ОТЗЫВ

официального оппонента
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

1) Соответствие диссертации специальности и отрасли науки, по которой она представлена к защите

Диссертационное исследование Ю. Д. Голяк посвящено разработке метода, алгоритма и лингвистического обеспечения предиктивного автодополнения русскоязычных запросов пользователей в системах информационного поиска и их реализации в виде промышленного прототипа, что предполагает решение задачи автодополнения пользовательских запросов на основе разработки, апробации и внедрения авторского лингвистического обеспечения и алгоритмов автоматического распознавания в полнотекстовых базах данных и базе данных подсказок, соответствующих базовым синтаксическим структурам основных типов пользовательских запросов.

Прикладное лингвистическое исследование включало создание и детальный анализ большого корпуса пользовательских запросов на русском и английском языках, а также корпуса русскоязычных текстовых документов общим объемом более 180 миллионов словоупотреблений, около 7,5 миллионов предложений и более 0,4 гигабайт текста в заархивированном виде, что позволило предложить детальную классификацию основных типов пользовательских запросов. Особым источником информации послужил специальный корпус для формирования базы подсказок как основы решения задачи автодополнения пользовательских запросов, а также для разработки и тестирования алгоритмов автоматического распознавания подсказок. Полученный на этом этапе результат определил возможность разработки принципиально новой схемы решения задачи автодополнения пользовательских запросов, ее реализацию и апробацию в рамках практической системы инженерии знаний.

В результате проведенного исследования разработан прототип оригинальной системы предиктивного, т.е. прогнозирующего автодополнения русскоязычных пользовательских запросов в системах информационного поиска, позволяющей предоставлять пользователю возможность более точно формулировать собственные информационные потребности, существенно минимизировать общее время решения поисковой задачи, решать задачу информационного поиска, не ориентируясь на его историю в конкретной системе, гарантированно получать релевантную реакцию поисковой системы. Тем самым в работе выполнен полный цикл построения воспроизводящей лингвистической модели, что позволяет сделать вывод о том, что работа Ю. Д. Голяк соответствует отрасли науки

филология и специальности 10.02.21 – прикладная и математическая лингвистика.

2) Актуальность темы диссертации

Многолетний опыт использования компьютерных технологий для решения различных типов задач показал, что получаемый результат во многом зависит от того, насколько корректно произведена автоматическая переработка текста на естественном языке (ЕЯ). Поскольку ЕЯ является основным средством формирования, хранения и передачи информации, то такая переработка текста может осуществляться на основе применения информационных технологий для создания систем поиска, генерации и поддержки многоязычной информации, для локализации данных и программного обеспечения. Эти задачи могут быть решены путем создания и внедрения практических систем информационного поиска, автоматического (машинного) перевода, компьютерных словарных и обучающих систем и т. д. Особую актуальность сегодня приобретает разработка не просто систем информационного поиска, а их особого класса, так называемых вопросно-ответных систем, то есть систем экспертных, в задачу которых входит не просто поиск текстов или фактов, но преобразование информации в соответствии с запросом пользователя.

Оценивая **актуальность** исследования, рассмотрим два его аспекта: актуальность области исследования в целом и актуальность конкретной проблемы. В этом плане необходимо констатировать, что

- теоретическое осмысление и практическое определение алгоритма анализа такой сложной лингвистической задачи как автоматическое дополнение пользовательских запросов без привлечения истории запросов к конкретной системе,
- разработка методов формирования поискового пространства на основе предварительного моделирования поиска в виде множества так называемых подсказок, автоматически распознаваемых в самой поисковой базе данных,
- установление особенностей, методов и принципов организации лингвистического обеспечения для такой сложной информационной системы, какой является любая система автоматической переработки текста,
- встраивание системы автоматического дополнения запроса в архитектуру промышленной системы

безусловно, **актуально** как с точки зрения прикладной и математической лингвистики, так и с позиций развития лингвистических информационных технологий.

В соответствии с поставленной целью в диссертации разработаны методы анализа, проблемно-ориентированное лингвистическое обеспечение и структурно-функциональная схема системы автоматического предиктивного дополнения запросов, которая в целом характеризует разработанную концепцию системы. При этом сам термин *концепция* понимается в работе как конструктивный принцип организации конкретного вида деятельности.

3) Степень новизны результатов, полученных в диссертации, и научных положений, выносимых на защиту

Новизна исследования, проведенного Ю. Д. Голяк, подтверждается

(1) обоснованием общей концепции решения задачи автоматического дополнения поискового запроса за счет прогнозирования возможных дополнений на основе полнотекстовых баз данных методом погружения уже набранной части пользовательского запроса в контекст, что предполагает возможность дополнения формулировки запроса в начале, в конце и внутри конструкции;

(2) разработкой классификации основных типов пользовательских запросов и их базовых синтаксических структур, что позволило установить 70 лингвистических правил (паттернов), применение которых основано на результатах базового лингвистического анализа текстовых документов в виде последовательности лексико-грамматических тегов. Такие правила описывают возможный контекст, в котором выделенные элементы могут быть удалены, а оставшиеся являются базой для возможного дополнения запроса.

