ПРОЕКТ, ПОДДЕРЖАННЫЙ РОССИЙСКИМ ФОНДОМ ФУНДАМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ (РФФИ) №19-07-00243 А

Создание методов автоматизированной обработки текстов с целью авторизации, анализа и построения ритмичных текстов на русском, французском и английском языках

Номер проекта	19-07-00243
Руководитель коллектива	**Бойчук Елена Игоревна**
Исполнители	**Лагутина Надежда Станиславовна**, кандидат физико-математических наук, доцент кафедры вычислительных и программных систем ЯрГУ им. П.Г. Демидова **Воронцова Инна Алексеевна**, кандидат филологических наук, доцент кафедры теории и практики перевода ЯГПУ им. К.Д. Ушинского **Шляхтина Елена Васильевна**, кандидат филологических наук, доцент кафедры теории и практики перевода ЯГПУ им. К.Д. Ушинского **Лагутина Ксения Владимировна**, ст. преподаватель кафедры вычислительных и программных систем ЯрГУ им. П.Г. Демидова **Беляева Ольга Васильевна**, ст. преподаватель кафедры теории и практики перевода ЯГПУ им. К.Д. Ушинского
Название проекта	Создание методов автоматизированной обработки текстов с целью авторизации, анализа и построения ритмичных текстов на русском, французском и английском языках
Код и название конкурса	А Проекты фундаментальных научных исследований
Область знания	07 ИНФОКОММУНИКАЦИОННЫЕ ТЕХНОЛОГИИ И ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ
Основной код (по классификатору РФФИ)	07-986 Системы искусственного интеллекта для текстового поиска, обработки и анализа естественного языка
Дополнительные коды (по классификатору РФФИ)	12-101 Русская литература 12-103 Литература народов стран зарубежья 12-104 Теория литературы. Текстология 12-201 Русский язык 12-204 Германские языки 12-205 Романские языки
Ключевые слова	Художественный текст, ритм, повтор, стилистические средства, переводческие стратегии, адекватность перевода, идиостиль, доминанта текста, средства создания ритма, коэффициент ритма, алгоритм анализа ритма
Аннотация проекта	Основной научной задачей, на решение которой направлено данное исследование, является создание новых

методов в компьютерной лингвистике для автоматизированной обработки текстов на русском, английском и французском языках с целью их авторизации, анализа и построения ритмичных текстов на этих языках. Данная задача интегрирует компьютерную лингвистику и информационные технологии с такими областями филологического знания, как лингвистика текста, текстология, стилистика текста, риторика, психолингвистика, культура речи, а также с квантитативной лингвистикой, суггестивной лингвистикой, лингвокультурологией.

Основополагающим параметром создания данных методов является ритм текста. Актуальность данной задачи состоит в том, что большое количество существующих теоретических исследований в области ритма, который требует статистического подхода, применения методов количественного анализа, автоматизированной обработки текста, не имеет выхода к осуществлению такого рода анализа. Анализ ритма осуществляется вручную, что снижает эффективность текстовой обработки и не всегда приводит к точным результатам.

В основу предложенных исследователями современных программных приложений для русского языка положены такие факторы определения ритмичности текста, как размер предложений и среднеарифметический показатель безударных слогов [Белоусов, К.И., Дусакова Г.Ф., Леонов Д.В.,2017; Кишалова Л.В., 2017; Андрусенко Т.В., 2017]. Для английского и французского языков существуют приложения, позволяющие анализировать текст с точки зрения его лексического состава, размера предложений и n-грамм [Text Analyzer, Sketch Engine, Stylo, Word2vec и др.]. Анализ ритма в рамках данного проекта направлен на определение авторства текстов, сопоставление переводов художественных произведений, а также на анализ и составление суггестивных текстов на основе определения их ритмических характеристик, складывающихся из употребления средств ритмизации, проявляющихся на фонетическом, лексико-грамматическом и структурно-композиционном уровнях. В состав данных средств входят стилистические средства выразительности речи, в основу выделения которых заложена повторяемость элемента (рифма, ассонанс, аллитерация, ономатопея, анаграмма, парономазия, таутацизм, деривация, полиптотон, анафора, эпифора, симплока, хиазм, однородные члены, эпаналепсис, анадиплозис, редупликация и др.).
Для реализации цели проекта необходимо создать автоматизированные алгоритмы, позволяющие проводить сравнительно-сопоставительное исследование текстов с точки зрения ритмических характеристик; определить статистические характеристики ритма текста и создать банк данных, включающий тексты с выявленными характеристиками ритма; разработать алгоритм анализа и

