Вычислительные технологии, 2023, том 28, № 3, с. 136–166. © ФИЦ ИВТ, 2023 ISSN 1560-7534
Computational Technologies, 2023, vol. 28, no. 3, pp. 136–166. © FRC ICT, 2023 eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2023.28.3.009

Информационные системы анализа поэтических текстов: история, методы и алгоритмы

О. Ю. Кожемякина
Федеральный исследовательский центр информационных и вычислительных технологий, 630090, Новосибирск, Россия
Контактный автор: Кожемякина Ольга Юрьевна, e-mail: olgakozhemyakina@mail.ru
Поступила 01 мая 2023 г., доработана 03 мая 2023 г., принята в печать 05 мая 2023 г.

Разработка информационных систем анализа поэтических текстов является сложной задачей, в решении которой используются методы и алгоритмы, как принадлежащие наследию классической математики, так и самые современные, связанные с машинным обучением. В статье представлен исторический обзор существующих систем, их компонентов и используемых методов, современных исследований, использующих наследие классических методов, но получивших новые возможности благодаря развитию информационных технологий. Изучение поэтических текстов с использованием методов искусственного интеллекта, методов теории информации является перспективным направлением в задачах обработки естественного языка.

Ключевые слова: обработка естественного языка, проектирование программной системы, информационная система анализа поэтического текста, методы анализа текста, алгоритмы анализа текста.

Цитирование: Кожемякина О.Ю. Информационные системы анализа поэтических текстов: история, методы и алгоритмы. 2023; 28(3):136–166. DOI:10.25743/ICT.2023.28.3.009.

> Математизация отнюдь не сводится к выражению явлений в числах, таблицах и графиках. Числа, таблицы и графики могут вообще отсутствовать. Главное в математизации — это создание такого описания явления, которое было бы безупречным с логической точки зрения, а математика выступает здесь в роли оценщика (и одновременно идеала) степени логической безупречности.
>
> *В.А. Успенский. Предварение для читателей "Нового литературного обозрения" к семиотическим посланиям Андрея Николаевича Колмогорова [1].*

Введение

В Федеральном исследовательском центре информационных и вычислительных технологий разработана программная система автоматизированного комплексного анализа русских поэтических текстов [2]. Этот проект является оригинальным, поскольку в настоящее время не существует подобных комплексных систем, связанных с автоматизированным анализом поэтических текстов на русском языке и апробированных

в научных публикациях. Однако в основе создания системы лежат разработки, методы, алгоритмы, как принадлежащие наследию классической математики, так и самые современные, связанные с машинным обучением.

В рамках наших исследований термин "информационная система" получает уточненное определение как соответствующий компонент программной системы, объединяющий разнородную информацию о результатах анализа поэтических текстов, полученных на каждом из уровней анализа. В статье [3] подробно описано концептуальное проектирование системы: процесс проектирования и реализации информационной системы представления результатов анализа поэтических текстов, включающий в себя формулировку задач, которые должна решать информационная система, а также изложение требований в порядке приоритета для общего проекта программной системы, полноценная реализация информационной системы обеспечивает существенное упрощение исследований поэтических текстов для филологов-экспертов.

Текст, даже в виде произвольного сообщения, — структура, иерархичная по своей природе. В поэтических текстах уровни структуры сообщения отображаются в уровни структуры стиха, которые также имеют устойчивую иерархию [4]: метрика, ритмика, фонетика, лексика, грамматика, литературный стиль, тематика, литературный жанр. Уровни структуры произвольного сообщения и стиха однозначно сопоставляются: к статистическому уровню относится фонетика, к синтаксическому уровню — метрика и ритмика, к семантическому — лексика и грамматика, на границе семантического и прагматического уровней находится тематика [3]. В целом, процесс анализа поэтического текста состоит из рассмотрения уровней его структуры как самостоятельных смысловых единиц с последующим связыванием полученных данных с другими ее элементами, однако выбор методов, так же как и разработка и реализация алгоритмов, должны быть ориентированы на специфику каждого уровня. Так, например, фонетика и метроритмика являются объектом, скорее, методов классических, в частности методов математической статистики, но очевидно, что семантика и прагматика текста в силу своей специфики подлежат анализу с применением методов машинного обучения.

Алгоритмы и системы анализа фонетического и метроритмического уровней поэтических текстов, предлагаемые в зарубежных работах, весьма зависят от особенностей конкретного языка и, что вполне ожидаемо, не могут быть успешно применены в анализе русских поэтических текстов. Разработка универсального инструмента для автоматического анализа поэтических текстов также затруднительна по причине особенностей строя конкретных языков, о чем мы писали в наших ранних работах [5]. Определенные исследования показывают потенциальную возможность анализа для языков с некоторыми идентичными признаками, однако инструменты, которые анализируют более одного языка, являются редким исключением и зависят от определенной близости языков. Такие особенности русского языка, как отсутствие общих правил морфологического изменения слов, их фонетических характеристик, в первую очередь акцентуации, делают автоматизацию анализа русских текстов нетривиальной задачей. Однако нами предприняты достаточно успешные попытки проверить алгоритмы разработанной системы на другом языке, для чего был выбран казахский язык, поскольку данному языку присущи соответствующая написанию фонетика, фиксированное ударение, а также отсутствие редукции (изменение звуковых характеристик, вызванное их безударным положением). В работах [6–8] предлагаются алгоритмы стемматизации и генерации существительных, прилагательных и глаголов казахского языка, что полностью решает задачу анализа и синтеза словоформ для научно-технических текстов на казахском языке.

Определяющим качеством является факт отсутствия ошибок при тестировании на словах, принадлежащих различным частям речи, что, в свою очередь, позволяет заявить о корректности предложенных алгоритмов [8]. Таким образом, алгоритмы продемонстрировали работоспособность даже для языка из другой языковой семьи, хотя это, подчеркнем еще раз, исключение, и важную роль в этом сыграла общность в структуре языка.

Исторический обзор систем анализа поэтических текстов и их компонентов с акцентами на методах и алгоритмах, использованных в предшествующих исследованиях, подчеркивает пионерный характер системы, разработанной в ФИЦ ИВТ. Работы, описанные в статье, разделены на группы соответственно языкам, на которых написаны поэтические тексты. Проанализированы также современные исследования, использующие наследие классических методов, но получившие новые возможности благодаря развитию информационных технологий. Настоящая статья дает представление об алгоритмах, методах и информационных системах анализа поэтических текстов, об исторических основах методологии современной обработки естественного языка, что может быть полезно для специалистов, работающих в области обработки текстов на естественном языке.

1. Становление и развитие применения количественных методов в исследованиях русской поэзии

1.1. А.А. Марков

Андрей Андреевич Марков (старший) (1856–1922) — русский математик, занимавшийся алгеброй и теорией вероятностей. Одним из его главных достижений стали так называемые цепи Маркова (1907). Модель цепей Маркова построена на основе изучения последовательностей испытаний, цепью Маркова называется любая система, в которой переход из одного состояния в другое не зависит от предыстории процесса, а зависит только от текущего состояния. Цепь задается в виде ориентированного графа; вершины показывают состояния системы, ребра — переходы из одного состояния в другое, при этом с каждым переходом фиксируется его вероятность. Ученый внес существенный вклад в становление количественных методов анализа русской поэзии: в 1913 г. вышла статья “Опыт статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь”, в которой ученый определил частотность повторения гласных и согласных звуковых единиц в тексте романа А.С. Пушкина [9]. Марков выполнил пересчет всех букв “Евгения Онегина”, чтобы понять частотность следования букв друг за другом.

Математическая модель была описана следующим образом: пусть есть система с дискретными состояниями, время задается таким образом, что на каждом дискретном шаге времени происходит переход в одно из существующих состояний; исключаются и рассматриваются только те итерации, в которых происходят скачки состояний; каждое состояние описывается множеством условных вероятностей перехода в любое состояние системы. Такая модель, выраженная с помощью геометрической схемы (ориентированного графа), получила впоследствии название цепей А.А. Маркова, и это исследование является исторической основой метода определения авторства.

1.2. А.Н. Колмогоров

Андрей Николаевич Колмогоров (1903–1987) — выдающийся математик XX в., проявлявший глубокий интерес к гуманитарным наукам. Среди его исследований в 60-е гг. XX в. особое место занимают работы по анализу статистики речи и стиховедению. Исследования Колмогорова в этом направлении тесно связаны, с одной стороны, с вероятностным и алгоритмическим подходами к теории информации и, с другой стороны, отражают его давний интерес к анализу закономерностей, свойственных форме и языку литературных произведений [10]. Сохранились черновые машинописные странички из его письма Ю.М. Смирнову в Курск, названные А.Н. Колмогоровым "Замечаниями об основах русского стихосложения" (1943), в которых он впервые сформулировал основной принцип ударности классического русского стиха. Отчасти объяснение истокам этого интереса можно найти в строках того же письма: "...В Комаровке, когда Вы туда вновь попадете, Вы найдете книжку Валерия Брюсова "Наука о стихе"; в ней перечислены случаи, в которых "ипостаси" стоп одного размера стопами других размеров называются правильными. Эти традиционные правила очень сложны..." [11]. Глобальная идея Колмогорова заключается в том, что "энтропия речи" (т. е. мера количества информации, передаваемой речью) раскладывается на две компоненты: внеречевую (смысловую, семантическую) и собственно речевую (лингвистическую) информацию [12]. По определению В.А. Успенского, "первая из этих компонент характеризует разнообразие, позволяющее передавать различную смысловую информацию; вторая компонента, названная Колмогоровым "остаточной энтропией", характеризует разнообразие возможных способов выражения одной и той же или равносильной смысловой информации, ее наличие обеспечивает звуковую выразительность речи при передаче смысловой информации" [13]. Колмогоров дал расчет "затрат энтропии" на отдельные приемы звуковой выразительности стиха [14], однако соответствующий раздел работы А.Н. Колмогорова и А.В. Прохорова "Статистические методы исследования ритма стихотворной речи" полностью опубликован не был [15].

1.3. М.Л. Гаспаров

Михаил Леонович Гаспаров (1935–2005) — крупнейший отечественный филолог, литературовед, переводчик, автор многочисленных трудов по античной литературе, поэтике и стиховедению. Создатель истории стиха как отдельной дисциплины, истории поэтики и риторики, новых норм редактирования переводов и популяризации литературы. Анализируя произведения В.В. Маяковского [16], Гаспаров вывел статистику употребляемых Маяковским выразительных средств языка с указанием функций для каждого из них. В процессе исследования произведений А.С. Пушкина Гаспаров также применял статистические методы, которые, однако, позволили сделать исследователю выводы о чертах авторского стиля, что является чисто литературоведческой задачей. Так, лирика у Пушкина более "статична", чем эпос; "Кавказский пленник" "статичнее", чем реалистические "Евгений Онегин" и "Домик в Коломне"; проза Пушкина более "динамична" (в ней используется больше глаголов), чем его лирика и поэмы; "Песни западных славян" и сказки Пушкина "еще более насыщены глаголами", чем проза — сжатость и динамичность казались поэту приметами "народного" стиля. Сравнивая "Пиковую даму" и рассказ Л.Н. Толстого "Хозяин и работник", исследователь приходит к выводу, что Толстой изображает свой мир в детализованных действиях, а Пушкин свой — в сум-

марных (статистически в произведении Толстого больше глаголов, чем в произведении Пушкина) [17].

1.4. Дж.Т. Шоу

Джозеф Томас Шоу (1919–2011) — один из наиболее известных пушкинистов США, его первые пушкиноведческие работы относятся к началу 1950-х гг. На протяжении почти полувека Шоу опубликовал ряд трудов, ставших классическими: перевод и издание с комментариями трехтомного собрания писем Пушкина, словарь рифм Пушкина (1974), в словарь-справочник всех стихотворных произведений Пушкина (1985), вызвавший многочисленные отклики в научной печати [18]. Кроме этих фундаментальных работ Дж.Т. Шоу принадлежит множество статей, часть которых вошла в книгу, представляющую собой первый том собрания трудов ученого [19]. Огромным вкладом в понимание и анализ произведений Пушкина стал конкорданс (расширенный словарь языка писателя), составленный Дж.Т. Шоу [20, 21], для разработки которого был применен компьютер. Первое издание конкорданса на русском языке вышло в 1984 г., в том же году, когда книга впервые была издана в США. В.Е. Холшевников, работая над словарем языка Пушкина, отозвался о конкордансе Шоу: “Прежде всего, надо сказать, что Дж.Т. Шоу проделал большую и добросовестную работу: план словаря хорошо продуман и серьезно аргументирован (хотя, естественно, не бесспорен); материал тщательно обработан. По количеству сообщаемой информации словарь превышает известные мне словари рифм писателей. . . это стало возможным потому, что, как сообщает автор, он пользовался компьютером. Несомненно, словарь, составленный Дж.Т. Шоу, является весомым вкладом в науку” [22]. При этом Холшевников подчеркивает ряд отличий его словаря от работы Шоу: наличие хронологической характеристики, связи с жанром, размером, строфой; иная методика составления фонической характеристики; большое количество статистических данных (суммирующие таблицы и др.); сопоставление суммарных данных с аналогичными данными по словарям рифм других поэтов, которые составлены не по полным собраниям стихотворений, а по избранным, но в статистически достоверных выборках. Различие в этих пунктах дало основание Холшевникову продолжать свою работу по составлению собственного словаря, при этом имея возможность сравнивать свои статистические данные с работами Шоу. Исследования Дж.Т. Шоу показывают, что вполне возможно на ограниченном материале прийти к серьезным выводам, позволяющим по-новому взглянуть на старые произведения. Появление прогрессивных вычислительных средств и их использование в филологическом анализе позволили исследователю добиться результатов, признанных современными учеными-филологами.

