Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Ярославский государственный университет им. П. Г. Демидова»

На правах рукописи

Лагутина Ксения Владимировна

АВТОМАТИЧЕСКИЙ АНАЛИЗ РИТМИЧЕСКИХ ХАРАКТЕРИСТИК ТЕКСТОВ НА ЕСТЕСТВЕННОМ ЯЗЫКЕ

РЕЗЮМЕ
диссертации на соискание учёной степени
кандидата компьютерных наук

Ярославль — 2021

Диссертационная работа выполнена в федеральном государственном бюджетном образовательном учреждении высшего образования «Ярославский государственный университет им. П. Г. Демидова»

Научный руководитель: Соколов Валерий Анатольевич, доктор физико-математических наук, профессор, Ярославский государственный университет им. П. Г. Демидова

Актуальность темы. Стилометрия — это раздел компьютерной лингвистики, который изучает количественную оценку языковых особенностей в текстах на естественном языке. Стилометрия тесно связана с определением индивидуального стиля и идиолекта автора, которые представляют собой систему языковых особенностей, используемых автором [1]. Отличительные черты стиля, в котором написан текст, можно формализовать, разработать алгоритмы их автоматического определения и использовать для задач верификации авторства, классификации текстов на естественном языке по времени публикации или жанру, а также для статистического анализа характеристик текстов. Поиск стилометрических характеристик текстов достаточно трудоёмок и требует значительного количества времени при обработке вручную, поэтому его требуется автоматизировать.

Выбор стилометрических характеристик текста является наиболее важным этапом исследования. Исследователи выделяют около тысячи средств на разных уровнях анализа: лексический (включая уровни символов и слов), синтаксический, семантический, структурный и предметно-ориентированный [2; 3]. Это свидетельствует о сложности и многогранности текста, поэтому необходимо оценивать текстовые единицы, отобранные для количественного анализа, и их способность выражать оригинальность авторского стиля.

Сегодня нет единого мнения о том, как подобрать оптимальные стилометрические характеристики для решения любой из задач классификации или анализа текстов. Большинство современных исследователей применяет практически один и тот же набор стандартных характеристик, моделирующих текст на уровне слов и символов, а для повышения качества решения задачи обращают внимание на методы уменьшения размерности характеристических векторов и подбор классификаторов и их параметров. Напротив, в классической лингвистике ученые при анализе стиля текста концентрируются на сложных лингвистических параметрах. Поэтому поиск и анализ новых средств стиля является актуальной задачей компьютерной лингвистики.

Одним из важных аспектов специфики стиля текста является ритм. Ритм — это регулярное повторение схожих и соизмеримых единиц речи, которое выполняет структурирующие, текстообразующие и выразительные функции [4]. В классической лингвистике основная цель анализа ритма — глубокое проникновение в творческий метод автора, в его замысел, оригинальность индивидуального творчества и мастерства, поэтому выявление специфики ритма писательских произведений позволит более успешно решить проблему определения индивидуального авторского стиля. Этот метод используется в анализе поэтических текстов, в то время как его применение для прозаической художественной литературы

почти не исследовано [5]. В частности, алгоритмы поиска ритмических характеристик трудно формализуются, а доступные программные средства для их явного определения отсутствуют. Поэтому разработка автоматизированных инструментов для анализа ритма в прозаическом тексте и их апробация в классификации и анализе текстов является одной из важных задач обработки естественного языка.

Целью диссертационной работы является разработка и исследование комплекса ритмических характеристик текста и их сравнение со стандартными стилометрическими характеристиками в решении задач классификации текстов по авторам и периодам публикации.

Для достижения поставленной цели необходимо решить следующие **задачи**.

1. Разработка алгоритмов и программного инструмента для автоматического поиска ритмических средств в прозаических текстах.

2. Статистический анализ динамики изменения ритмических средств в прозаических текстах XIX–XXI веков.

3. Классификация художественной литературы XIX–XXI веков по векам и полувекам их публикации на основе ритмических и стандартных стилометрических характеристик.

4. Верификация авторов художественной литературы XIX–XXI веков на основе ритмических и стандартных стилометрических характеристик.

Объектом исследования являются прозаические тексты на естественном языке.

Предметом исследования является комплекс ритмических характеристик текста.

Методология и методы исследования. Методология диссертационного исследования основана на постановке и формализации целей и задач, разработке моделей текстов, методов и алгоритмов анализа текстов, экспериментальной оценке при помощи статистических экспериментов, апробации и анализе результатов. Для решения поставленных задач были использованы методы автоматической предобработки текстов, поиска статистических и лексико-грамматических характеристик текста. Анализ динамики ритма текстов проводился с помощью статистических метрик и методов их визуализации. Классификация текстов осуществлялась на основе методов машинного обучения и нейросетей.

На защиту выносятся следующие **положения**, обладающие **научной новизной**.

1. Разработаны алгоритмы для автоматического поиска и визуализации лексико-грамматических ритмических средств в прозаических текстах для русского, английского, французского и испанского языков.

2. Разработан комплекс числовых ритмических характеристик для прозаического текста. Продемонстрирована его пригодность для проведения объемных экспериментов на примере статистического анализа динамики изменения ритмических средств в прозаических текстах XIX–XXI веков.

3. Показана эффективность применения ритмических характеристик для классификации художественной литературы XIX–XXI веков по векам и полувекам их публикации. Проведено сравнение ритмических и стандартных стилометрических характеристик для решения данной задачи.

4. Показано, что ритмические характеристики по качеству определения автора соответствуют стандартным характеристикам, а в комбинации с ними повышают эффективность верификации авторов художественной литературы XIX–XXI веков.

