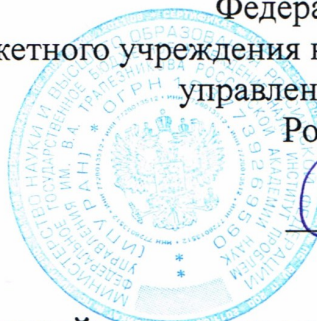


УТВЕРЖДАЮ
Директор
Федерального государственного
бюджетного учреждения науки «Институт проблем
управления им. В.А. Трапезникова
Российской академии наук»
Академик РАН
Д.А. Новиков
«29» января 2026



ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

на диссертационную работу Бутенко Юлии Ивановны «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе», представленную на соискание ученой степени доктора технических наук по специальности 2.3.8 «Информатика и информационные процессы»

Актуальность темы исследования обусловлена ускорением роста объема научно-технической информации, что требует разработки эффективных средств ее автоматизированной обработки. Современные нейросетевые методы существенно превосходят традиционные подходы к обработке текстов, основанные на правилах и статистических методах, прежде всего по качеству получаемых результатов. Вместе с тем успешность их применения зависит от наличия больших и качественно размеченных массивов текстовых данных. Существующие инструменты корпусной лингвистики лишь частично решают проблему подготовки таких массивов: как правило, средства автоматической обработки ограничены лемматизацией и морфологической разметкой, тогда как более сложные виды разметок выполняются специалистами вручную. Вместе с тем востребованность параллельных корпусов при решении широкого круга теоретических и прикладных задач подтверждает необходимость развития технологий их формирования. Таким образом, разработка моделей, методов и программных средств для автоматизированной разметки и выравнивания научно-технических текстов в параллельных корпусах является актуальной научной задачей.

Структура и содержание диссертации. Диссертация состоит из введения, шести глав, заключения и списка использованных источников.

Во введении обоснована актуальность темы исследования, определены цель и задачи исследования, раскрыта научная новизна и практическая значимость работы, а также представлены основные положения, выносимые на защиту.

В первой главе диссертации выполнен обзор и проведен анализ степени автоматизации процессов лингвистической разметки и выравнивания текстов в современных параллельных и специализированных корпусах, при этом особое внимание уделено корпусам текстов на русском языке. На основе проведенного анализа показано, что несмотря на значительные успехи в обработке естественного языка особенности научно-технических текстов обуславливают необходимость

разработки новых подходов, моделей, методов и программных средств для формирования параллельных корпусов научно-технических текстов. К таким особенностям отнесены композиционная структура научно-технических текстов, специальная терминология, а также стилистические маркеры антропогенного или машинного способов создания текстов.

Вторая глава посвящена анализу и разработке моделей композиционной структуры научно-технических текстов, являющихся источниками для наполнения корпуса. В качестве таких источников выбраны научно-технические статьи, учебные пособия и монографии, а также нормативные документы, имеющие переводные версии (русский-английский языки). Для каждого типа текста в предложенных моделях идентифицированы структурные элементы, определено их место в структуре текста, а также указаны их обязательные и факультативные элементы.

В третьей главе предложены структурные англо- и русскоязычные модели многокомпонентных терминов, которые основаны на общепринятом в терминоведении подходе, выделяющем в структуре термина ядерный элемент, левое и правое определения, что обеспечивает приведение термина к нормальной форме. Разработаны методы разметки и выравнивания многокомпонентных терминов, базирующиеся на грамматических признаках специальной лексики. Также выделены структурные модели номенклатурных наименований на русском и английском языках. В основу метода разметки номенклатурных наименований положены морфологическая и терминологическая разметки.

В четвертой главе показано, что лингвистические свойства языка являются основой для выявления машинных текстов. В русском языке к таким свойствам относится актуальное членение предложения, реализуемое в тема-рематическом делении. К классу машинных текстов отнесены машинно-сгенерированные и машинно-переведенные тексты. Автором предложены методы разметки машинно-сгенерированных и машинно-переведенных научно-технических текстов и их фрагментов, а также разработан статистический подход к их обработке.

Пятая глава посвящена описанию разработанной системы управления корпусными данными параллельного корпуса научно-технических текстов, включающей подсистему автоматической разметки текстов, базу данных размеченных текстов, корпусный менеджер и словарь корпуса. Также предложена информационная технология обработки научно-технических текстов, включающая этапы метатекстовой, структурной, морфологической, терминологической синтаксической и стилистической разметки, а также выравнивания структурных элементов текстов и многокомпонентных терминов. Детально описаны программные средства структурной разметки и выравнивания параллельных научно-технических текстов, разметки и выравнивания специальной лексики, а также разметки семантических ролей. Продемонстрированы способы использования параллельного корпуса при решении прикладных задач лингвистики и информатики.

В шестой главе рассмотрены аспекты применения разработанных моделей и методов для решения прикладных задач обработки научно-технических текстов в

учебной лексикографии, наукометрии, информационном поиске и обработке нормативных документов.

В заключении диссертации сформулированы основные результаты и выводы, полученные в результате проведенного диссертационного исследования.

Тема и содержание диссертации Бутенко Ю.И. «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» полностью соответствуют выбранной специальности – 2.3.8. Информатика и информационные процессы.

Основные результаты работы и их новизна заключаются в следующем.

1. Предложены модели иерархически структурированных научно-технических текстов, в которых учтены межуровневые связи между единицами текста, что обеспечивает более эффективную их автоматизированную обработку на разных уровнях языковой системы.