	составления ритмичного текста с целью активного суггестивного воздействия на реципиента; создать метод атрибуции на основе ритмических средств текста и программный прототип, позволяющий хранить и обрабатывать тексты, проводить сравнительно-сопоставительный анализ ритма в переводе и осуществлять атрибуцию. В данном исследовании впервые предлагается использование перечисленных выше ритмических средств в качестве основы создания автоматизированного метода атрибуции текстов. Разработанные в настоящее время алгоритмы составления ритмичных текстов основаны на абзацном членении и размере предложения. Предлагаемый метод позволит расширить возможности суггестивного воздействия на реципиента при помощи представленных ритмических характеристик. Новизна предлагаемого метода заключается также в том, что он позволяет осуществлять ритмический анализ текстов, оценивать адекватность перевода с точки зрения ритмических характеристик, осуществлять атрибуцию текстов на трех языках русском, английском и французском. В разработке задействованы новые методы: ритмический анализ текста, сравнительно-сопоставительный ритмический анализ, которые позволят получить более точные результаты обработки ритма текста. Автоматизированная обработка текста с точки зрения ритмических характеристик позволит более успешно реализовывать задачу авторизации. Областью практического применения автоматизированной обработки текста с позиций реализации ритмических характеристик может являться публицистика, в рамках которой эффективным с точки зрения воздействия на реципиента является создание суггестивных текстов. Ожидаемые результаты Основным результатом проекта должно стать создание программного прототипа, включающего банк текстов на трех языках (корпус электронных художественных текстов на русском, английском и французском языках для осуществления анализа ритма и корпус художественных текстов с выявленными статистическими характеристиками ритма), позволяющего осуществлять автоматизированный анализ ритма, авторизацию текста и проводить сравнительно-сопоставительный анализ переводов текстов.
Аннотация по результатам 1 года реализации проекта	В ходе работы над проектом по автоматизации обработки текстов с целью авторизации, анализа и построения ритмичных текстов на русском, французском и английском языках были найдены решения для реализации поставленных на первый год целей. В частности, для достижения цели по созданию автоматизированных алгоритмов, позволяющих проводить анализ текстов с точки зрения проявления ритмических характеристик, были собраны данные по всем существующим

на момент начала исследования инструментам, позволяющим реализовывать анализ текста с целью определения индивидуального авторского стиля, а также в той или иной степени ориентированным на осуществление авторизации текста и было выявлено, что в существующих исследованиях мало используются семантические и чисто лингвистические параметры (аспекты ритма, синонимы, фразеологизмы и т.д.). Применение «более языковых» параметров анализа текста, таких как лексические повторы, ритмические конструкции, особенности синтаксических структур, позволило бы корректно, не формально подойти к анализу языка автора. Проведенное исследование показало, что из работ 40 авторов по автоматизированному анализу текста лишь 10% исследователей касаются некоторых параметров ритма, в частности изучения слоговых структур и рифмы.

