Санкт-Петербургский государственный университет

Медведев Никита Дмитриевич

**Выпускная квалификационная работа**

**Создание инструмента для исследования метрики и ритмики поэтического текста**

Уровень образования: магистратура

Направление 45.04.02 «Лингвистика»

Основная образовательная программа ВМ.5805. «Компьютерная и прикладная лингвистика»

Профиль «Компьютерная лингвистика и интеллектуальные технологии»

Научный руководитель:

доцент, Кафедра математической лингвистики,

Митренина Ольга Владимировна

Рецензент:  
доцент, ФГБОУВО «Вятский государственный университет»,

Валова Ольга Михайловна

Санкт-Петербург

2022

[**Введение**](#_gjdgxs) **[3](#_gjdgxs)**

[**Глава 1. Теоретические аспекты русской метрики**](#_30j0zll) **[8](#_30j0zll)**

[1.1. Анализ становления и развития русской силлабо-тоники](#_1fob9te) [8](#_1fob9te)

[1.2. Исследования актуальных статистических моделей анализа стихотворного размера](#_3znysh7) [25](#_3znysh7)

[**Глава 2. Создание инструмента для автоматического анализа поэтической метрики**](#_2et92p0) **[47](#_2et92p0)**

[2.1. Имплементация классификатора поэтической метрики, основанного на правилах](#_tyjcwt) [48](#_tyjcwt)

[2.2. Имплементация классификатора поэтической метрики с применением методов машинного обучения](#_3dy6vkm) [56](#_3dy6vkm)

[2.3. Сравнительный анализ результатов работы классификатора поэтической метрики, основанного на строгих правилах, и классификатора поэтической метрики с применением методов машинного обучения](#_1t3h5sf) [66](#_1t3h5sf)

[**Заключение**](#_4d34og8) **[69](#_4d34og8)**

[**Библиографический список**](#_2s8eyo1) **[80](#_2s8eyo1)**

[**Приложения**](#_17dp8vu) **84**

# Введение

***Проблема***

Мировая тенденция к глобализации проникает во все сферы жизни. В условиях постоянного обновления и совершенствования технологий от современного специалиста требуется столь же постоянное обучение и поиск новых способов, методов и техник применения передовых наработок науки к своей сфере деятельности. Именно в этом русле лежит проблема создания инструмента для автоматического анализа поэтического текста. Мы не можем не признавать, что проблема определения размера стихотворения до сих пор остаётся не разрешённой в достаточной мере и вызывает множество затруднений и разночтений.

Между тем, мы должны отметить, что данная проблема не выпадала из внимания учёных, периодически вызывая дополнительный интерес, связанный с появлением какой-либо новой технологии или идеи, которую можно было бы применить к существующей проблеме и получить возможные положительные результаты. Так случилось в начале XX века, когда к проблеме определения размера были применены математические методы, потом то же самое произошло уже в последней трети XX века, когда компьютеры получили широкое распространение в науке. Так происходит и в последние несколько лет, когда компьютерных мощностей становится достаточно для применения алгоритмов машинного обучения и нейросетей.

***Актуальность*** изучения возможностей создания инструмента для автоматического определения поэтической ритмики текста заключается в том, что данная проблема представляет собой постоянное препятствие как для школьников, пытающихся определить размер стихотворения на уроке литературы, так и для исследователей, занимающихся анализом поэтического текста и его влиянием на психику и когнитивные способности человека. При этом новые наработки в области глубокого обучения и нейросетей позволяют надеяться, что их можно будет применить и к данной проблеме.

***Степень научной разработанности темы***

Поэтическая ритмика не раз выступала объектом исследования в разных гуманитарных науках. Однако попыток создания инструмента для автоматического анализа текста было не так много, более того, не все из этих попыток заканчивались рабочими прототипами, пригодными к использованию.

Следует отметить, что практические подходы к созданию данного инструмента наметились в двух вариациях. Во-первых, это наиболее часто используемый метод классификатора, основанного на строгих правилах, где в систему изначально заложены схемы размеров, и её задача состоит в сравнении подаваемой пользователем строки с этими размерными сетками и в выдаче наиболее подходящего варианта. Во-вторых, существует разновидность вышеупомянутого классификатора, в котором он не сравнивает заранее известный размер с размером, полученным для строки, а пытается на основе выявления закономерностей и слогов определить его с нуля. Иной подход связан с использованием алгоритмов машинного обучения и расширением способов применения задачи небинарной классификации, для которой сперва собирается объёмный корпус размеченных стихотворных строк и даётся машине на обучение, после чего происходит извлечение различных свойств и закономерностей, по которым машина старается предсказать размер строки, полученной на вход.

***Гипотеза***

Выдвигается предположение о том, что через применение алгоритмов машинного обучения возможно будет добиться точности определения размера стихотворения выше, чем при применении классификатора, основанного на строгих правилах, поскольку технологии машинного обучения являются более адаптивными и лучше смогут работать с такой непостоянной и многоаспектной структурой как стихотворный размер.

***Объект исследования*** — русская поэтическая метрика.

***Предмет исследовательской работы*** — поэтический размер текста, программы и библиотеки для создания инструмента анализа размера поэтического текста, способы разработки инструментов для автоматического анализа метрики и ритмики поэтического текста.

***Цель работы*** — создание и сравнительный анализ эффективности двух прототипов для анализа поэтического размера текста.

Поставленная цель определила следующие ***задачи***:

* 1. Проанализировать путь становления и развития русской поэтической метрики;
  2. Выявить наиболее подходящий и актуальный список размеров для анализа;
  3. Изучить предыдущий опыт создания инструментов для автоматического анализа метрики и ритмики поэтического текста как в русскоязычных исследованиях, так и в англоязычной среде;
  4. Разработать прототип классификатора, основанного на строгих правилах, для задачи автоматического определения стихотворного размера;
  5. Разработать прототип классификатора с применением алгоритмов машинного обучения для задачи автоматического определения стихотворного размера;
  6. Провести сравнительный анализ эффективности и точности определения поэтической метрики текста.

В работе над диссертацией используется комплекс ***методов*** современного языкознания и программирования, направленных на теоретическое и практическое осмысление основных аспектов исследования русской поэтической метрики в её современном осмыслении.

***Материал исследования***

Стоит обозначить, что исследование ограничено, прежде всего, силлабо-тоническими размерами как теми, которые до сих пор получают наибольшее распространение в образовательной среде. С учётом относительного постоянства размерной сетки на протяжении всего произведения нами были отобраны стихотворения общим размером более пяти тысяч стихотворных строк, по одной тысяче строк каждого размера соответственно. Стихотворения были отобраны из материалов поэтического подкорпуса НКРЯ, Полный список этих стихотворений с разбивкой по размерам можно увидеть в Приложении 1.

***Положения, выносимые на защиту***

Создание инструмента для автоматического анализа поэтической метрики и ритмики текста; сравнительный анализ точности классификатора, основанного на строгих правилах, с алгоритмом машинного обучения; применение алгоритмов машинного обучения для дальнейшего анализа поэтических текстов.

***Теоретическая новизна***

В данном исследовании впервые в отечественном гуманитарном знании представлен развернутый анализ современных англоязычных исследований, посвящённых проблематике применения алгоритмов машинного обучения и нейронных сетей к определению метрики и ритмики поэтического текста. Проанализированы материалы для языков как с подвижным, так и с фиксированным ударением. Материалы работы могут послужить хорошей отправной точкой для более масштабных и детальных исследований по данной теме.

***Практическая новизна***

Данное исследование может послужить как научно-методический материал для исследований, связанных с областью исследования русской поэтической метрики и способов применения алгоритмов машинного обучения для упрощения этой задачи, а также как основа для будущих исследований, расширения и углубления.

***Структура выпускной квалификационной работы***

Работа состоит из Введения, двух глав, Заключения, Списка использованной литературы и Приложения. Во Введении представлена проблема исследования, актуальность, степень разработанности темы, обозначены объект, предмет, цель и задачи исследования. В первой главе представлено рассмотрение становления и развития русской поэтической традиции, а также обзор современных исследований проблематики использования алгоритмов машинного обучения при решении задачи автоматического определения метрики и ритмики текста. Вторая глава посвящена созданию двух прототипов инструментов для анализа поэтического текста, а именно — классификатора, основанного на строгих правилах, и алгоритма машинного обучения. В заключении подведены основные выводы и итоги исследования. Общий объём работы — 85 страниц.

# Глава 1. Теоретические аспекты русской метрики

## 1.1. Анализ становления и развития русской силлабо-тоники

Прежде чем мы перейдём к вопросу об истории русского стиха и эволюции русской метрики, мы должны упомянуть отдельный аспект, кажущийся нам заслуживающим внимания. При обращении к истории русского стиха и стихотворной речи можно заметить определённую странность — её обсуждение начинается только с XVII века. Это кажется не вполне интуитивно понятным, ведь подводит нас к мысли о том, что на Руси до этого периода не существовало стихотворных изобразительных средств и самого стиха. Это, конечно, не совсем соответствует действительности. В своей статье «Оппозиция “Стих – Проза” и становление русского литературного стиха» известный литературовед и теоретик литературы Михаил Леонович Гаспаров пишет о причинах этого феномена: “Нужно присмотреться ближе, чтобы уточнить это впечатление: поэзия существовала, стихотворные средства выражения — ритм и рифма — существовали, но стиха, действительно, не существовало.

Все средства стихотворной речи, во главе с ритмом и рифмой, были доступны уже древнерусской литературе. Однако, существуя порознь и даже в совокупности, все эти средства не складывались в понятие «стих».

Противоположность «стих — проза», которая ныне кажется столь очевидной, для древнерусского человека не существовала. Она появилась только в начале XVII в. и была отмечена новым словом в русском языке, ранее неизвестным, а стало быть, ненужным: словом «вирши» — стихи. До этого вместо противоположности «стих — проза» в сознании древнерусского человека жила другая противоположность: «текст поющийся — текст произносимый». При этом в первую категорию одинаково попадали народные песни и литургические песнопения, а во вторую — деловые грамоты и риторическое «плетение словес», хотя бы и насквозь пронизанные ритмом и рифмами. Это противоположение не было единственным; одинаково отчетливо ощущалась в древней Руси, например, и противоположность «книжная словесность — народная словесность». Но и накладываясь друг на друга, такие противопоставления не давали привычных нам понятий «стих — проза». Вот почему ни появление ритма, ни появление рифмы в древнерусских текстах не означало для читателя, что перед ним — «стих»“ [1, с. 265-266]. Исходя из вышеприведённой цитаты мы считаем уместным в дальнейшем, используя термины “русский стих” или “русская поэтическая традиция”, ограничиваться рамками XVII-XXI веков и не обращаться к древнерусским текстам.

Также стоит определить, что именно мы считаем стихотворной речью и по каким признакам мы отделяем её от прозы. Здесь мы позволим себе сослаться на Бориса Викторовича Томашевского, крупного советского теоретика стиха и текстолога, который определял два признака, отличающих стихотворную речь от прозаической. Он писал:

“1) стихотворная речь дробится на сопоставимые между собой единицы (стихи), а проза есть сплошная речь;

2) стих обладает внутренней мерой (метром), а проза ею не обладает...” [2, с. 10]. Далее мы будем придерживаться именно такого определения в разговоре о поэтической речи и стихосложении.

Как известно, в стихотворной речи (как и в прозаической, но в данном вопросе мы позволим себе оставить её без внимания) речь разделена на предложения, которые, в свою очередь, поделены на синтагмы, связанные интонационно и по смыслу группы слов. Предложения и синтагмы отделены между собой логическими ударениями, паузами разной длительности, понижением или повышением тона голоса и так далее. Стих — это основная повторяющаяся ритмическая единица в любом стихотворении, однако он различается по внутренней структуре в зависимости от системы стихосложения, чьи особенности и различия, в свою очередь, связаны с фонетическими и культурными особенностями языка. Далее мы будем рассматривать три наиболее распространённые системы стихосложения и их развитие в контексте русского языка. Мы будем говорить о метрической, силлабической и тонической системах стихосложения.

В предыдущем параграфе мы упомянули о фонетических особенностях языка, которые могут непосредственно влиять на природу стихотворного ритма — в русском языке такую функцию играет подвижность ударения: иногда она приводит к изменению формы (*рукИ* и *рУки*) или значения (*зАмок* и *замОк*) слова. Как пишет Холшевников Владислав Евгеньевич, автор известного учебника «Основы стиховедения», хрестоматии по русскому стихосложению, “подобные случаи встречаются в речи постоянно. Поэтому ухо русского не может не слышать ударения, и в русском языке оно гораздо энергичнее, чем в польском или французском, благодаря чему различие между ударными и безударными гласными заметно увеличивается по сравнению с этими языками. В русском языке ударный слог качественно не равен безударному и значительно заметнее последнего. Поэтому основой ритма в русском языке (как в немецком или английском) является повторение ударных слогов; расположение безударных слогов может учитываться, но может и не приниматься во внимание. Такая система стихосложения называется тонической (от греч. zovos — тон, напряжение, ударение). Присуща она языкам с подвижным ударением” [3, с. 15]. Холшевников справедливо отмечает здесь, что расположение безударных слогов может учитываться, а может и не учитываться в тонической системе стихосложения, однако в русскоязычном стихосложении параллельно существовали две системы стихосложения, различавшиеся именно по этому критерию: русская силлабика в своё время сменилась на русскую силлабо-тонику благодаря реформе Ломоносова-Тредиаковского, параллельно с этим продолжала своё существование русская тоника (стихотворения, написанные, к примеру, дольником, популярным тоническим размером, мы можем встретить как у Пушкина [4, с. 135], так и у Тютчева [5]), которая получила своё активное распространение в период Серебряного века русской поэзии. В контексте нашего исследования мы считаем необходимым обратиться к силлабо-тонике, оставляя за собой право в дальнейшем считать русскую тонику перспективным направлением для расширения работы нашего инструмента в случае, если результат работы для силлабо-тонических размеров будет сочтён удовлетворительным.