Особо стоит отметить работу А. Миттманна (автора системы Aidos, о которой речь пойдет ниже) “Escansão automática de versos em português”, содержащую некоторый обзор существующих систем анализа стиха и представляющую собой, пожалуй, единственное подобное исследование [23]. В отношении русского языка А. Миттманн ссылается на работы И. Пильщикова [24–28], а также на наши исследования, проведенные в ФИЦ ИВТ [5, 29, 30], как на основные.

Классическая методология, применяемая учеными-теоретиками и практиками в области как математики, так и филологии, легла в основу разработанной в ФИЦ ИВТ системы комплексного анализа русских поэтических текстов и была дополнена современными методами компьютерной лингвистики. В сети Интернет в настоящее время можно найти программы, алгоритмы и сервисы, позволяющие производить анализ раз-

нородных текстов (в том числе поэтических) [31, 32], причем их число со временем только возрастает. Такая тенденция, прежде всего, связана с широким применением машинного обучения в задачах анализа текстов любого типа. Однако те инструменты анализа, которые находятся в свободном доступе в Интернете, не являются научными разработками, подтвержденными публикациями в научных изданиях.

2. Зарубежные системы по сбору, хранению и анализу результатов лингвистических исследований. Алгоритмы и методы систем анализа текста

2.1. Обзор работ по анализу поэтических текстов на латинском и греческом языках

Латинский и греческий языки обладают некоторой идентичностью, поэтому логично, что попытки создания универсального инструмента по работе с этими языками предпринимались учеными. Исследование Р. Мансиллы и Е. Буша [33] посвящено разработкам метода анализа увеличения сложности стиха от классической греческой поэзии к классической латинской поэзии путем сопоставления так называемых больших образцов этой поэзии с символическими временными рядами. Авторы определяют характеристики регулярной последовательности ритмов, т. е. структуру ударных и безударных слогов в стихе. Используя метод из теории информации, точнее энтропию Реньи (обобщение энтропии Шеннона), — меру количественного разнообразия, неопределенности или случайности некоторой системы [34], авторы демонстрируют, как ритмические паттерны в греческой поэзии эволюционируют к более сложным формам в латинской поэзии и делают вывод о том, что усложнение ритмической структуры от греческой поэзии к латинской наблюдается в гекзаметре. Метод, по утверждению авторов, позволяет различать греческие и латинские стихи, опираясь на различия в использовании дактиля и спондея, а также в положении и использовании цезур (ритмических пауз). Кроме этого, авторы подчеркивают, что полученные результаты могут помочь в решении так называемого Гомеровского вопроса — комплекса проблем, связанных с авторством и происхождением “Илиады” и “Одиссеи” [35], что имеет большую практическую ценность.

Д.А. Фуси в своих исследованиях [36, 37] создал базу для работы с классическим стихотворением и несколько версий специализированной системы Chiron, построенной на этой основе. Chiron построена на достаточном уровне абстракции, на котором возможна работа с несколькими различными языками (в данном случае латинским и греческим), метрами и текстами. В архитектуре системы заложен многокомпонентный подход: каждая функциональность реализована независимо и может быть заменена другой; компоненты могут быть разработаны любым пользователем для любой конкретной цели без изменения структуры; обеспечивается свобода выбора технологий для тех частей, которые с большей вероятностью могут измениться или которые обеспечивают максимальную совместимость с существующими. Сам анализ выполняется посредством прохода данных через цепочку слоев алгоритма, при этом “потомок” системы может использовать результаты всех его “предков” и чем выше уровень иерархической цепочки, тем более абстрактный (и более общий) анализ выполняется этим компонентом. Уровни иерархии в системе следующие: фонология (звуковой строй и функционирование звуков) и просодии (особенности произношения); аппозитивы (слова, служащие

для идентификации или переименования) [38] и клитики (фонологически зависимые слова, не имеющие собственного ударения); метрическое сканирование. Данные уровни схожи для латинского и греческого языков, что и дает возможность использовать систему для обоих языков. Анализ запроса идет по иерархии от фонологии до метрики и обратно, что позволяет системе иметь дело с любой цепочкой, от самой сложной до самой простой, без какой-либо разницы, как если бы существовал только один объект взаимодействия, при этом каждый уровень специализируется на своей собственной задаче.

2.2. Обзор работ по анализу поэтических текстов на английском языке

В английском языке один из принципиальных моментов в анализе стихов состоит в задаче определения, слабый слог или сильный. Работ много, анализируется объемный пласт литературы, от "Беовульфа" (конец VII — начало VIII в.), Джеффри Чосера (XIV в.), Уильяма Шекспира (XVI в.) и т. д. до творчества современных писателей. Хронологически и тематически работы можно разделить на группы.

Работы по анализу поэмы "Беовульф" проводились несколькими исследователями в разное время. Эксперимент Дж.М. Фоли [39] состоял в следующем: 3182 строки поэмы были внесены в базу данных в закодированном виде, и посредством статистического анализа данных некоторые гипотезы по поводу языка произведения получили подтверждение. Г.Р. Хидли [40] разработал систему для анализа стихов "Беовульфа". Система полуавтоматическая, пользователь проверяет результаты, система в некоторой степени учится на исправлениях, внесенных пользователем. Осуществляются слоговое деление на основе определенных закономерностей, встречающихся в словах, и синтаксический анализ, помогающий в слоговом делении и акцентуации. К.Р. Барквист и Д.Л. Ши [41] создали базу данных с характеристиками каждого стиха "Беовульфа" и произвели ряд статистических анализов аллитерации (основного звукового приема древнеанглийской поэзии).

Г.С. Донау [42] применил компьютер для стилистического анализа произведений Шекспира, однако количество слогов каждого слова, содержащегося в стихотворении, а также его шаблоны ударений требовалось вводить вручную в базу данных. К. Барбер и Н. Барбер [43, 44] также вручную создали базу данных из 15 942 стихов Чосера с определенными фонетическими характеристиками и применили компьютерный анализ, чтобы получить выводы о произношении и окончаниях слов. Э. Грин, Т. Бодрумлу и К. Найт [45] использовали машинное обучение с учителем в процессе создания системы анализа для перевода и автоматического создания стихов; корпус, используемый в качестве уобучающей выборки, — это сонеты Шекспира.

М. Хэйвард провел компьютерное исследование особенностей метрики в стихах различных поэтов [46, 47]. М.Р. Пламондон в 2005 г. предпринял попытку проанализировать английскую поэзию фонетическим способом, чтобы определить стихотворный ритм [48, 49]. Ч. Хартманн [50] представил Scandroid, инструмент с открытым исходным кодом, что встречается нечасто в исследованиях, способный анализировать стихи на английском языке. По выражению самого автора, создана "программа, которая сканирует английские стихи, написанные ямбом и анапестом" [51]. Х.Р. Тижуш и Р.А. Дара [52] искали способы автоматически отличить прозу от поэзии; для этого из текста и обученных классификаторов извлекаются такие особенности, как рифма и ритм, точность полученных результатов превысила 90 %. Д. Каплан и Д. Блей проанализировали стили американских поэтов на основе орфографических, синтаксических и фонетических

характеристик их стихов [53]. Аналогичное исследование с целью выявить особенности стиля профессиональных поэтов, отличающие его от стиля поэтов-любителей, проделано Дж. Као и Д. Джурафски [54].

Ф. Каванах [55] проанализировал рифмы и паттерны английских стихов с использованием правилового подхода. Я. Генцель, Д. Узскорейт и Ф. Ош [56] описали способ создания метрических стихов при переводе: переводы ограничены ритмическими формами. Х. Хирджи [57] предложил инструмент анализа, основываясь на своем опыте в изучении песенной лирики. Система выполняет фонетическую транскрипцию и слоговое деление на основе адаптированного внешнего модуля. Посредством итеративного процесса автор создал банк стихов с известными ритмическими паттернами и сравнил слоги слов с паттернами, таким образом связывая с каждым словом вероятность определенной акцентуации в стихах. Строчки стихотворения для анализа фонетически транскрибируются и преобразуются в вероятностные ритмические паттерны, которые можно сравнить с заранее определенными образцами. Результаты анализа отображаются в графическом интерфейсе в виде выделения ударных слогов.

Система ZeuScansion, разработанная коллективом авторов [58], использует словари для определения расположения ударного слога. Выполняется синтаксический анализ и далее применяются следующие правила: а) основной акцент слов открытого класса — сильные слоги в стихе; б) вторичные акценты полисиллабических (многосложных) слов и первичный акцент полисиллабических слов закрытого класса также сильны. Если слово отсутствует в словаре, то программа ищет и использует ближайшее слово. Авторы фокусируются на самом ритмическом паттерне, предполагается разделение слогов. Для оценки корпус с 759 строками был проанализирован вручную, но ритмический паттерн определен всего в 199 случаях.

Система SPARSAR, описанная в работе Р. Дельмонте [59, 60], предназначена для автоматического комплексного анализа поэтических текстов, выполняемого на уровне предложения, строки и строфы с целью изучения стиля. Структура системы следующая:

- Оценка синтаксических, семантических и грамматических характеристик (см. более раннюю работу [61]).
- Перевод стихотворения в фонетическую форму, при этом сохраняются визуальная структура и разделение на строки и строфы.
- Получение параметров: средняя длина стиха в миллисекундах и в количестве стоп. Последнее выводится линейным и строфическим представлением метрической структуры.
- Классификация ссылочных выражений с выделением конкретных и абстрактных существительных (используется корпус WordNet).
- Синтез проведенных исследований в модуль TextToSpeech для автоматизированного чтения текста (в идеале с правильной интонацией). Авторы используют акцентуацию из базы данных английских слогов для распознавания речи [62], чтобы определить предполагаемые слоги и ударение в каждой строке текста. Далее все данные используются для создания размеченной версии стихотворения, которая может быть озвучена машиной с соответствующей выразительностью.

Для визуализации фонетических характеристик стихов под руководством Н. Мак-Карди были созданы системы RhymeDesign [63] и Poemage [64].

В статье О. Калена [65] представлен количественный подход к поэзии, основанный на использовании нескольких статистических показателей (энтропия, информационная

энергия, n-граммы и т. д.), применяемых к нескольким знаковым произведениям английской литературы. Автор статьи определяет факт изменения энтропии английского языка с течением времени, и эта энтропия зависит от используемого языка и от автора. Для оценки информационной энтропии между двумя текстами применен статистический метод; разработан метод вычисления средней информации, передаваемой группой букв, о следующей букве в тексте. Кроме этого, автор утверждает, что найдена формула для вычисления языковой энтропии Шеннона [66], и вводит понятие n-граммной информационной энергии поэзии. В числе результатов заявлено также построение нейронной сети, способной генерировать поэзию, близкую к подлинной поэзии Байрона, и анализировать ее.

2.3. Обзор работ по анализу поэтических текстов на немецком языке

Структура немецкого стиха основана, прежде всего, на чередовании сильных и слабых слогов. Основные работы следующие.

Д. Чисхольм [67] предложил методы анализа немецкого стиха, в которых для определения стилистических особенностей используется фонетическая транскрипция.

Web-приложение Metricalizer^2 [68], разработанное К. Боббенхаузеном и Б. Хаммерихом, позволяет производить автоматический анализ метрических характеристик немецких стихов. Система включает в себя следующие компоненты [69]:

- Подсистема метрического анализа. Отображает фрагменты в анализируемом стихотворении, по которым определяется метрика стихотворения, а также ведет подсчет статистики.
- Подсистема анализа корпусов текстов. Включает в себя разбор текстов по акцентуации и рифме. Ведется статистика по отдельным метрическим формам. Используются корпусы текстов “Freiburger Anthologie” и “Textgrid” [70].
- Расширенные возможности системы для зарегистрированных пользователей: сохранение отдельных произведений и создание XML-документов по результатам анализа.
- Фонетический разбор слов. Позволяет производить разбор по одному слову на немецком языке в форматы X-SAMPA (Extended Speech Assessment Methods Phonetic Alphabet, “расширенный фонетический алфавит методов оценки речи”) и IPA (International Phonetic Alphabet, “международный фонетический алфавит”) [71]. Подсистема работает на основе правил.

Авторами был проведен эксперимент с 153 стихами. После устранения двух стихотворений из-за проблем, связанных с апострофами, Metricalizer^2 правильно определил ритмический рисунок 94 % стихов.

А. Эстес и К. Хенч [72] представили подход, основанный на методах машинного обучения с учителем, к классификации метрических слогов на средневысоком немецком языке (термин, обозначающий форму немецкого языка с условной датировкой между 1050 и 1350 гг.).

2.4. Обзор работ по анализу поэтических текстов на испанском и португальском языках

Испанский язык имеет близкое к фонетическому написание, что облегчает задачу ученым, так как произведениям на этом языке не нужно делать фонетическую транскрипцию, текст уже представлен в требуемом формате.

В работе П. Герваса [73] описан инструмент, выполняющий анализ стихов на испанском языке, — Gervás Prolog. Система делит слова на слоги, определяет ударный слог, передачу синалеф (слияние двух гласных, принадлежащих разным словам) между словами и извлекает рифмы. Программа начинает работу с деления слов на слоги без фонетической транскрипции. Создается серия групп для каждого слова, содержащих определенные комбинации графем, каждой группе присваивается одна из 14 возможных категорий (например, двойной согласный или высокий гласный), что используется не только при слоговом делении, но и позже, при идентификации синалеф. Затем происходит перегруппировка, при которой возникают согласные графемы и гласные графемы, чтобы различать диерезы (раздельное произношение двух соседних гласных) и синерезы (слияние двух смежных гласных в дифтонг), эта стадия учитывает диакритические знаки испанского языка. Окончательное слоговое разделение получается на основе идентифицированных гласных, вокруг которых распределены согласные группы. Оценка работы системы была сделана на основе 64 сонетов (все с шестью слогами) шести авторов так называемого испанского золотого века (приблизительно между 1492 и 1659 гг.), всего 896 стихов. Общая точность анализа составила 88.7 %, вариации точности для различных авторов — от 81.3 до 95.2 %, что может указывать на то, что правила синалеф и синерез работают для одних авторов лучше, чем для других. Процент ошибок связан с двумя случаями: с ситуацией, когда у стиха больше или меньше слогов, чем должно быть, и с современной трактовкой конъюнкции звука “y”; если не учитывать оба источника ошибок, точность возрастает до 99.3 %.