Практическая значимость. Результаты исследований по верификации авторов и классификации текстов по векам публикации показывают, что модель текста на основе ритмических характеристик может быть успешно использована для решения задач классификации художественных текстов. Программный инструмент на основе предложенных алгоритмов для поиска и визуализации ритмических характеристик, разработанный под руководством диссертанта, полезен экспертам-лингвистам для автоматизации их работы и сокращения времени на объемную рутинную работу при исследованиях.

Апробация работы. Основные результаты работы докладывались на международных научных конференциях:

1. «AIST 2019 — The 8th International Conference on Analysis of Images, Social Networks and Texts» (Казань, Россия, 2019);

2. «The 25th Conference of Open Innovations Association FRUCT» (Хельсинки, Финляндия, 2019);

3. «The 26th Conference of Open Innovations Association FRUCT» (Ярославль, Россия, 2020);

4. «The 27th Conference of Open Innovations Association FRUCT» (Тренто, Италия, 2020);

5. «The 28th Conference of Open Innovations Association FRUCT» (Москва, Россия, 2021);

6. «The 29th Conference of Open Innovations Association FRUCT» (Тампере, Финляндия, 2021).

Личный вклад. Содержание диссертации и основные положения, выносимые на защиту, отражают персональный вклад автора в опубликованные работы. Из работ, выполненных в соавторстве, в диссертацию включены результаты, которые соответствуют личному участию автора.

Публикации. Основные результаты по теме диссертации изложены в следующих печатных изданиях.

Публикации стандартного уровня:

1. Лагутина Н. С. Автоматизированный поиск средств ритмизации художественного текста для сравнительного анализа оригинала и перевода на материале английского и русского языков / Лагутина Н.С., Лагутина К.В., Бойчук Е.И., Воронцова И.А., Парамонов И.В. //Моделирование и анализ информационных систем. – 2019. – Т. 26. – №. 3. – С. 420-440. (список журналов, рекомендованных ВШЭ)

2. Лагутина К. В. Автоматизированный поиск и анализ стилометрических характеристик, описывающих стиль прозы 19-21 веков / Лагутина К. В., Манахова А. М. //Моделирование и анализ информационных систем. – 2020. – Т. 27. – №. 3. – С. 330-343. (список журналов, рекомендованных ВШЭ) — главный соавтор.

3. Lagutina N. S. Automated Rhythmic Device Search in Literary Texts Applied to Comparing Original and Translated Texts as Exemplified by English to Russian Translations / Lagutina, N. S., Lagutina, K. V., Boychuk, E. I., Vorontsova, I. A., Paramonov, I. V. //Automatic Control and Computer Sciences. – Springer, 2020. – Vol. 54. – №. 7. – pp. 697-711. (Scopus, Q3) [Автоматизированный поиск средств ритмизации художественного текста для сравнительного анализа оригинала и перевода на материале английского и русского языков]

4. Lagutina K. V. Comparison of Style Features for the Authorship Verification of Literary Texts / Lagutina K. V. //Modeling and Analysis of Information Systems. – 2021. – Vol. 28. – №. 3. – pp. 250-259. (список журналов, рекомендованных ВШЭ) [Сравнение стилистических характеристик для верификации авторов художественных текстов]

5. Лагутина К. В. Классификация текстов по жанрам на основе ритмических характеристик / Лагутина К.В., Лагутина Н.С., Бойчук Е.И. //Моделирование и анализ информационных систем. – 2021.

– Т. 28. – №. 3. – С. 280-291. (список журналов, рекомендованных ВШЭ) — главный соавтор.

Прочие публикации

6. Lagutina K. A Survey on Stylometric Text Features / Lagutina K., Lagutina N., Boychuk E., Vorontsova I., Shliakhtina E., Belyaeva O., Paramonov I. // Proceedings of the 25th Conference of Open Innovations Association FRUCT, IEEE, 2019 – Vol. 25. – № 1. – pp. 214-219. (Web of Science, Scopus) — главный соавтор. [Обзор стилометрических характеристик текста]

7. Boychuk E. Automated Approach to Rhythm Figures Search in English Text. /Boychuk E., Vorontsova I., Shliakhtina E., Lagutina K., Belyaeva O. // International Conference on Analysis of Images, Social Networks and Texts. CEUR Workshop Proceedings. Springer, Cham CCIS, Vol. 1086, 2020. pp. 107-119. (Web of Science, Scopus) [Автоматизированный подход к поиску ритмических средств в английском тексте]

8. Lagutina K. Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries / Lagutina K., Poletaev A., Lagutina N., Boychuk E., Paramonov I. // Proceedings of the 26th Conference of Open Innovations Association FRUCT. IEEE, 2020. – Vol. 26. – № 1. – pp. 247-255. (Web of Science, Scopus) — главный соавтор. [Автоматическое извлечение ритмических средств и анализ их динамики в прозе XIX–XXI веков]

9. Lagutina K. The Inuence of Different Stylometric Features on the Classification of Prose by Centuries / Lagutina K., Lagutina N., Boychuk E., Paramonov I. //Proceedings of the 27th Conference of Open Innovations Association FRUCT. – IEEE, 2020. – Vol. 27. – № 1. – С. 108-115. (Web of Science, Scopus) — главный соавтор. [Влияние различных стилометрических характеристик на классификацию прозы по векам]

10. Boychuk E. Evaluating the Performance of a New Text Rhythm Analysis Tool / Boychuk E., Lagutina K., Vorontsova I., Mishenkina E., Belyayeva O. // English Studies at NBU. – New Bulgarian University, 2020. – Vol. 6. – №. 2. – pp. 217-232. (Web of Science) [Оценка производительности нового инструмента для анализа ритма текста]