(3) разработкой структурно-функциональной схемы системы предиктивного автодополнения русскоязычных пользовательских запросов в системе информационного поиска и ее базового лингвистического процессора. Исследование реализаций синтаксических структур и паттернов в репрезентативном корпусе текстов позволило определить метод и алгоритм выявления лингвистических принципов свертки структур подсказок, сформировать совокупность эмпирически определенных параметров для установления их преференций;

(4) разработкой собственного лингвистического обеспечения как дополнения базового и алгоритмов автоматического распознавания в базе подсказок, соответствующих синтаксическим структурам основных типов пользовательских запросов для их автодополнения;

(5) разработкой и реализацией прототипа оригинальной системы предиктивного автодополнения русскоязычных пользовательских запросов в системе информационного поиска, предоставляющей пользователю возможности уточнять свои информационные потребности, существенно минимизировать общее время решения поисковой задачи, решать задачу автодополнения, не опираясь на историю информационного поиска, получать гарантированно релевантную реакцию поисковой системы. Тестирование и внедрение системы в промышленную эксплуатацию подтверждают новизну и достоверность принятых решений.

На защиту автором выносится 6 положений, отражающих концептуальные и инженерно-лингвистические положения и выводы исследования, соответствующие поставленной цели и решаемым задачам, эти положения также характеризуются существенной новизной.

Основной целью работы является разработка методов, алгоритмов и лингвистического обеспечения, необходимых для решения задачи предиктивного автодополнения русскоязычных запросов пользователей в системах информационного поиска и их реализации в виде промышленного

прототипа. Эта цель в диссертационном исследовании Ю. Д. Голяк достигается за счет

- последовательного рассмотрения типов пользовательских запросов, выявления их синтаксических и семантических моделей, определяющих возможность и целесообразность автоматического дополнения запроса на основе информации, извлекаемой их полнотекстовой базы данных;
- создания концепции промышленной системы автоматического дополнения запросов, основанной на комплексном учете синтаксических, семантических, контекстуальных и логических ограничений и преференций;
- разработки структурно-функциональной схемы системы автоматического дополнения запроса, включающей базовый лингвистический процессор и его расширение до лингвистического процессора задачи, а также функциональность лингвистического анализа корпоративной полнотекстовой базы данных и автоматического построения базы подсказок;
- создания и апробации модуля прогнозируемого автодополнения запроса как прототипа для промышленной системы информационного поиска.

В результате проведенного исследования установлена структура и состав лингвистического обеспечения подсистемы (модуля) автоматического автодополнения запросов на русском языке, сам модуль реализован и апробирован.

Следует отметить особые сложности решения поставленной задачи, связанные с условием внедрения разработанного прототипа в состав промышленного многоязычного лингвистического процессора системы инженерии и управления знаниями IHS Goldfire, а также с условием универсальности лингвистического обеспечения прототипа.

4) Обоснованность и достоверность выводов и рекомендаций, сформулированных в диссертации

Проведенный в диссертации анализ материала на основе четко определенных параметров позволил автору

(1) установить естественно-языковой интерфейс пользователя как необходимую функциональность любой современной системы информационного поиска, важную роль в которой играет автодополнение (автоматическое завершение, предиктивный ввод) пользовательского запроса,

(2) разработать корпус пользовательских запросов, обеспечивающий установление лексических, синтаксических и семантических характеристик наиболее частотных поисковых запросов на русском и английском языках общим объемом 13000 запросов,

(3) создать специальный корпус текстовых документов общим объемом более 180 миллионов словоупотреблений, включающий около 7,5 миллионов предложений и более 0,4 гигабайт текста в заархивированном виде. Этот корпус использовался для классификации основных типов документов, а также лексических и синтаксических структур пользовательских запросов,

для разработки алгоритма, его тестирования и оценки показателей полноты и точности его работы,

(4) установить особенности формулировки запросов в корпусе запросов и разработать классификацию типов запросов и их синтаксических структур,

(5) разработать общую процедуру определения структуры подсказок на основе корпуса текстовых документов и разработать методы их представления, а также

(6) разработать, реализовать и верифицировать в рамках промышленной системы модуль автоматического дополнения запроса.

Достоверность результатов исследования обеспечивается репрезентативностью лингвистического материала, а также корректностью использованных методов анализа и верификацией реализованных алгоритмов и лингвистического обеспечения.

