



УТВЕРЖДАЮ

проректор по науке и цифровому развитию
МГТУ им. Н.Э. Баумана, д.э.н., профессор

П.А. Дроговоз

«22» 01.10.25

ЗАКЛЮЧЕНИЕ

федерального государственного бюджетного образовательного учреждения
высшего образования «Московский государственный технический университет
имени Н.Э. Баумана (национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

Диссертация на тему «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» выполнена в МГТУ им. Н.Э. Баумана на кафедре теоретической информатики и компьютерных технологий.

В период подготовки диссертации соискатель Бутенко Юлия Ивановна работала в МГТУ им. Н.Э. Баумана в должности доцента кафедры теоретической информатики и компьютерных технологий.

Бутенко Юлия Ивановна, 01.03.1987 года рождения, гражданка РФ, в 2009 году окончила Национальный аэрокосмический университет им. Н.Е. Жуковского «Харьковский авиационный институт» по специальности «Прикладная лингвистика», диплом ХА № 37113061 от 30 июня 2009 г.

Ученая степень кандидата технических наук присуждена решением диссертационного совета при Национальном политехническом университете «Харьковский политехнический институт» 12 ноября 2015 года. Решением высшей аттестационной комиссии Министерства образования и науки Украины от 15 декабря 2015 года выдан диплом ДК № 033224.

По итогам обсуждения принято следующее заключение:

Тема диссертации окончательно утверждена на заседании Ученого совета НУК Информатика и системы управления МГТУ им. Н.Э. Баумана от 28 октября 2024, протокол № 7.

Актуальность выполненной соискателем работы. Диссертационная работа Бутенко Юлии Ивановны посвящена исследованию и разработке основ обработки композиционных, терминологических и стилистических особенностей научно-технических текстов как наиболее рутинных процедур при создании параллельного корпуса.

Актуальность теоретических и практических результатов диссертационной работы подтверждается востребованностью в отечественной науке моделей, методов и программных средств обработки русскоязычных научно-технических текстов при создании параллельных корпусов, которые представляют собой большие данные.

Анализ современных параллельных корпусов показал, что они создаются лингвистами вручную, что требует значительных временных затрат на выполнение рутинных процедур разметки и выравнивания параллельных текстов. Вместе с тем широкий спектр применения параллельных корпусов при решении ряда теоретических и практических задач свидетельствует о необходимости создания таких информационных ресурсов. Эта отрасль является недостаточно формализованной, слабо автоматизированной, существующие методы работы не универсальны, а операции по разметке научно-технических текстов выполняют лингвисты собственноручно отдельно для каждого вида разметки. Следовательно, с целью автоматизировать и ускорить процедуру автоматической обработки научно-технических текстов при создании параллельных корпусов необходима инструментальная поддержка процедуры обработки параллельных научно-технических текстов.

Личное участие соискателя ученой степени в получении результатов, изложенных в диссертации. В диссертацию включены результаты, полученные лично автором.

Научная новизна. Научной новизной проведенного исследования являются теоретические основы построения моделей и создания методов обработки англо- и русскоязычных научно-технических текстов, направленные на проектирование параллельного корпуса, что имеет важное хозяйственное значение в области информатики, а именно:

1. Усовершенствованы модели иерархически-структурированных научно-технических текстов, за счет добавления межуровневых элементов и оценки значимости каждого структурного элемента при создании параллельного корпуса, что позволяет более эффективно обрабатывать научно-технические тексты на разных уровнях языковой системы.

2. Получили дальнейшее развитие модели и методы разметки и выравнивания англо- и русскоязычных терминологических единиц из научно-технических текстов, отличающиеся от существующих возможностью извлечения терминов с правыми определениями, что позволяет использовать эти модели и методы при обработке текстов при создании параллельного корпуса.

3. Впервые разработаны модели и метод разметки номенклатурных наименований в научно-технических текстах на русском и английском языках, что позволяет повысить эффективность разметки научно-технических текстов за счет учета лексических единиц, в состав которых входят произвольные буквенно-числовые последовательности в том числе символы разных алфавитов.

4. Впервые предложены методы выявления машинно-сгенерированных и машинно-переведенных текстов на основе семантико-синтаксических особенностей русского языка.

5. Разработан прототип системы управления корпусными данными, который в отличие от существующих корпусных менеджеров позволяет управлять корпусными данными на разных этапах их обработки, а также формировать различные наборы данных для машинного обучения.

Практическая значимость работы. Практическая ценность работы заключается в создании программного средства, позволяющего использовать разработанные теоретические основы, модели и методы обработки англо- и

русскоязычных научно-технических текстов для создания параллельного корпуса. Данное программное средство позволяет автоматизировать рутинный процесс обработки англо-и русскоязычных параллельных текстов и увеличить объемы филологически компетентных баз данных размеченных научно-технических текстов.