11. Lagutina K. Authorship verification of literary texts with rhythm features. / Lagutina K., Lagutina N., Boychuk E., Larionov V., Paramonov I. // Proceedings of the 28th Conference of Open Innovations Association FRUCT, IEEE, 2021 – Vol. 28. – № 1. – pp 240–251. (Web of Science, Scopus) — главный соавтор. [Верификация авторства художественных текстов с помощью ритмических характеристик]

12. Lagutina K. A Survey of Models for Constructing Text Features to Classify Texts in Natural Language. / K. Lagutina, N. Lagutina. // Proceedings of the 29th Conference of Open Innovations Association FRUCT. – IEEE, 2021 – Vol. 29. – № 1. – pp. 222-233. (Scopus) — главный соавтор. [Обзор моделей построения характеристик текста для классификации текстов на естественном языке]

Свидетельства о регистрации программ ЭВМ:

1. 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 / Ратников Е.С., Туманова А. Д., Бойчук Е. И., Лагутина Н. С., Лагутина К. В. // Свидетельство о государственной регистрации программы для ЭВМ № 2019619380 от 16 июля 2019.

2. Программа для статистического анализа стилометрических и ритмических характеристик текстов на русском, английском, французском и испанском языках / Манахова А. М., Лагутина К. В., Лагутина Н. С. // Свидетельство о государственной регистрации программы для ЭВМ № 2020618648 от 30 июля 2020.

3. Программа для автоматического выделения из текстов стилометрических характеристик различных уровней и классификации текстов по авторам / Лагутина К. В. // Свидетельство о государственной регистрации программы для ЭВМ № 2021616718 от 26 апреля 2021.

4. Программный прототип для автоматического выявления качественных параметров стиля текстов / Лагутина К. В. // Свидетельство о государственной регистрации программы для ЭВМ № 2021664205 от 1 сентября 2021

5. Программный прототип для обработки ритма текстов, сравнительно-сопоставительного анализа ритмики в их переводе и авторизации текстов / Лагутина К. В., Лагутина Н. С., Бойчук Е. И. // Свидетельство о государственной регистрации программы для ЭВМ № 2021664248 от 2 сентября 2021.

Диссертационная работа была выполнена при поддержке гранта РФФИ №20-37-90045.

Содержание работы

Во **введении** обосновывается актуальность исследований, проводимых в рамках данной диссертационной работы, формулируется цель, ставятся задачи работы, излагается научная новизна и практическая значимость представляемой работы, приводятся новые научные результаты, выносимые на защиту.

Первая глава посвящена обзору и анализу стилометрических средств в тексте, используемых для атрибуции авторства, проверки авторства, составления профиля автора, обнаружения изменения стиля и классификации текстов по жанру и тональности. Методы решения данных задач основаны на предположении, что можно выявить особенности текста, которые подтверждают авторство [6].

Стилометрические средства можно разделить на две категории: простые статистические средства, для подсчёта которых текст рассматривается как набор символов или слов, и сложные лингвистические, поиск которых требует знаний о языке.

Простые статистические (или стандартные) средства включают в себя характеристики уровней символов и слов. Как показывает анализ современной литературы, они наиболее просто и быстро вычисляются и используются намного чаще других [7].

На уровне символов текст представляется в виде последовательности символов, тогда как сами стилометрические средства образуют простейшую структуру текста. Обычно в качестве средства берется n-грамма, определённая как непрерывная последовательность из n элементов из данного фрагмента текста.

На уровне слов текст часто рассматривается как «мешок слов», независимо от порядка слов, грамматики или контекста. В этом случае измеряются частота появления слова, средняя длина слова, n-граммы слова и статистические характеристики словарного запаса.

К стандартным стилометрическим средствам также относятся и эмбеддинги символов и слов, которые основываются на описанных выше простых статистических средствах.

Сложные лингвистические средства включают в себя синтаксические, ритмические, тематические, семантические и другие средства.

Синтаксические средства основаны на структуре предложения. Одни из самых простых и распространенных — частота появления знаков препинания, длина и средняя длина предложения и частота появления служебных слов. Более сложные характеристики включают в себя особенности синтаксического дерева предложения.

Для поиска ритма в тексте выделяются лексико-грамматические средства, например, анафора, эпифора или апозиопеза, основанные на

повторении слов или знаков препинания, или фонетические средства, например, аллитерация и ассонанс, основанные на повторении звуков.

Тематические средства базируются на выделении ключевых слов и анализе их встречаемости.

Семантические средства основываются на отношениях между словами: синонимических, ассоциативных и т. п.

Набор стилометрических средств, используемых в компьютерной лингвистике, очень велик и неоднороден. Однако исследователи уделяют недостаточно внимания систематизации этих средств, изучению их влияния на качество решения задач и обоснованию выбора средств для решения конкретной задачи. Большинство авторов сравнивают алгоритмические подходы экспериментально [8]. Гораздо реже исследователи ставят задачу изучения влияния различных стилометрических средств на качество классификации текста по авторскому стилю [9]. Почти никто из исследователей не изучает причины, по которым средства или группы средств релевантны и эффективны.

Сравнивая исследования с наибольшими показателями качества (около 90 % и выше) алгоритмов с различными категориями стилометрических средств, можно сделать вывод, что подобные результаты чаще всего достигаются при одном или нескольких из следующих условий:

— сравнительно небольшой корпус текстов (не более 200–250 текстов), где тексты достаточно большие;

— тексты принадлежат небольшому числу авторов, обычно 10 или меньше;

— анализируется большое количество текстов конкретного автора, затем для этого автора получается один из лучших результатов классификации;

— исследователи успешно отобрали стилистические средства, по которым классификатор принимает решения, и наборы средств могут отличаться для текстов с разными темами и жанрами.

