MODELING AND ANALYSIS OF INFORMATION SYSTEMS, VOL. 27, NO. 3, 2020
journal homepage: www.mais-journal.ru

THEORY OF DATA

Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th–21st Centuries

K. V. Lagutina[1], A. M. Manakhova[1]

DOI: 10.18255/1818-1015-2020-3-330-343

[1]P. G. Demidov Yaroslavl State University, Sovetskaya str., 14, Yaroslavl, 150003, Russia.

MSC2020: 68T50
Research article
Full text in Russian

Received May 14, 2020
After revision June 8, 2020
Accepted June 10, 2020

The article is devoted to comparison of stylometric features of several levels, which are markers of the style of the prose text and analysis of the stylistic changes in Russian and British prose of the 19th–21st centuries. Stylometric features include the low-level features based on the words and symbols and high-level based on rhythmic. These features model the style of a text and are the indicators of the time when the text was created.

Calculations of all the features are performed completely automatically, so it allows to conduct the large-scale experiments with artworks of a large volume and speeds up the work of a linguist. To calculate the stylometric features including ones based on the search results for rhythmic figures the ProseRhythmDetector program is used. As a result of its work, each text is presented as a set of the same features of three levels: characters, words, rhythm. Texts are combined by decades, for each decade there are found average values of stylometric features. The obtained models of decades are compared using standard similarity metrics, results of comparison are visualized in the form of the heat maps and dendrograms. Experiments with two corpora of Russian and British texts show that during the 19th–21st centuries there are general trends in style change for both corpora, for example, a decrease in the number of rhythmic figures per sentence, and also particular trends for each language, for example, dynamics of change of the word and sentence lengths. Stylometric features of all levels reveal the similarity in the style of texts published in one century. Also, features of three levels in the complex better demonstrate the uniqueness of each decade than features of a particular level. This study shows the importance of stylometric features as style markers of the different eras and allows us to identify trends in style during several centuries.

Keywords: text rhythm; rhythm analysis; natural language processing; stylometry; rhythm figures; automation.

INFORMATION ABOUT THE AUTHORS

Ksenia Vladimirovna Lagutina correspondence author	orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru postgraduate student.
Alla Mikhajlovna Manakhova	orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru student.

Funding: The study was funded by RFBR according to the research project №19-07-00243.

For citation: K. V. Lagutina and A. M. Manakhova, “Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th–21st Centuries”, *Modeling and analysis of information systems*, vol. 27, no. 3, pp. 330-343, 2020.

МОДЕЛИРОВАНИЕ И АНАЛИЗ ИНФОРМАЦИОННЫХ СИСТЕМ, ТОМ 27, № 3, 2020
сайт журнала: www.mais-journal.ru

THEORY OF DATA

Автоматизированный поиск и анализ стилометрических характеристик, описывающих стиль прозы 19–21 веков

К. В. Лагутина[1], А. М. Манахова[1]

DOI: 10.18255/1818-1015-2020-3-330-343

[1]Ярославский государственный университет им. П. Г. Демидова, ул. Советская, 14, г. Ярославль, 150003 Россия.

УДК 004.912
Научная статья
Полный текст на русском языке

Получена 14 мая 2020 г.
После доработки 8 июня 2020 г.
Принята к публикации 10 июня 2020 г.

Статья посвящена сравнению стилометрических характеристик нескольких уровней, являющихся маркерами стиля прозаического текста, и анализу стилистических изменений русской и британской прозы 19–21 веков. Стилометрические характеристики включают в себя низкоуровневые характеристики, основанные на словах и символах, и высокоуровневые — ритмические. Подобные характеристики моделируют стиль текста и являются индикаторами времени его создания.

Вычисление всех характеристик происходит полностью автоматически, что позволяет проводить крупные эксперименты с художественными произведениями большого объёма и ускоряет работу эксперта-лингвиста. Для подсчёта стилометрических характеристик, в том числе основанных на результатах поиска ритмических средств, используется программа ProseRhythmDetector. В результате её работы каждый текст представляется в виде набора одних и тех же характеристик трёх уровней: символов, слов, ритма. Тексты объединяются по десятилетиям, для каждого десятилетия находятся средние значения стилометрических характеристик. Полученные модели десятилетий сравниваются при помощи стандартных метрик близости, результаты сравнения визуализируются в виде тепловых карт и дендрограмм. Эксперименты с двумя корпусами русских и британских текстов показывают, что в течение 19–21 веков появляются как общие тенденции изменения стиля для обоих корпусов, например, уменьшение количества ритмических средств в расчёте на одно предложение, так и собственные для каждого языка, например, динамика изменения длин слов и предложений. Стилометрические характеристики всех уровней выявляют схожесть стиля текстов, опубликованных в одном веке. Также характеристики трёх уровней в комплексе лучше демонстрируют уникальность каждого десятилетия, чем характеристики конкретного уровня. Это исследование показывает значимость стилометрических характеристик как маркеров стиля различных эпох и позволяет выявить тенденции изменения стиля на протяжении нескольких веков.

Ключевые слова: ритм текста; анализ ритма; обработка естественного языка; стилометрия; ритмические средства; автоматизация.

ИНФОРМАЦИЯ ОБ АВТОРАХ

Ксения Владимировна Лагутина
автор для корреспонденции
orcid.org/0000-0002-1742-3240. E-mail: lagutinakv@mail.ru
аспирант.

Алла Михайловна Манахова
orcid.org/0000-0001-7429-3529. E-mail: al.mnkhv@yandex.ru
студент.

Финансирование: Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта №19-07-00243.