5) Научная, практическая, экономическая и социальная значимость результатов диссертации с указанием рекомендаций по их использованию

Работа Ю. Д. Голяк является серьезным интеллектуальным продуктом, значимость которого может быть оценена как в теоретическом, так и в практическом аспектах. **Научная значимость** работы связана с разработкой общей концепции системы автоматического предиктивного дополнения запроса, основанной на вопросно-ответной системе информационного поиска с естественно-языковым интерфейсом пользователя и русскоязычной базы текстов, с реализации процедуры предполагаемого дополнения на основе поиска в базе подсказок, заранее автоматически распознаваемых текстах, глубокий анализ текстовых документов, созданием специального лингвистического обеспечения системы.

Практическая ценность исследования заключается в том, что автором проведен комплексный анализ репрезентативного корпуса текстов запросов и текстовых документов, создана и апробирована система автоматического дополнения запросов. Результаты проведенного исследования могут найти широкое практическое применение в системах автоматической обработки текстов, при проектировании и реализации различных систем информационного поиска. Теоретическая и практическая значимость исследования также позволяют говорить **о целесообразности** включения полученных результатов в теоретические и практические курсы по прикладной лингвистике, лингвистике текста, лингвистическим автоматам.

6) Опубликованность результатов диссертации в научной печати

Широкое обсуждение процедуры и результатов исследования является важной частью научной дискуссии и верификации предложенных в исследовании процедур и методов. Поэтому особенно важен тот факт, что положения, выносимые на защиту, не только отличаются логичностью и продуманностью, но и прошли серьезную апробацию на научных конференциях, и нашли отражение в представительном списке печатных публикаций, общим числом 8, из которых 5 опубликованы в научных

журналах, включенных в Перечень научных изданий Республики Беларусь, утвержденный Высшей аттестационной комиссией, 3 – в сборниках материалов научных конференций. Общий объем опубликованных материалов составляет 4,21 авторского листа. Одна публикация посвящена проблеме выбора концепции автоматического предиктивного дополнения запросов (положение 1); в одной публикации рассмотрена классификация основных типов пользовательских запросов (положение 2); предлагаемый метод решения задачи предиктивного автодополнения (положение 3) рассмотрен в двух публикациях; структурно-функциональная схема системы, разработанная в соответствии с предложенной концепцией и методом решения целевой задачи (положение 4) рассмотрены в двух публикациях; в трех публикациях рассматривается дополнение базового лингвистического обеспечения задачи и алгоритмы автоматического распознавания подсказок (положение 5); сама система автодополнения русскоязычных пользовательских запросов отражена в одной публикации (положение 6).

Автореферат должным образом резюмирует итоги проведенного исследования.

7) Соответствие оформления диссертации требованиям ВАК

Оформление диссертационной работы соответствует стандартам ВАК. Работа оформлена в виде 1 книги, что позволяет максимально полно представить результаты экспериментальной апробации разработанной методики. В структуре диссертации представлены все обязательные составляющие: введение, общая характеристика работы, теоретическая глава и две главы с результатами практического исследования, выводы по главам, общее заключение и библиографический список, включающий 107 источников на русском и английском языках. В композиционном плане все части работы взаимосвязаны и служат реализации общего замысла исследования. Стиль изложения четок, приложения характеризуют особенности изученного материала запросов и текстовых документов.

8) Недостатки диссертации

Рецензируемая диссертационная работа характеризуется глубоким проникновением в исследуемую проблематику, широкой научной и исследовательской базой, сбалансированностью комплексного отображения множества лингвистических явлений, а также обоснованностью выводов, охватывающих все основные результаты работы. Вместе с тем в соответствии с жанром отзыва следует сделать несколько замечаний.

1. В исследовании подробно рассматриваются различные теории и классификации систем и процедур информационного поиска, что можно только приветствовать, однако следует учитывать, что терминологическая четкость является одним из условий корректного сопоставления теоретических подходов и систем, а также собственных решений. К сожалению, в некоторых случаях такой четкости в работе не хватает. Так, например, одним из важных терминов и рабочих инструментов работы является термин *подсказка*, описание систем информационного поиска и подходов к дополнению запросов осуществляется при использовании этого

инструмента как основного, но только на с. 27 появляется текст, который с некоторой натяжкой можно считать рабочим определением: множества P так называемых подсказок, автоматически распознаваемых в самой ПБД, как заранее «известного» поисковой системе поискового пространства. Собственного рабочего определения подсказки в работе нет.