Кроме того, исследователи чаще всего принимают во внимание только некоторые особенности идиолекта или лингвистическую специфику авторского стиля, которые состоят, как правило, в числовых значениях средств низкого уровня, таких как количество слов, слогов, размер предложения и т. д. Однако идиостиль выражается и в лингвистических средствах, которые довольно сложны для поиска и связаны с личностью автора. Дополнительная сложность заключается в том, что нет таксономии или контрольного списка элементов индивидуального стиля, поскольку любые слова, слововочетания или знаки препинания могут быть эле-

ментами индивидуального стиля, если они последовательно используется таким образом, чтобы способствовать выражению личности автора [10].

Таким образом, реализация комплексного анализа индивидуального стиля автора является довольно сложной задачей, для решения которой выявление и анализ ритмических средств изучены недостаточно. Формализация и автоматизация поиска этих средств — первый шаг к всестороннему пониманию стиля отдельного автора.

Вторая глава посвящена разработке алгоритмов поиска ритмических средств и их реализации в приложении ProseRhythmDetector. Программный инструмент предназначен для автоматического выявления повторяющихся лексико-грамматических фигур и их визуализации.

Существующие инструменты оказываются ориентированными на анализ ритма текста на фонетическом, лексическом и/или синтаксическом уровнях или на решение конкретной задачи практически без анализа промежуточных шагов и лингвистической интерпретации [11; 12]. Новизна инструмента ProseRhythmDetector заключается в его способности искать и обрабатывать стилистические фигуры на основе повторения, а также визуализировать их, предоставляя эксперту-лингвисту возможность изучать как ритм текста в целом, так и отдельные его аспекты.

В данной работе ритмические характеристики текста определяются на основе повтора слов и знаков препинания в определенной конфигурации, в определенной позиции, с определенным количеством повторяющихся элементов, в соответствии с их определениями в классической лингвистике. ProseRhythmDetector выделяет следующие ритмические средства:

1. Анафора — скрепление речевых отрезков (частей фразы, стихов) с помощью повтора слова или словосочетания в начальной позиции. Пример: «**Все было** так же. **Все было** в том же самом грозно-знакомом виде».

2. Эпифора — скрепление речевых отрезков (частей фразы, стихов) с помощью повтора слова или словосочетания в конечной позиции. Пример: «Ось бублики, маковники, вертычки, буханци **хороши**! Ей-богу, **хороши**!»

3. Симплока — фигура синтаксического параллелизма в смежных стихах или фразах, у которых одинаковые начало и конец при разной середине. Пример: «**Во поле** берёза **стояла**. **Во поле** кудрявая **стояла**».

4. Анадиплозис — риторическая фигура, в которой следующее предложение начинается теми же словами, которыми оканчивается предыдущее. Пример: «А **сам Обломов**? **Сам Обломов** был полным и

естественным отражением и выражением того покоя, довольства и безмятежной тишины».

5. Эпаналепсис — фигура речи, состоящая в повторении одного и того же слова или словосочетания с небольшими вариациями в конце и начале фразы. Пример: «**Видит** ли он это или не **видит**?»

6. Многосоюзие — стилистическая фигура, состоящая в намеренном увеличении количества союзов в предложении, обычно для связи однородных членов. Пример: «Тут обыкновенно говорилось обо всем: **и** о том, **кто** пошил себе новые шаровары, **и** что находится внутри земли, **и кто** видел волка».

7. Диакопа — риторический термин для повторения слова или фразы, разбитых на одно или несколько промежуточных слов. Пример: «Каждый вынул из кармана своего **деревянную** ложку, иные, за неимением, **деревянную** спичку».

8. Эпизевксис — фигура речи, которая обозначает повторение слов без разрыва между повторениями. Пример: «**Пошли! пошли!**»

9. Хиазм — фигура речи, состоящая в изменении грамматической структуры в последовательных фразах или предложениях с повторением слов. Пример: «**Ты забыл**, что **ты** хотел **помнить**, и **ты помнишь**, что **ты** хотел **забыть**».

10. Апозиопеза — фигура речи, состоящая во внезапном прерывании высказывания. Пример: «Подумать только, чтобы он**...** Но довольно! Чтобы я когда-нибудь**...**»

11. Повторяющиеся вопросительные предложения — фигура речи, состоящая в повторении вопросительных знаков в концах соседних предложений. Пример: «А он что**?** Что он сделал**?**»

12. Повторяющиеся восклицательные предложения — фигура речи, состоящая в повторении восклицательных знаков в концах соседних предложений. Пример: «Какой чудесный день**!** Какой чудесный я**!**»

Выбор данных средств для анализа ритма, а именно для их автоматизированного поиска и количественной обработки обусловлен тем, что это ритмические средства, употребляемые в прозаических текстах наиболее часто, и именно они выделяются в качестве ритмических средств на лексико-грамматическом уровне большинством лингвистов, проводящих исследования в области ритмизации текста.

Для того, чтобы проанализировать ритм прозаического произведения, был разработан комплекс алгоритмов, автоматически находящих в тексте ритмические средства, а именно, лексические и синтаксические.

Входными данными для всех алгоритмов является необработанный текст. Каждый текст представляется как набор упорядоченных предложений, состоящих из слов и знаков препинания. В процессе работы алгоритмов последовательно перебираются предложения, в них находятся повторения слов и знаков препинания. Повторения данных элементов, подходящие под определения ритмических средств, вносятся в списки аспектов. Эти списки возвращаются в качестве выходных данных.