Степень достоверности результатов. Достоверность научных результатов работы подтверждается непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, высокой степенью сходимости теоретических результатов с данными экспериментов и определяется применением теоретических и методологических основ разработок ведущих ученых в области обработки естественного языка, корректным и обоснованным использованием математического аппарата, экспериментальными исследованиями разработанных моделей и методов

Апробация результатов диссертации. Основные результаты работы докладывались и обсуждались на X Международной научно-практической конференции студентов, аспирантов и молодых ученых «Информационные технологии в науке, бизнесе и образовании» (Москва, 2018), Всероссийской научной конференции «Нейрокомпьютеры и их применение» (Москва, 2018, 2019, 2020, 2022), II Всероссийской национальной научной конференции студентов, аспирантов и молодых ученых «Молодежь и наука: актуальные проблемы фундаментальных и прикладных исследований» (Комсомольск-на-Амуре, 2019, 2020), Международной научно-практической конференции «Современное технологическое образование» (Москва, 2019), II Всероссийской научно-практической конференции «Системы управления полным жизненным циклом высокотехнологичной продукции в машиностроении: новые источники роста» (Москва, 2019), Международном форуме «Цифровые технологии в инженерном образовании: новые тренды и опыт внедрения» (Москва, 2019), XII Всероссийской конференции молодых ученых и специалистов (с международным участием) «Будущее машиностроения России» (Москва, 2019), Международной научной конференции «Фундаментальные и прикладные задачи механики»,

посвященная 100-летию со дня рождения Академика Константина Сергеевича Колесникова (Москва, 2019), Международном молодежном научном форуме «ЛОМОНОСОВ-2020» (Москва, 2020), Академических чтениях по космонавтике «Королевские чтения» (Москва, 2019, 2021, 2022), Международной конференции «Моделирование в инженерном деле» (Москва, 2019, 2020, 2022), II Всероссийской молодёжной научно-практической конференции с международным участием «LinguaNet» (Севастополь, 2020), Межвузовской заочной конференции аспирантов, соискателей и молодых ученых «Наука, технологии и бизнес» (Москва, 2020, 2022, 2024), VI Международном форуме «Instrumentation Engineering, Electronics and Telecommunications – 2020» (Ижевск, 2020), II Международной научно-практической конференции «Лингвистические и культурологические аспекты современного инженерного образования» (Томск, 2021), International Conference Aviation Engineering and Transportation (AviaEnT) (Иркутск, 2020), XXI Международной научно-технической конференции «Развитие информатизации и государственной системы научно-технической информации» (РИНТИ-2022) (Минск, 2022), Международной ИТ-конференции «Ключевые тренды развития искусственного интеллекта: наука и технологии» (Москва, 2023).

Публикации. По теме диссертации опубликовано 60 научных работ, из которых 27 статей в научно-технических журналах, входящих в перечень ВАК РФ, 19 – в изданиях, входящих в международные наукометрические базы Scopus и Web of Science. В трудах российских и международных конференций опубликовано 29 работ. Основные публикации по теме диссертации:

1. Бутенко Ю.И. Метод извлечения многокомпонентных терминологических единиц с правыми определениями из научно-технических текстов // Вестник НГУ. Серия: Информационные технологии. 2024. №3. С. 4-10. (К-1) (0,7 п. л.)

2. Бутенко Ю.И. Извлечение номенклатурных наименований из англо- и русскоязычных научно-технических текстов // Искусственный интеллект и принятие решений. 2024. №3. С. 95-103. (RSCI) (0,9 п. л.)

3. Бутенко Ю.И. Метод выравнивания многокомпонентных терминологических единиц в параллельном корпусе научно-технических текстов // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2024. №8. С. 29-38. (RSCI) (0,75 п. л.)

Бутенко Ю.И., Сидняев Н.И., Синева Е.Е. Стратегии поиска в пространстве состояний // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2024. №6. С. 25-39. (WoS) (1,1 / 0,5 п.л.)

Соискателем проработаны принципы информационного поиска в научно-технических текстах на английском и русском языках.

4. Бутенко Ю.И., Солошенко К.А. Лексический тренажер по иностранному языку для студентов технических специальностей МГТУ им. Н.Э. Баумана // Экономика. Информатика. 2024. 51(1). С. 189-200. (К-1) (0,85 / 0,7 п. л.)

Соискателем предложен подход к созданию учебных и специальных словарей на основе параллельного корпуса научно-технических текстов.

5. Бутенко Ю.И. Использование базы данных структурных трансформаций для извлечения многокомпонентных терминологических единиц // Системы и средства информатики. 2023. Т. 33. №1. С.35-44. (RSCI) (0,5 п. .л.)

6. Бутенко Ю.И., Галетка М.Л. Синева Е.Е. Создание системы разметки семантических ролей в научно-технических текстах по авиации и космонавтике // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2022. №10. С. 23-32. (WoS) (1,1 / 0,6 п.л.)

Соискателем предложен общий подход к установлению семантических ролей в научно-технических текстах.

7. Бутенко Ю. И. Модель научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. №3 (20). С. 5-13. (К-1) (0,6 п. л.)

