного Ю. Д. Голяк, не вызывает сомнений, так как оно находится в русле задач повышения эффективности информационного поиска, связанных с оптимизацией пользовательских запросов при работе с системами, снабженными дружественными естественно-языковыми интерфейсами. Одним из важнейших показателей эффективности

информационного поиска является полнота, качество которой напрямую зависит от релевантности поисковому запросу выданных в ответ на него документов или фактов, в зависимости от типа информационной системы. Некорректно или неточно сформулированный к информационно-поисковой системе запрос может привести к тому, что при высоком показателе релевантности выдач поисковому запросу может наблюдаться несоответствие полученных данных информационным потребностям пользователя и/или его информационным ожиданиям. К поисковым неудачам могут привести и недостатки лингвистического обеспечения, используемого в конкретной автоматизированной информационно-поисковой системе, что во многом объясняется сложностями автоматической обработки текста на естественных языках, характерной особенностью которых является принципиальная неоднозначность.

Новые технологические возможности позволяют осуществлять глубокую обработку текстовой информации на основе предварительного обучения систем автоматического анализа текстов на естественном языке, в том числе не соответствующих правилам действующей нормы. В этом свете исследование Ю. Д. Голяк приобретает особую актуальность, поскольку разработанный ею подход к предиктивному автодополнению запросов пользователей в системах информационного поиска служит существенным дополнением к лингвистическому обеспечению базового лингвистического процессора системы автоматизации инженерии знаний и управления знаниями IHS Goldfire. Апробация и внедрение в систему IHS Goldfire разработанной Ю. Д. Голяк подсистемы автодополнения русскоязычных пользовательских запросов в составе естественно-языкового интерфейса модуля информационного поиска, включая алгоритмическое и лингвистическое обеспечение данной подсистемы, метод и технологию решения задачи, подтверждается актом внедрения ООО «АйЭйчЭс Глобал» от 06.01.2022 г.

Таким образом, значимость диссертационного исследования Ю. Д. Голяк обусловлена разработкой метода, алгоритмов и лингвистического обеспечения решения задачи предиктивного автодополнения запросов пользователей в системах информационного поиска и их реализации в виде промышленного прототипа. Особую значимость имеет выбор именно русского языка как языка информационных запросов, поскольку для него характерна развитая и нерегулярная система словоизменения, что существенно осложняет задачу автоматической обработки текстов запросов на этом языке.

Конкретные научные результаты (с указанием их новизны и практической значимости), за которые соискателю может быть присуждена ученая степень

Проведенное Ю. Д. Голяк исследование позволило получить конкретные научные результаты, обладающие достаточной степенью новизны и высокой практической значимостью, а именно:

1. Разработана структурно-функциональная схема системы предиктивного автодополнения русскоязычных пользовательских запросов, согласно которой дополнение поисковых запросов осуществляется на этапе взаимодействия пользователя с естественно-языковым интерфейсом корпоративной информационно-поисковой системы, снабженной базой подсказок;

2. На базе свободно доступных специализированных интернет-ресурсов сформирован целевой корпус текстов пользовательских запросов, путем анализа которого впервые построена типология этого типа мини-текстов с учетом их лексических и семантико-синтаксических характеристик;

3. Разработан алгоритм расширения функционала лингвистического обеспечения базового лингвистического процессора путем включения в него оригинального русскоязычного ресурса в виде базы подсказок, которая, представляя собой специализированный машинный лексикографический ресурс (систему машинных словарей), создается автоматически на основе использования предварительно сформированной полнотекстовой базы, рассматриваемой в данном случае в качестве поискового массива корпоративной информационной системы;

4. С использованием лексико-грамматического классификатора и системы словарей, составляющих часть лингвистического обеспечения для русскоязычных задач базового лингвистического процессора IHS Goldfire, сформированы сопряженная с ними совокупность правил в виде пяти базовых процедур и множество паттернов, которые обеспечили эффективность алгоритмов автоматического построения базы подсказок. Предложенные правила позволили расширить функциональность используемого лингвистического процессора и являются важными составляющими алгоритмов автоматического построения списков подсказок, формулируемых с ориентацией на базовые синтаксические структуры русскоязычных запросов.

Полученные результаты являются новыми, имеют научное значение и могут быть использованы в таких областях прикладной лингвистики, как разработка систем понимания текста, автоматизация работ по информационному поиску, а также шире – в системах обработки вербальных и вербализованных данных на русском языке.