Прежде всего, при разговоре о силлабо-тоническом стихосложении, нужно разобраться с силлабическим стихосложении. Для начала дадим определение. “Силлабическое стихосложение (от греч. syllabe — слог) — система стихосложения, в которой длина стиха определяется только количеством слогов вне зависимости от количества ударений; стихи называются двух-, трёх-, четырёх-, пяти-, шести-сложными и т.д. В строках допускается и разное количество слогов; желательно только, чтобы четносложные стихи сочетались с четносложными, а нечетносложные — с нечетносложными. В 10-, 11- и более сложных стихах появляется цезура — обязательный словораздел, членящий стих на короткие полустишия” [6]. Этим определением мы будем оперировать в дальнейшем, хотя и опуская в ходе рассмотрения нашей задачи вопрос цезуры [28], поскольку её наличие или отсутствие существенным образом не влияет на определение поэтического метра.

В дискуссии о русской силлабике и в особенности о причинах её замены на силлабо-тонику достаточно часто в качестве одной из причин этого перехода приводится аргумент о “скучности” и “однообразии” силлабики, её чуждости русскому уху. Эта точка зрения не вполне соответствует действительности, вот что об этой позиции пишет уже упомянутый нами ранее М.Л. Гаспаров в своём «Очерке истории русского стиха»: “силлабика никоим образом не была случайной заемной модой, будто бы не соответствовавшей «духу» русского языка. Это был стих широко употребительный, хорошо разработанный и гибкий. Если он и отступил перед натиском силлабо-тоники, то только потому, что та своей однообразной строгостью лучше подчеркивала противопоставление стиха и прозы” [7, с. 33-34].

В XVIII в. русская культура постепенно секуляризуется, содержание различных произведений, в том числе и поэтических, становится светским, духовная поэзия теряет актуальность. Нарастает противоречие между искусственно поддерживаемой церковной распевной манерой декламации стихотворений, присущей силлабическому стихосложению, и содержанием этих творений. В этой связи в культурном русском обществе возникает запрос на поиск новых форм выражения стиха. Успех подобного поиска происходит в общей тенденции беспрецедентного роста русской культуры. Не является случайностью то, что авторами реформы русского стихосложения и его перехода к силлабо-тонике выступили В.К. Тредиаковский и М.В. Ломоносов. Василий Кириллович Тредиаковский в 1735 году выпустил трактат «Новый и краткий способ к сложению российских стихов», где изложил теоретическую подоплёку оснований русского стихосложения и предложил способы дальнейшего его улучшения. Тредиаковский писал: “...долгота и краткость слогов, в новом сем Российском Стихосложении, не такая, разумеется, какова у Греков и у Латин в сложении Стихов употребляется; но токмо тоническая, то есть в едином ударении голоса состоящая... в чем вся сила нового сего Стихосложения содержится” [8, с. 368]. Этот отрывок свидетельствует об осознании Тредиаковским качественного различия ударных и безударных слогов, которое и приведёт его к выводу о необходимости перехода к силлабо-тонике. Именно Василий Кириллович введёт в русскую поэзию понятие стопы как основной ритмической единицы, состоящей из ударного и безударного слога, чередующихся в определённом порядке, составляющем главный метр произведения.

Человеком, завершившим реформу российского стихосложения, начатую Тредиаковским, стал Михаил Васильевич Ломоносов, который в 1739 году прислал в Россию «Ода блаженныя памяти Государыне Императрице Анне Иоанновне на победу над Турками и Татарами и на взятие Хотина», в которой применил упорядоченное чередование ударных и безударных слогов. Также вместе с этой одой им был отправлен теоретический трактат «Письмо о правилах российского стихотворства», в котором была описана завершённая система силлабо-тонического метра. Основным положением Ломоносова, которое сподвигло его на создание и описание подобной системы, было убеждение в отсутствии пользы заимствований из зарубежных культур без осмысления их применимости к условиям русской культуры. В своём трактате М.В. Ломоносов писал: “Российские стихи надлежит сочинять по природному нашего языка свойству; а того, что ему весьма несвойственно, из других языков не вносить” [9, с. 486]. Учёный заметил в знакомой ему немецкой поэзии с похожим на русское акцентным подвижным ударением разработанную систему образов тонического стиха и осознал, что опыт этого стихосложения может быть перенесён и на русский стих. Далее в нашей работе мы также увидим, что наиболее приближенный к нашему опыт исследования метрики стиха был осуществлён также на материале средненемецкой поэзии.

Важное дополнение, сделанное Ломоносовым, к предложенному Тредиаковским положению о правилах российского стихосложения, состояло в расширении понятия стопы и добавлении к нему трёхсложных размеров — стихотворная стопа, по замыслу Ломоносова, могла состоять не только из одного ударного и одного безударного слогов, но также из одного ударного и двух безударных слогов, при этом расположение ударных и безударных слогов относительно друг друга играл важную роль, в разных вариациях определяя различные размеры стиха.

Помимо этого, новаторством Ломоносова в области преобразования российского стихотворства являлось введение системы метров, которая была основана на регулярном и упорядоченном чередовании безударных и ударных слогов в строке.

Справедливо будет заметить, что “правильные” схемы размеров, объяснённые ранее, принимаются в современной российской школе при изучении произведений, однако, как мы укажем далее, это определение не обладает достаточной точностью для трёхсложных размеров стиха и имеет больший процент исключений, чем совпадений, для двусложных размеров. Также стоит отметить, что современное литературоведение склоняется к иктовой теории и отказывается от стопной теории Тредиаковского и Ломоносова, которая выглядит справедливой лишь при беглом анализе стихотворения, однако мы будем придерживаться именно этой теории, поскольку основная задача нашей программы — создание инструмента для помощи школьникам, изучающим литературу, а в школьной программе, как мы упомянули, данная теория до сих пор является актуальной. Тем не менее, не упомянуть о современном взгляде литературоведческой науки на этот вопрос мы также не можем.

Даже М.В. Ломоносов при проработке положений теории силлабо-тонического стиха выявил особенность двусложных размеров для русского языка: далеко не всегда на месте ударного слога, который подразумевается по схеме размера, стоит ударный слог — это сперва показалось Ломоносову недостатком стиха. Однако стоит отметить, что если провести анализ большого объёма русскоязычных поэтических текстов, то можно заметить, что позиций, в которых ударные и безударные слоги совпадают с изначальными схемами размеров, меньше, чем тех, в которых данные позиции не совпадают. В случае, когда исключений из правила больше, чем случаев, описываемых этим правилом, едва ли можно говорить о системной ошибке, скорее — о некоем языковом свойстве, которое присуще русскому языку. Дело в том, что при анализе достаточного количества любого текста, написанного на русском языке, можно увидеть то, к чему ещё в середине XIX века пришёл Н.Г. Чернышевский в своей статье, посвящённой изданию сочинений А.С. Пушкина П.В. Анненским. Там он писал: “Чтобы видеть, в какие стопы всего естественнее должна ложиться рус­ская речь, попробуем сосчитать количество ударений, в ней нахо­дящихся. <...> Вот несколько строк из первой страницы повести г. Писемского «Виновата ли она?» <...> Всего 193 слога; ударений 66; 3x66=198. Итак, только пяти слогов не достает, чтобы количество слогов было втрое больше количества ударений. Вот начало рассказа «Голубые глазки» <...> Всего 75 слогов и 25 ударений. Совпадение чисел точно до странности. Вот последний отрывок — начало первой главы пере­водного романа «Часовщик» <...> 83 слога и 27 ударений; 3x27=81; следовательно, только два слога лишних против точного определения. Соединив итоги всех трех отрывков, получим 351 слог и в них 118 ударений; 3x118=354. Итак, уклонение от точного размера: на три слога одно ударение (дактиль, амфибрахий, анапест) равняется только трем слогам на 351 слог, или одной двадцатой. Близость удиви­тельная.

Нам кажется, что из этих цифр нельзя не извлечь заключе­ния, что ямб и хорей, требующие в 30 слогах 15 ударений, далеко не так естественны в русском языке, как дактили, амфибрахии, анапесты, требующие в 30 слогах 10 ударений” [10, c. 469-471].

Таким образом, можно прийти к выводу, что средняя длина русского слова — три слога. Этот факт важен для осознания особенностей русской метрики и поэтики, поскольку логическим выводом из него следует некая амбивалентность положения русских двусложных размеров в русской метрике: если поэт будет стремиться писать схематическими ямбом и хореем, то он будет вынужден существенно ограничить свой лексикон, поскольку будет вынужден специально подыскивать короткие слова. Даже при употреблении поэтом трёхсложного (естественного для русского языка, как мы упомянули чуть ранее) слова, ему нужно будет поставить рядом односложное. При этом для подобной схемы подойдут только те трёхсложные слова, ударение в которых стоит посередине, иначе они также будут выпадать из схемы. Подобные стихотворения будут выглядеть искусственно и, что важнее, совершенно напрасно и бесцельно ограничат поэтический лексикон. Исходя из этих размышлений, Ломоносов и Тредиаковский предположили, что пропуск ударения в случае двусложного размера означает собой появление особой стопы — пиррихия, которая состоит из двух безударных слогов. Из того, что при постоянно встречающихся пиррихиях в хорее и в ямбе и, более того, при их численном преимуществе по отношению к “правильным” стихам, укладывающимся в изначальную схему размера, может быть сделано логическое заключение о несоответствии данного схематического разделения двусложных размеров с реальностью, о представлении метрики стихотворения в виде чередования одинаковых стоп как об ошибочном и не вполне верном тезисе, который по причине поверхностного удовлетворения формальным критериям повторяется авторами школьных учебников до сих пор.

К подобному выводу пришли исследователи первой трети XX века, времени, отмеченного особым интересом литературоведов и филологов к теории стиха. Такой всплеск интереса к поэзии не был единственным в истории развития русской поэзии — поэзия на разных этапах становления выдвигала новые требования к поэтам, а значит, к этим же проблемам обращалась и теоретическая мысль стиховедения. Так, можно сказать, что сама реформа силлабо-тоники, проведённая Тредиаковским и Ломоносовым, была вызвана острой потребностью общества в создании национально-ориентированных, собственных форм стиха. В продолжении этого учёные уже XIX в. осознавали необходимость расширения форм метрики русскоязычной поэзии, что совпадает с усилением интереса как к народной культуре, так и к стихотворным опытам поэтов античности. Наиболее полно этот период был отражён в труде известного русского филолога и выдающегося стиховеда Александра Христофоровича Востокова «Опыт о русском стихосложении» [11], которая вышла в 1817 году. Данный труд, в свою очередь, интересен нам не только и не столько в качестве самостоятельной работы, сколько в контексте источника влияния на поэзию ряда современников автора и на общий поворот к звуковой стороне стиха в творчестве поэтов конца XIX века, прежде всего символистов. Исследователи того времени уделяли большее внимание вариациям ритмики стихотворных размеров.

На этом этапе развития исследований русской поэтики главными опытами в этом направлении становятся исследования поэта-символиста Андрея Белого, который выпустил книгу под названием «Символизм» [12], где в числе прочих содержатся статьи с его рассуждениями о русском четырёхстопном ямбе, где он принимает к рассмотрению метр не как абстракцию, распространённую на множестве произведений, а отдельную метрику конкретного произведения. Особенности ритма, как пишет А. Белый, определяются комбинацией и расположением стоп пиррихия в стихе. Автор также впервые широко применяет статистические методы исследования, которые в современном стиховедении впоследствии станут распространённой практикой.

Так, Белый заметил, что примерно половина всех стихов, рассмотренных им, имеют пиррихийную стопу в качестве третьей из четырёх (таким образом, наиболее распространённая схема четырёхстопного ямба выглядит в этом случае как 0x0x000x, где 0 обозначает безударный слог, а x обозначает ударный слог). Также было отмечено, что около 25% всех стихов являются полноударными или т.н. “правильными” (то есть, их ударная схема полностью соответствует четырёхстопному ямбу и выглядит как 0x0x0x0x), прочие формы, такие как пиррихий на первой, на второй или на четвёртой стопах, а также строки с разными вариациями двух пиррихиев на строку, встречаются существенно реже.

Что действительно смог показать статистический метод, использованный А. Белым, это историческая эволюция отдельно взятого стихотворного размера, в данном случае — четырёхстопного ямба. В среднем, у русскоязычных поэтов пиррихии чаще всего попадают на третью стопу, а на первых двух стопах их распределение зависит от индивидуальных поэтических особенностей, который также стоит принимать во внимание при изучении ритмического рисунка стихотворных произведений.