Качество алгоритмов поиска средств оценили эксперты-лингвисты. Четыре исследователя обработали в общей сложности 24 текста разных авторов, случайно выбранных из корпуса. Каждый эксперт работал 16 часов. Они вручную оценили точность поиска всех ритмических средств. Исключением является диакопа, поскольку для неё ProseRhythmDetector обнаружил несколько тысяч ритмических средств, поэтому эксперты проверяли из них только случайные 10 %. Эксперты заключили, что точность поиска средств достигает 80–95 % для всех ритмических средств.

Приложение ProseRhythmDetector реализовано на языке программирования Python с помощью библиотеки Stanza для обработки текста. Оно доступно по ссылке `https://github.com/text-processing/prose-rhythm-detector`.

Таким образом, данный инструмент позволяет быстро, достаточно точно и полностью автоматически выявить ритмические средства даже для текстов большого объёма. Это существенно ускоряет работу эксперта-лингвиста при сравнении авторского стиля текстов и позволяет ставить крупные эксперименты по анализу ритма крупных корпусов текстов, что было бы практически невозможно выполнить без подобной автоматизации.

Третья глава посвящена исследованию того, как комплекс ритмических средств может использоваться для автоматизированных экспериментов с анализом стиля автора в прозе. Исследователь осуществляет автоматический поиск этих средств в художественной литературе и статистический анализ их появления в XIX–XXI веках. Ритмические характеристики сравниваются со стандартными характеристиками уровней слов и символов. Эксперименты проводятся с алгоритмами поиска средств, реализованными в инструменте ProseRhythmDetector.

Для ритмических средств были выбраны следующие числовые стилометрические характеристики:

— количество появлений в тексте конкретного средства, делённое на количество предложений;

— количество появлений в тексте всех средств, делённое на количество предложений;

— доля уникальных слов среди всех, составляющих средства, в данном случае тех, которые повторяются только один раз;

— доли существительных, прилагательных, глаголов и наречий среди слов, составляющих средства.

Выбор данных средств для анализа ритма, а именно, для их автоматизированного поиска и количественной обработки, обусловлен тем, что они выделяются на лексико-грамматическом уровне в качестве ритмических средств большинством лингвистов, проводящих исследования в области ритмизации текста.

В качестве стилометрических характеристик на уровне символов и слов были выбраны нижеперечисленные характеристики.

На уровне символов:

— доли отдельных букв среди общего количества букв;

— доли отдельных знаков препинания среди общего количества знаков препинания;

— средняя длина предложения в символах.

На уровне слов:

— средняя длина предложений по количеству слов;

— средняя длина слова;

— частоты топ-40 n-граммов для $n = 1, 2, 3$. Для каждой униграммы, биграммы или триграммы вычисляется количество вхождений в корпус текстов, а затем выбирается 40 наиболее часто встречающихся униграмм, биграмм и триграмм. Для каждого текста также вычисляется их количество появлений и делится на общее количество вхождений данных 120 n-грамм в текст.

Выбор данных стилометрических средств на уровне символов и слов был обусловлен тем, что они являются наиболее показательными при определении авторского стиля во время исследования произведения.

Стилометрические характеристики трёх разных уровней вычисляются и визуализируются автоматически. Эксперименты с этими характеристиками были поставлены следующим образом.

— Сначала в текстах были выявлены ритмические средства.

— Для выявленных ритмических средств были подсчитаны стилометрические характеристики.

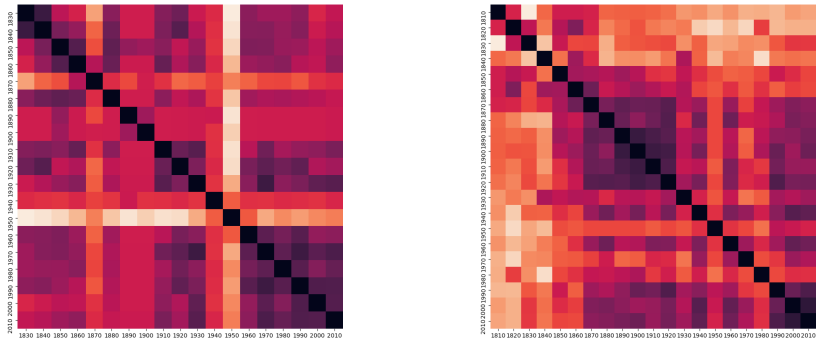
— Параллельно с подсчётом характеристик ритма для текстов были вычислены стилометрические характеристики уровня слов и символов.

— Стилометрические характеристики текстов были агрегированы по десятилетиям, десятилетия сравнивались между собой.

— На последнем этапе результаты сравения были визуализированы с помощью тепловых карт и графиков.

Были проведены эксперименты с корпусами текстов на английском, русском, французском и испанском языках.

Каждый корпус включает в себя по 240 произведений более 90 известных авторов. У каждого из текстов указана дата публикации с 1815 по 2019 год. Каждый текст содержит в себе до 425 000 слов.