Проекты LuCas и SAEP, созданные Н. Мамедом, включают в себя первоначальные исследования, проводимые П. Араужо, описаны в статье [74] и в диссертации [75]; реализация представлена в работах [76, 77] и в диссертации, написанной João Marques [78]. Программа LuCas, названная в честь де Камоэнса, реализована первой: она умеет считать слоги, но не может анализировать метаплазмы (различные преобразования слов). Далее была разработана SAEP (Sistema de Apoio à Escrita de Poemas — система поддержки написания стихов), уже с ограниченной опцией определенного анализа метаплазм. Обе системы схожи: предназначены только для одного языка; в качестве входных данных — стихотворение, закодированное в виде простого текста, которое делится на стихи и строфы в соответствии с пустыми строками. Слова фонетически транскрибируются через внешний модуль DIXI [79], который также производит грамматическое слоговое деление, пунктуация не учитывается. Есть отличия: в системе LuCas имеется немного возможностей для метрического анализа стихов, она не обнаруживает метаплазмы, соединяя фонетические транскрипции слов друг с другом; SAEP содержит некоторые механизмы для разрешения проблем с гласными, применяя правила к фонетической транскрипции, чтобы скорректировать слоговое деление. Однако обе системы не предоставляют полного анализа стихов, считая, в первую очередь, количество слогов. В интерфейсе SAEP отображается минимальное и максимальное количество слогов, которые может содержать анализируемый стих, причем минимум слогов рассчитывается со всеми синалефами и синерезисами, а максимум — с диалефами и диерезами. Учитывая другие стихи стихотворения, программа может представить вероятное количество слогов каждого стиха. Тестовая выборка для LuCas содержала около 200 строф, но тесты оценивали только время отклика системы, а не ее точность. Точность SAEP была оценена на основе набора из 12 поэм, насчитывающего 197 стихов. В критериях оценки классификации рифм и строф система не выдала каких-либо ошибок, за исключением точного подсчета слогов, в результате чего точность составила 82.2 %, количество

ошибок учитывает свободный стих, который, как правило, не является объектом автоматизированного анализа, или, как в системе ФИЦ ИВТ, получает в результате анализа определение дисметрического и рекомендацию обратиться к эксперту для дальнейшего анализа. При уменьшении тестовой выборки до 105 стихов из семи поэм без дисметрических или полиметрических стихов точность составила 77.1 %. Основное ограничение систем LuCas и SAEP состоит в том, внимание уделяется количеству слогов, а не их распределению в стихе.

Дж. Р. Робинсон разработал программу Colors of Poetry [80], которая анализирует стихи на испанском языке и отображает их в графическом виде. В работе нет информации о формате ввода и возможных этапах предварительной обработки, о том, как определяется ударение, автор просто утверждает, что программа следует грамматическим правилам испанского языка. Инструмент выполняет слоговое разделение, но в работе не указано, как именно. Colors of Poetry использует довольно простую схему фонетической транскрипции, включающую обязательные маркеры. Другие маркеры — синереза и диалефа (разделение смежных гласных на разные слоги) — указаны графически, но программа не устанавливает, существуют ли они на самом деле. В связи с этим автор предлагает альтернативу: сравнение стихов с фиксированными моделями и друг с другом для решения вопроса, какие синалефы или диерезы должны учитываться, чтобы получить правильную метрику. Система Colors of Poetry не выполняет автоматического ритмического анализа, вместо этого графически отображаются различные параметры, такие как ударный слог и возможные паузы, которые помогают пользователю найти ритм в стихотворении.

Б. Наварро-Колорадо предложил инструмент, изучающий метрики сонетов на испанском языке и производящий семантический анализ [81–83]. Эта система применяется к корпусу из 5078 сонетов XVI и XVII вв., который преобразуется в формат TEI [84], чтобы его можно было использовать в других исследованиях. На вход системе подается последовательность символов из одного стиха без дополнительной маркировки. Модуль, работающий на правиловом подходе, выполняет разделение слогов и обнаруживает ударный слог. На основании этих результатов слова из определенных категорий, таких как клитики, помечаются как безударные, независимо от правописания. Если слоговое разделение дает ровно 10 метрических слогов, система считает, что анализ завершен. Для нестандартных ситуаций применяется ряд правил, подробное описание которых, однако, не приводится авторами проекта. Результат, полученный системой для 100 сонетов в количестве 1400 стихов, сравнивался с результатами, полученными вручную двумя экспертами. Неверно было распознано 108 стихов, что соответствует точности в 92.3 %. Уровень точности при работе экспертов составил 96.2 %.

Система Aoidos [23], разработанная А. Миттманном, представляет собой инструмент анализа поэтических текстов на португальском языке. На вход системы подаются стихи в формате TEI, производится независимый анализ каждого стихотворения. Процесс включает в себя следующие шаги: приведение текста к формату TEI; извлечение слов из стихотворения; нахождение ударного гласного для каждого слова; деление слов на слоги; формулировка фонетической транскрипции (с использованием изолированного словаря); подбор вариантов транскрипции для каждого стихотворения (определение ритмической схемы); попытка определения метрики стихотворения; поиск совпадающей метрики с учетом наиболее подходящей ритмической схемы. Варианты фильтруются в соответствии с метрикой: если стихотворение было классифицировано как декасиллабическое (поэтический метр из десяти слогов), то рассматриваются только варианты

с десятью слогами. Далее происходит разбиение произведения на слоги в соответствии с метрикой. Результат проведенного анализа может отображаться как фонетически, так и помечаться на символах исходного текста, заявленная точность системы составляет более 95 %.

2.5. Обзор работ по анализу поэтических текстов на итальянском и провансальском языках

Разработка Д. Роби [85] содержит полуавтоматический инструмент, анализирующий "Божественную комедию" Данте Алигьери на итальянском языке (созданную между 1308 и 1321 гг.), при этом формат ввода текста не уточняется. Группы гласных графем анализируются для определения центра слога, но не уточняется, к какому слогу принадлежат согласные, что приводит к определению всех возможных синерез и синалеф. Система предполагает, что большинство слов имеют ударение на предпоследнем слоге, однако нахождение лексического ударения не означает определения ритма стиха. Разделение слогов, произведенное системой, не включает уточнение ударения, в этот момент система запрашивает вмешательство пользователя: отображается выполненный анализ, и пользователь может вносить изменения, которые сохраняются программой. Корректировки, внесенные пользователем, сохраняются в базе данных, далее система отображает для пользователя все варианты.

Т.М. Рейнсфорд и О. Скривнер в работе [86] описывают методы, используемые в процессе добавления метрической маркировки к корпусу стихов на провансальском (окситанском) языке, являющемся диалектом Южной Франции, — Lo poema de Boecis ("Боэций"), анонимному фрагменту, написанному около 1010 г. и содержащему 257 стихов, соответствующих метрической схеме из четырех и шести слогов. Авторы применяют алгоритм автоматического разделения стихов на слоги, однако определение ударного слога и контроль разделения на слоги — задача эксперта, как и синерезы и синалефы. Алгоритм перебирает все различные комбинации, пока не будет найден стих из десяти поэтических слогов, в которых акцентированы четвертый и шестой, если такой критерий не достигнут, стих считается некорректным. Формальная оценка точности анализа не сделана, из 257 стихов 25 считаются некорректным, однако авторы обнаружили, что такие стихи являются исключениями в тексте.

2.6. Обзор работ по анализу поэтических текстов на французском языке

Фонология французского языка существенно отличается от фонологии португальского, испанского и итальянского языков. Вопросы синалефы и синерезы упрощены, но принципиальна проблема классификации графемы "e" в конце слов. Первые работы появились в 1980-х гг., это иссследования Ж. Рубо [87, 88], но они не получили продолжения. Два самых известных инструмента для анализа поэзии на французском языке — это Métromètre и Anamètre.

Проект Métromètre авторства В. Бодуэн и Ф. Ивона [89] применялся исключительно для анализа классических французских александрийских стихов (1630–1830 гг.). Система выполняет фонетическую транскрипцию после этапа синтаксического анализа, что облегчает разметку ударных слогов; синтаксический анализ системы поддерживает фонетическую транскрипцию: Métromètre рассматривает ударения только в глаголах, существительных, прилагательных и наречиях. Чтобы разделить стих на 12 слогов

александрийского стиха, система рассматривает три ситуации, которые могут влиять на количество полных гласных в фонетической транскрипции: диереза, немое “e” и связывание; выбор между диерезой и синерезой соответствует разнице слога в делении на слоги. Métromètre предполагает не более двух вариантов для каждого стиха, используя свой набор правил для создания единого окончательного результата анализа, при этом заранее известно, что каждый нечетный стих рифмуется со следующим. Это, конечно, предпосылка французского классического александрийского стиха, которая не относится к другим случаям. В Métromètre создано четыре модуля для производства стилистической информации об авторах и произведениях, о чем написано в более поздней работе [90]: корпус текстов в определенном формате, позволяющий аккумулировать тексты, у которых более одного автора; сам Métromètre; корпус слов, рифмующихся друг с другом; модуль текстовой статистики, назначающий лексико-семантический класс. Точность анализа, выполняемого Métromètre, не указывается, однако на некоторых статистических данных достигается 99.7 %, что, безусловно, было бы весьма достойным результатом.

Проект Anamètre [91–93] включает в себя корпус, состоящий из более чем 500 000 стихов и инструментов для сканирования и автоматической маркировки стихов на французском языке без ограничения определенным типом стихотворения. На вход подается текст в формате XML, пунктуация игнорируется; фонетическая транскрипция не производится; ударные гласные определяются с помощью эвристики; нестандартные случаи проверяются с помощью словаря. Из каждого стихотворения извлекается вектор характеристик, чтобы с применением правил выполнить классификацию стихотворения по определенной метрической схеме. Однако между выходными данными и исходным текстом существует слабая связь, не позволяющая разделить его на слоги. Окончательный результат, который является определением размера поэмы и ритма каждого стиха, не может быть сопоставлен с оригинальным текстом. Система изначально не знает, какие метрики стихотворения изучаются, поэтому при восьми слогах в стихе или меньше метр стихотворения классифицируется как простой, последующий анализ не проводится. Если в стихах более восьми слогов, они считаются составными, система стремится определить размер, ищет цезуру в стихах. Набор свойств, указывающих слабость слога в позиции, оценивается для каждого слога каждого стиха, образуя матрицу с шаблонами. Таким образом, цезура определяется по столбцу матрицы, в котором минимум указанных свойств, что показывает правильное положение в стихах, где встречается ударный слог. Если алгоритм находит слоговую позицию, в которой нет свойства для всех проанализированных стихов, то, фактически, матрица определяет стандарт метрики для стихотворения, каким бы необычным оно ни было. Если хотя бы одна позиция в матрице содержит какое-либо свойство, то подключаются предопределенные размеры. Сравнивая матрицу с предопределенными стихотворными размерами, система классифицирует стихотворение в целом. Однако не проводилось никаких экспериментов для оценки качества анализа, кроме того, у инструмента, видимо, нет средств для повторного анализа для каждого стиха: после синерезы он невозможен.

2.7. Обзор работ по анализу поэтических текстов на арабском языке и санскрите

Некоторые разработки в области автоматизированного анализа текста существуют для арабского языка и санскрита. Х.Е. Айек, А. Махфуф и А. Зриби использовали ней-

ронные сети для распознавания метрического строя стихов на арабском языке [94]. В работе Д.Э. Кулугли [95] представлены перспективы анализа поэзии на арабском языке без конкретизации инструментов. Коллектив авторов, возглавляемый А. Альмухаребом, описал некоторые методы определения поэтических шаблонов на арабском языке для извлечения стихов [96]. В работе А. Курта и М. Кары [97] предложен алгоритм распознавания и анализа стихотворений, написанных в особой, характерной для восточной (арабской, персидской, турецкой) поэзии, системе стихосложения аруд. М.А. Алнагдави разработал алгоритм определения поэтической метрики с помощью контекстно-свободных грамматик [98].

Д. Вуджастик [99] создал компьютерную программу, написанную на языке программирования SNOBOL 4 (предназначенном преимущественно для обработки текстовых данных), анализирующую стихи на санскрите. Работа К.М. Майрхофера [100] описывает авторский подход к анализу стихов на классическом санскрите. И.М. Усака и М.М. Ямазаки [101, 102] использовали вычислительные методы для анализа канонических стихов, написанных на одной из версий санскрита. Н. Рама и М. Лакшманн [103] представили метод метрической классификации стихов на санскрите. Коллектив под руководством Г. Ракшита [104] применил стилистический анализ с последующим присвоением авторства к стихотворным текстам на бенгальском языке, происходящем из санскрита.

2.8. Обзор работ по анализу поэтических текстов на чешском языке

Чешский исследователь К. Сгаллова [105] собрала ритмические и метрические данные 6466 стихов Франтишека Яромира Рубеша, разработала метод их кодирования, провела вручную статистический и стилистический анализ. Позднее она опубликовала тезаурус [106] на основе почти 5000 чешских стихов.