Конкретные рекомендации по использованию результатов диссертации

Полученные Ю. Д. Голяк результаты могут найти применение в сфере автоматической обработки текстовой информации, в т.ч. для расширения возможностей информационно-поисковых систем, снабженных интерфейсом на русском языке, путем введения модуля предиктивного автодополнения запросов пользователей, основанного на использовании выделенных Ю. Д. Голяк базовых синтаксических структур, соответствующих выявленным диссертантом типам пользовательских запросов. Разработанные в диссертации метод и алгоритмы предиктивного автодополнения могут использоваться при проектировании и

реализации корпоративных систем информационного поиска, снабженных интерфейсом пользователя, использующим другие языки, кроме русского.

Построенная система предиктивного автодополнения русскоязычных запросов пользователей внедрена в систему информационного поиска, входящую в многопрофильную многоязычную информационно-поисковую платформу IHS Goldfire, используемую для решения задач автоматизации инженерии и управления знаниями крупнейшими компаниями мира.

Не вызывает сомнения возможность использования результатов диссертации Ю. Д. Голяк при подготовке специалистов, чья профессиональная деятельность предполагает работу с интеллектуальными системами и/или разработкой лингвистического обеспечения для такого рода систем, причем полученные Ю. Д. Голяк результаты уже внедрены на кафедре информатики и прикладной лингвистики Минского государственного лингвистического университета, где используются при проведении занятий по дисциплинам «Компьютерные технологии обработки естественного языка» и «Информационная лингвистика», что подтверждается актом внедрения от 04.06.2022 г.

Предложенная Ю. Д. Голяк классификация основных типов поисковых запросов и соответствующих им базовых синтаксических структур русского языка может использоваться в дидактике русского языка как иностранного, а также для формирования цифровой грамотности у иностранцев, обучающихся в вузах Республики Беларусь, путем освоения ими основ информационного поиска, осуществляемого на русском языке.

Соответствие научной квалификации соискателя ученой степени, на которую он претендует

Диссертационное исследование Ю. Д. Голяк обеспечено должной теоретической базой, основанной в том числе на анализе патентов, отражающих состояние целевой для данного исследования предметной области. Автор хорошо ориентируется в своей предметной области, несмотря на ее интердисциплинарность и относительную новизну.

Исходя из поставленной задачи диссертантом был сформирован необходимый для исследования текстовый материал, состоящий из корпуса пользовательских запросов, обеспечивающий формирование представления о лексической, синтаксической и семантической природе наиболее частотных поисковых запросов, и полнотекстовый корпус как прототип поискового массива корпоративной информационно-поисковой системы. Основу корпуса поисковых запросов составили открытые источники и общедоступная система Wordstat.Yandex, которая позволяет формировать массивы русскоязычных поисковых запросов любого объема с информацией об их частотности для различных синтаксических структур и синонимической вариативности. Тестовый поисковый массив был сформирован на основе выборки из полнотекстовых ресурсов транснациональной IT-компании IHS Markit, дополненной текстами из

доступных интернет-источников.

На основе анализа корпуса поисковых запросов определены их основные типы: одно или несколько несогласованных ключевых слов; грамматически согласованные словосочетания; вопросительные предложения; утвердительные предложения; комбинация двух последних типов, нередко без знака препинания между ними при наличии / отсутствии вопросительного знака в конце поискового запроса.

Каждому из выделенных типов запросов была сопоставлена определенная синтаксическая структура русского языка и выявлены базовые составляющие таких структур: именные группы (простые именные группы, расширенные именные группы с предложно-падежными зависимыми конструкциями), глагольные группы (инфинитив с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом; форма 3 лица единственного или множественного числа глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом), грамматическая основа предложения (предикативный центр).