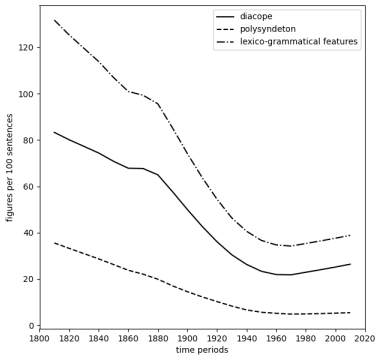


а) б)

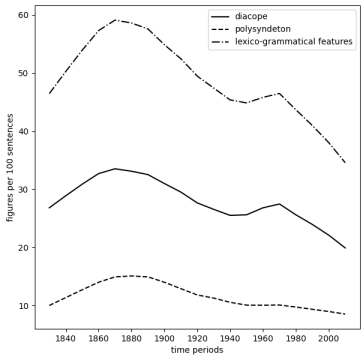
Рис. 1 — Все ритмические средства по десятилетиям, метрика Чебышёва, для а) русского языка, б) английского языка

Тепловые карты показывают большие кластеры десятилетий с похожим ритмом для текстов XIX-го века и конца XX-го — начала XXI-го века. Кроме того, можно выделить маленькие кластеры с 2–3 десятилетиями с близкими ритмическими характеристиками.

Если сравнивать графики и тепловые карты за десятилетия, можно сделать вывод, что расстояние Чебышёва работает хорошо и позволяет выделять кластеры, только когда количество средств достаточно велико.



а)



б)

Рис. 2 — Ритмические средства по десятилетиям: все, диакопа, многосоюзие на а) английском языке, б) русском языке

Для XXI-го века, когда средства становятся более редкими, эта метрика бесполезна.

Таким образом, тепловые карты и графики показывают тенденции использования средств на протяжении десятилетий и веков, поэтому ритмические средства могут быть полезными индикаторами изменения стиля.

Количественный анализ позволяет разделить набор ритмических средств на две группы по частоте встречаемости: частые (диакопа, многосоюзие) и редкие (анафора, эпифора, анадиплозис). Выявленное уменьшение общего количества средств наблюдается только в первой группе. Количества редких средств этой тенденции не соответствуют. Поэтому можно сделать вывод, что наиболее распространенные ритмические средства являются наиболее полезными для определения времени написания текста.

Эксперименты показали, что, хотя десятилетия можно успешно кластеризовать по близости друг к другу, каждое из них является уникальным по совокупности ритмических и простых стилометрических характеристик. Это значит, что на основе модели, построенной на данных характеристиках, тексты можно успешно классифицировать по векам и десятилетиям создания/публикации.

Четвертая глава посвящена двум исследовательским задачам: (1) автоматической классификации художественной литературы XIX–XXI веков по периодам их публикации с использованием ритмических харак-

теристик и (2) сравнению качества классификации трех типов стилометрических характеристик: уровня символов, слов и ритма. Такая классификация может дать объяснение изменению и развитию стиля письменных текстов [13; 14].

Модель текста, то есть набор его стилометрических характеристик, была взята та же, что и в предыдущей главе: характеристики уровней символов, слов и ритма.

Стилометрические характеристики образуют модель стиля текста, которую можно использовать для классификации текстов. Тексты классифицируются по трём классам по дате их публикации: XIX, XX и XXI-й века.

Сначала из текстов на естественном языке извлекаются ритмические средства, для чего используются алгоритмы из второй главы. Точность поиска составляет 80–95 %. Затем вычисляются ритмические характеристики отдельно для каждого текста.

Параллельно для текстов вычисляются характеристики уровней символов и слов. Алгоритм находит самые часто встречающиеся в корпусе n-граммы, затем независимо для каждого текста вычисляет частоту встречаемости топа n-грамм и другие низкоуровневые характеристики.

После расчета характеристик для каждого текста результаты объединяются в общую матрицу. Таким образом каждый текст представляется как вектор числовых стилометрических характеристик.

Векторы стилометрических характеристик применяются в качестве входных данных для четырёх классификаторов вида «обучение с учителем»: AdaBoost, RandomForest, двунаправленная LSTM, нейронная сеть GRU. Эти четыре алгоритма часто демонстрируют высокое качество классификации текста [15; 16], поэтому они были выбраны для экспериментов.

Все алгоритмы обучаются на половине корпусов текстов. Размеры корпусов невелики, поэтому классификаторы тестируются на значительной части выборки. Для обучения нейронных сетей применяется категориальная кросс-энтропия как функция потерь и алгоритм оптимизации Adam.

Результаты тестовой фазы классификации оценивались по четырём общим метрикам: доля правильных ответов, точность, полнота и F-мера.

Диссертант провела эксперименты с четырьмя корпусами текстов на английском, русском, французском и испанском языках из 3 главы.

Для всех языков нейронные сети превосходят мета-классификаторы Random Forest и AdaBoost. Точность и F-мера для мета-классификаторов составляет менее 80 %, в то время как нейронные сети обеспечивают от 82 до 89 % в лучших случаях.

В таблицах 1, 2, 3 и 4 представлены результаты классификации по векам русских и французских текстов для отдельных типов характеристик и комбинаций типов характеристик с лучшими результатами. Столбцы A, P, R, F содержат значения следующих метрик качества: доля правлильных ответов (accuracy), точность (precision), полнота (recall) и F-мера (F-measure). «Символы» означает признаки на уровне символов, «Слова» — на уровне слов, «Ритм» — ритмические, + обозначает комбинацию двух типов признаков, Все — комбинацию трех типов признаков. Точность, полнота и F-мера рассчитываются как среднее арифметическое для всех авторов. Полужирным отмечены строки с лучшим качеством верификации и лучшие значения F-меры.