2. Продолжением этого замечания является констатация того факта, что в наборе приложений, прекрасно характеризующих разные типы текстовых документов, наполнение лингвистического обеспечения, систему лексико-грамматических кодов и т.п., не нашлось места для фрагмента базы подсказок, что позволило бы более ясно представить себе процедуру их выявления и использования. В то же время утверждается, что автодополнение пользовательского запроса осуществляется на этапе взаимодействия пользователя с Базой подсказок, которая создается предварительно и автоматически (с.52). В работе явно не хватает лингвистической структуры подсказки, поскольку именно на результате ее применения пользователь принимает решение либо о ее использовании, либо о продолжении ввода запроса вручную.
3. Остался без толкования и термин *префикс*, есть только информация о рассмотрении в качестве префикса строки вводимого запроса (с.5, с. 26, 30), на с. 83 появляется вопросительный префикс, см. также Таблицу 2.4, в которой префиксом оказываются начала слов запросов или начальные слова запросов, эта информация явно недостаточна, как и в случае с термином *подсказка* требуется рабочее определение.
4. В работе используется специально созданный автором корпус текстовых документов, который недоопределен именно как корпус: не определены принципы его балансировки (есть только информация о типах документов в корпусе на с. 36), нет информации о разметке, даже о типах метаразметки, связанной с типологией исследуемых текстовых документов, если частеречная разметка не применялась.
5. Текст диссертации не свободен от прямых повторов, далеко не всегда оправданных, так, например, цель и задачи диссертации дословно рассматриваются трижды на с. 6, 28, 30-31.
6. Список использованной литературы, а также анализ практических систем автоматического анализа текстов, безусловно, обеднены. В нем нет лингвистических работ на русском языке, опубликованных после 2002 года, в результате автору приходится самому составлять списки сложных предлогов и союзов, а также стоп-слов, давно описанных для русского языка.

9) Соответствие научной квалификации соискателя ученой степени, на которую он претендует

Междисциплинарность проведенного исследования подтверждается сочетанием анализа в области автоматизированного лингвистического

анализа текста, а также разработкой и реализацией процедуры объективизации такого анализа. Информационные технологии в области анализа естественного языка и текстовых документов (лингвистические технологии), реализующие алгоритмы автоматической переработки текста, являются необходимым условием решения многих задач, относящихся к информационным технологиям в целом. Именно в русле этого актуального подхода и выполнено исследование Ю. Д. Голяк. Автором проведено действительно сложное исследование, позволившее получить новые результаты при исследовании сложных проблем

Соответственно, можно утверждать, что диссертационное исследование Юлии Дмитриевны Голяк "Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)", представленное на соискание ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика, соответствует всем требованиям, предъявляемым к кандидатским диссертациям, а его автор Ю. Д. Голяк заслуживает присуждения ей ученой степени кандидата филологических наук.

10) Заключение

Диссертация Ю. Д. Голяк "Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)", является квалификационной научной работой, содержание которой соответствует специальности 10.02.21 – прикладная и математическая лингвистика и отрасли науки (филология).

В ходе проведенного исследования, подтвердившего зрелость научного мышления автора и аргументированность избранной позиции, были доказательно решены все поставленные задачи и сформулированы отличающиеся новизной выводы, намечающие дальнейшую проекцию исследований при создании лингвистических процессоров для всего многообразия систем автоматической обработки текстовых документов.

В соответствии с п. 19 и 20 Положения о присуждении ученых степеней и присвоении ученых званий автор диссертации Ю. Д. Голяк заслуживает присуждения ей искомой ученой степени кандидата филологических наук по специальности 10.02.21 – прикладная и математическая лингвистика и отрасли науки (филология) за:

– разработку концепции системы предиктивного автодополнения запросов пользователей, основанной на лексических, синтаксических, семантических характеристиках текстовых документов и запросов;

– разработку и верификацию структурно-функциональной схемы системы предиктивного автодополнения запросов пользователей;

– создание лингвистического обеспечения системы предиктивного автодополнения запросов пользователей;

– разработку эффективной архитектуры открытой системы предиктивного автодополнения запросов пользователей, базирующейся на разделении программного кода и лингвистических ресурсов;

– создание и верификацию прототипа промышленной системы предиктивного автодополнения запросов пользователей к базе текстовых документов на русском языке.

Выражаю свое согласие на размещение моего отзыва о диссертации Голяк Ю. Д. «Предиктивное автодополнение запросов пользователей в системах информационного поиска (на материале русского языка)» на сайте учреждения образования «Минский государственный лингвистический университет».

Официальный оппонент
профессор кафедры образовательных технологий в филологии
филологического факультета РГПУ им. А. И. Герцена,
доктор филологических наук,
профессор
заслуженный деятель науки РФ Л.Н. Беляева
13 февраля 2023 г.

РГПУ им. А.И. ГЕРЦЕНА
подпись Беляевой Марины Николаевны
удостоверяю «17» февраля 2023 г.
Отдел кадров управления по работе с кадрами
и организационно-контрольному обеспечению

ВЕДУЩИЙ СПЕЦИАЛИСТ ПО КАДРАМ ЗЕВЛОВА Н. М.