Для создания нового альтернативного инструмента, позволяющего осуществлять анализ идиолекта писателя с точки зрения ритмических параметров, были собраны данные по определению основных параметров ритма на лексико-грамматическом уровне. Данный языковой уровень включает употребление стилистических средств, содержащих повтор, то есть соответствующий структуре основной элемент+повторяющиеся элементы. Среди ритмических средств, которые соответствуют данной структуре, были определены следующие: анафора эпифора, симплока, анадиплозис, эпаналепсис, редупликация, полисиндетон, апозиопеза. Для данных средств были определены условия их употребления для английского, французского, испанского и русского языков, сформированы списки исключений и специфического употребления тех или иных форм. Также были составлены списки стоп-слов, позволяющих при возможности исключать из поиска и количественной обработки определенные части речи и конкретные служебные слова, подсчет которых нарушает статистику. При помощи созданного инструмента ProseRhythmDetector (PRD), позволяющего осуществлять поиск указанных выше средств в английском и русском текстах, проведен анализ произведений более 100 авторов и переводов их произведений на русский язык (Ч. Диккенса, Ш. Бронте, К. Аткинсон, Д. дю Морье, Дж. К. Роулинг, Дж. Остин, Э. Гаскелл, Дж. Джойс, А. Мердок, Ф.С. Фицджеральда и многих других). В ходе работы с инструментом выяснилось, что PRD значительно сокращает время работы с текстом, облегчает сам процесс анализа и позволяет выявить довольно широкий круг средств одновременно. Помимо этого, инструмент помогает не упустить те средства, которые можно было бы не заметить в рамках «ручного»/ «механического» анализа. Степень совпадения оригинала и перевода с точки зрения реализации ритмических средств являлась маркером адекватности перевода. Результаты продемонстрировали низкую степень совпадения, что вероятно обусловлено спецификой лексико-грамматической структуры языков.

	В рамках задачи по исследованию ритмической структуры художественных прозаических текстов 20-21 вв. в трех неблизкородственных языках (русском, английском и французском) на основе диахронического сравнительно-сопоставительного исследования ритма художественных произведений русских, французских и английских авторов была несколько расширена: для исследования ритмической структуры использовались также произведения 19 века, кроме того, к перечисленным выше языкам был добавлен испанский язык. В результате бы проведен анализ текстов в четырех неродственных языках с опорой на приложение Rythmanalyse, позволяющее осуществлять анализ ритма французских текстов, инструмента ProseRhythmDetector, используемый для анализа лексико-грамматического аспекта англоязычного и русскоязычного текстов и веб-приложения, позволяющего осуществлять поиск и подсчет повторов в испаноязычном тексте. В настоящее время осуществляется работа по объединению задач анализа ритма на материале четырех языков (французского, английского, русского и испанского) внутри одного инструмента ProseRhythmDetector. Для получения результатов по определению статистических характеристик была определена модель текста, описывающая его ритмическую специфику. Модель включает в себя сам текст, разделённый по главам и абзацам, а также номера слов, входящих в ритмические характеристики, и контекстов этих характеристик. На основе модели был вычислен набор статистических показателей, который включал в себя как метрики, описывающие ритм текста в целом, так и опирающиеся на конкретные ритмические средства. Согласно полученным на данный момент результатам статистический анализ ритмических средств показывает четкое разделение произведений 19 и 20 веков согласно ритмическим параметрам на лексико-грамматическом уровне. Также по ритмическим характеристикам удалось выделить три группы русскоязычных переводов англоязычных текстов: тексты 19 века, начала и середины 20 века, конца 20 — начала 21 века.
Публикации за 1 год реализации проекта (ссылки на публикации на странице руководителя проекта)	**Монографии** 1. Бойчук Е. И. Анализ ритма прозы (на материале французского языка): монография / Е. И. Бойчук. – Ярославль: Канцлер, 2019. – 232 с. 2. Идиостиль и ритм текста: коллективная монография / Бойчук Е.И., Воронцова И.А., Шляхтина Е.В., Беляева О.В., Ярославль: РИО ЯГПУ, 2019. **Публикации в SCOPUS:** 1. Lagutina K., Lagutina N., Boychuk E., Vorontsova I., Shliakhtina E., Belyaeva O., Paramonov I. A Survey on Stylometric Text Features FRUCT: Helsinki, Finland, 5-8 November 2019.