2. Разработаны модели и методы разметки и выравнивания англо- и русскоязычных многокомпонентных терминов и номенклатурных наименований в научно-технических текстах, что позволяет повысить эффективность разметки текстов за счет учета лексических единиц разной формальной структуры, которые могут состоять из более чем пяти компонентов.

3. Предложены методы идентификации машинно-сгенерированных и машинно-переведенных текстов, принципиальным отличием которых от существующих является использование семантико-синтаксических особенностей русского языка, а не машинного обучения.

4. Разработан прототип системы управления корпусными данными, который в отличие от существующих корпусных менеджеров позволяет формировать различные наборы данных с многоуровневой разметкой для машинного обучения.

Достоверность полученных результатов подтверждается результатами сравнительного анализа предложенных в диссертации моделей и методов с существующими аналогами, их апробацией и успешным применением при решении широкого круга практических задач автоматической обработки научно-технических текстов, а также соответствием полученных выводов известным научным результатам в области компьютерной лингвистики. Дополнительным свидетельством достоверности представленных в диссертации результатов являются публикации в профильных ведущих научных изданиях.

Основные результаты диссертации Бутенко Ю.И. были представлены на всероссийских и международных научных конференциях, а также отражены в 60 печатных работах, опубликованных в рецензируемых научных журналах, входящих в перечень ВАК и/или индексируемых в международных наукометрических базах Scopus и Web of Science.

Значимость полученных результатов

В диссертационной работе предложен комплекс моделей, методов и программных средств, разработанных на основе фундаментальных теоретических положений лингвистики, что обеспечивает возможность эффективного решения широкого круга задач автоматической обработки научно-технических текстов. Практическая значимость полученных результатов заключается в разработке программных средств для обработки научно-технических текстов в составе параллельного корпуса, применение которых позволяет автоматизировать трудоемкие процедуры по разметке и выравниванию текстов при формировании корпусов. Помимо непосредственного использования в корпусной лингвистике, предложенные решения могут быть востребованы при решении ряда прикладных задач в области терминографии, перевода, наукометрии, машинного обучения и инженерии знаний.

Замечания по диссертационной работе

1. В диссертации под термином «параллельный корпус» подразумеваются тексты на разных языках (переводные параллельные корпуса). Вместе с тем, за рамками рассмотрения находится важная область параллельных корпусов, основанных на моновязычных данных, например корпусов перефразирования, которые также активно используются в задачах автоматической обработки текстов. Желательно зафиксировать границы применимости методов.

2. В работе предложен метод извлечения многокомпонентных терминов, основанный на выделении левого и правого определений ядерного элемента термина. Следовало бы привести развернутые примеры типовых корректных/ошибочных извлечений и описать ограничения метода на специфических для научно-технических текстов конструкциях, что позволило бы точнее очертить область его применимости.

3. В методе выравнивания многокомпонентных терминов в параллельных текстах используется подход SimAlign, основанный на достаточно вычислительно затратной нейросетевой модели архитектуры BERT. Желательно было бы оценить его вычислительную сложность.

4. В методе выявления машинно-сгенерированных и машинно-переведенных научно-технических текстов, основанном на изменении порядка слов при переводе с английского языка на русский, автором продемонстрирована значительная статистическая разница показателей, что позволяет выявлять класс машинных текстов. Вместе с тем изложение экспериментальной части по данному методу можно усилить за счет более детального описания состава выборок, схемы разбиения данных и представления оценок вариативности результатов.

Перечисленные замечания не снижают общей высокой оценки диссертационного исследования.

Заключительная оценка

Диссертация Бутенко Юлии Ивановны на тему «Модели и методы автоматической обработки научно-технических текстов в параллельном корпусе» является завершённой научно-квалификационной работой, в которой решается научная задача, имеющая важное практическое значение.

Представленные в работе результаты и выводы корректно обоснованы, обладают научной новизной и достоверностью. Результаты работы своевременно и полно опубликованы в 60 печатных трудах, а также докладывались автором на международных и всероссийских конференциях. Автореферат верно и полно отражает содержание диссертационной работы.

Диссертационная работа соответствует всем требованиям пунктов 9–14 «Положения о порядке присуждения ученых степеней», утвержденного постановлением Правительства Российской Федерации от 24.09.2013 г. № 842, а ее автор, Бутенко Юлия Ивановна, заслуживает присуждения ученой степени доктора технических наук по специальности 2.3.8 – «Информатика и информационные процессы». Диссертационная работа и отзыв на работу рассмотрены на заседании научного семинара Института проблем управления им. В.А. Трапезникова Российской академии наук (протокол № 1 от 11 сентября 2025 года).

доктор технических наук,
ведущий научный сотрудник лаб. № 25

ПОДПИСЬ

ЗАВЕРЯЮ

ЗАВ. ОБЩИМ ОТДЕЛОМ

Сведения о составителе отзыва

Фамилия, имя, отчество: Губанов Дмитрий Алексеевич

Ученая степень: доктор технических наук

Наименования отрасли наук, научных специальностей, по которым защищена диссертация: 05.13.10 – Управление в социальных и экономических системах (технические науки)

Должность: ведущий научный сотрудник лаборатории № 25

Контактный телефон: +7 495 198-17-20, доб 1402

Адрес электронной почты: gubanov@ipu.ru

Сведения о ведущей организации

Федеральное государственное бюджетное учреждение науки «Институт проблем управления им. В.А. Трапезникова Российской академии наук» (ИПУ РАН)

Сайт: www.ipu.ru

Почтовый адрес: 117997, ГСП-7, В-342, г. Москва, ул. Профсоюзная, д.65

Телефон: +7 495 334-89-10

E-mail: dan@ipu.ru