Для цитирования: K. V. Lagutina and A. M. Manakhova, “Automated Search and Analysis of the Stylometric Features that Describe the Style of the Prose 19th–21st Centuries”, *Modeling and analysis of information systems*, vol. 27, no. 3, pp. 330-343, 2020.

Введение

Ритм текста определяется как регулярное повторение схожих единиц речи [1]. В литературе отмечается, что ритм прозы отличается от ритма поэзии и требует собственных методик определения, в том числе таких, которые бы моделировали его в количественном виде и позволяли бы сравнивать прозаические тексты между собой [2].

Количественные характеристики ритма текста входят в обширную группу стилометрических характеристик, которые на разных уровнях описывают стиль текста и составляют его числовую модель [3], на основе которой можно проводить статистический и сравнительный анализ текстов различных авторов, жанров, временных эпох и т.п.

Несмотря на широкую распространённость стилометрических характеристик в сфере обработки естественного языка, они, как правило, используются для решения конкретных задач классификации, например, определения автора или жанра. Более того, из всего спектра характеристик наиболее изученными являются низкоуровневые, которые моделируют текст на уровне слов или символов. Высокоуровневые или лингвистические характеристики, включая ритмические, являются малоизученной частью стиля текста [4].

Авторы поставили перед собой задачу сравнить высокоуровневые и низкоуровневые стилометрические характеристики в прозаических текстах 19–21 веков по десятилетиям и статистически проанализировать динамику изменения стиля русской и британской прозы. Для этого была использована программа ProseRhythmDetector, которая была представлена в предыдущих исследованиях как инструмент для поиска ритмических средств в прозаических текстах. Авторы добавили в ProseRhythmDetector модуль, который вычисляет стилометрические характеристики текста уровня символов, слов и ритма. Данная статья описывает эксперименты с этим модулем по анализу текстов по векам и десятилетиям.

1. Обзор смежных работ

Стилометрия – научная дисциплина, занимающаяся измерением стилевых характеристик текстов с целью их упорядочивания, диагностики, идентификации, параметризации, таксономии, атрибуции и периодизации [5]. Стилометрические характеристики прозаических текстов изменяются с течением времени для литературы на разных языках, поэтому они могут служить индикаторами эпохи создания произведений [6]. Кумар и другие исследуют лексические маркеры стиля для 18–21 веков, по которым можно определить дату публикации. Используя характеристики текста уровня слов и фраз, они достигают средней ошибки в 32 года, т.е. достаточно точно определяют век публикации текста.

Для решения подобных задач обычно берутся простые характеристики уровня слов и символов. Авторы статьи [7] строили именно такую модель, но по результатам исследования рекомендовали расширять модели текстов более сложными лингвистическими характеристиками.

Контест Semeval 2015 [8] был посвящён определению временного периода публикации статьи в диапазоне от 18 до 21 века. Лучших результатов до 86.8 % точности достигли участники, которые применили широкий спектр стилометрических характеристик: от лексических до грамматических, включая даже мета-свойства документа.

Гопиди и Алам [9] показали, что числовые грамматические характеристики, а также характеристики, основанные на ударениях и рифме, отличаются для разных 50-летий для прозы и поэзии.

Эти результаты доказывают, что стилометрические характеристики различных уровней хорошо моделируют стиль текста и могут указывать на конкретную эпоху его создания.

Стилометрические характеристики отличаются не только для разных эпох, но и для разных языков. Авторы работы [10] кластеризовали тексты при помощи алгоритма k-средних, основываясь на встречаемости слов и символов. F-мера для такого алгоритма классификации получилась не выше

53 %. С применением нейросетей классификация текстов по языкам на основе характеристик уровня символов и слов может достигать более высоких значений F-меры 70–80 % [11]. Но такие исследования затрагивают только низкоуровневые характеристики, оставляя открытым вопрос значимости лингвистических характеристик для моделирования и анализа стиля.

Авторы в предыдущей работе [12] исследовали вариативность ритмических характеристик для различных периодов времени (19–21 веков) и языков (русского и английского), где показали, что все три века отличаются по ритму. В этой статье описываются результаты сравнительного анализа стилометрических характеристик нескольких уровней.

2. Стилометрические характеристики

Стилометрический анализ текста включает в себя поиск и подсчёт различных стилометрических характеристик. Среди таких характеристик можно выделить несколько категорий:

1. Уровень символов;
2. Уровень слов;
3. Уровень ритма.

Ритмические характеристики текста определяются исходя из употребления средств создания ритма, в основе которых находится повтор в определенной конфигурации, в определенной позиции, с определенным количеством повторяющихся элементов. Для данной работы были выбраны следующие ритмические средства:

1. Анафора — скрепление речевых отрезков (частей фразы, стихов) с помощью повтора слова или словосочетания в начальной позиции.
2. Эпифора — скрепление речевых отрезков (частей фразы, стихов) с помощью повтора слова или словосочетания в конечной позиции.
3. Симплока — фигура синтаксического параллелизма в смежных стихах или фразах, у которых одинаковые начало и конец при разной середине или наоборот, разные начало и конец при одинаковой середине.
4. Анадиплозис — риторическая фигура, в которой следующее предложение начинается теми же словами, которыми оканчивается предыдущее.
5. Эпаналепсис — фигура речи, состоящая в повторении одного и того же слова или словосочетания с небольшими вариациями.
6. Многосоюзие — стилистическая фигура, состоящая в намеренном увеличении количества союзов в предложении, обычно для связи однородных членов.
7. Диакопа — риторический термин для повторения слова или фразы, разбитых на одно или несколько промежуточных слов
8. Эпизевксис — фигура речи, которая обозначает повторение слов без разрыва между повторениями.