В то же время у статей А. Белого были и слабые стороны (как, к примеру, сомнения, исходившие из статистических исследований, в самом факте существования схематических двусложных размеров), которые отмечали другие исследователи, такие как, например, В. М. Жирмунский в работе «Теория стиха» [13] и Б. В. Томашевский в работе «Русское стихосложение. Метрика» [14]. Хотя исследователи расходились во взглядах на определённые вопросы, они оба были уверены в существовании классических размеров, объясняя это наличие эмпирических свидетельств того, как и поэты, и читатели предыдущих веков могли отличать ямб от хорея без особой путаницы, однако они поддерживали тезис Андрея Белого относительного того, что традиционное определение данных размеров нуждалось в дополнении, расширении и уточнении. Как Жирмунский, так и Томашевский приходили к выводу, что пиррихий не должен рассматриваться в качестве отдельной стопы наравне с ямбом и хореем, поскольку он не является ритмообразующей структурой в той же степени, в какой являются вышеупомянутые стопы, так как ритмичность, основа стиха, создаётся повторением одинаковых величин, а пиррихий представляет собой отсутствие какой-либо ударной величины слога, следовательно, не может считаться стихотворной стопой, равной другим. Отсюда следует уточнение определения силлабо-тонических размеров стиха, основанное на положении об искусственности деления стиха на стопы.

Данное положение содержит в себе тезис о том, что стихотворные стопы слышны только при скандовке текста, то есть, при таком способе декламации стиха, в котором скандирующий ставит сильное и отчётливое ударение на каждый чётный или нечётный слог (в зависимости от размера), при этом не обращая внимания на то, есть это ударение в самом слове или нет. Это особенно заметно на примере многосложных слов, когда в пределах одного и того же слова может быть поставлено два, три или даже четыре ударения. Для иллюстрации этого способа можно привести первые два стиха из «Евгения Онегина» [15], написанные ямбическим размером, и выделить ударные слоги в соответствии со способом скандовки:

*“Мой дЯдя сАмых чЕстных прАвил,  
КогдА не в шУтку зАнемОг…”*

Как мы видим, если первая строка полностью соответствует схеме четырёхстопного ямба, то вторая строка, по строгой силлабо-тонической теории содержала бы в себе пиррихий на третьей стопе: *“шу /* ***тку за*** */ немог”*, однако способ скандовки стихотворения позволяет выделить эту стопу именно как ямбическую, таким образом, не отступая от выбранного метра, хотя при этом в слове “занемог” получается ударение как на первый, так и на третий слог, что помогает определить размер, но, тем не менее, не является нормой в естественной речи.

В связи с этим в современном литературоведении распространено деление стиха на сильные места (также называемые слогами или иктами) и слабые места, также именуемые слогами. Реальные ударения в словах совпадают с сильными местами, но не на всех сильных слогах обязательно присутствует реальное ударение слова. Слабые слоги, за относительно редкими исключениями, такими как сверхсхемные ударения, безударны. В рамках нашего исследования мы ограничимся кратким упоминанием сверхсхемных ударений, поскольку они не связаны напрямую с выбранным нами направлением, однако не упомянуть об этом явлении мы также не можем.

Краткая литературная энциклопедия даёт следующее определение: “Сверхсхемное ударение (внеметрическое) — ударение, стоящее на [слабом слоге](http://feb-web.ru/feb/kle/Kle-abc/ke6/ke6-9062.htm) стиха в силлабо-тонич. размерах, напр. в ямбе, — на нечетном слоге <...>. В рус. ямбах С. у. образуется только односложными словами; чаще всего стоит на первом слоге стиха, после [цезуры](http://feb-web.ru/feb/kle/Kle-abc/ke8/ke8-3812.htm) или сильной паузы. С. у. в ямбе встречается двух типов: т. н. [спондей](http://feb-web.ru/feb/kle/Kle-abc/ke7/ke7-1242.htm), т. е. столкновение двух ударений, схемного и С. у. (, и [хориямб](http://feb-web.ru/feb/kle/Kle-abc/ke8/ke8-3212.htm), т. е. сочетание С. у. с пропуском схемного ударения <...>. Аналогичные явления наблюдаются и в хорее. В [трехсложных размерах](http://feb-web.ru/feb/kle/Kle-abc/ke7/ke7-6142.htm) С. у. встречаются чаще, чем в двусложных, употребляются свободнее и могут образоваться как односложными, так и двусложными словами; наиболее частый случай — С. у. на первом слоге анапеста <...>. Расположение С. у. в [дольниках](http://feb-web.ru/feb/kle/Kle-abc/ke2/ke2-7342.htm) подчиняется тем же закономерностям, что и в классич. трехсложных размерах” [16, c. 699].

На данном этапе добиться постоянства в отделении сверхсхемных или внеметрических ударений от пропуска стоп или от наличия в отдельных местах логического, иктового ударения, не представляется возможным, поэтому данное направление остаётся перспективой для дальнейших исследований.

Таким образом, наши рассуждения логически подводят нас к следующему этапу, а именно к размышлению и анализу тонической системы стихосложения, чьи основные размеры (дольник, тактовик, акцентный стих) также видятся нам перспективным направлением для дальнейшего изучения, соответственно, и тоническое стихосложение нас будет интересовать в основном в качестве потенциала для дальнейших исследований.

При первом приближении, тонические и силлабо-тонические размеры являются противоположностью друг другу, поскольку в силлабо-тонике присутствует относительно строгий порядок в чередовании сильных и слабых мест, однако в расположении ударных и безударных слогов в тонике порядок отсутствует. При этом подобная разница не так существенна, если мы обратимся к деталям. Во-первых, обе системы стихосложения в основе своей построены на качественном различении между собой ударных и безударных слогов. Во-вторых, между чистой схематикой силлабо-тоники и полной тоникой акцентного стиха существует целый спектр размеров и переходных форм, каждый из которых может иметь элементы как одной, так и другой системы. К примеру, между ямбом и хореем и акцентным стихом располагаются условный дольник (размер стиха с переменным интервалом безударных слогов между ударными с ограничением в разнице на один или два слога) и тактовик (более свободный тонический размер с тем же интервалом до трёх слогов). Более того, стоит отметить, что с учётом всех изложенных нами ранее особенностей и специфики стихосложения как такового, можно сказать, что между упомянутыми размерами также есть ряд переходных форм.

Также ясно, что нет достаточных оснований для чёткого разделения на различные системы стихосложения для современного русского языка, поскольку все они, как показано в предыдущих исследованиях, являются вариантами единой тонической системы русского стихосложения, органически свойственной нашему языку. Чёткие различия между системами стихосложения на данном этапе развития стиховедения можно проводить только в том случае, если мы не считаем основой фонетических особенностей русского языка качественное различие между ударными и безударными слогами, в противном случае логичным будет являться вывод о равноправии метрической, силлабической и тонической систем стихосложения. В первом случае слоги различаются долготой, во втором случае они считаются равноценными. поэтому учитывается только их количество, а в третьей системе важна именно сила произношения (ударные и безударные). При таком варианте рассмотрения силлабо-тонические и тонические размеры русского стиха будут являться равноценными вариантами единой системы, поскольку оппозиция ударных и безударных слогов будет составлять основу ритма во всех из них.

Таким образом, можно сказать, что вопрос о выделении силлабо-тоники и чистой тоники в качестве отдельных самостоятельных систем является исключительно вопросом дефиниции терминов, однако при подобном подходе следовало бы выбрать другой термин, обозначающий ту ритмическую их общность, о которой мы говорили ранее. До тех пор наиболее адекватным нашему исследованию подходом будет обращение к традиционным силлабо-тоническим стиха с целью наиболее аккуратного и точного определения поэтической метрики в стихотворениях поэтов, обращавшихся к строгим формам силлабо-тоники.

Исходя из всего вышесказанного, мы считаем вопрос о теории стиха рассмотренным в достаточной степени, чтобы перейти к рассмотрению актуальных статистических моделей анализа стихотворного размера.

## 1.2. Исследования актуальных статистических моделей анализа стихотворного размера

Прежде чем приступить к анализу актуальных статистических моделей определения стихотворного размера, требуется остановиться на отдельных теоретических аспектах исторического развития построения подобных моделей, в частности, на проблематике вероятностного подхода к анализу поэтической метрики.

В русле этого направления главными теоретическими наработками для русского языка, на наш взгляд, можно считать отдельные места из третьего тома избранных трудов М. Л. Гаспарова «О стихе» [17], в частности, статьи из этого тома, посвящённые как рассуждениям о потенциале создания вероятностных моделей анализа стиха, так и возможным практическим применениям подобных моделей.

Как пишет М. Л. Гаспаров в статье «Вероятностная модель стиха», “хотя для решения отдельных проблем хронологии и атрибуции стихотворных памятников (преимущественно античных и средневековых) подсчеты такого рода (*имеются в виду точные подсчёты при исследовании стиха — прим. авт.*) широко использовались еще в XIX в., но в основу систематической разработки теории и истории стиха в целом подсчеты впервые были положены лишь в XX в., и первым решающим шагом было исследование Андрея Белого о русском 4-ст. ямбе” [17, с. 9]. В приведённом фрагменте М. Л. Гаспаров упоминает ту попытку точного исследования ямбического стиха Андреем Белым [12], о которой мы говорили в предыдущей главе.

Гаспаров также отмечал, что после этого исследования разработка статистических методов применительно к стиховедению имела место в трудах в основном учёных-славистов, прежде всего, русских учёных, а также польских, чешских [19], сербских [20] и болгарских. За пределами славистики данное направление исследования особой популярности не возымело до определённого момента.

При разговоре о вероятностном подходе к стиховедению обычно под этим термином понимается подсчёт каких-либо метрических явлений в стихах определённых эпох, авторов, в отдельных произведениях и их последующий сравнительный анализ. Так, к примеру, уже упомянутые нами исследователи русского ямба как правило подсчитывали частоту появления пиррихиев и спондеев в каждом стихе. Такие исследования существуют и на других языках, в частности, и для английского ямба [21]. Подобный подход хотя и позволяет составить достаточную характеристику индивидуальных особенностей ритмики отдельных поэтов, равно как и провести обширное исследование общей исторической эволюции ритма, однако этим возможные способы приложения математических методов к стиховедению не исчерпываются.

Другим направлением стиховедения, где математические методы находят своё применение, является сравнительный анализ поэтической речи с естественным ритмом языка, что позволяет выделить те особенности ритма произведения, которые были созданы поэтом намеренно, и отделить их от возникших стихийно. Именно это направление исследования позволяет нам выделить два типа теоретических моделей анализа стихотворного размера: статистическую и вероятностную.

Статистическая модель образуется путём выделения в произвольно взятом отрывке прозаического текста всех словосочетаний, метрически совпадающих с указанным заранее ритмом (например, с четырёхстопным ямбом). Подобная выборка производится либо механическим путём, либо с ограничениями по семантике, чтобы выделяемые словосочетания имели какую-либо смысловую законченность и синтаксическую связность. Таким образом образуется выборка, принимаемая исследователями за образец естественного проявления заданного ритма в речи. Произведя подсчёты частотности появления пиррихия и спондея на каждой стопе стиха, можно сопоставить полученные результаты с реальным поэтическим текстом и при обнаружении отклонений от разброса значений можно говорить об определённой специфике стихотворного ритма. Очевидным недостатком данной модели является необязательность истинности предпосылки, на которой она строится, а именно — предположение, что выделяемый прозаический текст лишён собственного ритма. Именно поэтому в стиховедческой науке предпочтение отдаётся вероятностной модели стиха.

Вероятностная модель стиха строится на предпосылке ритмической независимости слов естественного языка друг от друга, а значит, вероятность встретить в тексте определённое ритмическое словосочетание равна произведению вероятностей употребления этих слов по отдельности. Если следовать подобной логике, можно высчитать вероятность словосочетаний, подходящих под ритмику определённого размера, после чего сопоставить распределение пропорции вероятностей с их распределением в реальном поэтическом тексте. Подобный подход к построению стихотворной модели был предложен уже упоминавшимся нами ранее филологом Б. Томашевским [14], после чего был расширен и дополнен А. Н. Колмогоровым в 1960-ых годах [22]. Стоит отметить также, что результаты, полученные в ходе использования двух приведённых выше моделей, во многом оказываются сходными между собой, а основное различие в результатах, по-видимому, может быть объяснено уже упомянутым специфическим внутренним ритмом прозы, который на сегодняшний момент остаётся не до конца изученным. На данном этапе нашего исследования объём вышепредставленной теоретической информации позволяет составить полную картину о наработках в исследуемой области стиховедения, поэтому мы можем перейти к анализу современных исследований компьютерных моделей, созданных для определения ритмики поэтического текста.

Нельзя не отметить небольшой академический интерес к данному направлению стиховедения как в русскоязычной литературе, так и в иностранных исследованиях. Так, нам удалось найти всего пять публикаций на русском языке, в той или иной степени связанных с данной темой за последние 15 лет. Это может быть связано с тем, что объективная сложность в решении задачи создания компьютерной модели определения ритма поэтического текста в научном сообществе отдельных стран продиктована языковыми особенностями. Так, в англоязычной сфере подобная задача ставится и решается в рамках одного исследования, поскольку английский язык обладает фиксированным ударением [23], соответственно, в нём не возникает дополнительных задач по разграничению, например, ямбического стиха с пиррихийной стопой и анакрузой от хорейного стиха со спондеем, каковая проблема есть при работе с русским стихом. Также в аналитических языках, таких как английский, не возникает проблемы в работе со словарём ударений, поскольку присутствующие словоформы в данном языке не меняют ударение, в то время как в русском языке, относящемся к синтетической группе, есть значительное количество случаев, когда одно слово может обладать разным ударением в зависимости от падежа (например, водА (ед.ч., И.п.) — вОду(ед.ч., В.п.)), более того, есть дополнительная сложность в различении омографов (например, водЫ (ед.ч., Р.п.) — вОды(мн.ч., И.п.)).

