

ОТЗЫВ

официального оппонента на диссертацию

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

На отзыв представлены диссертационная работа и автореферат. Диссертационная работа состоит из введения, 6 глав, заключения и списка литературы из 298 наименований. Общий объем диссертации 304 страницы, включая 103 рисунка и 41 таблицу. Объем автореферата 36 страниц.

Актуальность исследования обусловлена тем, что в настоящее время существует значительное количество параллельных корпусов с разными языковыми парами, однако параллельные корпуса научно-технических текстов имеют незначительный объемы размеченных данных, поскольку разметка и выравнивание, как правило, проводятся вручную, что резко ограничивает применение этих корпусов в интеллектуальных системах обработки научно-технических текстов. Поэтому для автоматизации процедуры автоматической обработки научно-технических текстов при создании параллельных корпусов необходима инструментальная поддержка процедуры обработки параллельных научно-технических текстов. Таким образом, разработка информационных моделей и методов решения задач автоматической обработки научно-технических текстов при создании параллельного корпуса с применением методов структурного, терминологического и стилистического моделирования является актуальной задачей, требующей для своего решения новых научных подходов, которые предлагаются в представленной диссертации.

Анализ и общая оценка основного содержания диссертационной работы. Основные материалы диссертации скомпонованы в 6 глав.

Первая глава содержит анализ современного состояния исследований в области обработки научно-технических текстов при создании и наполнении корпусов текстов, а также применимости существующих моделей, методов и программных средств обработки научно-технических текстов для создания параллельного корпуса.

Во второй главе содержится описание моделей научно-технических текстов. В качестве источников наполнения параллельного корпуса использованы научно-технические статьи, учебники, нормативная документация. Эти виды текстов обладают строгой иерархической структурой с небольшим числом структурных элементов, притом структура является упорядоченной.

В третьей главе изложена разработка моделей и методов разметки и выравнивания многокомпонентных русско- и англоязычных терминов с правыми определениями и номенклатурных наименований в параллельном корпусе научно-технических текстов.

Четвертая глава содержит описание методов разметки стилистических и переводческих особенностей параллельных текстов, прежде всего, выявления и разметки машинно-сгенерированных и машинно-переведенных русскоязычных научно-технических текстов и их фрагментов.

В пятой главе приведено описание исследовательского прототипа системы управления корпусными данными параллельного корпуса научно-технических текстов, в которой реализованы предложенные в диссертации модели и методы обработки русско- и англоязычных научно-технических текстов.

В шестой главе представлено использование разработанных моделей и методов для решения практических задач в области лингвистики и

информатики. В частности, показаны подходы к практическому использованию предложенных в диссертации методов обработки научно-технических текстов в сферах цифровой и учебной терминографии, к выявлению тенденций развития научных направлений на основе публикаций, к формированию нормативных профилей при сертификации продукции, к анализу студенческих работ на предмет использования систем генерации текстовых последовательностей, выявлению использования машинных переводчиков при переводе научно-технических текстов и др.

В заключении обобщены теоретические и практические результаты работы, представлены рекомендации и перспективы дальнейшей разработки темы.

Таким образом, в диссертации разработаны информационные модели и методы решения задач автоматической обработки научно-технических текстов при создании параллельного корпуса с применением методов структурного, терминологического и стилистического моделирования, позволяющие повысить эффективность автоматической разметки и выравнивания лингвистических единиц разной формальной структуры в параллельном корпусе путем автоматизации процесса обработки научно-технических текстов.

Научная новизна. Результаты, выводы и положения диссертации теоретически обоснованы, имеют практическую значимость и обладают научной новизной. Отметим следующие результаты:

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

2. Новые 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. Оригинальные модель и метод разметки номенклатурных наименований в научно-технических текстах на русском и английском языках, позволяющие повысить эффективность разметки научно-технических текстов за счет учета лексических единиц, включающих в себя и произвольные буквенно-числовые последовательности в том числе символы разных алфавитов.

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

5. Прототип системы управления корпусными данными, позволяющий управлять корпусными данными на разных этапах их обработки, а также формировать различные наборы данных для машинного обучения.

Основные научные результаты получены автором самостоятельно.

Практическая значимость результатов диссертации. В процессе работы над диссертацией создано программное средство, позволяющее использовать разработанные в диссертации теоретические основы, модели и методы обработки англо- и русскоязычных научно-технических текстов для создания параллельного корпуса. Созданное программное средство дает возможность автоматизировать рутинный процесс обработки англо-и русскоязычных параллельных текстов и увеличить объемы филологически компетентных баз данных размеченных научно-технических текстов. Предлагаемые в диссертации методы и программные средства автоматической обработки текстов внедрены в ООО «Агентство переводов «Эдельвейс»» и в ООО «Инженерное бюро ВССТ».