2. Boychuk E., Vorontsova I., Shliakhtina E., Lagutina K., Belyaeva O. Automated Approach to Rhythm Figures Search in English Text // AIST, 2019. (в печати)

Публикации в журналах, рецензируемых ВАК:

1. Лагутина Н.С., Лагутина К.В., Бойчук Е.И., Воронцова И.А., Парамонов И.В. Автоматизированный поиск средств ритмизации художественного текста для сравнительного анализа оригинала и перевода на материале английского и русского языков // Моделирование и анализ информационных систем. Т. 26, №3 (2019), с. 420–440. DOI: 10.18255/1818-1015-2019-3-420-440
2. Бойчук Е.И., Джонсон М.А. Специфика ритмической структуры испанских прозаических текстов // Верхневолжский филологический вестник= VerhnevolzhskiPhilologicalBulletin:научный журнал.– Ярославль: РИО ЯГПУ, 2019.–No2(17)– С. 119-129. DOI: 10.24411/2499-9679-2019-10322
3. Бойчук Е.И. Количественные наречия как средство авторизации художественных текстов (на материале французской прозы 19в.) // Верхневолжский филологический вестник= VerhnevolzhskiPhilologicalBulletin:научный журнал.– Ярославль: РИО ЯГПУ, 2019.–No1(16)– С. 123-130. DOI: 10.24411/2499-9679-2019-10394
4. Шляхтина Е.В. Передача ритмических средств при переводе художественной прозы с английского на русский язык (на материале романа К. Аткинсон «Жизнь после жизни») // Вестник МГОУ: Серия Лингвистика, 2019. - С. 149-161. DOI: 10.18384/2310-712X-2019-5-149-161

Публикации в сборниках РИНЦ и тезисы докладов на конференциях:

1. Бойчук Е.И. К проблеме автоматизированной обработки ритма текста // Романские языки в синхронии и диахронии : межвузовский сборник научных трудов. - М.: ИИУ МГОУ, 2019. - С. 52-59.
2. Бойчук Е.И., Трофимова Е.А. Лексико-грамматический аспект реализации ритма романа Шарлотты Бронте «Джейн Эйр» // Язык и общество: Диалог культур и традиций: сборник материалов международной научной конференции «Чтения Ушинского». – Вып. 17. – Ярославль: РИО ЯГПУ, 2019. - С. 40-52.
3. Бойчук Е.И. Особенности лексико-грамматического аспекта ритмизации романа А. Гавальда «Просто вместе» // Материалы Всероссийской научной конференции с международным участием «Европейское литературное наследие в

кросскультурном пространстве». – Ярославль: Канцлер, 2019. - С. 199-204.

4. Бойчук Е.И. Специфика ритма современной французской прозы (на материале романа А. Нотомб «Кодекс принца») // Материалы конференции Филологические чтения: Человек. Текст. Дискурс. - Изд. Ярославский государственный университет им П.Г. Демидова, 2019г. - С.6-13.
5. Boychuk E., Belyaeva O. La téchnique de stylométrie réalisée à la base de l'analyse informatique du rythme du texte // 10-ièmes Journées Internationales de Linguistique de Corpus (JLC). Université Grenoble-Alpes, 26-28.11.19
6. Лагутина Н.С., Лагутина К.В. Обзор инструментов для анализа ритма текста // Заметки по информатике и математике, Вып. 11. - Ярославль, ЯрГУ, 2019.
7. Туманова А.Д., Лагутина Н.С. Алгоритмы автоматического поиска лексических аспектов анализа ритма
8. Ратников Е.С., Лагутина Н.С. Программное приложение для автоматического определения лексических аспектов ритмики текста
9. Шляхтина Е. В., Туленцова Е. А. Особенности перевода художественного текста в жанре «хоррор» с английского на русский язык (на материале Р. Л. Стайна «Месть садовых гномов»)

Охранные документы на результаты интеллектуальной деятельности

Свидетельство о государственной регистрации программы для ЭВМ № 2019619380 «Программа, реализующая автоматизированный алгоритм анализа ритма текста на основе фонетических, лексико-грамматических и структурно-композиционных параметров ритма для текстов на русском, английском и французском языках». Правообладатели: Ратников Егор Сергеевич, Туманова Анастасия Дмитриевна, Бойчук Елена Игоревна, Лагутина Надежда Станиславовна, Лагутина Ксения Владимировна. Дата регистрации 16.07.2019г.