Тем не менее, при всех упомянутых сложностях, работа над данным направлением ведётся, предлагаются различные подходы к решению проблемы автоматического определения стихотворного размера в различных поэтических традициях. Учитывая всё вышеописанное, мы сочли необходимым провести анализ исследований как зарубежных, так и российских учёных в исследуемой области для того, чтобы строить свои модели на основании их положительного и негативного опыта.

Прежде чем перейти к непосредственному анализу исследований, стоит отметить, что большое количество найденного нами научного материала включало в себя задачу автоматического определения стихотворного размера не как конечную, но как один из этапов разработки инструментов для комплексного анализа поэтического произведения, что, в свою очередь, естественным образом связано с объективной простотой решения данной задачи в языках, обращающихся к метрической системе стихосложения. Именно поэтому не стоит воспринимать наличие публикаций, обращающихся к задаче автоматического определения метрики как к конечной и достаточно комплексной цели, в разрезе общей неразвитости данного направления, поскольку решение этой задачи для языков с подвижным ударением и большим количеством вольностей при обращении с размерами стиха составляет, повторимся, целый комплекс алгоритмов, разветвлённой категоризации и поиска ключевых решений под конкретные ситуации.

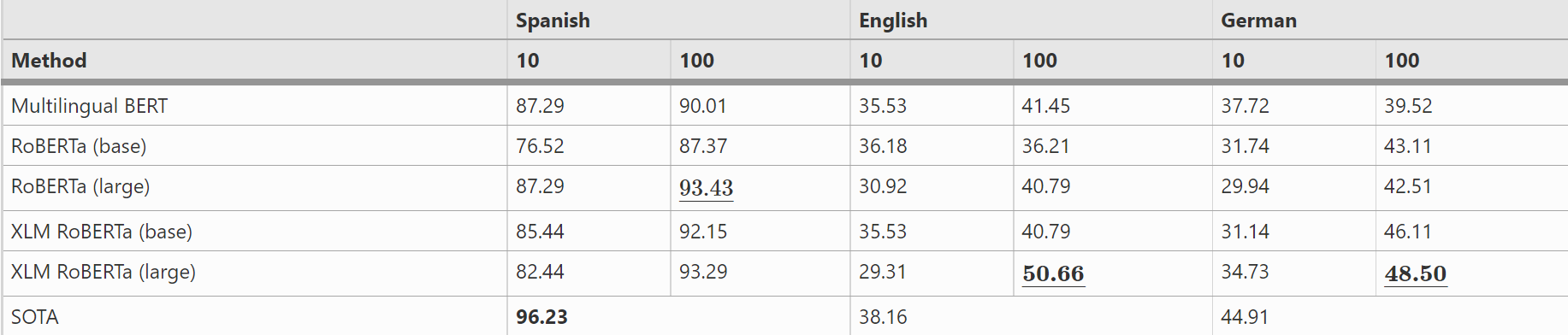
Принимая во внимание всё вышеупомянутое, первое исследование, на котором мы бы хотели остановиться, это «Using finite-state machines to automatically scan Ancient Greek hexameter» [24], проведённое в марте 2022 года, поскольку оно показывает применение современных нашему исследованию технологий к задаче, схожей с нашей, но применительно к древнегреческому языку. В этом исследовании представлен полностью автоматический подход к сканированию древнегреческого гекзаметрического стиха. В частности, в статье описывается алгоритм, который использует детерминированные конечные автоматы и локальные лингвистические правила для реализации целевого поиска допустимых паттернов спондея и, кроме того, взвешенный конечный преобразователь (WFST) для исправления и завершения частичного анализа и отклонения недопустимых кандидатов.

В статье также подробно описываются результаты эмпирической оценки качества аннотаций, полученной в результате такого подхода, на данных с ручными аннотациями. Показано, что подход с конечным состоянием обеспечивает быстрый и лингвистически обоснованный анализ стихов гекзаметра, а также эффективную формализацию лингвистических знаний. Авторы утверждают, что смогли добиться 84.07% точности определения размера в своём исследовании. К сожалению, полная версия исследования и код проекта находятся в платном доступе, поэтому у нас не было возможности протестировать их алгоритмы на практике, тем не менее, стоит отметить, что принципиальное достижение подобного результата для нашей задачи с использованием алгоритмов машинного обучения возможно, а значит, что даже если в рамках нашего исследования показатели метрик будут признаны не вполне удовлетворительными, это оставляет простор для дальнейших исследований применительно к русскому языку, поскольку принципиальное и стабильное решение данной задачи открывает огромные перспективы для дальнейшего исследования поэтической речи и её технических аспектов.

Следующее исследование, «Transformers analyzing poetry: multilingual metrical pattern prediction with transfomer-based language models» [25], проведённое в 2021 году, предлагает способ решения схожей задачи для испанского языка с использованием языковых моделей на базе нейросетевой архитектуры трансформер. Авторы отмечают, что сложные языковые правила и их исключения, а также поэтические вольности авторов делают вычисление этих шаблонов нетривиальной задачей. Некоторые риторические приемы сокращают метрическую длину, в то время как другие могут ее увеличить. Это открывает возможности для интерпретации и еще больше усложняет создание автоматизированных алгоритмов сканирования, полезных для автоматического анализа корпусов при дистанционном считывании.

В этом исследовании приводится сравнение автоматизированных систем идентификации метрических шаблонов, доступных для испанского, английского и немецкого языков, с точно настроенными одноязычными и многоязычными языковыми моделями, обученными одной и той же задаче. Несмотря на то, что изначально они были задуманы как модели, подходящие для семантических задач, результаты исследования показывают, что модели на основе трансформеров сохраняют достаточно структурной информации, чтобы достаточно хорошо работать с испанским языком в одноязычной среде, и превосходят как английский, так и немецкий при использовании модели, обученной на трех языках, что свидетельствует о преимуществах многоязычного перевод между языками.

Чтобы проверить переносимость языка применительно к структурным задачам, во второй серии экспериментов исследователи решили объединить наборы данных для трех языков. Затем они точно настроили модели на объединенном наборе данных и провели оценку на тестовых наборах для каждого отдельного языка. Учитывая хорошую производительность предположительно англоязычных моделей RoBERTa, было решено сохранить их и в этом наборе экспериментов. Как видно из таблицы 1, результаты по испанскому языку составили 93,43%, что говорит о том, что исследователи достигли пределов набора данных. С другой стороны, мы смогли превзойти предыдущее современное состояние для английского языка с увеличением точности на 12,5% до 50,66% и увеличением на 3,6% для немецкого языка до 48,50%.



*Таблица 1: Точность различных методов, настроенных для трёх языков в 10 и 100 эпох, и независимо оценённая на каждом наборе языковых тестов. Лучшие результаты по языку выделены жирным шрифтом, лучшие результаты собственной модели подчеркнуты.*

В данном исследовании была исследована возможность применения моделей на основе BERT при обучении задаче прогнозирования метрического рисунка стиха. В предположении, что модели на основе преобразования были способны выполнять задачи структурного характера, помимо задач семантического характера, было показано, что модели BERT достаточно хорошо работают для испанского языка, в то время как превосходят предыдущие современные для английского и немецкого языков.

Поскольку наиболее эффективные модели — это те, которые обучаются на комбинированных корпусах, есть свидетельства эффективного межъязыкового переноса. Это говорит о том, что дальнейшее обучение специализированной многоязычной предварительно обученной модели на поэтических корпусах может помочь улучшить задачу прогнозирования метрических паттернов. Этот результат в области языковой переносимости открывает путь для многоязычных моделей на основе трансформеров для прогнозирования метрических шаблонов, способных работать на языках, для которых существует очень мало аннотированных корпусов, как в случае с немецким языком. Традиционно автоматизированные системы метрических шаблонов создаются вручную для каждого отдельного языка, что является дорогостоящим предприятием, которое могло бы значительно выиграть от использования многоязычных подходов, подобных нашему.

Более того, подобные многоязычные отлаженные модели могут также помочь в создании стихотворений, анализируя метрическую структуру каждого стиха, сгенерированного сторонней системой. Аналогичным образом, можно было бы также протестировать целый ряд задач: метрическую длину, обнаружение сближения, обнаружение и положение цезуры, а также положения синалеф, диерезиса и синерезиса среди других. Также может быть интересно применить модели, зависящие от предметной области, на уровне строфы или даже целого стихотворения, чтобы выяснить, могут ли модели BERT предсказывать структуру или поэтический жанр.

Подводя итоги этого исследования, хотелось бы отметить, что заявленная точность модели для испанского языка существенно превосходит заявленные точности для английского и немецкого языков, хотя и заявляется, что метрика точности для этих языков превысила результаты предыдущих исследований. Мы уже упоминали, что решение задачи автоматического определения метрики стиха для английского языка представляется значительно проще, чем для русского языка, и на основании этого стоит отметить, что если ключевой процент точности для английского языка сейчас составляет ~51%, то мы не ошибёмся, если скажем, что для русского языка процент точности предсказуемо будет ниже.

Ещё одним исследованием, проанализированным нами в ходе нашего исследования, было «Supervised machine learning for hybrid meter» [26], исследовавшее способы решения задачи автоматического определения метрики стиха для средненемецкого языка. Данное исследование нам показалось особенно важным, поскольку языковые особенности средненемецких языковых диалектов представляют некоторую степень схожести с русским языком, особенно в аспекте подвижности ударений в словах. Таким образом, результаты именно этого исследования во многом сподвигли нас на разработку второй модели решения поставленной нами цели.

Авторы этого исследования отмечают, что в немецкоязычной литературе, посвященной компьютерному анализу поэтического текста, преобладают два подхода к исследованию размера. Один из подходов использует известный размер и присваивает слогам шаблоны ударения на основе таких параметров. Второй подход не предполагает ничего о размере и стремится определить его путем обозначения слогов и выявления закономерностей.

Предыдущие исследования, по словам авторов, были сосредоточены на относительно простых системах размера и использовали основанные на правилах, статистические или неконтролируемые подходы. Гибридная природа размера средненемецкого языка и других сложных систем, развивающихся из фонологических особенностей, затрудняет сканирование поэзии с использованием этих методологий. Строго основанный на правилах подход был предпринят Фридрихом Димпелом (2004). Хотя подход Димпела точен, это трудная задача, ограничительная и чрезвычайно специфичная для языка. Более того, идентифицируя только ударные слоги, он не охватывает всю сложность метра средненемецкого языка.

Исходя из этих размышлений, авторы сосредоточили свои усилия на построении модели условных случайных полей (CRF) для прогнозирования сканирования метра поэзии на средненемецких диалектах. 17 признаков применяются к исследуемому слогу, а также к десяти предшествующим слогам и десяти последующим слогам строки, если таковые имеются. В этой статье представлено новое применение моделей машинного обучения к поэзии, в частности к традициям с гибридным метром. Это обещает внести вклад в другие литературные интересы в области компьютерной лингвистики, такие как авторский и жанровый анализ.

Исследования показали, что правильный учет метра должен учитывать его вариации по всему тексту; автоматическое сканирование делает это реалистичным предприятием. Применяя этот метод, учёные могут исследовать крупномасштабные различия внутри и между текстами, чтобы обнаружить закономерности, характеризующие авторов и жанры. Действительно, тонкие различия в метре могут оказаться отчетливым авторским голосом или выявить значительный стилистический выбор. Эта статья также прокладывает путь для дальнейшей работы, такой как: кластерный анализ метра по большому корпусу текстов, тематическое моделирование ритма по жанрам и составление графика литературного влияния метра на текстологические особенности.

Ещё одним исследованием, на которое мы обратили своё внимание, стало «Automatic classification of stanzas in Spanish poetry» [27], в котором была поставлена проблема классификации строф в испаноязычной поэзии. Хотя испанский язык и наследует иную поэтическую традицию, нежели русский, методики, использованные в данном исследовании, показались нам достойными внимания.

Для решения означенной проблемы в данном исследовании были реализованы три различных подхода. Сначала была смоделирована экспертная система, основанная на правилах и экспертных знаниях. Эта модель учитывает три основные характеристики, необходимые для определения строфы в испанской поэзии, а именно: длина стихов, структура рифмы и тип рифмы. Результаты показали, что у этого подхода есть две основные проблемы. Во-первых, не все экспертные знания можно перевести на машинный язык — необходимые знания настолько обширны, что их невозможно приобрести и применить в компьютерной системе. Во-вторых, сообщается, что точность составляет 78,76%, что даёт слишком большой разброс для подхода, допускающего фактор человеческой ошибки.

Модели второго типа, основанные на деревьях решений и наборе правил, показали лучшее поведение (т.е. диапазон точности от 81,11 до 88,51%). но все еще имеют недостаток обобщения и зависят от знаний эксперта.

Наконец, были реализованы модели нейронных сетей, предполагающие, что свойства этих систем будут преимуществом для решения проблемы. Вопреки ожиданиям, эти системы получили худшую точность (т.е. 42,12 и 67,72%). Эти результаты подтвердите, что проблема классификации строф сложна и нетривиальна и что методы нейронных сетей далеки от применения в этом контексте, поэтому необходимы дополнительные эксперименты и дополнительные исследования, особенно в том, как кодировать входные данные сетей, поскольку семантическая информация вложений выглядит недостаточно богатой, чтобы помочь в решение. Точно так же способность запоминать информацию нейронов у двунаправленных LSTM сетей выглядит недостаточной для усиления структурной информации строфы.

Как и в предыдущих исследованиях, многие исследователи сталкиваются с тем, что машинные алгоритмы, вопреки ожиданиям, дают точность существенно ниже или около 50%. Подобное происходит, судя по всему, вне сильной корреляции с языковыми особенностями, поэтической традицией и объёмом использованного корпуса, поэтому данная тема и на сей день остаётся перспективным направлением для исследования с далеко идущими результатами.