8. Бутенко Ю. И. Строганов Ю. В. Сапожков А. М. Система извлечения многокомпонентных терминов и их переводных эквивалентов из параллельных научно-технических текстов // Научно-техническая информация: Серия 2.

Информационные процессы и системы. 2022. №9. С. 12-21. (RSCI) (0,7 / 0,4 п. л.)

Соискателю лично принадлежит общий подход к извлечению англо- и русскоязычных многокомпонентных терминов на основе синтаксических шаблонов, подкрепленных морфологической информацией о каждой словоформе.

9. Бутенко Ю.И., Тельнова И.Н., Гаража В.В. Метод выявления тенденций развития научных направлений (на материале анализа публикаций по газовому топливу) // Научно-техническая информация: Серия 2. Информационные процессы и системы. 2022. №1. С. 10-24. (WoS) (0,8 / 0,4 п. л.)

Соискателем проработаны общие принципы информационного поиска в коллекции сложно-структурированных научно-технических текстов.

10. Бутенко Ю.И. Использование онтологий для автоматизации формирования нормативного профиля при сертификации программного обеспечения // Искусственный интеллект и принятие решений. 2021. №2. С. 55-65. (Scopus, WoS) (1 п. л.)

11. Бутенко Ю.И. Модель учебно-научного текста для разметки корпуса научно-технических текстов // Экономика. Информатика. 2021. № 48 (1). С. 123–129. (К-1) (0,5 п. л.)

12. Бутенко Ю.И., Строганов Ю.В., Сапожков А.М. Метод извлечения русскоязычных многокомпонентных терминов в корпусе научно-технических текстов // Прикладная информатика. 2021. №6. С. 21-27. (RSCI) (0,6 / 0,3 п. л.)

Соискателю лично принадлежит общий подход к извлечению англо- и русскоязычных многокомпонентных терминов.

13. Бутенко Ю.И., Сидняев Н.И., Болотова Е.Е. Использование падежной грамматики при информационном поиске в базе знаний о конструкции летательных аппаратов // Системы и средства информатики. 2021. Т.31, №3. С. 75-82. (RSCI) (0,5 / 0,25 п. л.)

Соискателем предложено использовать семантические роли для информационного поиска базе знаний о конструкции летательных аппаратов.

14. Бутенко Ю. И. Модель текста стандарта при информационном поиске в коллекции документов нормативной базы // Вестник компьютерных и

информационных технологий. 2020. Т. 17, № 11. С. 23 – 32. (К-2) (0,7 п. л.)

15. Бутенко Ю.И. Метод разрешения лексической многозначности поискового запроса на основе онтологий // Прикладная информатика. 2020. Т. 15. №5. С. 103-110. (RSCI) (0,5 п. л.)

16. Бутенко, Ю.И. Онтологический подход к формированию нормативного профиля при сертификации программного обеспечения // Онтология проектирования. 2020. Т. 10, №2(36). С.190-200. (RSCI) (0,9 п. л.)

Публикации полностью соответствуют теме диссертационного исследования и раскрывают ее основные положения.

В диссертации разработаны теоретические положения, совокупность которых можно квалифицировать как научное достижение, а также содержит новые научно обоснованные технические решения, внедрение которых вносит значительный вклад в развитие средств обработки естественного языка.

Диссертация Бутенко Юлии Ивановны на тему «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» является самостоятельной, завершенной научно-квалификационной работой, которая соответствует требованиям пунктов 9-14 «Положения о порядке присуждения ученых степеней», утвержденное постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, а также паспорту специальности 2.3.8. Информатика и информационные процессы, в частности пунктам:

2. Техническое обеспечение информационных систем и процессов, в том числе новые технические средства сбора, хранения, передачи представления информации. Комплексы технических средств, обеспечивающих функционирование информационных систем и процессов, накопления и оптимального использования информационных ресурсов;

5. Лингвистическое обеспечение информационных систем и процессов. Методы и средства проектирования словарей данных, словарей индексирования и поиска информации, тезаурусов и иных лексических комплексов. Методы семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных и организации интерфейсов

информационных систем с пользователями;

11. Разработка принципов организации и технологий реализации систем управления базами данных и знаний, создание специализированных информационных систем управления текстовыми, графическими и мультимедийными базами данных. Создание языков описания данных, языков манипулирования данными, языков запросов.

Тема и содержание диссертации «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» полностью соответствуют выбранной специальности – 2.3.8. Информатика и информационные процессы. Диссертация Бутенко Юлии Ивановны на тему «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» рекомендуется к защите на соискание ученой степени доктора технических наук по специальности 2.3.8. Информатика и информационные процессы.

Заключение принято на заседании кафедры теоретической информатики и компьютерных технологий МГТУ им. Н.Э. Баумана. Присутствовало на заседании 12 человек. Результаты голосования: «за» - 12 чел., «против» - нет чел., «воздержалось» - нет чел., протокол № 9 от 14 ноября 2024 года.

Заведующий кафедрой
теоретической информатики и
компьютерных технологий
МГТУ им. Н.Э. Баумана,
д.т.н., доцент



И.П. Иванов