Для данных ритмических средств были выбраны следующие числовые стилометрические характеристики:

1. Количество появлений в тексте конкретного средства, делённое на количество предложений;
2. Количество появлений в тексте всех средств, делённое на количество предложений;
3. Доля уникальных слов среди всех, составляющих средства, в данном случае тех, которые повторяются только один раз;
4. 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 это наиболее частотные ритмические средства, употребляемые в прозаических текстах. Именно они выделяются в качестве ритмических средств на лексико-грамматическом уровне большинством лингвистов, проводящих исследования в области ритмизации текста [12].

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

На уровне символов:

1. Количество букв, как отдельных, так и их общее количество;
2. Количество символов, как отдельных, так и их общее количество;
3. Средняя длина предложения в символах.

На уровне слов:

1. Количество слов;
2. Количество предложений;
3. Средняя длина предложений по количеству слов;
4. Средняя длина слова.

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 [4].

3. Постановка экспериментов

3.1. Основные этапы экспериментов

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

- Сначала в текстах были выявлены ритмические средства. Алгоритмы поиска ритмических средств были взяты из статьи [12].
- 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 сравнения были визуализированы с помощью тепловых карт и иерархической кластеризации.

На первом этапе ритмические средства выявляются с точностью 80–95 %. Таким образом авторы получают качественную модель ритма текста.

Ритмические и простые стилометрические характеристики вычисляются на основе текста и модели его ритма по точным правилам, описанным в предыдущем разделе. В результате каждый текст представляется как вектор числовых характеристик. Вектора сравниваются при помощи мер близости, на основе которых организована визуализация результатов.

3.2. Визуализация стилометрических характеристик

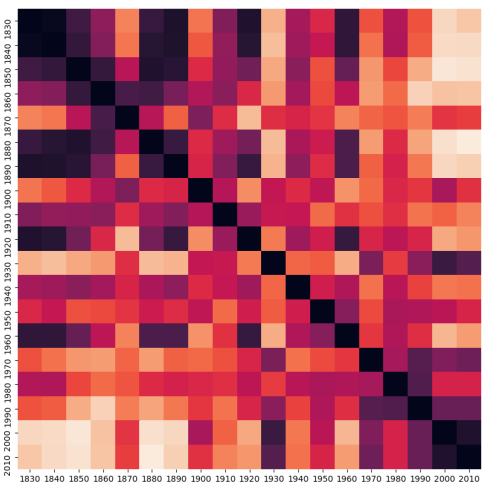
После того как вычислены стилометрические характеристики для текстов, считаются стилометрические характеристики для десятилетий. Для каждого десятилетия берутся средние значения характеристик текстов, опубликованных в этот период. Таким образом получаются вектора характеристик десятилетий такого же типа, как и вектора для отдельных текстов.

Стилометрические характеристики десятилетий и их сравнение визуализируются тремя способами:

- В виде тепловых карт, которые описывают близость десятилетий по стилю. Это квадратные тепловые карты, на осях которых расположены десятилетия, а оттенок в ячейке обозначает степень близости пары десятилетий: чем темнее оттенок, тем ближе объекты друг к другу. В качестве меры близости использовались четыре популярные метрики: расстояние Чебышёва, коэффициент корреляции, расстояние Евклида и манхэттенское расстояние.
- В виде тепловых карт, которые описывают диапазоны значений стилометрических характеристик. На горизонтальной оси располагаются названия конкретных характеристик, на вертикальной — десятилетия. Ячейки карты содержат значение характеристики, а также имеют цвет, оттенок которого обозначает величину значения относительно других. Самые большие значения обозначаются светлыми оттенками, самые маленькие — тёмными. Справа на карте отображается столбик с диапазоном значений и оттенками для разных значений.
- В виде дендрограмм, полученных в результате кластеризации. Листья дендрограммы — это десятилетия, они размещены по горизонтали. По вертикали отмечаются расстояния между кластерами в виде горизонтальных отрезков на определённом уровне. Дендрограмма строится с помощью агломеративного подхода, от листьев к стволу. Метрики близости используются те же, что и для тепловых карт. В качестве функций расстояния между кластерами применяются три метода: одиночной, средней и полной связи.

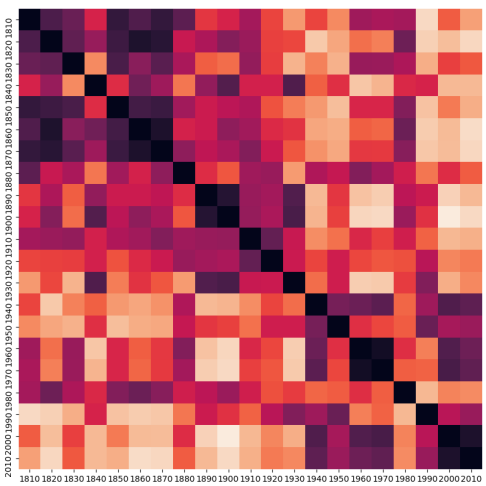
Для сравнения десятилетий стилометрические характеристики предварительно нормализуются: из конкретного значения вычитается среднее значение данной характеристики по всему корпусу текстов, полученная разность делится на среднеквадратическое отклонение данной характеристики.

Все три способа визуализации достаточно наглядны и позволяют проанализировать как динамику изменения стилометрических характеристик на протяжении десятилетий, так и близость десятилетий друг к другу с точки зрения стиля.