Следующим исследованием, которое было нами проанализировано в рамках сбора материала для построения нашей модели, было «Deep neural network based poetic meter classification using musical texture feature fusion» [29]. В этой статье предложена схема классификации метров с использованием музыкальных текстурных признаков (MTF) с использованием глубокой нейронной сети (DNN) и гибридной модели гауссовой смеси-deep neural network (GMM-DNN) framework. Производительность предлагаемой системы оценивается с использованием недавно созданного поэтического корпуса на малаялам, одном из известных языков Индии, и сравнивается с производительностью с помощью машины опорных векторов (SVM). классификатор. Проводится эксперимент на основе mel-frequency cepstral coefficient (MFCC). Позже MTF объединяется с MFCC. В то время как система MFCC сообщает об общей точности 78,33%, система fused сообщает о точности 86,66% в гибридной системе GMM-DNN. Общая точность, полученная для DNN и GMM-DNN, составляет 85,83%, а 86,66% соответственно. Архитектурный выбор классификатора на основе DNN с использованием функций, полученных из GMM, в парадигме feature fusion показал улучшение производительности. Предлагаемая система демонстрирует перспективность методологий глубокого обучения и эффективность MTF в распознавании метров из прочитанных стихотворений.

В этом исследовании рассматривается классификация поэтических метров с использованием слияния MFCC и MTF с использованием методологий глубокого обучения. В первом подходе используется MFCC, позже проводится эксперимент с MTF, за которым следует полный набор методики слияния. Систематическая оценка проводится с использованием шести метров поэтического корпуса малаялам. Эксперимент по слиянию привел к общей точности 86,66% в новой гибридной системе GMM-DNN. Результаты показывают потенциал MTF и методологий глубокого обучения в задаче классификации поэтических метров.

Данное исследование привлекло наше внимание обещанием высокой точности своих результатов и опоры на музыкальные особенности текста, однако после прочтения у нас остались серьёзные сомнения в достоверности и перспективности данного исследования в связи с тем, что в нём не представлено достаточное количество сведений относительно того, как именно была достигнута подобная степень точности, а также не указано, насколько сильно данная модель потеряет в точности определения размера при переносе на другой язык, что даёт основание считать это исследование успешным только в рамках одного из многочисленных языков Индии.

Ещё одной интересной работой, проанализированной нами в рамках сбора информации об актуальных исследованиях по нашей теме, стала работа «Learning meters of Arabic and English poems with Recurrent Neural Networks» [30]. В этой статье авторы строят модели рекуррентных нейронных сетей (RNN), которые могут классифицировать стихи в соответствии с их метрами из обычного текста. Вводимый текст кодируется на уровне символов и передается непосредственно в модели без создания элементов вручную. Авторы называют это шагом вперёд в машинном понимании и синтезе языков в целом и арабского языка в частности. Среди 16 стихотворных метров арабского языка и 4 метров из английского языка сети смогли правильно классифицировать метр стихотворения с общей точностью 96,38% и 82,31% соответственно. Корпус данных стихотворений, использованных для проведения этого исследования, составил более 1,5 миллионов стихотворных строк, и были собран из разных нетехнических источников, из арабских и английских литературных сайтов, в разнородных и неструктурированных форматах. По словам авторов, это исследование является первым, в котором рассматривается классификация счетчиков стихотворений в подходе машинного обучения в целом и в подходе, основанном на RNN без свойств, в частности. Кроме того, этот корпус является первым общедоступным набором данных, готовым для целей будущих вычислительных исследований.

Эта работа была направлена на обучение рекуррентных нейронных сетей (RNN) на уровне символов на стихотворения, написанные на арабском и английском языках, чтобы выучить и распознать их метры, которые делают стихотворение звучащим риторически или фонетически при произнесении. Использование машинного обучения (ML) в целом и глубоких нейронных сетей (DNN), в частности, для изучения стихотворных метров и фонетического стиля из письменного текста, наряду с наличием такого набора данных, являются относительно новым направлением в академической литературе.

Как говорят авторы исследования, из-за вычислительной трудоемкости и временной сложности обучения RNN их сетевые конфигурации не были исчерпывающими, чтобы охватить очень широкий диапазон конфигураций параметров обучения (например, количество слоев, размер ячейки и т.д.).

Тем не менее, точность классификации, полученная на арабском наборе данных, достойна упоминания, особенно по сравнению с тем, что получено из детерминированных и созданных человеком алгоритмов на основе правил, доступных в академической литературе. Авторы видят перспективу дальнейшего исследования и обсуждения некоторых дополнительных вопросов, как то: как повысить точность набора данных на английском языке, почему эффект диакритических знаков не согласуется и почему некоторые счетчики обладают низкой точностью для каждого класса. В дополнение к этому исследованию мы бы также хотели отметить «Meter classification of arabic poems using deep bidirectional recurrent neural networks» [31], где также используются рекуррентные нейронные сети для определения размера арабских стихотворений.

В этой статье представлен новый метод классификации стихотворных метров арабских стихотворений с использованием глубокого обучения на основе RNN. Это позволяет избежать необходимости преобразования стихотворения в звуковую форму, а также необходимости явного кодирования сложных правил, которым обычно следуют для определения метра. Представленный метод был оценен на большом наборе данных, собранном специально для этой цели. Авторы утверждают, что на независимом тестовом наборе точность их модели достигла 94,32%.

Мы не стали разделять предыдущие два исследования, поскольку они оба были нам интересны в формате применения рекуррентных нейронных сетей, безотносительно общих результатов и языковой традиции, поскольку, как можно понять из того объёма текста об этих исследованиях, который мы представили, решение задачи автоматического определения метрики стихотворения для арабском языке не является затруднением, что связано, на наш взгляд, как с фонетическими и метрическими особенностями языка, так и с общей ригидностью поэзии в конкретной традиции.

Следующее исследование, которое, на наш взгляд, заслуживает отдельного внимания, это «Generating homeric poetry with deep neural networks» [32]. Оно показалось нам интересным, поскольку это одно из немногочисленных современных исследований задачи генерации поэзии, которое обращается к ней не только с лексической, сочетательной стороны, но также и с фонетической, метрической стороны, пытаясь имитировать не только подбор слов Гомера, но и его общий поэтический слог. В этой статье исследуется создание метрически точной гомеровской поэзии и оценивается качество сгенерированных стихотворных строк с помощью количественного метрического анализа и экспертной оценки. Эта оценка показывает, что созданная модель неявно улавливает сложный поэтический метр, но недостаточно эффективна с точки зрения семантики и соответствия контексту. Это исследование подчеркивает важность экспертной оценки и показывает, что будущие исследования должны быть сосредоточены на поэтической семантике.

В этой статье авторы исследовали использование кодера-декодера RNNs для создания метрически точной гомеровской поэзии. Подобные сгенерированные машиной поэтические строки были оценены классицистами как метрически, так и в их первоначальном контексте. Результаты опроса показали, что модель недостаточно эффективна с точки зрения создания семантически связных стихотворных строк, которые соответствуют более широкому контексту. Тем не менее, эта модель весьма успешна в создании метрически идентичной гомеровской поэзии.

Наконец, прежде чем перейти к разговору о найденных нами русскоязычных исследованиях, мы бы хотели отметить большое исследование под названием «Computational analysis of the historical changes in poetry and prose» [33], где авторы говорят о том, что эзотерические определения поэзии недостаточны для того, чтобы охватить изменения в поэзии, свидетелями которых стала эпоха механического воспроизведения с широким распространением использования цифровых медиа и искусственного интеллекта. Они также недостаточны для проведения различия между прозой и поэзией, поскольку содержание как прозы, так и поэзии может быть поэтическим. Используя цитаты в качестве прозы, учитывая их поэтическую, контекстно-свободную и прославленную природу, рассматриваются стилистические различия между поэзией и прозой.

Грамматика и метр оправданы как отличительные черты. Наборы данных популярных были созданы проза и поэзия, охватывающие 1870-1920 и 1970-2019 годы, и было проведено множество экспериментов, чтобы доказать, что проза и поэзия в последний период более похожи, чем в первый. Точность классификации поэзии и прозы 1970-2019 годов значительно ниже, чем у 1870-1920, тем самым доказав сближение поэзии и прозы.

В результате проведенных экспериментов было доказано, что поэзия 1970-2019 годов больше похожа на прозу своего периода, чем на поэзию 1870 года — 1920 год для прозы того же периода. Изменения в прозе двух периодов в отношении стилистических особенностей минимальны, но в поэзии они значительны. Сближение поэзии и прозы и отсутствие изменений в прозе доказывают, что поэзия не обладает теми ограничительными границами, которыми обладает проза. Важность новой эпохи, вызывающей — таким образом, определение поэзии устанавливается с учетом изменений в поэзии как форме искусства. Помимо обоснования исторических изменений в поэзии и прозе, в этой статье также достигается высокая точность классификации стихотворений и прозы без использования семантических признаков. Это важный показатель того, что семантическое содержание поэзии и прозы может быть очень похожим и что их все еще можно дифференцировать с помощью стилистических особенностей, не принимая во внимание очевидную визуальную разницу в разрывах строк.

Будущая работа в этой статье заключается в использовании этих функций при создании персонализированного помощника по поэзии, который изучает стилистические предпочтения пользователя в инверсиях и метре, на основе пользовательского ввода креативного текста. Этот персонализированный характер помощника будет адаптироваться к пожеланиям пользователя стать ‘современным’ или ‘классическим’ поэтом.

Хотя подобная работа в строгом смысле не вполне связана с нашими задачами, нам показалось важным отметить её наличие, поскольку как подобное исследование, как многие из работ перед этим, так и наша работа могут видеть конечной целью своих усилий две большие задачи, означенные выше — прослеживание эволюционных динамических изменений в характере, ритме, тематиках и оппозиции стиха к прозе и создание условного персонального помощника, с помощью которого каждый интересующийся человек сможет выстроить свою личную траекторию понимания литературных аспектов.

При разговоре о современных русскоязычных исследованиях различных подходов к проблеме автоматического определения метрики стиха прежде всего стоит обратить внимание на работу «Автоматическое распознавание метра: проблемы и решения» [36] И. А. Пильщикова и Старостина А. С. Старостина, поскольку именно в этой работе были сформулированы положения, которые авторами данной работы рассматриваются как ключевые для решения задачи автоматического определения стихотворной метрики с помощью классификатора, основанного на правилах, а именно такой подход, при котором “для каждого метра вычисляются усредненные меры. Усреднение производится по всем возможным акцентуациям исходной строки. Эти средние величины становятся мерами метричности всей строки. Далее для каждой строки выбирается максимальное значение меры метричности и строке приписываются те метры, для которых метрические меры достигли максимального значения. Таким образом, в некоторых случаях строке может быть приписано более одного размера, что отражает ее подлинную или мнимую метрическую амбивалентность” [36, c. 494]. Хотя данное исследование было проведено десять лет назад, мы считаем основные идеи и положения, изложенные в нём, актуальными и применимыми с учётом сегодняшних технологических изменений. Более того, так считает и О. М. Аншаков, проведший в 2017 году исследование «Проблема автоматического определения метра и ритма русского стиха: математические модели и алгоритмы» [37], в котором он также опирался на работы этих учёных.

В своём исследовании Аншаков выдвигает предложение о представлении метрических схем стихотворения в битовом варианте. Он намеренно отказывается от понятия стопы и придерживается иктовой теории, попутно выдвигая несколько определений, помогающих, по задумке автора, наиболее простым способом добиться удовлетворительного результата в плане определения метра стихотворения, однако в ходе подбора подобных определений и моделей автор не принимает во внимание сверхсхемные ударения — проблема, о которой мы также говорили в предыдущем параграфе.

Тем не менее, отметим, что алгоритм определения т.н. “обеднённого” метра стихотворения, предложенный автором в данной работе, во многом схож с тем, на который опиралось и наше практическое исследование ещё до того, как мы ознакомились с работой Аншакова.

Также мы проанализировали работу «Автоматизация комплексного анализа русского поэтического текста: модели и алгоритмы» [34], в которой авторы систематизируют накопленный опыт в разработке автоматического анализа различных составляющих поэтического текста, в том числе и метрики текста, однако в данном вопросе работа берёт за основание методику, предложенную в работе «Автоматический анализ стиха в системе Starling» [39] 2006 года, в которой с незначительными отличиями использован такой же алгоритм, что и в работе Аншакова.

Наконец, наиболее полное и комплексное решение данной задачи, на наш взгляд, представлено в работе «Формально-языковая модель стиха и регулятивов его метра для автоматизированной метрической идентификации» [38], в которой не только рассмотрены примеры классических схем силлабических, силлабо-тонических и чисто тонических размеров, но и их переходные формы, вариации, в определённой мере обсуждён вопрос сверсхемных ударений и стоп, решён вопрос о разноударных словах и омографах, также выводятся формулы и способы решения для неоднозначных случаев. Более того, авторы предоставляют возможность воспользоваться собственной программой и выкладывают её код в открытый доступ, поэтому мы смогли также протестировать её и убедиться в её работоспособности и высокой точности.