Таблица 1

Классификация англоязычной прозы по векам

Классификатор	Уровень	A	P	R	F
LSTM	Символы	74.1	75.4	73.5	74.4
LSTM	Слова	70.7	69.2	69.2	69.2
LSTM	Ритм	70.0	70.5	70.9	70.7
LSTM	Слова + Ритм	**86.0**	**85.9**	**85.7**	**85.8**
LSTM	Все	**89.5**	**89.8**	**89.5**	**89.6**

Таблица 2

Классификация русскоязычной прозы по векам

Классификатор	Уровень	A	P	R	F
GRU	Символы	66.1	63.3	65.9	64.6
GRU	Слова	74.6	74.3	74.3	74.3
GRU	Ритм	68.3	69.1	70.6	69.8
GRU	Все	**88.1**	**88.1**	**88.7**	**88.4**

Подводя итог, можно сказать, что обнаружены одинаковые тенденции для классификации по векам. Для русского и английского языков чем больше типов характеристик мы используем, тем выше качество классификации. Для французского и испанского языков комбинация характеристик уровня символов и ритма даёт лучшие результаты, чем другие пары типов характеристик. Кроме того, среди отдельных типов характеристик уровни символов и слов обеспечивают более высокие результаты, чем уровень ритма. Но и сами ритмические характеристики достигают довольно хороших результатов с долей правильных ответов от 65 до 91 % для разных языков.

Таблица 3

Классификация франкоязычной прозы по векам

Классификатор	Уровень	A	P	R	F
LSTM	Символы	78.9	75.5	75.2	75.4
LSTM	Слова	76.3	78.4	74.9	76.6
LSTM	Ритм	65.8	61.2	60.3	60.7
LSTM	Символы + Ритм	**86.8**	**86.9**	**85.6**	**86.3**
LSTM	Все	**84.2**	**82.6**	**82.6**	**82.6**

Таблица 4

Классификация испаноязычной прозы по векам

Классификатор	Уровень	A	P	R	F
LSTM	Символы	94.1	91.5	94.2	92.8
LSTM	Слова	92.6	89.4	93.2	91.3
LSTM	Ритм	92.6	89.7	91.2	90.4
LSTM	Символы + Ритм	**95.6**	**90.6**	**97.3**	**93.8**
LSTM	Все	**95.6**	**90.6**	**97.3**	**93.8**

Пятая глава посвящена верификации авторства художественных текстов. В ней анализируются ритмические средства и популярные низкоуровневые характеристики, основанные на статистике элементов текста. Сравнение проводится на корпусах литературных текстов на английском, русском, французском и испанском языках.

Анализ современных исследований показывает, что сравнение различных типов характеристик с лингвистическими осуществляется редко, особенно для художественных текстов [17]. Авторы обычно полагаются на стандартные статистические характеристики, основанные на словах и символах, и пытаются расширить их за счет относительно небольшого числа синтаксических, тематических или других лингвистических средств. Глубинные языковые характеристики остаются малоизученными, скорее всего, из-за сложности поиска. Хотя такие стилистические средства напрямую определяют авторский стиль [4] и могут интерпретироваться лучше остальных, поэтому оказываются перспективными для исследования.

Диссертант сравнивает три типа характеристик: уровней символов, слова и ритма, — те же, что и в предыдущих главах.

После извлечения параметров стиля из текстов получается матрица, в которой строки — это тексты конкретных авторов, а столбцы — это

конкретные характеристики. Каждый автор верифицируется отдельно, с использованием всей матрицы для авторов текстов на конкретном языке. Его тексты помечаются как принадлежащие или не принадлежащие ему. Затем выполняется бинарная классификация.

Сравниваются два классификатора: AdaBoost и BiLSTM. Они уже доказали свое качество в решении современных задач классификации текстов, как показано в предыдущей главе. Для оценки устойчивости классификаторов применяется техника пятикратной кросс-валидации. Тексты делятся на пять частей, 80% составляют обучающую выборку, 20% — тестовую.

Диссертант сравнивала художественные тексты на четырех языках: английском, русском, французском и испанском. Корпуса были созданы вручную из фрагментов произведений известных авторов, написанных на их родном языке.

Для того, чтобы тексты были одинаковыми по размеру, из каждого прозаического текста было извлечено от 1 до 4 фрагментов размером около 50 000 знаков, включая пробелы. Таким образом, каждый автор представлен 40 фрагментами текста. Корпуса на английском, русском и французском языках содержат тексты 20 известных авторов XIX–XXI веков, по 800 текстов в корпусе. В испанском корпусе есть тексты 8 авторов XIX–XX веков, всего 320 текстов.

Сравнение двух классификаторов показало, что AdaBoost превосходит нейронную сеть на 10–15% по точности, полноте и F-мере. Скорее всего, это связано с тем, что обучающая выборка имеет недостаточный размер для более эффективной работы нейронной сети LSTM. Таким образом, таблицы в этом разделе содержат качество классификации для алгоритма AdaBoost.

В таблице 5 описывается качество верификации авторства для всех типов характеристик и их комбинаций. Характеристики уровня ритма обеспечивают хорошее качество классификации. В большинстве случаев оно ниже на 3–11% F-меры, но имеет довольно высокие значения 78–87%. Кроме того, количество ритмических характеристик в несколько раз меньше, чем характеристик уровней символов и слов, поэтому относительно небольшое количество специфических параметров стиля позволяет добиться значительно высокого качества верификации авторства. Комбинация типов характеристик улучшает качество на 2–14%.