Р. Ибрагим и П. Плехач [107] разработали систему KVĚTA для анализа стихотворений, написанных на чешском языке. Она помогла в разметке поэтического корпуса, в котором собрано более 2.5 миллионов стихов [108, 109]. В качестве входных данных система получает стихотворение, слова которого должны содержать разметку по лемматизации и морфологии и к которому применяется ряд правил, преобразующих поэтический текст в фонетическую транскрипцию. В определенных орфографических контекстах существуют правила, которые с большой точностью определяют правильность транскрипции. Если применение правил невозможно, система запрашивает словарь, в котором указан правильный вариант для каждой схемы. Из фонетической транскрипции создается двумерная структура, которая соответствует слогам стихов и стихам стихотворения. Слоги ищутся в фонетической транскрипции, для каждого вычисляется ряд булевых атрибутов, которые помогают характеризовать его, например, по позиции в начале слова или в конце, или по факту, является ли слово предлогом. Основываясь на количестве слогов в стихах стихотворения, система генерирует ряд ритмических возможностей и сравнивает шаблоны, найденные в самих стихотворениях, со сгенерированными вариантами. Первоначально использовалась идея метрического индекса из работы И.А. Пильщикова и А.С. Старостина [24], который рассчитывается с применением правил, но далее он был заменен метрическим коэффициентом, использующим вероятностные параметры. Наилучший индекс или коэффициент, найденный между генерируемыми ритмическими паттернами, определяет классификацию стиха. Если автоматическая классификация невозможна, система определяет стихотворение

как подлежащее экспертной оценке. Точность измерялась с помощью эталонного корпуса, размеченного полуавтоматически: KVĚTA использовалась для разметки метра, более сложные случаи размечались экспертом. Вручную было размечено 300 корпусов, состоящих из 25 779 стихов, в 99.97 % случаев метр был определен верно. При работе с корпусом из 2 336 435 стихов неверные результаты были в 108 979 случаях, точность составила 95.34 %, однако при использовании первоначального метрического индекса точность составляла 94.88 %.

2.9. Обзор работ по анализу поэтических текстов на русском языке

Отдельные работы, описывающие комплексный подход к автоматизации характеристик русских поэтических текстов, в частности исследование М.Л. Гаспарова [110], затрагивают, как правило, весьма специфические жанры поэзии: например фольклорные стихи, структурные характеристики которых, например метрика, тематика и т. д., значительно отличаются от соответствующих структур “литературного” стиха. Широко известны частотные словари языка поэтов-классиков. Проводились многочисленные исследования статистики типов русской рифмы (в том числе и применительно ко временной динамике), обобщенные в книге Д. Самойлова [111].

Работы Е.М. Брейдо связаны с автоматическим анализом метрики русского стиха, чему посвящена диссертация [112], в которой “построена формальная модель (названная интервальной), описывающая русскую метрику от силлабо-тоники до границы с прозой”, и “основанные на интервальной модели алгоритмы компьютерного анализа метрики позволяют автоматически и полуавтоматически распознавать любые метрические формы русского стиха и определять основные понятия метрики, исходя из алгоритмов анализа”. В статье [113], по утверждению автора, “дается формальное описание стиховой метрики, пригодное для компьютерного анализа текста”. Стоит отметить фундаментальную мысль: “. . . компьютерная модель, построенная из логики стиховедческой науки, — редкая возможность строгого и точного описания языковой системы” [113]. Однако в работе 2021 г. [114] автор признает “четвертьвековой перерыв”, подчеркивает “переосмысление” и сосредоточивает свое внимание на применении “интервального подхода” уже для разграничения стихотворных размеров, таким образом сужая задачу до исследования “строгого тонического стиха” — нового стихотворного размера, открытого с помощью “интервального подхода”.

Известной в своем кругу системой является проект коллектива авторов (А.Е. Поляков, И.А. Пильщиков, М.Б. Бергельсон) “Конкорданс к текстам Ломоносова” [115]. Данный конкорданс строится на основе корпуса авторских текстов, снабженных структурной, филологической и грамматической разметкой. Проект предусматривает создание открытого интернет-ресурса, который будет включать в себя [116]: корпус текстов Ломоносова, построенный на основе наиболее авторитетных изданий; биографические, литературно-критические и историко-научные работы о Ломоносове; полный алфавитно-частотный конкорданс к текстам Ломоносова. Конкорданс строится на основе электронного корпуса текстов М.В. Ломоносова, представляющего собой филологически корректную цифровую версию академического Полного собрания сочинений и писем Ломоносова в 11 томах[1] и ряда дополнительных изданий [117]. Исследователями принято решение учитывать в алфавитной части все словоупотребления, зафиксированные в полном собрании сочинений и в дополнительных источниках. Вопрос о том, как

[1]Ломоносов М.В. Полное собрание соч.: в 11 т. М.;Л.: Изд-во АН СССР; 1950.

учитывать вариативность в частотном словаре, остается открытым. Также авторами было принято решение не учитывать вариативность буквенного состава слов, слитного/раздельного написания и написания с прописной/строчной буквы при сведении словоформ в лексему (вокабулу). Подготовка корпуса включает первичную разметку текста для представления в электронной библиотеке; дополнительную структурную разметку и сегментацию текста для корпуса; грамматическую разметку и ее ручную постобработку (снятие омонимии, исправление разборов); преобразование в базу данных, построение конкордансов и других производных. Первоначально корпус подготавливается в формате HTML со специальной разметкой, ориентированной на представление в электронной библиотеке, он включает метатекстовую, структурную и форматную разметку, необходимую для точного воспроизведения содержания и внешнего вида текста, но недостаточную для корпуса. Далее в текст вносится дополнительная структурная разметка, маркирующая фрагменты текста, требующие специальной обработки (заголовки, цитаты, примеры, комментарии, иноязычный текст и т. д.), текст пропускается через морфологический анализатор (парсер), который выдает для каждого слова множество вариантов разбора. После этого необходимо вручную проверить и исправить ошибки разбора, удалить неправильные варианты и добавить недостающие. Размеченный конкорданс представляет собой словарную базу данных, из которой путем различной проекции и группировки данных можно получать различные виды словарей и проводить объективные исследования авторского языка. Форма базы данных открывает целый ряд возможностей, недоступных в традиционных бумажных словарях: динамический выбор примеров по любым параметрам, динамическая сортировка и группировка, быстрый переход из словаря в корпус текстов, просмотр и выдача словарной информации в различных форматах, генерация печатных словарей.

Несомненно, важной является работа А.В. Козьмина "Автоматический анализ стиха в системе Starling" [118], посвященная автоматизированному определению метроритмических характеристик русских поэтических текстов. Она опирается на проект "Автоматизированный лингвостиховедческий анализ русских поэтических текстов", или "Вавилонская башня". "Вавилонская башня" — название международного интернет-проекта, посвященного сравнительно-историческому языкознанию. Информационная система включает в себя компоненты:

- Базы данных этимологий (происхождение слов): глобальная, Ностратическая, Индоевропейская, Алтайская, Уральская и др. Каждая база данных доступна для просмотра, осуществления поиска (для всех пользователей) и редактирования (для авторизованных пользователей).
- База данных "Тематическая классификация и распределение фольклорно-мифологических мотивов по ареалам" Ю.Е. Березкина. Содержание и структура базы данных обусловлены исследовательскими задачами, определенными ее составителем. Это выявление древних миграций и культурных контактов на основе использования материалов фольклора и мифологии, выявление возможной функциональной зависимости между отдельными явлениями в сфере фольклора и мифологии и особенностями природной среды и социальной организации.
- База данных "Квантитативно-реализационный грамматический словарь современного монгольского языка" С.А. Крылова.
- Словари русского языка (Ожегова, Зализняка, Этимологический словарь Фасмера, Справочник по метаязыку русских грамматистов первой половины XX в. С.А. Крылова).

- Средство управления базами данных Starling, с помощью которого осуществляется работа с имеющимися базами данных (составление и выполнение сложных поисковых запросов, добавление новых записей, организация ссылок). Предусмотрена поддержка иероглифов и разных текстовых кодировок.
- Морфологический анализатор. Предоставляет возможность проанализировать любое русское слово и получить его полную акцентуированную парадигму.

В рамках системы созданы следующие программные модули [119]: модуль определения метра и размера, подмодуль определения рифмовки, модуль преобразования орфографической записи в фонетическую транскрипцию, модуль выявления анаграмматических эффектов. Однако после смерти руководителя проекта С.А. Старостина работы в этом, несомненно перспективном, направлении были прекращены.

Одной из принципиальных работ по автоматизации процесса распознавания метрических характеристик текста является исследование И.А. Пильщикова и А.С. Старостина [25], в котором авторы используют компьютерный морфолого-акцентологический анализ [24]. Проблему подвижного ударения, характерного для русского языка, авторы предлагают решить с помощью алгоритма, который позволяет распознавать дольники и тактовики. Авторы представляют свой программный модуль автоматического определения классических силлабо-тонических размеров, усовершенствованный по сравнению с предыдущими разработками этого же коллектива исследователей [26, 120, 121]. Авторы поставили вопрос о неоднозначности акцентуации, предложили алгоритм, но в их работах нет подробностей реализации и результатов его использования.

Коллектив авторов (В.Н. Бойков, В.Е. Захаров, М.С. Каряева, В.А. Соколов) разработал открытый сетевой ресурс Wikipoetics (сегодня не работающий), который был представлен компонентами: проблемно-ориентированным “Тезаурусом по поэтологии” и “Блоком анализа и спецификации” текстовых объектов [122]. В блоке анализа и спецификации выделяются два комплекса задач: спецификация терминологических статей тезауруса и спецификация поэтического произведения. Структура комплекса включает в себя такие группы решений задач, как: метроритмическую разметку текста, заполнение полей спецификации произведения, идентификацию метра.

В статье другого коллектива авторов (В.Н. Бойков, М.С. Каряева, В.А. Соколов, И.А. Пильщиков) [28] рассматриваются автоматические процедуры спецификации поэтического текста — метроритмическая разметка и идентификация стихового метра. В работе И.А. Пильщикова и А.С. Старостина [27] для автоматизации метроритмической разметки предложены следующие процедуры: нумерация строк произведения, токенизация слов, акцентуация произведения, выделение рифмованных строк, выделение слоговой схемы и др. Однако морфологический анализ является весьма сложной задачей, при этом авторы не приводят результаты тестирования точности предложенных ими алгоритмов, практическая реализация исследования в явном виде, скорее всего, не производилась.

В системе автоматизированного комплексного анализа русских поэтических текстов, разработанной в ФИЦ ИВТ, алгоритмы, созданные коллективом авторов под руководством И.А. Пильщикова, применяются после проведенных модификаций.

3. Новейшие исследования

В начале XXI в. развитие количественных методов, в том числе при исследовании русской поэзии, претерпело большие изменения. Этому в большой степени способствовало

бурное развитие информационных технологий. Автоматизированные алгоритмы, большие вычислительные ресурсы, развитые языки программирования, машинное обучение — далеко не полный перечень технологий, используемых в настоящее время. Возвращаясь к цепям Маркова, стоит отметить, что их развитие не остановилось, а наоборот, приобрело популярность. Так, в статье И.И. Дроздовой и А.Д. Обуховой [123] рассматривается тема определения авторства анонимного текста за счет частотных характеристик. Тема является достаточно актуальной на сегодняшний день и охватывает большой спектр целей: от отыскания автора необходимой вам статьи в интернете или запоминающегося отрывка художественного произведения до достаточно серьезных военных целей. К исследованиям, проводимым в смежных сферах с использованием формальных методов работы с текстом, можно отнести метод определения неестественного происхождения документа, основанный на изучении статистики встречаемости пар соседних слов, описанный в работе коллектива авторов (Е.А. Гречникова, Г.Г. Гусева, А.А. Кустарева и А.М. Райгородского) [124]. Такие исследования становятся наиболее популярными из-за большого количества автоматически генерируемого текста, в частности спама.

В ряде исследований ФИЦ ИВТ (начатых в 2012 г.) намечены (а в дальнейших работах применены с получением результатов) основные подходы к автоматизации процесса статистического анализа низших структурных уровней (метр, ритм, фонетика, лексика, грамматика) русских поэтических текстов [5, 125]. Что касается применения методов машинного обучения, интересные результаты, в частности, дало применение нейросетей в рамках задачи определения характеристик авторского стиля [126].

Описание моделей и алгоритмов системы автоматизированного комплексного анализа русских поэтических текстов, созданной в ФИЦ ИВТ [2], представлено в [127].

Заключение

Алгоритмы и системы автоматизированного анализа поэтических текстов, предлагаемые в зарубежных работах, сильно зависят от особенностей конкретного языка и практически неприменимы для поэтических текстов на русском языке. Разработка универсального инструмента также весьма затруднительна, хотя некоторые исследования показали возможность разработки инструмента анализа для языков с некоторыми идентичными признаками, однако это, скорее, исключение. Особенности строя русского языка делают автоматизацию анализа русских текстов весьма нетривиальной задачей. Разработанная в ФИЦ ИВТ программная система автоматизированного комплексного анализа русских поэтических текстов является оригинальным проектом. Апробированных в научных публикациях комплексных систем до наших разработок создано не было, однако в основе создания системы ФИЦ ИВТ лежат разработки, методы, алгоритмы, принадлежащие наследию классической математики, и современные, связанные с машинным обучением. Подходы к исследованию поэтических текстов с использованием методов искусственного интеллекта, методов теории информации, несомненно, являются перспективным направлением в задачах обработки естественного языка.

Благодарности. Академику Юрию Ивановичу Шокину с благодарностью за создание направления обработки текстов на естественных языках в Институте (ФИЦ ИВТ), за идеи, советы и поддержку.

Список литературы

[1] **Успенский В.А.** Предварение для читателей “Нового литературного обозрения” к семиотическим посланиям Андрея Николаевича Колмогорова. Новое литературное обозрение. 1997; (5):128.

[2] Анализ поэтических текстов онлайн. Адрес доступа: `http://poem.ict.nsc.ru`.

[3] **Kozhemyakina O.Yu.** Conceptual design of the software system for automated complex analysis of poetic texts. Computational Technologies. 2022; 27(2):122–137. DOI:10.25743/ICT.2022.27.2.010.

[4] **Магомедова Д.М.** Филологический анализ лирического стихотворения. М.: Академия; 2004: 187.