Мы также обращались и к разработкам отдельных энтузиастов в данной области, в частности, на известном интернет-ресурсе, посвящённом программированию, Habr, были найдены наработки по нашей теме. В этой публикации автор рассказывает о собственном опыте в создании инструмента для работы с автоматическим определением силлабо-тонических размеров [49]. Хотя алгоритмическое ядро инструмента в сути своей повторяет наш алгоритм, нельзя не отметить некоторые интересные находки автора, которые могут быть использованы в рамках дальнейшей работы с нашим инструментом, основанным на правилах. В частности, автор ввёл определённую рейтинговую систему, где начислял или убирал очки того или иного размера у строки за наличие или отсутствие определённых признаков, к примеру, 5 очков штрафа к выбору размера строке начисляется, если встречается слово с уточнённым ударением, которое ни в какой из своих орфоэпических вариаций не может попадать в сильную долю размерной сетки. Мы считаем подобный подход более утончённым и в перспективе обладающим более высоким потенциалом для точного определения стихотворных размеров. При этом мы также не можем не отметить, что на данный момент степень разработки этой темы автором далека от завершения, а программа представляет сырой прототип, который не учитывает множества особенностей русской фонетики, учтённых в нашем классификаторе.

С учётом всего вышеперечисленного, мы считаем вопрос о текущем состоянии проработанности темы автоматического определения поэтической метрики рассмотренным в достаточной степени, чтобы обратиться к практической части нашего исследования.

# Глава 2. Создание инструмента для автоматического анализа поэтической метрики

В качестве практической части исследования нами было разработано два прототипа классификаторов поэтической метрики, в которых использовались два принципиально различных подхода.

Нами был разработан автоматический анализатор, основанный на строгих правилах и сравнении вероятности строк. После этого мы взяли за основу вероятностную модель размера и создали на её основе алгоритм машинного обучения с задачей мультиклассификации и несколькими дополнительными метриками для повышения точности определения размера. В качестве материала для обучения был создан корпус из 5062 стихотворных строк, категоризированных на традиционные силлабо-тонические размеры (Приложение 1), после чего корпус был разделён на тренировочную и тестовую выборку в соотношении 80\20%.

Далее мы опишем работу, проделанную над созданием каждого из алгоритмов, подробнее.

## 2.1. Имплементация классификатора поэтической метрики, основанного на правилах

При разработке инструмента анализа поэтической метрики нашей первоначальной идеей было создание RBC(Rule-Based Classificator) или классификатора, основанного на правилах. Наиболее удобным и современным языком программирования для выполнения поставленной задачи, на наш взгляд, является Python, поэтому был выбран именно он. В качестве среды разработки нами был выбран Google Colab, поскольку он, помимо необходимого нам своевременного онлайн-доступа к результатам и ходу разработки, также предоставлял доступ к собственным мощностям компании Google, использованных нами для тренировки алгоритма машинного обучения.

Базовый алгоритм, лежавший в основе нашего классификатора, заключался в следующем: мы брали “идеальные” схемы традиционных силлабо-тонических размеров (для ямба и хорея — в промежутке от трёх до семи стоп, для дактиля, анапеста и амфибрахия — в промежутке от двух до пяти стоп, поскольку именно в этом диапазоне находится большинство стихотворений русской классической поэзии), затем мы обозначали безударные слоги нулями, а ударные — единицами, и сводили получившиеся последовательности нулей и единиц в один словарь, разбитый на пять категорий для каждого размера соответственно. Так, например, схема трёхстопного ямба выглядела как “010101”, а четырёхстопного хорея — как “10101010”.

После выполнения данного этапа основной нашей проблемой стал подбор ударений для каждого слова. Нами было подобрано несколько разных вариантов решения данной задачи, наиболее удачной из которых при первом рассмотрении показался оцифрованный и переведённый в необходимый нам вариант Грамматический словарь русского языка А.А. Зализняка [40], однако от этого нам впоследствии пришлось отказаться, поскольку даже наиболее облегченная и уменьшенная версия этого словаря занимала слишком много места для того, чтобы быть использованной на постоянной основе.

Далее мы обратились к библиотекам Python, предоставляющим возможность парсинга сайтов, в частности, нас интересовал парсинг требуемых страниц русской версии Викисловаря [41]. Нами были рассмотрены возможности и механики имплементации применения таких библиотек, как Natasha, Spacy и Text Accentizer. После анализа подходящих библиотек, наиболее подходящим вариантом нам показалась библиотека wikitextparser[42], которая, хотя и не обладала поддержкой специфически русскоязычного Викисловаря, тем не менее, содержала в своём коде алгоритм, позволявший путём определённых манипуляций перенастроить её под русскую версию.

После внедрения данной библиотеки мы получали точность определения ударения в требуемом слове в 38%. Данный результат был сочтён неудовлетворительным. Проанализировав механику работы выбранной библиотеки, мы пришли к выводу, что низкая точность правильности получаемых ударений объясняется тем, что в ходе работы данной библиотеки во многих случаях программе не удавалось сопоставить предлагаемое слово с падежной формой, представленной в таблице Викисловаря, в случае чего она была запрограммирована брать ударение по начальной морфеме, что иногда приводило к различию в ударных слогах и, соответственно, меняло числовую картину строки. Более того, нами было определено, что при отключении данной функции, процент точности определения ударений падал до 27%, что давало нам основания считать, что изначальная точность была на 11% случайной. В результате вышеописанных тестов было принято решение отказаться от использования данной библиотеки.

Решение, к которому мы обратились после, имело существенно более высокую точность определения ударений, оно было взято нами из материалов к соревнованию по генерации поэтических текстов классических авторов от Сбербанка. В материалах был представлен облегченный и упрощённый вариант словаря, созданный на основе вышеупомянутого нами Грамматического словаря А.А. Зализняка, в группе файлов Phonetic Baseline [43]. Данное решение показалось нам наиболее простым, удобным и занимающим сравнительно небольшой объём (3Мб в сравнении с 800Мб изначального парсинга словаря Зализняка, составленного нами). Более того, оно показало точность определения ударений в 53%, что, хотя и далеко от идеала, однако превышает пятидесятипроцентный барьер, а значит, превосходит расстановку ударений наугад.

Однако окончательное решение было взято нами из библиотеки для Python под названием russtress [47], которая явилась результатом работы группы российских учёных в области применения рекуррентных нейронных сетей с механизмом LSTM для расстановки ударений в русских словах [48]. Заявленная точность данного инструмента, 91%, проведённая на собранном нами поэтическом корпусе, подтвердилась. Главная задача, которую мы решили в процессе имплементации, это процесс расстановки ударений в односложных словах. Для этого нами был создан список из односложных непроизводных союзов, предлогов а также частиц, на которые статистически почти никогда не падает логическое ударение в стихе, они по умолчанию считались безударными, а остальные семантически значимые односложные слова (к примеру, “день”, “прыть” и др.) по умолчанию считались ударными. Это позволило снизить общую неточность выстраивания схемы размера инструментом на 2.5 процента.

После решения вышеизложенных задач мы прописали функцию, где на вход принимался стихотворный текст, делился построчно, из каждой строки убирались все ненужные символы, строка делилась на слова, для каждого слова в строке подбиралось ударение из словаря и возвращалось в виде комбинации единицы и нулей, после чего эта комбинация складывалась в последовательность для целой строки.

Далее с помощью модуля Collections функция проходилась по словарю с “идеальными” схемами размеров, сравнивала получившуюся числовую комбинацию со всеми записями в словаре и записывала процент схожести строк. Запись с наивысшим процентом схожести с анализируемой строкой записывалась в функцию. После итерации по всем строкам предложенного текста программа выдаёт пользователю конечный результат в формате: “*Хорей наиболее вероятен. Он встречается в 46 строках. Также программа выделила Ямб, который встречается в 29 строках.*”. Как видно, сперва указывается наиболее вероятный размер и количество строк, где программа выделила именно его, после чего идёт второй по вероятности размер и количество строк, где программа предположила именно этот размер.

Для проведения численной оценки качества работы инструмента нами был составлен корпус в размере 500 стихотворных строк для каждого из пяти традиционных силлабо-тонических размеров, общим объёмом в 2500 стихотворных строк. Данные о метрике стихотворных текстов были взяты из поэтического подкорпуса НКРЯ [44]. Оценка результатов работы была проведена для каждого размера отдельно по четырём метрикам: точность (Accuracy), точность (Precision), полнота (Recall) и f-мера. Во избежание возможной путаницы в дальнейшем мы будем использовать латинские названия этих терминов.

Accuracy измеряется по формуле P\N, где P — это количество правильных определений программы, а N — размер выборки.

Precision для какого-либо класса измеряется как отношение количества фактически принадлежащих к этому классу данных к количеству данных, отнесённых к этому классу программой.

Recall измеряется как отношение количества правильно определённых программой данных какого-либо класса к общему количеству данных этого класса в выборке.

Для наглядности представим возможные варианты определения размеров программой следующим образом:

True Positive (TP) — программа определила тот размер, к которому строка на самом деле принадлежит.

True Negative (TN) — программа определила тот размер, к которому на самом деле строка не принадлежит.

False Positive (FP) — программа определила не тот размер, к которому строка на самом деле принадлежит.

False Negaitve (FN) — программа определила, что размер не относится к тому классу, к которому он на самом деле не относится.

Исходя из этого, Precision рассчитывается по формуле P = TP/(TP + FP). Recall при этом рассчитывается по формуле R = TP/(TP+FN). F-мера, таким образом, является гармоническим средним из Precision и Recall и рассчитывается по формуле F = 2 \* ((Precision \* Recall)/(Precision + Recall)).

Проведя испытание программы, мы получили следующие результаты:

Для хорея: Первое: Ямб, 214, Второе: Хорей, 144, Третье: Дактиль, 89, Четвёртое: Амфибрахий, 38, Пятое: Анапест, 16.

Для ямба: Первое: Ямб, 247, Второе: Хорей, 102, Третье: Дактиль, 79, Четвёртое: Амфибрахий, 60, Пятое: Анапест, 13.

Для дактиля: Первое: Дактиль, 229, Второе: Ямб, 128, Третье: Хорей, 100, Четвёртое: Амфибрахий, 24, Пятое: Анапест, 20.

Для амфибрахия: Первое: Дактиль, 148, Второе: Амфибрахий, 136, Третье: Ямб, 128, Четвёртое: Хорей, 73, Пятое: Анапест, 16.

Для анапеста: Первое: Ямб, 142, Второе: Дактиль, 129, Третье: Амфибрахий, 86, Четвёртое: Хорей, 76, Пятое: Анапест, 48.

Таким образом, ключевые метрики по каждому размеру получаются следующие:

Хорей:

Accuracy = 144/500 = 0.288

Precision = 144/495 ~ 0.291

Recall = 144/144+253+271+364+452=144/1484 ~ 0.098

f-мера = 2\*((0.291\*0.098)/(0.291+0.098)) ~ 2\*(0.003/0.389) ~ 0.015

Ямб:

Accuracy = 247/500 = 0.494

Precision = 247/859 ~ 0.288

Recall = 247/356+247+271+364+452 = 247/1690 ~ 0.146

f-мера = 2\*((0.288\*0.146)/(0.288+0.146)) ~ 2\*(0.042/0.434) = 0.194

Дактиль:

Accuracy = 229/500 = 0.458

Precision = 229/674 ~ 0.34

Recall = 229/356+253+229+364+452 = 229/1654 ~ 0.138

f-мера = 2\*((0.34\*0.138)/(0.34+0.138)) ~ 2\*(0.047/0.478) ~ 0.197

Амфибрахий:

Accuracy = 136/500 = 0.272

Precision = 136/344 ~ 0.395

Recall = 136/356+253+271+136+452 = 136/1468 ~ 0.093

f-мера = 2\*((0.395\*0.093)/(0.395+0.093) ~ 2\*(0.037/0.488) ~ 0.151

Анапест:

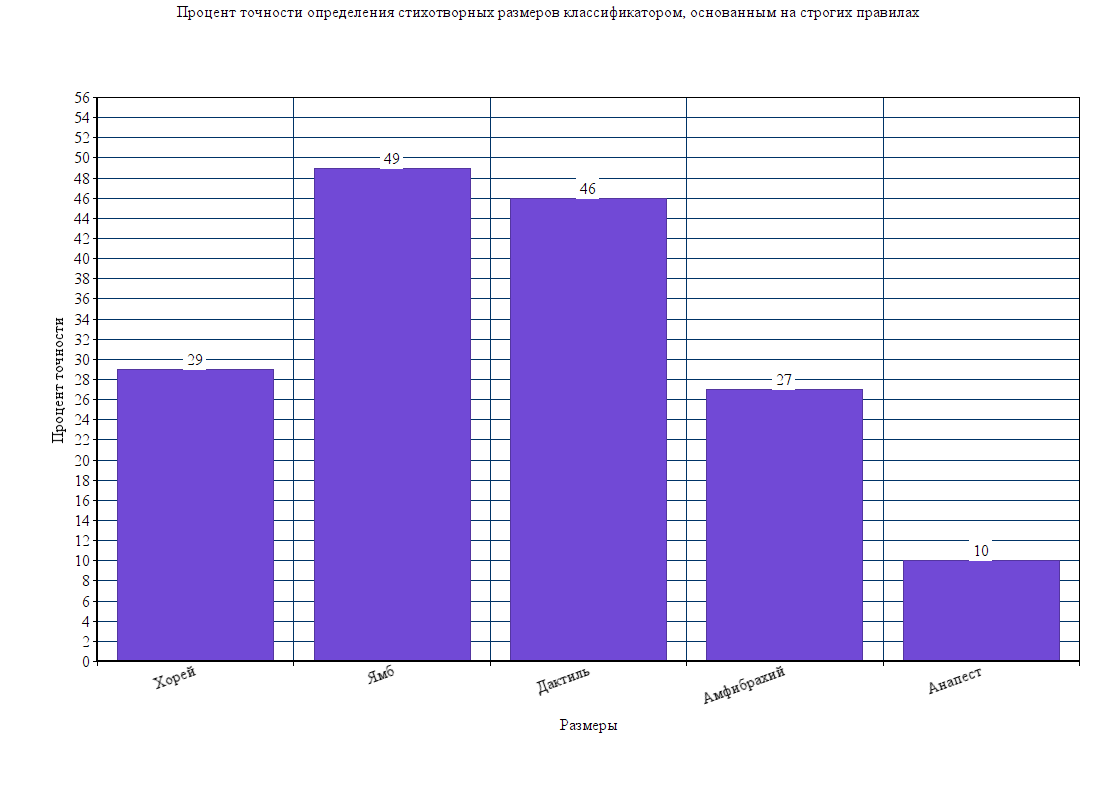
Accuracy = 48/500 = 0.096

Precision = 48/113 ~ 0.425

Recall = 48/356+253+271+364+48 = 48/1292 ~ 0.037

f-мера = 2\*((0.425\*0.037)/(0.425+0.037)) = 2\*(0.016/0.462) ~ 0.07

Как видно из полученных результатов, программа лучше справляется с двусложными размерами, чем с трёхсложными. При этом различение между ямбом и хореем на больших данных проявляет себя лучше, чем ожидалось.. При определении трёхсложных размеров, при этом, программа показывает лучшие результаты для дактиля и худшие результаты при определении анапеста, поскольку она практически не обращается к схемам этого размера при рассмотрении каких-либо поэтических строк. В более удобной форме результаты представлены в таблице 2.