Таким образом, все типы характеристик могут обеспечить хорошее качество верификации авторов. Специфические лингвистические характеристики — ритмические — во многих случаях обеспечивают высокую точность, полноту и F-меру с небольшими стандартными отклонениями. Таким образом, они являются такими же полезными и стабильны-

Таблица 5

Средние значения мер для верификации авторства

Язык	Тип характеристики	Точность	Полнота	F-мера
Английский	Символы	87.8	80.7	84.1
Английский	Слова	85.8	78.2	81.8
Английский	Ритм	82.0	74.2	77.9
Английский	Все	**94.7**	**85.4**	**89.8**
Русский	Символы	91.2	81.4	86.0
Русский	Слова	92.0	81.9	86.7
Русский	Ритм	84.7	76.7	80.5
Русский	Все	**96.9**	**87.4**	**91.9**
Французский	Символы	93.7	86.5	90.0
Французский	Слова	91.8	80.1	85.6
Французский	Ритм	83.5	75.9	79.5
Французский	Все	**97.5**	**90.0**	**93.6**
Испанский	Символы	89.9	85.0	87.4
Испанский	Слова	92.3	87.9	90.1
Испанский	Ритм	88.5	86.3	87.4
Испанский	Все	**94.1**	**90.0**	**92.0**

ми маркерами стиля, как и стандартные статистические характеристики уровня символов и слов.

Верификация конкретных авторов показывает, что многие авторы используют один и тот же стиль в разных фрагментах текстов. Их можно успешно отделить от других, используя только один тип характеристик или комбинацию стандартных и ритмических характеристик. Тем не менее, тексты некоторых авторов верифицируются с очень высокими стандартными отклонениями, поэтому требуются другие лингвистические средства для надежной верификации их текстов.

В **заключении** приведены основные результаты работы:

1. Разработаны алгоритмы и программный инструмент для автоматического поиска и визуализации лексико-грамматических ритмических средств в прозаических текстах для русского, английского, французского и испанского языков. Инструмент позволил автоматизировать работу эксперта-лингвиста по анализу ритма авторского текста, а также провести эксперименты по автоматическому выявлению ритма текста для построения модели ритма текста.

2. Показано, что ритмические характеристики могут быть индикатором стиля эпохи на основе статистических экспериментов с прозаическими текстами XIX–XXI века большого объема.

3. Продемонстрирована высокая эффективность модели ритма текста для решения задачи классификации художественной литературы XIX–XXI веков по векам и полувекам их публикации. Показано, что ритмические характеристики в комбинации со стандартными характеристиками повышают качество решения данной задачи.

4. Показано, что ритмические характеристики являются самостоятельными маркерами индивидуального авторского стиля и по качеству определения автора близки к стандартным характеристикам, а в комбинации с ними повышают эффективность верификации авторов.

Список литературы

1. *Bergman D. J.*, *Bergman C. C.* Elements of stylish teaching: Lessons from Strunk and White // Phi Delta Kappan. — 2010. — Vol. 91, no. 4. — P. 28–31.
2. *Neal T.* [et al.]. Surveying stylometry techniques and applications // ACM Computing Surveys (CSUR). — 2018. — Vol. 50, no. 6. — P. 86.
3. *Rudman J.* The state of authorship attribution studies: Some problems and solutions // Computers and the Humanities. — 1997. — Vol. 31, no. 4. — P. 351–365.
4. *Boychuk E.* [et al.]. Automated approach for rhythm analysis of French literary texts // Proceedings of 15th Conference of Open Innovations Association FRUCT. — IEEE. 2014. — P. 15–23.
5. *Freyermuth S.* Poétique de la prose ou prose poétique? Le rythme contre le prosaısme // Questions de style, Vous avez dit prose? — 2009. — P. 67–80. — (in French).
6. *Stamatatos E.* A survey of modern authorship attribution methods // Journal of the American Society for information Science and Technology. — 2009. — Vol. 60, no. 3. — P. 538–556.
7. *Lagutina K.* [et al.]. A Survey on Stylometric Text Features // Proceedings of the 25th Conference of Open Innovations Association FRUCT. — IEEE. 2019. — P. 184–195.
8. *Kestemont M.* [et al.]. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection // Working Notes Papers of the CLEF 2018 Evaluation Labs. CEUR Workshop Proceedings. Vol. 2125. — 2018. — P. 1–25.

9. *Plecháč P.*, *Bobenhausen K.*, *Hammerich B.* Versification and authorship attribution. A pilot study on Czech, German, Spanish, and English poetry // Studia Metrica et Poetica. — 2018. — Vol. 5, no. 2. — P. 29–54.

10. *Robinson J.* General and individual style in literature // The Journal of aesthetics and art criticism. — 1984. — Vol. 43, no. 2. — P. 147–158.

11. *Balint M.*, *Trausan-Matu S.* A critical comparison of rhythm In music and natural language // Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information. — 2016. — Vol. 9, no. 1. — P. 43–60.

12. *Niculescu I.-D.*, *Trausan-Matu S.* Rhythm analysis in chats using Natural Language Processing // Proceedings of the 14th International Conference on Human-Computer Interaction RoCHI'2017. — 2017. — P. 69–74.

13. *Rubino R.* [et al.]. Modeling Diachronic Change in Scientific Writing with Information Density // Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. — 2016. — P. 750–761.

14. *Rowlett J. L.* Ralph Cohen on Literary Periods: Afterword as Foreword // New Literary History. — 2019. — Vol. 50, no. 1. — P. 129–139.

15. *Kowsari K.* [et al.]. Text classification algorithms: A survey // Information. — 2019. — Vol. 10, no. 4. — 150 (1–68).

16. *Nowak J.*, *Taspinar A.*, *Scherer R.* LSTM recurrent neural networks for short text and sentiment classification // International Conference on Artificial Intelligence and Soft Computing. — Springer. 2017. — P. 553–562.

17. *Lim C.-G.*, *Jeong Y.-S.*, *Choi H.-J.* Survey of Temporal Information Extraction. // Journal of Information Processing Systems. — 2019. — Vol. 15, no. 4. — P. 931–956.