Выделение типов поисковых запросов и базовых синтаксических структур позволило предложить метод решения целевой задачи, основанный на автоматическом распознавании этих структур в поисковом массиве для разработки базы подсказок как компонента лингвистического обеспечения, необходимого для реализации предиктивного автодополнения поисковых запросов. Предложенная и реализованная Ю. Д. Голяк постадийная процедура построения базы подсказок показала свою эффективность в качестве инструмента управления наполнением этой базы при принятии решений относительно включения в ее состав тех или иных типов подсказок в зависимости от предпочтений пользователя или особенностей предметной области. В результате Ю. Д. Голяк разработан прототип системы предиктивного автодополнения русскоязычных пользовательских запросов, внедрение которого в промышленную эксплуатацию показало, что он, в силу использования разработанных концептуальных и алгоритмических решений, а также лингвистических ресурсов, обладает следующими особенностями: предоставляет пользователю возможность более точно формулировать свою информационную потребность; минимизирует общее время решения поисковой задачи; способен решать поисковые задачи, не ориентируясь на историю информационного поиска, что в значительной степени повышает вероятность гарантированно релевантной реакции поисковой системы.

Закономерно, таким образом, что реализация предложенной Ю. Д. Голяк структурно-функциональной модели системы предиктивного автодополнения русскоязычных поисковых запросов доказала ее эффективность с точки зрения

традиционно используемых для оценки качества работы информационно-поисковых систем показателей полноты и точности поисковых выдач.

Текст диссертации должным образом структурирован, написан грамотным научным языком. Аргументация автора и выводы сопровождаются примерами, визуализированы посредством 17 рисунков. Оформление диссертации и автореферата Ю. Д. Голяк в целом соответствует требованиям «Инструкции о порядке оформления квалификационной научной работы (диссертации) на соискание ученых степеней кандидата и доктора наук, автореферата и публикаций по теме диссертации» ВАК Республики Беларусь.

По теме диссертации опубликовано 8 работ, из них 4 – статьи в научных рецензируемых журналах, соответствующих п. 19 Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь. 6 работ подготовлены Ю. Д. Голяк единолично и 2 статьи – в соавторстве с научным руководителем.

Все вышеизложенное убедительно свидетельствует о том, что Ю. Д. Голяк обладает научной квалификацией, соответствующей квалификационным требованиям, предъявляемым к соискателю ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика.

Замечания по диссертации

Внимательное ознакомление с текстом диссертации не позволило выявить недостатков, которые бы касались новизны, научной и практической значимости полученных Ю. Д. Голяк результатов. При общей высокой оценке содержания диссертации считаем необходимым остановиться на некоторых содержательных и технических моментах, требующих внимания и дополнительных пояснений.

1. Представляется слишком лаконичным изложение положений, выносимых на защиту, которые носят исключительно констатирующий характер.

2. Вызывает вопросы выбор использованных источников. Так, в списке из 107 позиций лишь 5 (п.п. 34, 37, 38, 55, 85) – это издания, вышедшие за последние 5 лет. В списке представлены 2 издания одного и того же энциклопедического лингвистического словаря (п.п. 88 и 89) и Словарь-справочник лингвистических терминов Д. Э. Розенталя и М. В. Теленковой (п. 91); учебники и учебные пособия (п.п. 63, 64, 90, 93), причем в п. 63, по сути, объединены 2 источника, и оба библиографических описания оформлены не по правилам; в список включены источники, положенные в основу формирования материала исследования (п.п. 58–62), в том числе Научная электронная библиотека «КиберЛенинка». В список включен ряд патентов (п.п. 14, 31, 32), при этом в таблицу 1.1 (стр. 22–26 диссертации), озаглавленную как «Обзор патентов по теме решаемой задачи», эти патенты не включены. Наблюдаются единичные случаи несовпадения нумерации источников в списке и в ссылках в тексте диссертации. На стр. 5 автореферата и стр. 9 диссертации отмечено, что «научный руководитель принимал участие в

выборе направления исследования, постановке задач, обсуждении теоретических и практических результатов, полученных автором», но нигде не отмечена степень участия научного руководителя в подготовке двух опубликованных в соавторстве с ним статей в научном журнале «Вестник МГЛУ. Сер. 1. Филология».

3. В качестве материала исследования, помимо русскоязычных ресурсов, определены англоязычные. Не понятна необходимость их применения для решения целевой задачи, ориентированной на русский язык. Можно было бы ожидать, что их привлечение необходимо для обеспечения выдач англоязычных данных в ответ на русскоязычные запросы, но это предположение не нашло подтверждения в тексте работы. Типологические отличия между английским и русским языками, особенно значимые на уровне синтаксиса, не позволяют также предположить необходимость привлечения англоязычных запросов для выявления базовых синтаксических структур таких запросов на русском языке.