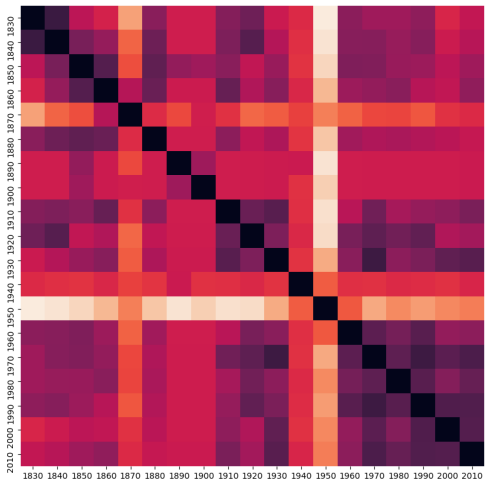
a)

Fig. 1. All levels, correlation metric, a) Russian, b) English



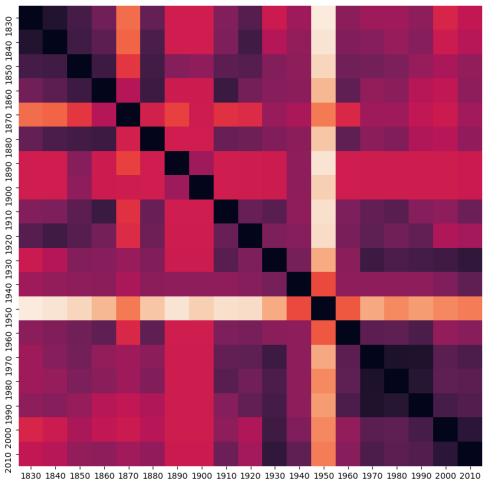
b)

Рис. 1. Все уровни вместе, метрика корреляции, a) русский язык, b) английский язык



a)

Fig. 2. a) All levels, b) symbol level, Chebyshev metric, Russian



b)

Рис. 2. a) Все уровни вместе, b) уровень символов, метрика Чебышёва, русский язык

4. Эксперименты

4.1. Программная реализация и корпус

Инструмент ProseRhythmDetector [1], позволяющий выявлять и подсчитывать стилометрические характеристики, был разработан на языке Python. При разработке также использовалась библиотека textblob, которая была особенно полезна при подсчёте слов и предложений в тексте.

После завершения разработки был проведен ряд экспериментов на основе двух корпусов текстов. Один из них на английском языке, а другой — на русском. Каждый корпус включает в себя по 243 произведения более 90 известных авторов. У каждого из текстов указана дата публикации: для текстов на английском языке с 1815 по 2019 гг., а для текстов на русском языке с 1832 по 2019 гг. Каждый текст содержит в себе до 425 000 слов.

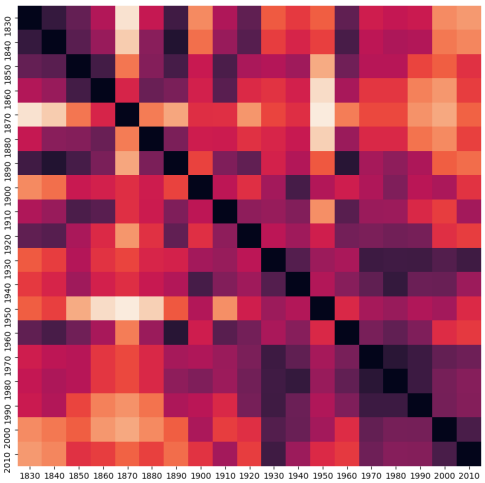
4.2. Тепловые карты близости

На основе получившихся векторов характеристик по корпусу текстов были построены 4 набора тепловых карт по метрикам близости Чебышёва, Евклида, корреляции и манхэттенского расстояния. Метрики Евклида и манхэттенское расстояние показали практически идентичные результаты на всех рассматриваемых уровнях. Относительно заметные различия наблюдаются только при объединении всех уровней.

Метрика корреляции не сильно помогла в анализе литературы за указанный период, однако она продемонстрировала близость 00-х и 10-х годов 21-ого века на обоих корпусах (см. Рис. 1). Кроме того на тепловой карте, построенной на основе корпуса русских текстов, можно отметить особую близость 1830-х и 1840-х годов. Также по карте видно, что 21 век более далёк от 19-го века, чем от 20-го, а ещё можно увидеть, что 1960-е и 1920-е годы близки к началу 19-го столетия. Наконец, по карте можно сделать вывод, что литература конца 19-го века была весьма похожа на литературу начала 19-го века, в то время как произведения середины 19-го столетия сильно отличаются от них.

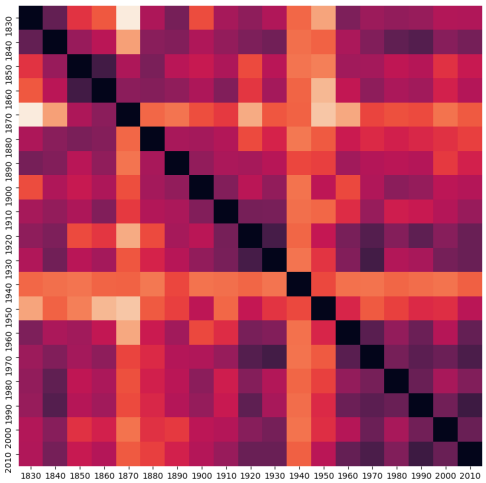
По тепловой карте, построенной на основе корпуса английских текстов с расчётом по метрике корреляции можно понять, что произведения большей части 19-го века близки между собой. Но,

[1]Инструмент доступен по ссылке: https://github.com/text-processing/prose-rhythm-detector