Апробация результатов работы. Содержание диссертации в достаточной мере отражено в имеющихся публикациях автора, включая 36

статей в изданиях, рекомендованных ВАК РФ для публикации работ соискателей, 20 из которых входит в перечень индексируемых Web of Science и Scopus. Результаты диссертации неоднократно докладывались и обсуждались на научных конференциях и семинарах.

Замечания. Необходимо отметить ряд недостатков диссертационной работы.

1. В главе 2 модель текста научно-технической статьи, модель учебно-научного текста и модель текста стандарта разработаны независимо друг от друга, хотя они имеют единый общий вид: $\langle E^L, R \rangle$, где E^L – структурный элемент, R – отношения между структурными элементами. Следовало бы разработать единую модель текста научно-технического документа, конкретизировав ее в зависимости от вида документа.

2. Иногда соискатель недостаточно внимательно работает с библиографией. Так, на стр.17 диссертации применительно к польско-русскому параллельному корпусу Варшавского университета сказано: «Выравнивание параллельных текстов осуществляется при помощи программы ABBY Aligner, морфологическая разметка польских текстов выполнена с помощью программного средства TAKIPi, а для морфологической разметки русскоязычных текстов использован программный продукт Pantera [60]». Однако в работе: «60. Куратчик М. Параллельные корпуса русского и польского языков и их использование в сопоставительной лингвистике и лингводидактике // Русский язык и литература в пространстве мировой культуры: Материалы XIII Конгресса МАПРЯЛ: В 15 т., Гранада, Испания, 13–20 сентября 2015 года / Составители: Н. М. Марусенко, М. С. Шишков. Том 11. – Гранада, Испания: Международное некоммерческое партнерство преподавателей русского языка и литературы "МАПРЯЛ", 2015. С. 152-157.» – указанной информации нет, она содержится в близкой по содержанию работе

Лазинский М., Куратчик М. Польско-русский параллельный корпус Варшавского университета. *Language and Method*, 2016. С. 85-93.

3. В таблицах 1.1–1.4 для некоторых записей отсутствует единица измерения в столбце «Объем».

4. Имеются отдельные опечатки. Так, на стр. 112 второй параграф главы 3 пронумерован как 2.2.

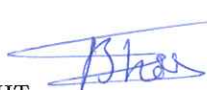
Заключение. Перечисленные выше недостатки не меняют общей положительной оценки диссертационной работы. Диссертация Бутенко Юлии Ивановны представляет собой законченную научно-квалификационную работу, в которой автором проведены новые исследования, разработаны информационные модели и методы решения задач автоматической обработки научно-технических текстов при создании параллельного корпуса с применением методов структурного, терминологического и стилистического моделирования, позволяющие повысить эффективность автоматической разметки и выравнивания лингвистических единиц разной формальной структуры в параллельном корпусе путем автоматизации процесса обработки научно-технических текстов.

Диссертация соответствует паспорту научной специальности 2.3.8 — «Информатика и информационные процессы», а именно: пп. 2, 5, 11, поскольку содержит создание комплекса технических средств, обеспечивающих функционирование информационных систем и процессов, накопления и оптимального использования информационных ресурсов; разработку методов семантического, синтаксического и прагматического анализа текстовой информации для представления в базах данных; создание специализированных информационных систем управления текстовыми базами данных.

Таким образом, диссертационное исследование Бутенко Юлии Ивановны на тему «Модели и методы автоматической обработки научно-

технических текстов в параллельном корпусе» соответствует требованиям ВАК Минобрнауки России, предъявляемым к диссертациям на соискание ученой степени доктора технических наук, а ее автор – Бутенко Юлия Ивановна – заслуживает присуждения искомой ученой степени доктора технических наук по специальности 2.3.8 — «Информатика и информационные процессы» (технические науки).

Официальный оппонент:

ведущий научный сотрудник лаборатории информационных ресурсов
Федерального исследовательского центра информационных и
вычислительных технологий,
доктор технических наук, доцент  Владимир Борисович Барахнин

28.01.2026 г.

Подпись официального оппонента В.Б. Барахнина заверяю:

Ученый секретарь Федерального исследовательского центра
информационных и вычислительных технологий,
кандидат физико-математических наук  Н.В. Киланова



630090, Новосибирск, пр. Академика Лаврентьева, 6, ФИЦ ИВТ,

+7-913-911-12-89

bar@ict.nsc.ru