4. Не эксплицированы критерии, положенные в основу формирования диссертантом как корпуса текстов поисковых запросов, так и тестового поискового массива, именуемого в автореферате и в диссертации «корпусом русскоязычных текстовых документов»; не указан формат их представления в базовом лингвистическом процессоре, а также не ясно, каким именно образом использовались эти ресурсы для построения базы подсказок. Лишь однажды в диссертации (стр. 54) упоминается использование инструмента биграмм и триграмм при анализе пользовательских запросов из соответствующего корпуса. Остается неясным и объем «корпуса русскоязычных текстовых документов», поскольку на стр. 3 автореферата и стр. 7 диссертации указан объем 7,5 миллионов предложений, а на стр. 9 автореферата и на стр. 51 диссертации – 75 миллионов предложений. К тому же, если указание на количество текстов / документов в корпусе является общепринятой практикой в корпусной лингвистике, то измерение объема корпуса в предложениях вызывает вопросы, особенно для русского языка, в котором, в силу свободного порядка слов и развитой системы флексий, может наблюдаться существенная вариативность предложений по количеству использованных в них словоупотреблений (токенов).

5. База подсказок строится на основе предложенного «корпуса русскоязычных текстовых документов». В связи с этим хотелось бы знать, во-первых, что происходит с запросами, для которых в базе не будут найдены подсказки, и, во-вторых, предполагается ли автоматическое пополнение базы подсказок при пополнении «корпуса русскоязычных текстовых документов».

6. В тексте диссертации слабо представлена связь с приложениями. Остается неясным, какова роль приложений А, Б, Г: они представляются абсолютно излишними, тем более, что, во-первых, они не являются разработкой диссертанта, а, во-вторых, их фрагменты даются и в самом тексте диссертации, в том числе в форме таблиц.

7. Не ясно, для чего цель и задачи исследования, сформулированные в общей характеристике работы, полностью повторяются в выводах по первой главе (стр. 30 и 31 диссертации).

8. На стр. 11 диссертант пишет о двух основных типах информационно-поисковых систем – документальных и фактографических и грамотно дифференцирует выдачи, которые могут обеспечивать эти системы. В то же время в диссертации явно не указано, на какой(ие) именно тип(ы) информационно-поисковых систем ориентировано разработанное им лингвистическое обеспечение.

9. На стр. 6 в разделе общей характеристики **Связь работы с крупными научными программами и темами** среди названных программ и тем указано, что «диссертационное исследование выполнялось на кафедре прикладной лингвистики БГУ в рамках госбюджетной НИР «Комплексное научно-методическое обеспечение преподавания русского языка как иностранного в контексте межкультурной коммуникации» в соответствии с пунктом 11 «Общество и экономика» приоритетных научных исследований БГУ на 2016–2020 годы», но нигде в тексте диссертации не отмечена абсолютно, на наш взгляд, очевидная применимость полученных результатов в практике обучения русскому языку как иностранному.

10. Вполне обоснованно при разработке лингвистического обеспечения, ориентированного на русский язык, много внимания уделено диссертантом предлогам и предложным конструкциям. В этой связи возникает вопрос, почему не были использованы такие фундаментальные труды, как: Всеволодова, М. В. Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 1: Введение в объективную грамматику и лексикографию русских предложных единиц. Изд. 2-е / М. В. Всеволодова, О. В. Кукушкина, А. А. Поликарпов. – М. : ЛЕНАНД, 2018. – 304 с.; Всеволодова, М. В. Русские предлоги и средства предложного типа. Материалы к функционально-грамматическому описанию реального употребления. Книга 2: Реестр русских предложных единиц: А - В (объективная грамматика) / М. В. Всеволодова, Е. Н. Виноградова, Т. Е. Чаплыгина. – М. : УРСС, 2018. – 798 с.

В целом в весьма грамотно оформленных текстах автореферата и диссертации можно отметить лишь некоторые недочеты: обилие сокращений, затрудняющих восприятие содержания; пропуск буквы на стр. 30 диссертации; излишнее использование деепричастных оборотов (стр. 6 автореферата и стр. 10 диссертации); не совсем удачный авторский термин *языконезависимый метод*, встречающийся на стр. 28 диссертации; явное влияние английского языка на используемую диссертантом терминологию из сферы информационного поиска (*система информационного поиска – СИП* вместо *информационно-поисковая система – ИПС*; *пользовательский запрос* вместо *поисковый запрос*, *поисковое пространство* вместо *поисковый массив*).