*Таблица 2. Сравнительный график точности работы классификатора размеров, основанного на строгих правилах.*

По результатам проведённого нами исследования с классификатором, основанным на строгих правилах, мы считаем данное направление разработки перспективным и видим в продолжении работы над совершенствованием инструмента несколько возможных путей развития, которые могут привести к появлению лёгкой, удобной и высокоточной версии этой программы. В качестве главного направления дальнейшей работы мы можем выделить дальнейшие поиски вариантов усовершенствования и расширения словаря, поскольку основное снижение точности связано именно с нестабильной и не до конца разработанной версией словаря. Одним из основных вариантов решения данной проблемы, на наш взгляд, может являться одновременное использование нескольких орфоэпических словарей параллельно и построение вероятностных моделей контекста, по которым программы будет выбирать наиболее подходящее ударение в каждом неоднозначном случае. Подобный подход позволит существенно увеличить точность определения размера, а также практически полностью решить проблему омонимов и разноударных слов.

Ещё одним перспективным направлением дальнейшей разработки является расширение размерной сетки, добавление возможности определения пиррихия и спондея, а также тонических размеров, таких как дольник, логаэд и акцентный стих.

Однако в ходе нашего исследования нами было принято решение использовать алгоритмы машинного обучения для решения поставленной задачи и провести сравнительный анализ конечной точности, простоты использования и дальнейших перспектив работы. О процессе, результатах и направлении работы в этом аспекте мы расскажем в следующем параграфе.

## 2.2. Имплементация классификатора поэтической метрики с применением методов машинного обучения

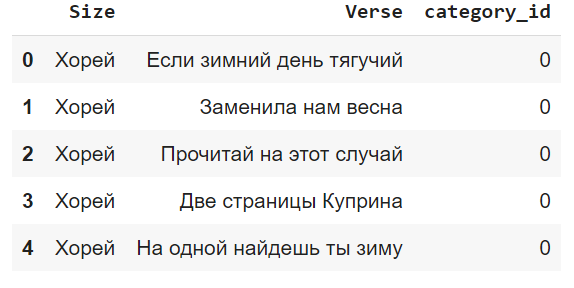
Для того, чтобы с достаточной степенью точности настроить алгоритм машинного обучения, сперва мы должны были определить класс выполняемой нами задачи. Она была определена нами как задача мульти-классификации текста. Основная сложность с решением данной задачи методами машинного обучения заключалась в том, что большинство научного материала по задаче классификации текста обращается к задаче бинарной классификации текста, которая используется при решении таких проблем, как фильтрация спама по электронной почте и анализ настроений. Однако в реальном мире наши проблемы в большинстве случаев гораздо сложнее этого. Таким образом, задача, которая была сформулирована нами при решении этой проблемы, это классификация строк поэтического текста по 5 заранее определённым классам.

Как и в случае с классификатором, основанным на строгих правилах, для разработки алгоритма машинного обучения под нашу задачу мы используем Python, в качестве среды разработки также используется Google Colab, который в случае с машинным обучением представляет особую ценность для нас, поскольку даёт доступ к мощностям компании Google, которые используются для ускорения обучения алгоритма. В качестве библиотеки для создания алгоритма мы использовали Scikit-Learn [45]. В качестве материала для обучения был создан корпус из 5062 стихотворных строк, категоризированных на традиционные силлабо-тонические размеры (Приложение 1), после чего корпус был разделён на тренировочную и тестовую выборку в соотношении 80\20%.

Наша цель при решении задачи мульти-классификации текста — выяснить, какие методы контролируемого машинного обучения лучше всего подходят для её решения.

При поступлении на вход новой стихотворной строки, мы хотим отнести её к одному из 5 размеров. Классификатор исходит из предположения, что каждая новая строка относится к одной и только одной категории.

Для начала нам требуется обработать данные. Для этого мы использовали библиотеку Pandas и сформировали данные в таблицу, где присвоили числовое значение каждой категории размера. Помимо этого мы обработали данные библиотекой NLTK, которая позволила избавиться от ненужных символов в тексте. Пример размеченных данных представлен в таблице 3.



*Таблица 3. Пример данных корпуса стихотворных строк, размеченного под использование при тренировке алгоритма машинного обучения.*

Классификаторы и алгоритмы машинного обучения не могут напрямую обрабатывать текстовые документы в их исходном виде, поскольку большинство из них ожидают числовых векторов по признакам с фиксированным размером, а не необработанных текстовых документов с переменной длиной. Поэтому на этапе предварительной обработки тексты преобразуются в более управляемое представление.

Одним из распространенных подходов к извлечению признаков из текста является использование модели bag of words (“мешок слов”): модели, в которой для каждого документа, в нашем случае стихотворной строки, учитывается наличие (и иногда частота) слов, но порядок, в котором они встречаются, игнорируется.

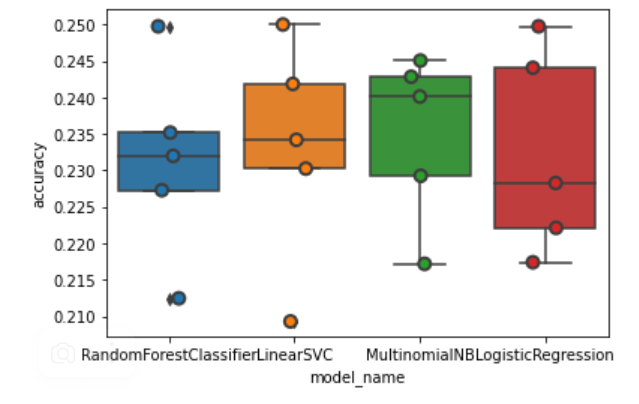
В частности, для каждого термина в нашем наборе данных мы рассчитали показатель, называемый Term Frequency, Inverse Document Frequency (частота термина, обратная частоте документов), сокращенно tf-idf. Это позволило нам получить 1271 признак для каждой из 5062 стихотворных строк, по которым наш алгоритм и будет оценивать стихотворную строку, получаемую на вход.

После всего вышеописанного преобразования данных, теперь, когда у нас есть все функции и метки категорий, мы можем перейти к обучению классификатора. Существует ряд алгоритмов, которые мы можем использовать для решения такого рода задач. Далее в тексте параграфа мы кратко расскажем о каждом.

Мы провели сравнительный анализ следующих четырех моделей:

* Logistic Regression (Логистическая регрессия)
* (Multinominal) Naive Bayes (Наивный Байес)
* Linear Support Vector Machine (Линейная машина с опорными векторами)
* Random Forest (Лес случайных деревьев)

Результаты тестирования представлены в таблице 4.



*Таблица 4. Результаты сравнительного анализа моделей машинного обучения для решения задачи автоматического определения поэтического размера текста.*

В первую очередь нужно упомянуть наивный байесовский классификатор как наиболее подходящий для решения задачи определения размера. В тестовых запусках мы использовали мультиноминальную его разновидность. В статистике наивные байесовские классификаторы представляют собой семейство простых "вероятностных классификаторов", основанных на применении теоремы Байеса с сильными (наивными) предположениями о независимости между признаками.

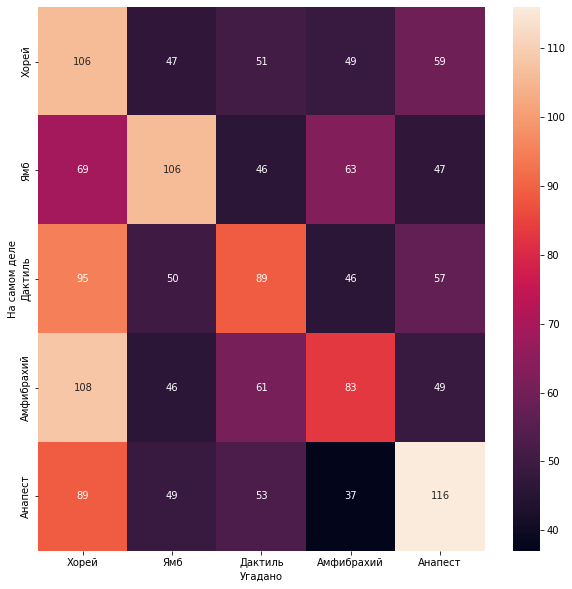
Логическая регрессия (Logistic Regression) в статистике представляет собой статистическую модель, которая моделирует вероятность одного события (из двух альтернатив), высчитывая логарифмические коэффициенты у события для того, чтобы предсказать, будет ли оно линейной комбинацией одной или нескольких независимых переменных ("предсказателей").

В машинном обучении машины с опорными векторами (Support Vector Machines, SVM) представляют собой контролируемые обучающие модели со связанными алгоритмами обучения, которые анализируют данные для классификации и регрессионного анализа. SVM — один из самых надежных методов прогнозирования. Учитывая набор обучающих примеров, каждый из которых помечен как принадлежащий к одной из двух категорий, алгоритм обучения SVM строит модель, которая присваивает новые примеры той или иной категории, что делает его вероятностным бинарным линейным классификатором. SVM сопоставляет обучающие примеры с точками в пространстве, чтобы максимально увеличить ширину разрыва между двумя категориями. Затем новые примеры отображаются в том же пространстве и прогнозируются как принадлежащие к категории в зависимости от того, на какую сторону разрыва они попадают.

Наконец, случайные леса (Random Forest) или леса случайных решений — это метод коллективного обучения для классификации, регрессии и других задач, который работает путем построения множества деревьев решений во время обучения. Для задач классификации результатом случайного леса является класс, выбранный большинством деревьев. Для задач регрессии возвращается среднее или среднее предсказание отдельных деревьев. Леса случайных решений корректируют привычку деревьев решений перестраиваться под их обучающий набор.  Случайные леса обычно превосходят деревья принятия решений, но их точность ниже, чем у деревьев с градиентным усилением. Однако характеристики данных могут повлиять на их производительность.

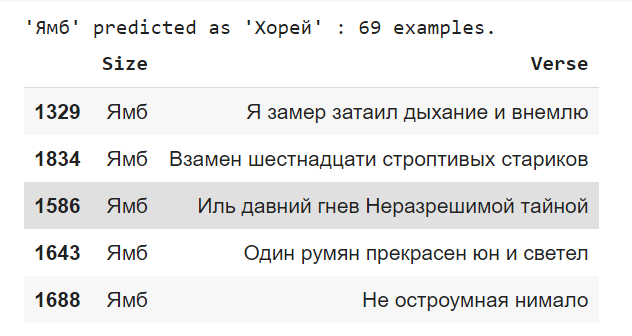
Как уже упоминалось выше, мы провели сравнительный анализ четырёх этих моделей. Результат немного выше прочих показал Наивный Байесовский классификатор, поэтому в качестве основной модели машинного обучения мы выбрали именно его.

Продолжая работу с нашей лучшей моделью (MultinominalNB — мультиноминальный Наивный Байес), мы рассмотрели матрицу путаницы и показали расхождения между прогнозируемыми и фактическими метками. Матрицу запутанности можно посмотреть в таблице 5.



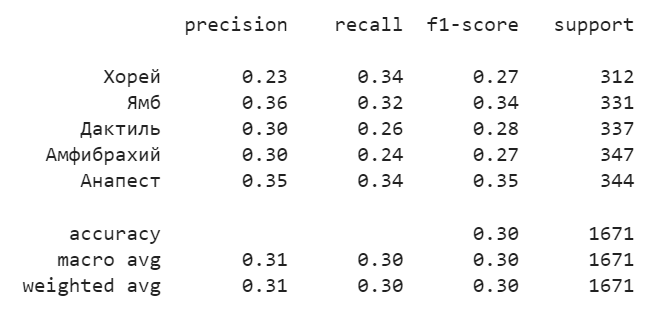
*Таблица 5. Матрица запутанности меток размеров и с чем чаще всего модель путала тот или иной размер.*

Пример наиболее распространённых ошибок модели представлен в таблице 6.