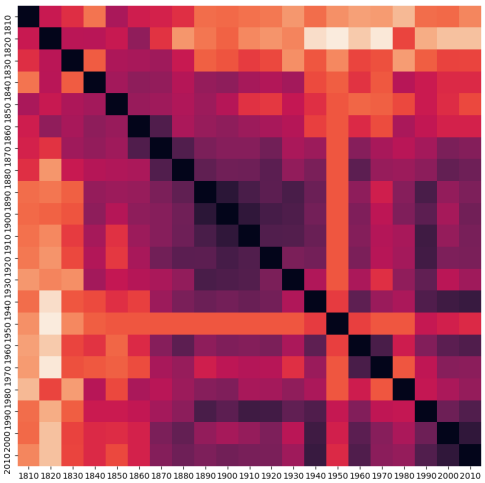
a)

Fig. 3. a) Word level, b) rhythm level, Chebyshev metric, Russian



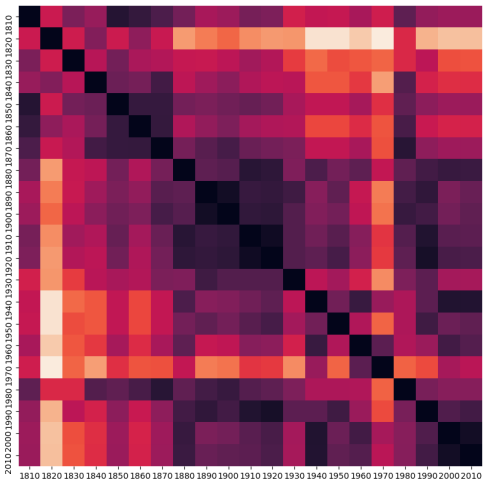
b)

Рис. 3. a) Уровень слов, b) уровень ритма, метрика Чебышёва, русский язык



a)

Fig. 4. a) Symbol level, b) word level, Chebyshev metric, English

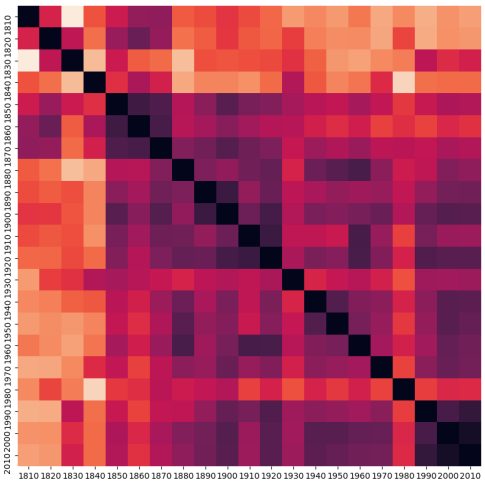


b)

Рис. 4. a) Уровень символов, b) уровень слов, метрика Чебышёва, английский язык

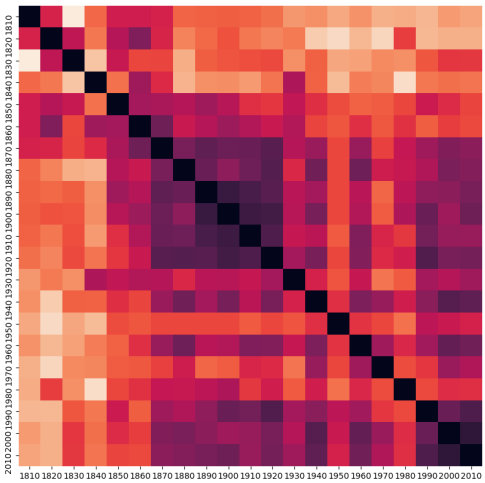
как и в случае с русской литературой, английская литература середины 19-го века выделяется на фоне всего остального столетия, пусть и не так сильно.

Наиболее показательными оказались результаты, полученные при расчёте по метрике Чебышёва. На картах, построенных по корпусу русского языка, отчётливо видно, как выделяются 1950-е годы (см. Рис. 2a)). Они отличаются от остальных периодов в литературе на всех уровнях, но особый вклад в это различие вносит уровень символов (см. Рис. 2b)) — все произведения рассматриваемой эпохи имеют сходство на уровне символов, в то время как 1950-е годы разительно отличаются от



a)

Fig. 5. a) Rhythm level, b) all levels, Chebyshev metric, English



b)

Рис. 5. a) Уровень ритма, b) все уровни вместе, метрика Чебышёва, английский язык

литературы всех периодов. На уровне слов (см. Рис. 3a)) можно выделить отличие текстов 21-го века от 19-го, а также стоит отметить различие 1950-х от второй половины 19-го века. Кроме того, на всех уровнях можно выделить схожесть 30-х и 40-х годов 19-го века. 1870-е года выделяются на фоне всей литературы 19-го века на всех уровнях, а уровень ритма (см. Рис. 3b)) придаёт им отличие от всей литературы за рассматриваемый период. В заключение, на уровнях символов и слов можно выделить близость второй половины 20-го века с началом 21-го века, однако уровень ритма слегка нивелирует это сходство.