Указанные вопросы, замечания и рекомендации касаются лишь отдельных аспектов выполненного Ю. Д. Голяк диссертационного исследования и не влияют на его квалификационную ценность и общую высокую оценку полученных ею результатов.

Заключение

Диссертация Голяк Юлии Дмитриевны «Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)», подготовленная под научным руководством доктора технических наук, профессора И. В. Совпеля, является самостоятельно выполненной квалификационной научной работой и соответствует специальности 10.02.21 – прикладная и математическая лингвистика и отвечает требованиям ВАК Республики Беларусь, предъявляемым к кандидатским диссертациям. Диссертация содержит новые научные результаты в области актуального в настоящее время направления прикладной лингвистики – лингвистическое обеспечение автоматизированных информационных систем. В соответствии с п. 19 и п. 20 «Положения о присуждении ученых степеней и присвоении ученых званий в Республике Беларусь» Голяк Юлия Дмитриевна заслуживает присуждения ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика за:

1) разработку оригинальной концепции системы предиктивного автодополнения информационных запросов пользователей, которая, в отличие от существующих систем автодополнения не только ориентирована на завершение уже набранной пользователем части поискового запроса, но и способна осуществлять дополнение в начале, в конце и внутри текста запроса на русском языке;

2) исчисление основных типов поисковых запросов: одно или несколько несогласованных ключевых слов; грамматически согласованные словосочетания; вопросительные предложения; утвердительные предложения; комбинация двух последних типов с учетом вариативности наличия / отсутствия знаков препинания между ними и вопросительного знака в конце запроса;

3) выявление базовых синтаксических структур для основных типов информационных запросов, включающих именные группы (простые именные группы, расширенные именные группы с предложно-падежными зависимыми конструкциями), глагольные группы (инфинитив глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом; форма 3 лица единственного или множественного числа глагола с прямым дополнением, с косвенным дополнением с предлогом, с прямым и косвенным дополнениями, с прямым дополнением и причастным оборотом), грамматическая основа предложения (предикативный центр);

4) разработку в виде расширения функциональности базового лингвистического процессора собственного лингвистического обеспечения и алгоритмов автоматического распознавания в полнотекстовом поисковом массиве подсказок, соответствующих базовым синтаксическим структурам основных типов поисковых запросов, с целью их автодополнения;

5) создание прототипа оригинальной системы предиктивного автодополнения русскоязычных пользовательских запросов и его внедрение в промышленную эксплуатацию.

Ю. Д. Голяк выступила с докладом на научном семинаре, за которым последовала дискуссия. Соискатель ответила на все заданные вопросы.

Отзыв о диссертации Ю. Д. Голяк «Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)», представленной на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика, согласно приказу проректора по научной работе ГрГУ им. Янки Купалы от 16.02.2023 № 183, рассмотрен и утвержден на научном семинаре «Актуальные проблемы романо-германского и славянского языкознания» филологического факультета ГрГУ им. Янки Купалы 03.03.2023, протокол заседания №1.

В работе семинара приняли участие 15 человек из 18: кандидаты филологических наук (И. И. Бубнович, В. Л. Воронович, С. А. Горская, О. И. Ковальчук, Л. Е. Ковалёва, С. С. Масленникова, И. Д. Матько, Л. В. Рычкова, Е. С. Садовская, Т. И. Скоробогатая, З. З. Сидорович, Ж. С. Сипливеня, Е. О. Шейко, С. А. Янковская, Е. Н. Ясюкевич).

Результаты открытого голосования: «за» – 15, «против» – нет, «воздержались» – нет.

Руководитель научного семинара:
кандидат филологических наук, доцент,
доцент кафедры английской филологии
ГрГУ им. Янки Купалы Л.Е. Ковалева

Эксперт оппонирующей организации:
кандидат филологических наук, доцент,
профессор кафедры перевода
и межкультурной коммуникации
ГрГУ им. Янки Купалы Л.В. Рычкова

Секретарь научного семинара:
кандидат филологических наук, доцент,
доцент кафедры русской филологии
ГрГУ им. Янки Купалы С.А. Янковская

ПОДПИСЬ УДОСТОВЕРЯЮ
Начальник отдела управления персоналом
учреждения образования «Гродненский
государственный университет имени Янки Купалы»
персонала
С.В. Нестерович
03 03 2023 г.