[5] **Барахнин В.Б., Кожемякина О.Ю., Забайкин А.В., Хаятова В.Д.** Автоматизация комплексного анализа русского поэтического текста: модели и алгоритмы. Вестник НГУ. Сер.: Информационные технологии. 2015; 13(3):5–18.

[6] **Барахнин В.Б., Федотов А.М., Бакиева А.М., Бакиев М.Н., Тажибаева С.Ж., Батура Т.В., Кожемякина О.Ю., Тусупов Д.А., Самбетаева М.А., Лукпанова Л.Х.** Алгоритмы генерации стемматизации словоформ казахского языка. Cloud of Science. 2017; 4(3):434–449.

[7] **Barakhnin V.B., Kozhemyakina O.Yu., Bakiyeva A.M., Sodboev M.K.** The algorithms for complex analysis of the corpuses of poetic texts in the Kazakh language. Journal of Physics: Conference Series. 2018; (1117):7. DOI:10.1088/1742–6596/1117/1/012003.

[8] **Barakhnin V.B., Fedotov A.M., Bakiyeva A.M., Bakiyev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tussupov D.A., Sambetbaiyeva M.A., Lukpanova L.Kh.** The software system for the study the morphology of the Kazakh language. The European Proceedings of Social and Behavioural Sciences. 2017; (XXXIII):18–27. DOI:10.15405/epsbs.2017.12.3.

[9] **Марков А.А.** Пример статистического исследования над текстом “Евгения Онегина”, иллюстрирующий связь испытаний в цепь. Известия Императорской академии наук. 1913; 7(3):153–162.

[10] **Кириллов И.А.** О поэтической информации. Информационная безопасность и межкультурная коммуникация в контексте цифровой трансформации. М.; 2022: 408.

[11] Колмогоров в воспоминаниях. М.: Физматлит; 1993: 736.

[12] **Прохоров А.В.** О работах А.Н. Колмогорова по стиховедению. Колмогоров А.Н. Труды по стиховедению. М.: Издательство МЦМНО; 2015: 256.

[13] **Успенский В.А.** Предварение для читателей “Нового литературного обозрения” к семиотическим посланиям Андрея Николаевича Колмогорова. Новое литературное обозрение. 1997; (5):129.

[14] **Колмогоров А.Н., Прохоров А.В.** К основам русской классической метрики. Содружество наук и тайны творчества. М.: Искусство; 1968: 397–432.

[15] **Колмогоров А.Н.** Труды по стиховедению. М.: Издательство МЦМНО; 2015: 256.

[16] **Гаспаров М.Л.** Владимир Маяковский. Очерки истории языка русской поэзии XX века: опыты описания идиостилей. М.: Наследие; 1995: 557.

[17] **Гаспаров М.Л., Скулачева Т.В.** Статьи о лингвистике стиха. М.: Языки славянской культуры; 2005: 132.

[18] **Левин Ю.Д.** Конкордация поэзии Пушкина. Русская литература. 1987; (1):212–214.

[19] **Shaw J.Th.** Pushkin: poet and man of letters and his prose. L.A.; 1995: 273.

[20] **Шоу Дж.Т.** Конкорданс к стихам А.С. Пушкина. (А–Н). Т. 1. М.: Языки русской культуры; 2000: 674.

[21] **Шоу Дж.Т.** Конкорданс к стихам А.С. Пушкина. (О–Я). Т. 2. М.: Языки русской культуры; 2000: 641.

[22] **Холшевников В.Е.** О словаре рифм Пушкина. Временник Пушкинской комиссии. 1973. Л.: Наука; 1975: 135.

[23] **Mittmann A.** Escansão automática de versos em português. Tese (doutorado). Universidade Federal de Santa Catarina, Centro Tecnológico. Programa Pós–Graduação em Ciência da Computação. Florianópolis; 2016: 303. Available at: `https://repositorio.ufsc.br/bitstream/handle/123456789/175819/345411.pdf?sequence=1`.

[24] **Пильщиков И.А., Старостин А.С.** Основные проблемы автоматизации базовых процедур ритмико-синтаксического анализа силлабо-тонических текстов. Национальный корпус русского языка: 2006–2008. Новые результаты и перспективы. СПб.: Нестор-История; 2009: 502.

[25] **Пильщиков И.А., Старостин А.С.** Автоматическое распознавание метра: проблемы и решения. Славянский стих. М.: Рукописные памятники Древней Руси; 2012; (9):568.

[26] **Pilshchikov I., Starostin A.** Automated analysis of poetic texts and the problem of verse meter. Current trends in metrical analysis. Littera: Studies in Language and Literature. Berlin; 2011; (2):368.

[27] **Pilshchikov I., Starostin A.** Reconnaissance automatique des mètres des vers russes: une approche statistique sur corpus. Langages; 2015; 3(199):89–106.

[28] **Бойков Н.В., Каряева М.С., Соколов В.А., Пильщиков И.А.** Об автоматической спецификации стиха в информационно-аналитической системе. Труды XVII Международной Конференции "Аналитика и управление данными в областях с интенсивным использованием данных". Обнинск: ИАТЭ НИЯУ МИФИ; 2015: 144–151. Адрес доступа: `http://ceur-ws.org/Vol-1536/paper22.pdf`.

[29] **Барахнин В.Б., Кожемякина О.Ю.** Об автоматизации комплексного анализа русского поэтического текста. CEUR Workshop Proceedings. 2012; (934):167–171. Адрес доступа: `http://ceur-ws.org/Vol-934/paper27.pdf`.

[30] **Барахнин В.Б., Кожемякина О.Ю., Забайкин А.В.** Алгоритмы комплексного анализа русских поэтических текстов с целью автоматизации процесса создания метрических справочников и конкордансов. CEUR Workshop Proceedings. 2015; (1536):138–143. Адрес доступа: `http://ceur-ws.org/Vol-1536/paper21.pdf`.

[31] MyStem. Available at: `https://tech.yandex.ru/mystem`.

[32] Rifmofed.ru. All about rhyme and versification. Available at: `http://rifmoved.ru`.

[33] **Mansilla R., Bush E.** Increase of complexity from classical Greek to Latin poetry. Complex Systems. 2003; 14(3). Available at: `https://arxiv.org/pdf/cond-mat/0203135.pdf`.

[34] **Rényi A.** On measures of information and entropy. Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability. 1960: 547–561.

[35] **Ярхо В.Н.** Гомеровский вопрос. Литературная энциклопедия. М.: Советская энциклопедия; 1987: 78.

[36] **Fusi D.** A multilanguage, modular framework for metrical analysis: IT patterns and theorical issues. Langages. 2015; 3(199):41–66.

[37] **Fusi D.** An expert system for the classical languages: metrical analysis components. Lexis. 2008; (27):25–45.

[38] **Мишина Ю.Е.** Аппозитивные конструкции и смежные синтаксические явления английского языка: проблема разграничения. Филологические науки. Вопросы теории и практики. Тамбов: Грамота; 2018; 1(1):155–158. DOI:10.30853/filnauki.2018-1-1.40.

[39] **Foley J.M.** A computer analysis of metrical patterns in Beowulf. Computers and the Humanities. 1978; (12):71–80.

[40] **Hidley G.R.** Some thoughts concerning the application of software tools in support of old English poetic studies. Literary and Linguistic Computing. 1986; 1(3):156–162.

[41] **Barquist C.R., Shie D.L.** Computer analysis of alliteration in Beowulf using distinctive feature theory. Literary and Linguistic Computing. 1991; 6(4):27–280.

[42] **Donow H.S.** Prosody and the computer. A text processor for stylistic analysis. Spring Joint Computer Conference. 1970: 712.

[43] **Barber C., Barber N.** The versification of The Canterbury Tales: a computer-based statistical study. Pt I. Leeds Studies in English. 1990; (21):81–103.

[44] **Barber C., Barber N.** The versification of The Canterbury Tales: a computer-based statistical study. Pt II. Leeds Studies in English. 1991; (22):57–83.

[45] **Greene E., Bodrumlu T., Knight K.** Automatic analysis of rhythmic poetry with applications to generation and translation. Conference on Empirical Methods in Natural Language Processing. 2010: 524–533.

[46] **Hayward M.** A connectionist model of poetic meter. Poetics. 1991; (20):303–317.

[47] **Hayward M.** Analysis of a corpus of poetry by a connectionist model of poetic meter. Poetics. 1996; 24(1):1–11. Available at: `http://www.english.iup.edu/mhayward/Metrics/Cormetrics.htm`.

[48] **Plamondon M.R.** Computer-assisted phonetic analysis of English poetry: a preliminary case study of Browning and Tennyson. Text Technology. 2005; 14(2):153–175.

[49] **Plamondon M.R.** Virtual verse analysis: analysing patterns in poetry. Literary and Linguistic Computing. 2006; 21(1):127–141.

[50] **Hartman C.** The Scandroid. New London; 2005. Available at: `http://charlesohartman.com/verse/scandroid/ScandroidManual.pdf`.

[51] The Scandroid. Available at: `http://charlesohartman.com/verse/scandroid/index.php`.

[52] **Tizhoosh H.R., Dara R.A.** On poem recognition. Pattern Analysis and Applications. 2006; (9):325–338.

[53] **Kaplan D.M., Blei D.M.** A computational approach to style in American poetry. 7th IEEE International Conference on Data Mining (ICDM 2007). 2007: 553–558.

[54] **Kao J., Jurafsky D.** A computational analysis of style, affect, and imagery in contemporary poetry. NAACL Workshop on Computational Linguistics for Literature. 2012. Available at: `https://nlp.stanford.edu/pubs/kaojurafsky12.pdf`.

[55] **Kavanagh F.** Analysis of a phonetic and rule based algorithm approach to determine rhyme categories and patterns in verse. Diss. (Mestrado). Open University; 2007.

[56] **Genzel D., Uszkoreit J., Och F.** "Poetic" statistical machine translation: rhyme and meter. Conference on Empirical Methods in Natural Language Processing. 2010: 158–166.

[57] **Hirjee H.** Rhyme, rhythm, and rhubarb: using probabilistic methods to analyze hip hop, poetry, and misheard lyrics. University of Waterloo. 2010. Available at: `https://uwspace.uwaterloo.ca/bitstream/handle/10012/5419/Hirjee_Hussein.pdf`.

[58] **Agirrezabal M., Arrieta B., Astigarraga A., Hulden M.** ZeuScansion: a tool for scansion of English poetry. 11th International Conference on Finite State Methods and Natural Language Processing. The Gateway, St Andrews, Scotland (UK), July 15–17, 2013. 2013: 18–24.

[59] **Delmonte R.** Computing poetry style. CEUR Workshop Proceedngs. 2013; (1096):148–155. Available at: `http://ceur-ws.org/Vol1096/paper11.pdf`.

[60] SPARSAR. Available at: `https://sparsar.wordpress.com`.

[61] **Delmonte R., Tonelli S., Boniforti M.A.P., Bristot A., Pianta E.** VENSES — a linguistically-based system for semantic evaluation. Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising

Tectual Entailment. 2005: 344–371. DOI:10.1007/11736790_20. Available at: https://www.researchgate.net/publication/225240840_VENSES_-_A_Linguistically-Based_System_for_Semantic_Evaluation.

[62] **Bacalu C., Delmonte R.** Prosodic modeling for speech recognition. Atti del Workshop AI*IA, "Elab.Ling.e Ric.". IRST Trento; 1999: 45–55.

[63] **McCurdy N., Srikumar V., Meyer M.** RhymeDesign: a tool for analyzing sonic devices in poetry. 4th Workshop on Computational Linguistics for Literature. 2015: 12–22. Available at: https://sci.utah.edu/~vdl/papers/2015_clfl_rhymedesign.pdf.

[64] **McCurdy N., Lein J., Coles K., Meyer M.** Poemage: visualizing the sonic topology of a poem. IEEE Transactions on Visualization and Computer Graphics. 2016; 22(1):439–448.

[65] **Calin O.** Statistics and machine learning experiments on English and Romanian poetry. Applied Sciences. 2020; 2(4):92. DOI:10.3390/sci2040092.

[66] **Shannon C.E.** A mathematical theory of communication. Bell System Technical Journal. 1948; 27(3):379–423.

[67] **Chishlom D.** Phonology and style: a computer-assisted approach to German verse. Computers and the Humanities. 1981; (15):199–210.

[68] Metricalizer^2. Available at: https://metricalizer.de.

[69] **Bobenhausen K., Hammerich K.** Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer^2. Traitement automatique des textes versifiés: problématiques et pratiques. Languages. 2015; 199(3):67–87.

[70] Freiburger anthologie. Textgrid. Available at: https://metricalizer.de/en/about.

[71] **Wells J.C.** Computer-coding the IPA: a proposed extension of SAMPA. Available at: https://www.phon.ucl.ac.uk/home/sampa/ipasam\T2A\textendashx.pdf.

[72] **Estes A., Hench C.** Supervised machine learning for hybrid meter. 5th Workshop on Computational Linguistics for Literature. 2016: 1–8. Available at: https://aclanthology.org/W16-0201.pdf.

[73] **Gervás P.** A logic programming application for the analysis of Spanish verse. 1st International Conference on Computational Logic. London, UK, July 24–28, 2000: 1399. Available at: https://link.springer.com/chapter/10.1007/3-540-44957-4_89.

[74] **Araújo P.A., Mamede N.J.** Classificador de poemas. Conferência Cientifíca e Tecnológica em Engenharia. Lisboa, Portugal, 2002.

[75] **Araújo P.A.M.** Classificação de poemas e sugestão das palavras finais dos versos. Diss. (Mestrado). Universidade Técnica de Lisboa; 2004.

[76] **Mamede N., Trancoso I., Arauújo P., Viana C.** Poetry assistant. Proceedings of the 8th International Conference on Spoken Language Processing. 2004: 3088.