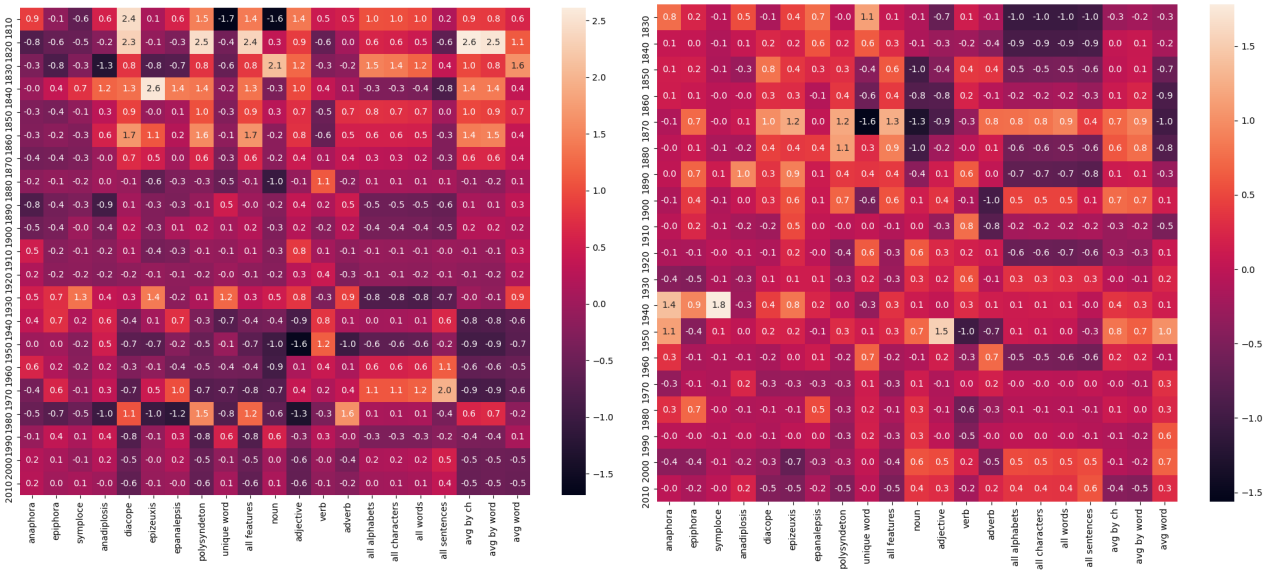
Эксперименты на корпусе английских текстов показали следующие результаты. На уровне символов (см. Рис. 4a)) можно выделить близость 1890-х и 1900-х годов, а также 2000-х и 2010-х. Также на этом уровне особенно отличаются 1950-е годы по сравнению со всеми другими десятилетиями. На уровне слов (см. Рис. 4b)) уже более ярко выделяется период с 1890-х по 1920-е годы, а также 21-й век. Уровень ритма (см. Рис. 5a)) способствует выделению периода с 1990-х по 2010-е годы, а также периода с 1850-х по 1970-е годы и с 1890-х по 1900-е годы. Таким образом, на карте, отражающей все характеристики сразу (см. Рис. 5b)), мы четко можем выделить интервал с 1870-х по 1920-е годы, а также 21-й век и 1990-е годы.

4.3. Тепловые карты диапазонов

Второй тип тепловых карт (Рис. 6) отображает нормализованные значения характеристик (разность между реальным и средним значением по всему корпусу, делённую на среднеквадратическое отклонение) для обоих языков.

Первые 14 столбцов обозначают ритмические характеристики: 9 конкретных средств, общее количество появлений средств, доля слов, повторяющихся только один раз (unique word), доли частей речи. Остальные 7 — это несколько характеристик низкого уровня: количество букв, всех символов, слов и предложений, а также средние длины предложений в символах (avg by ch) и словах (avg by word) и средние длины слов (avg by word).

Для ритмических средств обоих языков повторяется тенденция, выявленная в предыдущем исследовании на меньших корпусах текстов: на протяжении веков общее число ритмических характеристик снижается. Это видно по оттенкам цветов на карте: для диакопы и многосоюзия



a) b)

Fig. 6. Heatmap for a) English, b) Russian texts with normalized features

Рис. 6. Тепловая карта для a) английских, b) русских текстов с нормализованными значениями характеристик

(polysyndeton), как наиболее частых средств, и суммарного числа средств в 19 веке оттенки более светлые, а в конце 20-го — начале 21-го — более тёмные. Это значит, что в 19-м веке эти характеристики имеют значения выше среднего, а ближе к нашему времени — ниже среднего.

В британских текстах эта тенденция коррелирует с употреблением прилагательных: доля их появлений в ритмических средствах также снижается к 21-му веку. Доли остальных частей речи колеблются не так значительно. Для русских текстов закономерность другая: доли существительных и прилагательных в 20–21 веках возрастают.

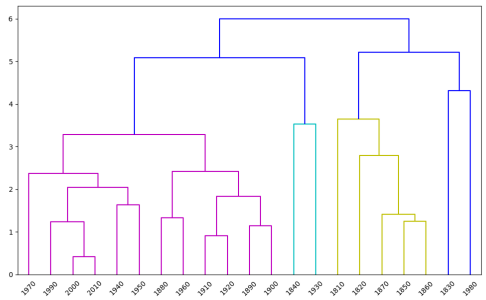
Что касается простых стилометрических характеристик, то для британских текстов они показывают, как изменяется средний размер художественных произведений: постепенно уменьшается к середине 21-го века, затем снова возрастает. Для русских текстов тенденция обратная. Эти закономерности можно скорее отнести к особенностям формирования корпуса: были выбраны известные художественные произведения популярных авторов. При увеличении корпуса и добавлении более разнообразных произведений тенденции могут измениться.

Средние длины предложений и слов представляют собой характеристики, лучше отражающие стиль текстов, чем абсолютные количества элементов текста. Для британских текстов все они уменьшаются в течение практически всех десятилетий. Средняя длина предложений в русских текстах увеличивается к концу 19-го века — началу 20-го века, немного уменьшается в первой половине 20-го века, увеличивается к 1950-м годам, затем снова уменьшается. Средняя длина слова увеличивается на протяжении всех десятилетий.

Как и по тепловым картам близости, по картам диапазонов можно видеть десятилетия, выделяющиеся на фоне остальных. Причём на картах диапазонов дополнительно можно обнаружить, по каким стилометрическим характеристикам десятилетия отличаются. Для британских текстов это 1930 и 1980 годы. Для русских — 1870 и 1940–1950 гг.

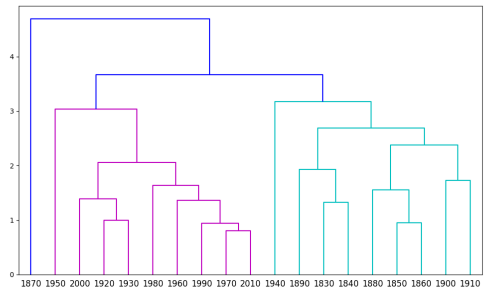
Таким образом, тепловые карты диапазонов позволяют как выявлять общие тенденции в изменении стиля произведений для языка в целом, так и обнаруживать отдельные десятилетия, выделяющиеся среди остальных.

4.4. Дендрограммы



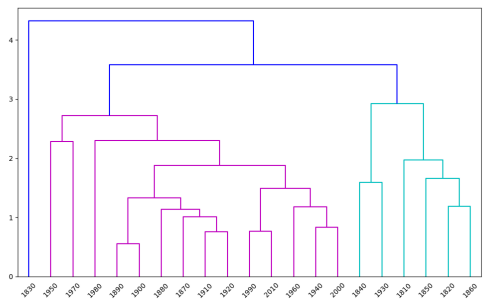
a)

Fig. 7. Dendrogram for a) English, b) Russian texts based on rhythm features, the Euclidean distance and complete-linkage method



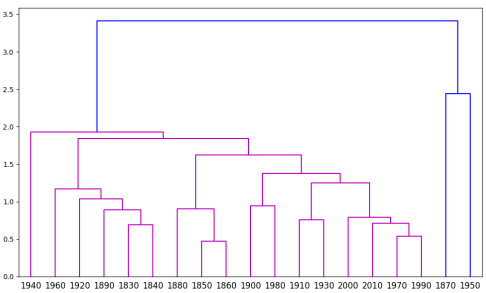
b)

Рис. 7. Дендрограмма для a) английских, b) русских текстов на основе ритмических средств, расстояния Евклида и метода полной связи



a)

Fig. 8. Dendrogram for a) English, b) Russian texts based on all features, the Chebyshev distance and complete-linkage method



b)

Рис. 8. Дендрограмма для a) английских, b) русских текстов на основе всех средств, расстояния Чебышёва и метода полной связи

Дендрограммы строились для обоих языков отдельно. Для каждого языка тексты кластеризовались иерархически как на основе отдельных типов стилометрических характеристик, так и на основе всех типов характеристик, чтобы сравнить разделение текстов только по ритму с разделением по всем стилометрическим характеристикам.

Среди функций расстояния между кластерами самые наглядные результаты показал метод полной связи, метод средней связи показал близкие к нему результаты. Метод одиночной связи выявил меньше кластеров, чем остальные.

Что касается метрик близости между элементами, то манхэттенское расстояние и расстояние Евклида показали близкие результаты, как и в случае с тепловыми картами. Коэффициент корреляции показал более хаотичное разбиение на кластеры, чем другие методы.

Для ритмических средств наиболее полезной оказалась метрика Евклида (см. Рис. 7).

Для британских текстов по ритму явно выделяются несколько небольших кластеров, содержащих соседние десятилетия: 1990–2010 гг., 1940–1950 гг., 1890–1920 гг., 1850–1870 гг. 1830 и 1980 десятилетия оказались самыми далёкими по ритму от остальных. 21 век наиболее похож на середину 20-го века: 1940–1950 гг. и 1970-й год.

Для русских текстов дендрограмма показывает меньшие расстояния по ритму между десятилетиями, чем для британских текстов. По ней видны два больших кластера, первый содержит большую часть десятилетий 20–21 веков, второй – 19-й век и начало 20-го: 1900–1910 гг. 21-й век не так явно выделяется по ритму, как в британских текстах. Самыми далёкими от остальных оказываются 1870-е и 1940-е десятилетия.

Для всех стилометрических средств наиболее наглядные результаты показала метрика Чебышёва (см. Рис. 8).

Для британских текстов по стилю некоторые пары соседних десятилетий снова оказываются близки друг к другу: 1890 и 1900, 1880 и 1870, 1910 и 1920 и т.д. В целом расстояния между десятилетиями оказываются меньше, в кластере 19-го века оказываются начало и середина этого столетия. 21-й век снова близок по стилю к середине 20-го века.

Для русских текстов по стилю 21-й век выделяется в отдельный кластер, но в этот кластер также попадают 1970-е годы. 19 век оказывается разбит на 2 более мелких кластера и более похож на 20-й век. 1940 и 1870 снова оказываются самыми далёкими от остальных, и к ним присоединяются 1950 гг.

Таким образом, дендрограммы показывают, что по ритмическим характеристикам века отличаются сильнее, чем по совокупности стилометрических характеристик. 19-й век и 1990–2010 годы могут выделяться в отдельные кластеры, 20-й век оказывается куда менее однородным как по ритму, так и по более простым стилометрическим характеристикам.