[77] **Mamede N., Trancoso I., Araújo P., Viana C.** An electronic assistant for poetry writing. Advances in Artificial Intelligence — IBERAMIA 2004, 9th Ibero-American Conference on AI. Puebla, México, 2004: 286–294. DOI:10.1007/978-3-540-30498-2_29. Available at: https://www.researchgate.net/publication/220943156_An_Electronic_Assistant_for_Poetry_Writing.

[78] **Marques J.A.D.** Sistema de apoio á escrita de poemas. Diss. Universidade Técnica de Lisboa; 2008: 89.

[79] **Oliveira L.C., Viana M.C., Trancoso I.M.** A rule-based text-to-speech system for Portuguese. International Conference on Acoustics, Speech, and Signal Processing. 1992; (2):73–76. Available at: https://ieeexplore.ieee.org/document/226117.

[80] **Robinson J.R.** Colors of poetry: computational deconstruction. Georgia State University; 2006. Available at: https://getd.libs.uga.edu/pdfs/robinson_jason_r_200605_ma.pdf.

[81] **Navarro-Colorado B.** A computational linguistic approach to Spanish Golden Age sonnets: metrical and semantic aspects. Fourth Workshop on Computational Linguistics for Literature. USA: Denver; 2015: 105–113. DOI:10.3115/v1/W15-0712. Available at: https://www.researchgate.net/publication/316284856_A_computational_linguistic_approach_to_Spanish_Golden_Age_Sonnets_metrical_and_semantic_aspects.

[82] **Navarro-Colorado B., Lafoz M.R., Sánchez N.** Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation. 9th International Conference on Language Resources and Evaluation. 2016: 5. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf.

[83] **Navarro-Colorado B., Lafoz M.R., Trigueros S.J., Sánchez N.** Compilación y anotación métrica de un corpus de sonetos del Siglo de Oro. II Congreso Internacional Humanidades Digitales Hispánicas: "Innovación, Globalizació e Impacto". Madrid, España, 5–7 Octobre, 2015. Available at: https://hispanismo.cervantes.es/congresos-y-cursos/ii-congreso-internacional-humanidades-digitales-hispanicas-innovacion-0.

[84] Text encoding initiative. Available at: https://tei-c.org.

[85] **Robey D.** Scanning Dante's the Divine Comedy. A computer-based approach. Literary and Linguistic Computing. 1993; 8(2):81–84.

[86] **Rainsford T.M., Scrivner O.** Metrical annotation for a verse treebank. The 13th International Workshop on Treebanks and Linguistic Theories (TLT13). Germany: Tübungen; 2014: 149–159. Available at: https://www.researchgate.net/publication/269410991_Metrical_Annotation_For_a_Verse_Treebank_wwwoldoccitancorpusorg.

[87] **Roubaud J.** DYNASTIE: études sur le vers Français, sur l'alexandrin classique. Cahiers de poétique comparée. Premiére Partie. 1986; (13):47–109.

[88] **Roubaud J.** DYNASTIE: études sur le vers Français, sur l'alexandrin classique. Cahiers de poétique comparée. Deuxiéme Partie. 1988; (16):41–60.

[89] **Beaudouin V., Yvon F.** The Metrometer: a tool for analysing French verse. Literary and Linguistic Computing. 1996; 11(1):23–31.

[90] **Beaudouin V.** Métre en régles. Revue Française de Linguistique Appliquée. 2004; IX(1):119–137. DOI:10.3917/rfla.091.0119. Available at: https://www.researchgate.net/publication/268150694_Metre_en_regles.

[91] **Delente É., Renault R.** Annotation automatique des textes versifiés. Schedae. 2011: 39–52.

[92] **Delente É., Renault R.** Projet anamétre: le calcul du métre des vers complexes. Langages. 2015; 3(199):125–148. Available at: https://www.cairn.info/revue-langages-2015-3-page-125.htm&wt.src=pdf.

[93] **Delente É., Renault R.** Traitement automatique des formes métriques des textes versifiés. Actes de la 22e Conférence sur le Traitement Automatique des Langues Naturelles. Caen, France. ATALA2015:116–122. Available at: https://aclanthology.org/2015.jeptalnrecital-court.18.pdf.

[94] **Ayech H.E., Mahfouf A., Zribi A.** Reconnaissance de la métrique des poémes arabes par les réseaux de neurones artificiels. 13éme Conférence sur le Traitement Automatique des Langues Naturelles. 2006: 462–472. Available at: https://aclanthology.org/2006.jeptalnrecital-poster.10.pdf.

[95] **Kouloughli D.E.** Traitement automatique de la métrique arabe: réalisations et perspectives. Bulletin D'études Orientales. 2010; (LIX):17–31. Available at: https://www.cairn.info/revue-bulletin-d-etudes-orientales-2010-1-page-17.htm.

[96] **Almuhareb A., Alkharashi I., Saud L.AL., Altuwaijri H.** Recognition of classical Arabic poems. Proceedings of the Second Workshop on Computational Linguistics for Literature. Atlanta, Georgia, June 14, 2013. 2013: 9–16. Available at: https://aclanthology.org/W13-1402.pdf.

[97] **Kurt A., Kara M.** An algorithm for the detection and analysis of arud meter in Diwan poetry. Turkish Journal of Electrical Engineering and Computer Sciences. 2012; 20(6):948–963.

[98] **Alnagdawi M.A., Rashideh H., Aburumman A.F.** Finding Arabic poem meter using context free grammar. Journal of Communications and Computer Engineering. 2013; 3(1):52–59.

[99] **Wujastyk D.** Automatic scansion of sanskrit poetry for authorship criteria. Association for Literary and Linguistic Computing Bulletin. 1978; 6(2):122–135.

[100] **Mayrhofer C.M.** Scansion and analysis of Prakrit verses by text-processing programs. Revue Informatique et Statistique dans les Sciences Humaines. 1987; (XXIII):99–110.

[101] **Ousaka Y.M., Yamazaki M.M.** Automatic analysis of the Canon in Middle Indo-Aryan by personal computer. Literary and Linguistic Computing. 1994; 9(2):125–136.

[102] **Ousaka Y.M., Yamazaki M.M.** Automatic analysis of the Canon in Middle Indo-Aryan by personal computer II. Literary and Linguistic Computing. 1996; 11(1):9–17.

[103] **Rama N., Lakshmanan M.** A computational algorithm for metrical classification of verse. International Journal of Computer Science Issues. 2010; 7(2):46–53.

[104] **Rakshit G., Ghosh A., Bhattacharyya P., Haffari G.** Automated analysis of Bangla poetry for classification and poet identification. 12th International Conference on Natural Language Processing. 2015: 247–253. Available at: `https://aclanthology.org/W15-5937.pdf`.

[105] **Sgallová K.** Využití moderní techniky při rozboru verše. Česká Literatura. 1964; 12(2):158–168.

[106] **Sgallová K.** Thesaurus českých meter. Česká Literatura. 1999; 47(3):286–289.

[107] **Ibrahim R., Plecháč P.** Towards the automatic analysis of Czech verse. Formal methods in poetics. Lüdenscheid: RAM-Verlag; 2011: 295–305.

[108] **Plecháč P.** Czech verse processing system KVĚTA — phonetic and metrical components. Ğlottotheory. 2016; 7(7):159–174.

[109] **Plecháč P., Kolár R.** The corpus of Czech verse. Studia Metrica et Poetica. 2015; 2(1):107-118.

[110] **Гаспаров М.Л.** Эволюция русской рифмы. Проблемы теории стиха. Л.: Наука; 1984: 255.

[111] **Самойлов Д.** Книга о русской рифме. М.: Художественная литература; 1982: 351.

[112] **Брейдо Е.М.** Автоматический анализ метрики русского стиха. Автореферат диссертации по филологии. М.: Институт русского языка РАН им. В.В. Вшюградова; 1996: 26.

[113] **Брейдо Е.М.** Интервальная модель русской метрики. Вопросы языкознания. 1996; (4):85–94.

[114] **Брейдо Е.М.** Интервальная модель русской метрики и строгий тонический стих. Вопросы языкознания. 2021; (5):106–136.

[115] **Поляков А.Е., Пильщиков И.А., Бергельсон М.Б.** Конкорданс к текстам Ломоносова. ФЭБ; 2009. Адрес доступа: `http://feb-web.ru/feb/lomoconc/abc`.

[116] **Поляков А.Е., Пильщиков И.А., Бергельсон М.Б.** Конкорданс к текстам Ломоносова — концепция и реализация. Адрес доступа: `https://www.dialog21.ru/digests/dialog2009/materials/html/61.htm`.

[117] Электронное научное издание “Ломоносов”. Адрес доступа: `http://febweb.ru/feb/lomonos/default.asp`.

[118] Вавилонская Башня. Проект этимологической базы данных. Русские словари и морфология. Адрес доступа: `http://starling.rinet.ru/indexru.htm`.

[119] **Крылов С.А., Старостин С.А.** Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING. Международная конференция “Диалог”: Компьютерная лингвистика и интеллектуальные технологии. Архив. 2003. Адрес доступа: `https://www.dialog-21.ru/media/2655/krylov.pdf`.

[120] **Пильщиков И.А., Старостин А.С.** Автоматическое распознавание стихотворных размеров: теория и практика. Поэтика и фоностилистика: Бриковский сборник. Вып. 1. Материалы Международной научной конференции "I-е Бриковские чтения: Поэтика и фоностилистика". М.; 2010: 41–49.

[121] **Пильщиков И.А., Старостин А.С.** Проблема автоматического распознавания метра: силлаботоника, дольник, тактовик. Отечественное стиховедение: 100-летние итоги и перспективы развития. Материалы Международной научной конференции. 25–27 ноября 2010 г. СПб.; 2010: 397–406.

[122] **Бойков В.Н., Захаров В.Е., Каряева М.С., Соколов В.А.** Тезаурус по поэтологии как инструмент для информационного поиска и коллекции знаний. Моделирование и анализ информационных систем. 2013; 20(4):125–135. Адрес доступа: `http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=mais&paperid=327&option_lang=rus`

[123] **Дроздова И.И., Обухова А.Д.** Определение авторства текста по частотным характеристикам. Труды VII Международной научной конференции "Технические науки в России и за рубежом". М.: Буки-Веди; 2017: 18–21.

[124] **Гречников Е.А., Гусев Г.Г., Кустарев А.А., Райгородский А.М.** Поиск неестественных текстов. Труды XXI Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Петрозаводск: Транскрипт; 2009: 306–308.

[125] **Барахнин В.Б., Кожемякина О.Ю., Забайкин А.В.** Технология создания метрических справочников и конкордансов русских поэтических текстов. Труды Международной конференции "Вычислительные и информационные технологии в науке, технике и образовании". Алма-Ата; 2015: 244–245.

[126] **Barakhnin V., Kozhemyakina O., Grigorieva I.V.** Determination of the features of the author's style of A.S. Pushkin's poems by machine learning methods. Applied Sciences. 2022; (12):1674. DOI:10.3390/app12031674.

[127] **Кожемякина О.Ю.** Программная система комплексного анализа русских поэтических текстов: модели и алгоритмы. Дис. ... доктора техн. наук: 05.13.17 — Теоретические основы информатики. Новосибирск; 2022: 288.

Вычислительные технологии, 2023, том 28, № 3, с. 136–166. © ФИЦ ИВТ, 2023 ISSN 1560-7534
Computational Technologies, 2023, vol. 28, no. 3, pp. 136–166. © FRC ICT, 2023 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2023.28.3.009

Information systems for the analysis of poetic texts: history, methods and algorithms

Kozhemyakina Olga Yu.

Federal Research Center for Information and Calculated Technologies, 650090, Novosibirsk, Russia
Corresponding author: Kozhemyakina Olga Yu., e-mail: `olgakozhemyakina@mail.ru`
Received May 01, 2023, revised May 03, 2023, accepted May 05, 2023.

Abstract

The development of information systems for the analysis of poetic texts is a complex task, in the solution of which methods and algorithms are used, both belonging to the heritage of classical mathematics, and the most modern ones related to machine learning. The article presents a historical overview of existing systems, their components and methods used, modern research using the legacy of classical methods, but gained new opportunities due to the development of

information technology. The study of poetic texts using artificial intelligence methods, information theory methods is a promising direction in the tasks of natural language processing.

Keywords: natural language processing, software system design, poetic text analysis information system, text analysis methods, text analysis algorithms.

Citation: Kozhemyakina O.Yu. Information systems for the analysis of poetic texts: history, methods and algorithms. 2023; 28(3):136–166. DOI:10.25743/ICT.2023.28.3.009. (In Russ.)

Acknowledgements. To Academician Yuri Ivanovich Shokin with gratitude for the creation of the direction of natural language processing (NLP) at the Institute (FRC ICT), for ideas, advices and support.

References

1. **Uspensky V.A.** Introduction for readers of the "New Literary Observer" to the semiotic epistles of Andrei Nikolaevich Kolmogorov. New Literary Observer. 1997; (5):128. (In Russ.)
2. Analysis of poetic texts online. Available at: `www.poem.ict.nsc.ru`. (In Russ.)
3. **Kozhemyakina O.Yu.** Conceptual design of the software system for automated complex analysis of poetic texts. Computational Technologies. 2022; 27(2):122–137. DOI:10.25743/ICT.2022.27.2.010.
4. **Magomedova D.M.** Filologicheskiy analiz liricheskogo stikhotvoreniya [Philological analysis of a lyrical poem]. Moscow: Academiya; 2004: 187. (In Russ.)
5. **Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V., Khayatova V.D.** Automation of complex analysis of Russian poetic text: models and algorithms. Bulletin of NSU. Ser.: Information Technologies. 2015; 13(3):5–18. (In Russ.)
6. **Barakhnin V.B., Fedotov A.M., Bakiev A.M., Bakiev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tusupov D.A., Sambetaeva M.A., Lukpanova L.H.** Algorithms for generating stemmatization of word forms of the Kazakh language. Cloud of Science. 2017; 4(3):434–449. (In Russ.)
7. **Barakhnin V.B., Kozhemyakina O.Yu., Bakiyeva A.M., Sodboev M.K.** The algorithms for complex analysis of the corpuses of poetic texts in the Kazakh language. Journal of Physics: Conference Series. 2018; (1117):7. DOI:10.1088/1742–6596/1117/1/012003.
8. **Barakhnin V.B., Fedotov A.M., Bakiyeva A.M., Bakiyev M.N., Tazhibayeva S.Zh., Batura T.V., Kozhemyakina O.Yu., Tussupov D.A., Sambetbaiyeva M.A., Lukpanova L.Kh.** The software system for the study the morphology of the Kazakh language. The European Proceedings of Social and Behavioural Sciences. 2017; (XXXIII):18–27. DOI:10.15405/epsbs.2017.12.3.
9. **Markov A.A.** Example of statistical research on the text of "Eugene Onegin" illustrating the connection of tests in the chain. News of Imperial Academy of Sciences. 1913; 7(3):153–162. (In Russ.)
10. **Kirillov I.A.** O poeticheskoy informatsii. Informatsionnaya bezopasnost' i mezhkul'turnaya kommunikatsiya v kontekste tsifrovoy transformatsii [On poetic information. Information security and intercultural communication in the context of digital transformation]. Moscow; 2022: 408. (In Russ.)
11. Kolmogorov v vospominaniyakh [Kolmogorov in memoirs]. Moscow: Fizmatlit; 1993: 736. (In Russ.)
12. **Prokhorov A.V.** O rabotakh A.N. Kolmogorova po stikhovedeniyu. Kolmogorov A.N. Trudy po stikhovedeniyu [About the works of A.N. Kolmogorov on poetry. Kolmogorov A.N. Works on poetry]. Moscow: Izdatel'stvo MTsMNO; 2015: 256. (In Russ.)
13. **Uspensky V.A.** Introduction for readers of the "New Literary Observer" to the semiotic epistles of Andrei Nikolaevich Kolmogorov. New Literary Observer. 1997; (5):129. (In Russ.)
14. **Kolmogorov A.N., Prokhorov A.V.** K osnovam russkoy klassicheskoy metriki. Sodruzhestvo nauk i tayny tvorchestva [On the basics of Russian classical metrics. The Commonwealth of Sciences and the secrets of creativity]. Moscow: Iskusstvo; 1968: 397–432. (In Russ.)
15. **Kolmogorov A.N.** Trudy po stikhovedeniyu [Works on poetry]. Moscow: Izdatel'stvo MTsMNO; 2015: 256. (In Russ.)
16. **Gasparov M.L.** Vladimir Mayakovskiy. Ocherki istorii yazyka russkoy poezii KhKh veka: opyty opisaniya idiostiley [Vladimir Mayakovsky. Essays on the history of the language of Russian poetry of the twentieth century: experiments describing idiostyles]. Moscow: Nasledie; 1995: 557. (In Russ.)
17. **Gasparov M.L., Skulacheva T.V.** Stat'i o lingvistike stikha [Articles on the linguistics of verse]. Moscow: Yazyki slavyanskoy kul'tury; 2005: 132. (In Russ.)
18. **Levin Yu.D.** Concordance of Pushkin's poetry. Russian Literature. 1987; (1):212–214. (In Russ.)
19. **Shaw J.Th.** Pushkin: poet and man of letters and his prose. L.A.; 1995: 273.

20. **Shaw J.Th.** Konkordans k stikham A.S. Pushkina. (A–N). T. 1. [Concordance to the poems of A.S. Pushkin. (A–N). Vol. 1]. Moscow: Yazyki russkoy kul'tury; 2000: 674. (In Russ.)
21. **Shaw J.Th.** Konkordans k stikham A.S. Pushkina. (O–Ya). T. 2. [Concordance to the poems of A.S. Pushkin. (O–Ya). Vol. 2]. Moscow: Yazyki russkoy kul'tury; 2000: 641. (In Russ.)
22. **Kholshevnikov V.E.** O slovare rifm Pushkina. Vremennik Pushkinskoy komissii [About the Dictionary of Pushkin's rhymes. The Periodical of the Pushkin Commission]. 1973. Leningrad: Nauka; 1975: 135. (In Russ.)
23. **Mittmann A.** Escansão automática de versos em português. Tese (doutorado). Universidade Federal de Santa Catarina, Centro Tecnológico. Programa Pós–Graduação em Ciência da Computação. Florianópolis; 2016: 303. Available at: `https://repositorio.ufsc.br/bitstream/handle/123456789/175819/345411.pdf?sequence=1`.
24. **Pilshchikov I.A., Starostin A.S.** Osnovnye problemy avtomatizatsii bazovykh protsedur ritmiko-sintaksicheskogo analiza sillabo-tonicheskikh tekstov. Natsional'nyy korpus russkogo yazyka: 2006–2008. Novye rezul'taty i perspektivy [The main problems of automation of basic procedures of rhythmic-syntactic analysis of syllabotonic texts. National corpus of the Russian language: 2006–2008. New results and perspectives]. St. Petersburg: Nestor-Istoriya; 2009: 502. (In Russ.)
25. **Pilshchikov I.A., Starostin A.S.** Avtomaticheskoe raspoznavanie metra: problemy i resheniya. Slavyanskiy stikh [Automatic meter recognition: problems and solutions. Slavic verse]. Moscow: Rukopisnye pamyatniki Drevney Rusi; 2012; (9):568. (In Russ.)
26. **Pilshchikov I., Starostin A.** Automated analysis of poetic texts and the problem of verse meter. Current trends in metrical analysis. Littera: Studies in Language and Literature. Berlin; 2011; (2):368.
27. **Pilshchikov I., Starostin A.** Reconnaissance automatique des mètres des vers russes: une approche statistique sur corpus. Langages; 2015; 3(199):89–106.
28. **Boikov N.V., Karyaeva M.S., Sokolov V.A., Pilshchikov I.A.** On automatic verse specification in the information and analytical system. Proceedings of the XVII International Conference "Analytics and Data Management in Areas with Intensive Data Usage". Obninsk: OINPE NRNU MEPhI; 2015: 144–151. Available at: `http://ceur-ws.org/Vol-1536/paper22.pdf`. (In Russ.)
29. **Barakhnin V.B., Kozhemyakina O.Yu.** On automation of complex analysis of the Russian poetic text. CEUR Workshop Proceedings. 2012; (934):167–171. Available at: `http://ceur-ws.org/Vol-934/paper27.pdf`. (In Russ.)
30. **Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V.** Algorithms of complex analysis of Russian poetic texts in order to automate the process of creating metric reference books and concordances. CEUR Workshop Proceedings. 2015; (1536):138–143. Available at: `http://ceur-ws.org/Vol-1536/paper21.pdf`. (In Russ.)
31. MyStem. Available at: `https://tech.yandex.ru/mystem`.
32. Rifmofed.ru. All about rhyme and versification. Available at: `http://rifmoved.ru`.
33. **Mansilla R., Bush E.** Increase of complexity from classical Greek to Latin poetry. Complex Systems. 2003; 14(3). Available at: `https://arxiv.org/pdf/cond-mat/0203135.pdf`.
34. **Rényi A.** On measures of information and entropy. Proceedings of the 4th Berkeley Symposium on Mathematics, Statistics and Probability. 1960: 547–561.
35. **Yarkho V.N.** Gomerovskiy vopros. Literaturnaya entsiklopediya [The Homeric question. Literary encyclopedia]. Moscow: Sovetskaya Entsiklopediya; 1987: 78. (In Russ.)
36. **Fusi D.** A multilanguage, modular framework for metrical analysis: IT patterns and theorical issues. Langages. 2015; 3(199):41–66.
37. **Fusi D.** An expert system for the classical languages: metrical analysis components. Lexis. 2008; (27):25–45.
38. **Mishina Yu.E.** Appozitivnye konstruktsii i smezhnye sintaksicheskie yavleniya angliyskogo yazyka: problema razgranicheniya [Appositive constructions and related syntactic phenomenas of the English language: the problem of differentiation]. Filologicheskie Nauki. Voprosy Teorii i Praktiki. Tambov: Gramota; 2018; 1(1):155–158. DOI:10.30853/filnauki.2018-1-1.40. (In Russ.)
39. **Foley J.M.** A computer analysis of metrical patterns in Beowulf. Computers and the Humanities. 1978; (12):71–80.
40. **Hidley G.R.** Some thoughts concerning the application of software tools in support of old English poetic studies. Literary and Linguistic Computing. 1986; 1(3):156–162.
41. **Barquist C.R., Shie D.L.** Computer analysis of alliteration in Beowulf using distinctive feature theory. Literary and Linguistic Computing. 1991; 6(4):27–280.

42. **Donow H.S.** Prosody and the computer. A text processor for stylistic analysis. Spring Joint Computer Conference. 1970: 712.
43. **Barber C., Barber N.** The versification of The Canterbury Tales: a computer-based statistical study. Pt I. Leeds Studies in English. 1990; (21):81–103.
44. **Barber C., Barber N.** The versification of The Canterbury Tales: a computer-based statistical study. Pt II. Leeds Studies in English. 1991; (22):57–83.
45. **Greene E., Bodrumlu T., Knight K.** Automatic analysis of rhythmic poetry with applications to generation and translation. Conference on Empirical Methods in Natural Language Processing. 2010: 524–533.
46. **Hayward M.** A connectionist model of poetic meter. Poetics. 1991; (20):303–317.
47. **Hayward M.** Analysis of a corpus of poetry by a connectionist model of poetic meter. Poetics. 1996; 24(1):1–11. Available at: `http://www.english.iup.edu/mhayward/Metrics/Cormetrics.htm`.
48. **Plamondon M.R.** Computer-assisted phonetic analysis of English poetry: a preliminary case study of Browning and Tennyson. Text Technology. 2005; 14(2):153–175.
49. **Plamondon M.R.** Virtual verse analysis: analysing patterns in poetry. Literary and Linguistic Computing. 2006; 21(1):127–141.
50. **Hartman C.** The Scandroid. New London; 2005. Available at: `http://charlesohartman.com/verse/scandroid/ScandroidManual.pdf`.
51. The Scandroid. Available at: `http://charlesohartman.com/verse/scandroid/index.php`.
52. **Tizhoosh H.R., Dara R.A.** On poem recognition. Pattern Analysis and Applications. 2006; (9):325–338.
53. **Kaplan D.M., Blei D.M.** A computational approach to style in American poetry. 7th IEEE International Conference on Data Mining (ICDM 2007). 2007: 553–558.
54. **Kao J., Jurafsky D.** A computational analysis of style, affect, and imagery in contemporary poetry. NAACL Workshop on Computational Linguistics for Literature. 2012. Available at: `https://nlp.stanford.edu/pubs/kaojurafsky12.pdf`.
55. **Kavanagh F.** Analysis of a phonetic and rule based algorithm approach to determine rhyme categories and patterns in verse. Diss. (Mestrado). Open University; 2007.
56. **Genzel D., Uszkoreit J., Och F.** "Poetic" statistical machine translation: rhyme and meter. Conference on Empirical Methods in Natural Language Processing. 2010: 158–166.
57. **Hirjee H.** Rhyme, rhythm, and rhubarb: using probabilistic methods to analyze hip hop, poetry, and misheard lyrics. University of Waterloo. 2010. Available at: `https://uwspace.uwaterloo.ca/bitstream/handle/10012/5419/Hirjee_Hussein.pdf`.
58. **Agirrezabal M., Arrieta B., Astigarraga A., Hulden M.** ZeuScansion: a tool for scansion of English poetry. 11th International Conference on Finite State Methods and Natural Language Processing. The Gateway, St Andrews, Scotland (UK), July 15–17, 2013. 2013: 18–24.
59. **Delmonte R.** Computing poetry style. CEUR Workshop Proceedngs. 2013; (1096):148–155. Available at: `http://ceur-ws.org/Vol1096/paper11.pdf`.
60. SPARSAR. Available at: `https://sparsar.wordpress.com`.
61. **Delmonte R., Tonelli S., Boniforti M.A.P., Bristot A., Pianta E.** VENSES — a linguistically-based system for semantic evaluation. Machine Learning Challenges. Evaluating Predictive Uncertainty, Visual Object Classification, and Recognising Tectual Entailment. 2005: 344–371. DOI:10.1007/11736790_20. Available at: `https://www.researchgate.net/publication/225240840_VENSES_-_A_Linguistically-Based_System_for_Semantic_Evaluation`.
62. **Bacalu C., Delmonte R.** Prosodic modeling for speech recognition. Atti del Workshop AI*IA, "Elab.Ling.e Ric.". IRST Trento; 1999: 45–55.
63. **McCurdy N., Srikumar V., Meyer M.** RhymeDesign: a tool for analyzing sonic devices in poetry. 4th Workshop on Computational Linguistics for Literature. 2015: 12–22. Available at: `https://sci.utah.edu/~vdl/papers/2015_clfl_rhymedesign.pdf`.
64. **McCurdy N., Lein J., Coles K., Meyer M.** Poemage: visualizing the sonic topology of a poem. IEEE Transactions on Visualization and Computer Graphics. 2016; 22(1):439–448.
65. **Calin O.** Statistics and machine learning experiments on English and Romanian poetry. Applied Sciences. 2020; 2(4):92. DOI:10.3390/sci2040092.
66. **Shannon C.E.** A mathematical theory of communication. Bell System Technical Journal. 1948; 27(3):379–423.