5. Обсуждение результатов

Автоматическое определение комплекса низкоуровневых и высокоуровневых стилометрических характеристик даёт возможность быстро проанализировать большое количество объёмных произведений и сделать качественные выводы об изменении стиля с течением времени. Этот подход позволяет эксперту за короткое время получить детализированную модель стиля художественного 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 потенциально можно успешно классифицировать по векам и десятилетиям создания/публикации, а также вычислять год публикации текста.

Кроме того, анализ текстов средствами кластеризации на основе тепловых карт и дендрограмм позволяет выявлять тенденции изменения стиля в литературе в целом, а также сравнивать литературы на разных языках между собой. В частности, для русской и британской литератур детектируется уменьшение числа ритмических средств на единицу текста (в данном случае предложение). Для русской литературы обнаружено, что средние длины предложений изменяются волнообразно в течение рассматриваемого периода, в то время как средние длины слов увеличиваются. В британской литературе средние длины как слов, так и предложений существенно уменьшаются. Если сравнивать между собой века обеих литератур, то 20-й оказывается самым разнородным с точки зрения стиля, а 21-й и 19-й отличаются между собой, но десятилетия внутри них достаточно похожи.

Помимо поиска общих тенденций, тепловые карты и дендрограммы позволяют обнаружить конкретные периоды времени, которые значительно отличаются от других. Это может быть интерпретировано как то, что в данный период попали один или несколько текстов, которые особенно сильно отличаются по стилю от современников. Таким образом в крупном корпусе текстов можно выявлять произведения с уникальным стилем.

Если сравнить между собой значимость стилометрических характеристик разных уровней, то можно сделать вывод, что и низкоуровневые, и ритмические достаточно полезны и могут обнаруживать одни и те же крупные кластеры десятилетий. Однако ритмические характеристики более разнородны, поэтому являются лучшими индикаторами уникальности стиля.

Таким образом, автоматизированное моделирование текстов при помощи стилометрических характеристик различных уровней позволяет анализировать и успешно сравнивать между собой литературы разных языков и эпохи их развития.

Заключение

Авторы провели исследование стилометрических характеристик разных уровней: символов, слов и ритма на двух корпусах художественных произведений русской и английской литературы 19–21 веков. Стилометрические характеристики считались полностью автоматически при помощи инструмента ProseRhythmDetector[2]. Он позволяет автоматически находить и статистически обрабатывать ритмические средства в комплексе с простыми стилометрическими характеристиками текста, что даёт возможность проводить анализ стиля прозы с разных сторон и исследовать большие корпуса текстов.

Анализ стилометрических характеристик позволил выявить как основные тенденции изменения стиля на протяжении 19–21 веков, так и обнаружить периоды времени, наиболее отличающиеся от других по ритму и стилю текстов. Кроме того, исследование показало значимость ритмических характеристик как маркеров особенностей стиля прозы.

Следующим этапом исследования стилометрических характеристик текста, в том числе и ритмических, может быть их использование для классификации текстов по веку или эпохе публикации, определения авторства, а также анализ и сравнение стилей литератур других языков.

References

[1] E. Boychuk, I. Paramonov, N. Kozhemyakin, and N. Kasatkina, “Automated approach for rhythm analysis of French literary texts”, in *Proceedings of 15th Conference of Open Innovations Association FRUCT*, IEEE, 2014, pp. 15–23.

[2] N. Golubeva-Monatkina, “On the Problem of Prose Rhythm”, *The Bulletin of the Russian Academy of Sciences: Studies in Literature and Language*, vol. 76, no. 2, pp. 16–27, 2017, In Russian.

[3] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, “Surveying stylometry techniques and applications”, *ACM Computing Surveys (CSUR)*, vol. 50, no. 6, p. 86, 2018.

[4] K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, and I. Paramonov, “A Survey on Stylometric Text Features”, in *Proceedings of the 25th Conference of Open Innovations Association FRUCT*, IEEE, 2019, pp. 184–195.

[5] Martynenko G. Ya., “Metody matematicheskoj lingvistiki v stilisticheskih issledovaniyah”, In Russian, Nestor-Istoriya, 2019.

[6] A. Kumar, M. Lease, and J. Baldridge, “Supervised language modeling for temporal resolution of texts”, in *Proceedings of the 20th ACM international conference on Information and knowledge management*, 2011, pp. 2069–2072.

[7] A. Jatowt and R. Campos, “Interactive system for reasoning about document age”, in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, 2017, pp. 2471–2474.

[8] O. Popescu and C. Strapparava, “Semeval 2015, task 7: Diachronic text evaluation”, in *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015, pp. 870–878.

[2]Инструмент опубликован по ссылке: https://github.com/text-processing/prose-rhythm-detector

[9] A. Gopidi and A. Alam, "Computational Analysis of the Historical Changes in Poetry and Prose", in *Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change*, 2019, pp. 14–22.

[10] H. Lan and J. Huang, "Chinese-English Cross-Lingual Text Clustering Algorithm based on Latent Semantic Analysis", *Proceedings of Science*, pp. 1–7, 2017.

[11] A. Esuli, A. Moreo, and F. Sebastiani, "Funnelling: A New Ensemble Method for Heterogeneous Transfer Learning and Its Application to Cross-Lingual Text Classification", *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 3, pp. 1–30, 2019.

[12] K. Lagutina, A. Poletaev, N. Lagutina, E. Boychuk, and I. Paramonov, "Automatic Extraction of Rhythm Figures and Analysis of Their Dynamics in Prose of 19th-21st Centuries", in *26th Conference of Open Innovations Association (FRUCT)*, IEEE, 2020, pp. 247–255.