67. **Chishlom D.** Phonology and style: a computer-assisted approach to German verse. Computers and the Humanities. 1981; (15):199–210.
68. Metricalizer^2. Available at: https://metricalizer.de.
69. **Bobenhausen K., Hammerich K.** Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer^2. Traitement automatique des textes versifiés: problématiques et pratiques. Languages. 2015; 199(3):67–87.
70. Freiburger anthologie. Textgrid. Available at: https://metricalizer.de/en/about.
71. **Wells J.C.** Computer-coding the IPA: a proposed extension of SAMPA. Available at: https://www.phon.ucl.ac.uk/home/sampa/ipasam-x.pdf.
72. **Estes A., Hench C.** Supervised machine learning for hybrid meter. 5th Workshop on Computational Linguistics for Literature. 2016: 1–8. Available at: https://aclanthology.org/W16-0201.pdf.
73. **Gervás P.** A logic programming application for the analysis of Spanish verse. 1st International Conference on Computational Logic. London, UK, July 24–28, 2000: 1399. Available at: https://link.springer.com/chapter/10.1007/3-540-44957-4_89.
74. **Araújo P.A., Mamede N.J.** Classificador de poemas. Conferência Cientifíca e Tecnológica em Engenharia. Lisboa, Portugal, 2002.
75. **Araújo P.A.M.** Classificação de poemas e sugestão das palavras finais dos versos. Diss. (Mestrado). Universidade Técnica de Lisboa; 2004.
76. **Mamede N., Trancoso I., Arauújo P., Viana C.** Poetry assistant. Proceedings of the 8th International Conference on Spoken Language Processing. 2004: 3088.
77. **Mamede N., Trancoso I., Araújo P., Viana C.** An electronic assistant for poetry writing. Advances in Artificial Intelligence — IBERAMIA 2004, 9th Ibero-American Conference on AI. Puebla, México, 2004: 286–294. DOI:10.1007/978-3-540-30498-2_29. Available at: https://www.researchgate.net/publication/220943156_An_Electronic_Assistant_for_Poetry_Writing.
78. **Marques J.A.D.** Sistema de apoio á escrita de poemas. Diss. Universidade Técnica de Lisboa; 2008: 89.
79. **Oliveira L.C., Viana M.C., Trancoso I.M.** A rule-based text-to-speech system for Portuguese. International Conference on Acoustics, Speech, and Signal Processing. 1992; (2):73–76. Available at: https://ieeexplore.ieee.org/document/226117.
80. **Robinson J.R.** Colors of poetry: computational deconstruction. Georgia State University; 2006. Available at: https://getd.libs.uga.edu/pdfs/robinson_jason_r_200605_ma.pdf.
81. **Navarro-Colorado B.** A computational linguistic approach to Spanish Golden Age sonnets: metrical and semantic aspects. Fourth Workshop on Computational Linguistics for Literature. USA: Denver; 2015: 105–113. DOI:10.3115/v1/W15-0712. Available at: https://www.researchgate.net/publication/316284856_A_computational_linguistic_approach_to_Spanish_Golden_Age_Sonnets_metrical_and_semantic_aspects.
82. **Navarro-Colorado B., Lafoz M.R., Sánchez N.** Metrical annotation of a large corpus of Spanish sonnets: representation, scansion and evaluation. 9th International Conference on Language Resources and Evaluation. 2016: 5. Available at: http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf.
83. **Navarro-Colorado B., Lafoz M.R., Trigueros S.J., Sánchez N.** Compilación y anotación métrica de un corpus de sonetos del Siglo de Oro. II Congreso Internacional Humanidades Digitales Hispánicas: "Innovación, Globalizació e Impacto". Madrid, España, 5–7 Octobre, 2015. Available at: https://hispanismo.cervantes.es/congresos-y-cursos/ii-congreso-internacional-humanidades-digitales-hispanicas-innovacion-0.
84. Text encoding initiative. Available at: https://tei-c.org.
85. **Robey D.** Scanning Dante's the Divine Comedy. A computer-based approach. Literary and Linguistic Computing. 1993; 8(2):81–84.
86. **Rainsford T.M., Scrivner O.** Metrical annotation for a verse treebank. The 13th International Workshop on Treebanks and Linguistic Theories (TLT13). Germany: Tübungen; 2014: 149–159. Available at: https://www.researchgate.net/publication/269410991_Metrical_Annotation_For_a_Verse_Treebank_wwwoldoccitancorpusorg.
87. **Roubaud J.** DYNASTIE: études sur le vers Français, sur l'alexandrin classique. Cahiers de poétique comparée. Premiére Partie. 1986; (13):47–109.
88. **Roubaud J.** DYNASTIE: études sur le vers Français, sur l'alexandrin classique. Cahiers de poétique comparée. Deuxiéme Partie. 1988; (16):41–60.

89. **Beaudouin V., Yvon F.** The Metrometer: a tool for analysing French verse. Literary and Linguistic Computing. 1996; 11(1):23–31.
90. **Beaudouin V.** Métre en régles. Revue Française de Linguistique Appliquée. 2004; IX(1):119–137. DOI:10.3917/rfla.091.0119. Available at: `https://www.researchgate.net/publication/268150694_Metre_en_regles`.
91. **Delente É., Renault R.** Annotation automatique des textes versifiés. Schedae. 2011: 39–52.
92. **Delente É., Renault R.** Projet anamétre: le calcul du métre des vers complexes. Langages. 2015; 3(199):125–148. Available at: `https://www.cairn.info/revue-langages-2015-3-page-125.htm&wt.src=pdf`.
93. **Delente É., Renault R.** Traitement automatique des formes métriques des textes versifiés. Actes de la 22e Conférence sur le Traitement Automatique des Langues Naturelles. Caen, France. ATALA2015:116–122. Available at: `https://aclanthology.org/2015.jeptalnrecital-court.18.pdf`.
94. **Ayech H.E., Mahfouf A., Zribi A.** Reconnaissance de la métrique des poémes arabes par les réseaux de neurones artificiels. 13éme Conférence sur le Traitement Automatique des Langues Naturelles. 2006: 462–472. Available at: `https://aclanthology.org/2006.jeptalnrecital-poster.10.pdf`.
95. **Kouloughli D.E.** Traitement automatique de la métrique arabe: réalisations et perspectives. Bulletin D'études Orientales. 2010; (LIX):17–31. Available at: `https://www.cairn.info/revue-bulletin-d-etudes-orientales-2010-1-page-17.htm`.
96. **Almuhareb A., Alkharashi I., Saud L.AL., Altuwaijri H.** Recognition of classical Arabic poems. Proceedings of the Second Workshop on Computational Linguistics for Literature. Atlanta, Georgia, June 14, 2013. 2013: 9–16. Available at: `https://aclanthology.org/W13-1402.pdf`.
97. **Kurt A., Kara M.** An algorithm for the detection and analysis of arud meter in Diwan poetry. Turkish Journal of Electrical Engineering and Computer Sciences. 2012; 20(6):948–963.
98. **Alnagdawi M.A., Rashideh H., Aburumman A.F.** Finding Arabic poem meter using context free grammar. Journal of Communications and Computer Engineering. 2013; 3(1):52–59.
99. **Wujastyk D.** Automatic scansion of sanskrit poetry for authorship criteria. Association for Literary and Linguistic Computing Bulletin. 1978; 6(2):122–135.
100. **Mayrhofer C.M.** Scansion and analysis of Prakrit verses by text-processing programs. Revue Informatique et Statistique dans les Sciences Humaines. 1987; (XXIII):99–110.
101. **Ousaka Y.M., Yamazaki M.M.** Automatic analysis of the Canon in Middle Indo-Aryan by personal computer. Literary and Linguistic Computing. 1994; 9(2):125–136.
102. **Ousaka Y.M., Yamazaki M.M.** Automatic analysis of the Canon in Middle Indo-Aryan by personal computer II. Literary and Linguistic Computing. 1996; 11(1):9–17.
103. **Rama N., Lakshmanan M.** A computational algorithm for metrical classification of verse. International Journal of Computer Science Issues. 2010; 7(2):46–53.
104. **Rakshit G., Ghosh A., Bhattacharyya P., Haffari G.** Automated analysis of Bangla poetry for classification and poet identification. 12th International Conference on Natural Language Processing. 2015: 247–253. Available at: `https://aclanthology.org/W15-5937.pdf`.
105. **Sgallová K.** Využití moderní techniky při rozboru verše. Česká Literatura. 1964; 12(2):158–168.
106. **Sgallová K.** Thesaurus českých meter. Česká Literatura. 1999; 47(3):286–289.
107. **Ibrahim R., Plecháč P.** Towards the automatic analysis of Czech verse. Formal methods in poetics. Lüdenscheid: RAM-Verlag; 2011: 295–305.
108. **Plecháč P.** Czech verse processing system KVĚTA — phonetic and metrical components. Ǧlottotheory. 2016; 7(7):159–174.
109. **Plecháč P., Kolár R.** The corpus of Czech verse. Studia Metrica et Poetica. 2015; 2(1):107-118.
110. **Gasparov M.L.** Evolyutsiya russkoy rifmy. Problemy teorii stikha [Evolution of Russian rhyme. Problems of the theory of verse]. Leningrad: Nauka; 1984: 255. (In Russ.)
111. **Samoylov D.** Kniga o russkoy rifme [The book on Russian rhyme]. Moscow: Khudozhestvennaya Literatura; 1982: 351. (In Russ.)
112. **Breido E.M.** Avtomaticheskiy analiz metriki russkogo stikha [Russian verse metric automatic analysis]. Avtoreferat dissertatsii po filologii. Moscow: Institut Russkogo Yazyka RAN im. V.V. Vshyugradova; 1996: 26. (In Russ.)
113. **Breido E.M.** Interval model of Russian metrics. Questions of Linguistics. 1996; (4):85–94. (In Russ.)
114. **Breido E.M.** Interval model of Russian metrics and strict tonic verse. Questions of Linguistics. 2021; (5):106–136. (In Russ.)

115. **Polyakov A.E., Pilshchikov I.A., Bergelson M.B.** Konkordans k tekstam Lomonosova [Concordance to the texts of Lomonosov]. FEB; 2009. Available at: http://feb-web.ru/feb/lomoconc/abc. (In Russ.)
116. **Polyakov A.E., Pilshchikov I.A., Bergelson M.B.** Konkordans k tekstam Lomonosova — kontseptsiya i realizatsiya [Lomonosov concordance — concept and implementation]. Available at: https://www.dialog21.ru/digests/dialog2009/materials/html/61.htm. (In Russ.)
117. Elektronnoe nauchnoe izdanie "Lomonosov" [Electronic scientific publication "Lomonosov"]. Available at: http://febweb.ru/feb/lomonos/default.asp. (In Russ.)
118. Vavilonskaya Bashnya. Proekt etimologicheskoy bazy dannykh. Russkie slovari i morfologiya [The Tower of Babel. Etymological database project. Russian dictionaries and morphology]. Available at: https://starlingdb.org/main.html. (In Russ.)
119. **Krylov S.A., Starostin S.A.** Aktual'nye zadachi morfologicheskogo analiza i sinteza v integrirovannoy informatsionnoy srede STARLING [Actual problems of morphological analysis and synthesis in the integrated information environment of STARLING]. International Conference "Dialogue": Computational Linguistics and Intelligent Technologies. Archive. 2003. Available at: https://www.dialog-21.ru/media/2655/krylov.pdf.
120. **Pilshchikov I.A., Starostin A.S.** Avtomaticheskoe raspoznavanie stikhotvornykh razmerov: teoriya i praktika [Automatic recognition of poetic dimensions: theory and practice]. Poetics and Phonostylistics: Brik's Collection. Is. 1. Proceedings of the International Scientific Conference "I-st Briks' readings: Poetics and Phonostylistics". Moscow; 2010: 41–49. (In Russ.)
121. **Pilshchikov I.A., Starostin A.S.** Problema avtomaticheskogo raspoznavaniya metra: sillabotonika, dol'nik, taktovik [The problem of automatic meter recognition: syllabotonics, dolnik, taktovik]. Russian Poetry: 100-year Results and Prospects of Development. Materials of the International Scientific Conference. November 25–27, 2010. St. Petersburg; 2010: 397–406. (In Russ.)
122. **Boykov V.N., Zakharov V.E., Karyaeva M.S., Sokolov V.A.** Thesaurus on poetologie as a tool for information retrieval and knowledge collection. Modeling and Analysis of Information Systems. 2013; 20(4):125–135. Available at: http://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=mais&paperid=327&option_lang=eng. (In Russ.)
123. **Drozdova I.I., Obukhova A.D.** Opredelenie avtorstva teksta po chastotnym kharakteristikam [Determining the authorship of the text by frequency characteristics]. Proceedings of the VII International Scientific Conference "Technical Sciences in Russia and Abroad". Moscow: Buki-Vedi; 2017: 18–21. (In Russ.)
124. **Grechnikov E.A., Gusev G.G., Kustarev A.A., Raygorodsky A.M.** Poisk neestestvennykh tekstov [Search for unnatural texts]. Proceedings of the XXI All-Russian Scientific Conference "Electronic Libraries: Perspective Methods and Technologies, Electronic Collections". Petrozavodsk: Transkript; 2009: 306–308. (In Russ.)
125. **Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V.** Tekhnologiya sozdaniya metricheskikh spravochnikov i konkordansov russkikh poeticheskikh tekstov [Technology of creating metric reference books and concordances of Russian poetic texts]. Proceedings of the International Conference "Computing and Information Technologies in Science, Technology and Education". Alma-Ata; 2015: 244–245. (In Russ.)
126. **Barakhnin V., Kozhemyakina O., Grigorieva I.V.** Determination of the features of the author's style of A.S. Pushkin's poems by machine learning methods. Applied Sciences. 2022; (12):1674. DOI:10.3390/app12031674.
127. **Kozhemyakina O.Yu.** Programmnaya sistema kompleksnogo analiza russkikh poeticheskikh tekstov: modeli i algoritmy [Software system for complex analysis of Russian poetic texts: models and algorithms]. Dis. ... Doctor of Technical Sciences: 05.13.17 — Theoretical Foundations of Computer Science. Novosibirsk; 2022: 288. (In Russ.)