*Таблица 6. Пример наиболее распространённых ошибок модели в определении размера.*

Наконец, мы предоставляем метрики отчётности классификации по каждому размеру. Полный отчёт можно просмотреть в таблице 7, здесь мы обратимся к метрикам полноты, точности и f-меры, по которым мы оценивали и классификатор, основанный на правилах.



*Таблица 7. Сравнительный анализ ключевых метрик по каждому размеру для модели машинного обучения.*

Так, для хорея precision (точность) составила 0.23, recall (полнота) составила 0.34, а f-мера составила 0.27. Общая точность (accuracy) составила 23%.

Для ямба precision (точность) составила 0.36, recall (полнота) составила 0.32, а f-мера составила 0.34. Общая точность (accuracy) составила 36%.

Для дактиля precision (точность) составила 0.30, recall (полнота) составила 0.26, а f-мера составила 0.28. Общая точность (accuracy) составила 30%.

Для амфибрахия precision (точность) составила 0.30, recall (полнота) составила 0.24, а f-мера составила 0.27. Общая точность (accuracy) составила 30%.

Для анапеста precision (точность) составила 0.35, recall (полнота) составила 0.34, а f-мера составила 0.35. Общая точность (accuracy) составила 35%.

Как мы видим, результаты, полученные нами в результате использования методов машинного обучения, также далеки от идеала, при этом в отдельных категориях, например, при определении анапеста и амфибрахия они на несколько процентов превосходят классификатор, основанный на строгих правилах.

Отметив это, мы должны также привести аргументы относительно того, почему данное направление разработки и применения алгоритмов машинного обучения для автоматического определения размера стихотворного текста кажется нам более перспективным, чем дальнейшее совершенствование классификатора, основанного на строгих правилах.

Прежде всего стоит упомянуть о дополнительном тесте, который мы провели в рамках эксперимента с уже обученной моделью, описанной ранее. Как видно из наших данных, мы не предоставляли модели машинного обучения никакой дополнительной информации о тексте, включённом в обучающий корпус, кроме самой стихотворной строки и пометки об её размере. Методом случайной выборки мы собрали вспомогательный корпус из 20 строк каждого из пяти размеров, и указали у них дополнительную числовую информацию, которая, по нашей гипотезе, могла бы помочь в повышении точности модели. Так, например, мы прописали количество гласных для каждой строки, количество стоп у каждой строки, а также при помощи одного из алгоритмов нашей модели из параграфа 2.1. придали каждой строке числовую картину, состоящую из единиц и нулей, после чего обучили модель заново и оценили точность (Accuracy) на случайных стихотворных строках, не включённых в изначальный корпус.

Точность была выше на 11% для хорея и на 13% для ямба, для дактиля повысилась на 8%, для амфибрахия и анапеста — на 6% и 9% соответственно. Мы считаем, что дополнительные числовые данные могут существенно поднять точность определения размера, а в случае нашего эксперимента рост от их введения был в определённой степени нивелирован крайне малой выборкой в масштабах машинного обучения. Мы убеждены, что если придать несколько дополнительных метрик для всего нашего корпуса из 5062 строк, то точность повысится более существенно. Мы также уверены, что существуют другие ключевые метрики, способные повлиять на точность определения стихотворного размера, которые не столь очевидны на первый взгляд, но могут проявить себя при дальнейшем исследовании. Таким образом, мы считаем попытку применения алгоритмов машинного обучения для решения данной задачи перспективным направлением для исследования, способным в дальнейшем в достаточной степени влиять на процесс анализа и исследования поэтических текстов, так и на процесс их преподавания, изучения, понимания и написания.

## 2.3. Сравнительный анализ результатов работы классификатора поэтической метрики, основанного на строгих правилах, и классификатора поэтической метрики с применением методов машинного обучения

В данной части исследования мы проведём сравнительный анализ результатов работы двух прототипов классификатора метрики стиха и подытожим наши размышления о перспективах работы над каждым.

Сперва отметим общую для обоих инструментов тенденцию к более высокому проценту точности при работе с двусложными размерами, чем с трёхсложными. Это может быть связано как с тем, что программа, определяя размер строки как двусложный, имеет больший шанс угадать при выборе из двух вариантов, так и с тем, что строки, написанные двусложными размерами, имеют больше общего между собой ритмически, чем каждая из них в сравнении с трёхсложными размерами, поэтому программа может брать во внимание такие признаки как, например, длина строки, которая у трёхсложных размеров будет в силу объективных причин больше, и средняя длина слова в строке, которая также будет больше у стихов, написанных трёхсложными размерами.

Продолжая анализ двусложных размеров, также отметим, что у обоих инструментов точность определения ямба несколько выше, чем точность определения хорея: 49% точности ямба против 29% точности определения хорея у классификатора, основанного на правилах, и 36% точности ямба против 23% точности определения хорея у классификатора с применением методов машинного обучения. Это с большой долей вероятности связано с тем, что схематически стопа ямба имеет ударение на чётный слог, которое также имеет лишь стопа амфибрахия, в то время как хорейная стопа имеет ударение на нечётный слог, которое также имеют в себе стопы дактиля и анапеста, соответственно, инструмент сталкивается здесь с таким же выбором, какой он принимает между двусложными и трёхсложными размерами, когда с одной стороны — выбор из двух вариантов, а с другой — из трёх. Дополнительным фактором предпочтения инструмента в пользу ямба может служить традиционная для русских двусложных размеров пиррихийная стопа на предпоследнем слоге стиха, в сочетании с которой числовая схема ямба всё ещё воспринимается инструментом как ямбическая, в то время как в сочетании с хореем инструмент уже рассматривает её также скорее как ямбическую.

Стоит также отметить довольно сильные расхождения у классификаторов в проценте точности определения анапеста. Классификатор, основанный на строгих правилах, в принципе крайне редко когда определяет любую строку как анапест: всего 113 строк из 2500 корпуса было отнесено к анапесту, из них 48 на самом деле являлись анапестом. Общая точность классификатора по данному размеру составила всего лишь 10%. При этом классификатор с применением методов машинного обучения показал точность определения анапеста в 35%, что является лучшим результатом для трёхсложных размеров и только на процент отстаёт от ямба. Подобным результатам мы не можем дать правдоподобного объяснения.

Также стоит отметить, что классификатор с применением методов машинного обучения в целом показывает более стабильные результаты, он также обладает меньшим разбросом между худшим и лучшим результатом определения точности. Более того, мы убеждены, что результаты работы данного классификатора могут быть существенно улучшены дополнительной работой с данными без изменения кода программы, в то время как дальнейшее повышение точности классификатора, основанного на правилах, без дополнительных надстроек в алгоритме, практически недостижимо. Мы отдельно останавливаемся на этом моменте, поскольку любое усложнение и разветвление алгоритма программы может существенно повлиять как на скорость её работы, так и в принципе на повышение потенциального риска нарушения работы программы в целом.

Нельзя не подчеркнуть, что приводимые результаты работы классификаторов рассматриваются в рамках анализа каждой строки в отдельности. При этом общая точность работы обеих программ в определении размера стихотворения целиком существенно выше для всех силлабо-тонических размеров, кроме анапеста. Особенно хорошо это заметно при обращении к стихотворениям, написанным двусложными размерами. Общая точность определения размера целого стихотворения для ямба и хорея составила 75% для классификатора, основанного на правилах, и 65% для классификатора с применением методов машинного обучения. Для дактиля и амфибрахия этот процент колеблется в районе 50%.

Подводя итоги сравнительного анализа работы двух классификаторов, стоит отметить, что с учётом скорости развития технологий и увеличения вычислительных мощностей, доступных для исследования, более перспективным направлением работы кажется классификатор метрики поэтического текста с применением методов машинного обучения, поскольку даже при первых тестах именно он показывает наиболее стабильные результаты. Следующим шагом к усовершенствованию данного инструмента, как нам кажется, должна являться имплементация библиотеки rustress [48], использующей рекуррентные нейронные сети с механизмом LSTM для расставления ударений, в качестве предобработки сырых данных обучающего корпуса. Наличие явно и высокоточно выделяемых ударных и безударных символов должно существенно увеличить точность определения размеров поэтической метрики стиха.

На этом практическую часть нашего исследования мы считаем завершённой, а её цели и задачи — выполненными.

# Заключение

Итогом настоящего магистерского диссертационного исследования явилось достаточно полное представление о возможностях, недостатках, особенностях и перспективах применения алгоритмов машинного обучения для решения задачи автоматического определения поэтической метрики текста. В ходе диссертационного исследования были получены следующие выводы:

1. *Проанализирован путь становления и развития русской поэтической метрики;*

Путь становления и развития русской поэтической метрики прослеживается от XVII века и до сегодняшнего дня. Русская силлабика явилась заимствованием из Франции эпохи Нового времени, в то время как русская тоника существовала в неформализованном виде ещё в устном народном творчестве. Благодаря реформе Ломоносова-Тредиаковского возникло представление о русских силлабо-тонических размерах, в которых имело значение как расположение и ритмическое чередование ударных и безударных слогов, так и общее количество слогов в строке.

XX век отметился в русской поэзии как переход от силлабо-тоники к формализованной тонике, разработкой тонических размеров, появлением попыток применения статистических методов к русскому стихосложению. Наблюдалось общее расширение границ допустимого в поэзии.

1. *Выявить наиболее подходящий и актуальный список размеров для анализа;*

Наиболее подходящий и актуальный список поэтических размеров для анализа был выявлен на основании анализа потребностей основной массы предполагаемых пользователей. В связи с этим было решено остановиться на традиционных школьных силлабо-тонических размерах стиха, поскольку большая часть предполагаемой аудитории — ученики школ, которым нужна помощь в определении размеров классической русской поэзии XIX века. Также, ограничиваясь рамками этих размеров, было существенно проще продумать и имплементировать изначальный алгоритм, который впоследствии можно было бы модифицировать и дополнить тоническими размерами и другой информацией об анализируемом стихотворении.

1. *Изучить предыдущий опыт создания инструментов для автоматического анализа метрики и ритмики поэтического текста как в русскоязычных исследованиях, так и в англоязычной среде:*

Большинство исследований, в той или иной мере обращающихся к автоматическому анализу метрики поэтического текста, воспринимают данную задачу как побочную для задачи генерации поэтического текста. Результат анализа исследований показывает, что подобное отношение может напрямую зависеть от языка, для которого проводится исследование. Так, для языка с неразвитой морфологией и фиксированным ударением (как, например, английский язык) подобная задача действительно не составляет особого труда и удовлетворительно решается простым классификатором, основанным на строгих правилах. При этом для языков с развитой морфологией и подвижным ударением (как, например, русский или отдельные языки хинди) подобная задача становится центральной проблемой исследования и почти никогда не решается в достаточной степени удовлетворительно, оставляя простор для дальнейших исследований. При этом опыт разработки данного направления в англоязычной сфере интересен скорее с методологической точки зрения, поскольку именно там на данный момент в первую очередь появляются и внедряются передовые технологии.

1. *Разработать прототип классификатора, основанного на строгих правилах, для задачи автоматического определения стихотворного размера;*

Прототип классификатора, основанного на строгих правилах, для задачи автоматического определения стихотворного размера был успешно разработан и протестирован на корпусе из 2500 стихотворных строк. В основе классификатора лежит алгоритм, присваивающий числовую комбинацию ударных и безударных слогов для каждой строки и сравнивающий её со словарём “идеальных” размерных сеток, после чего выдающий наиболее схожий с полученной комбинацией размер. Точность классификатора в построчном анализе существенно ниже, чем для стихотворений целиком, и сильно колеблется в зависимости от размера: для строк анапеста точность составляет всего 10%, в то время как ямбические строки определяются в половине случаев.

1. *Разработать прототип классификатора с применением алгоритмов машинного обучения для задачи автоматического определения стихотворного размера;*

Прототип классификатора с применением алгоритмов машинного обучения для задачи автоматического определения стихотворного размера был успешно разработан и натренирован на корпусе из 5062 стихотворных строк. С помощью библиотеки для машинного обучения scikit-learn был разработан механизм извлечения признаков из необработанного текста, затем был проведён сравнительный анализ результатов предварительной работы моделей машинного обучения для решения задачи мультиклассификации, после чего наиболее подходящая модель была обучена на корпусе размеченных и предобработанных данных. Данные в корпусе были собраны на материале поэтического подкорпуса НКРЯ и дополнительно предобработаны библиотекой NLTK, после чего сведены в одну таблицу с помощью библиотеки pandas.

1. *Провести сравнительный анализ эффективности и точности определения поэтической метрики текста;*

Точность классификатора с применением методов машинного обучения для отдельных размеров была несколько ниже, однако у него не наблюдалось такого разброса между лучшим и худшим результатами, как у классификатора, основанного на строгих правилах, поэтому средние значения точности получились несколько выше. Учитывая этот фактор, а также возможности для дальнейшего повышения процента точности и расширения определяемых признаков стихотворной строки в частности и стихотворения в целом, разработка данного направления классификации в рамках нашего исследования была признана более перспективной и целесообразной.

Таким образом, обозначенные во введении задачи выполнены. В настоящем исследовании был проведён анализ литературы, посвящённой проблеме автоматического определения стихотворного размера, а также были созданы классификаторы поэтической метрики с применением разных подходов и методик, после чего был проведён сравнительный анализ результатов работы. Цель работы была достигнута.

# Библиографический список

1. Гаспаров М. Л. Оппозиция «стих – проза» и становление русского литературного стиха // Русское стихосложение: Традиции и проблемы развития. – 1985. – С. 264-277.
2. Томашевский Б. В. Стилистика и стихосложение. – 1959.
3. Холшевников В. Е. Основы стиховедения. – Academia;, 2004.
4. Лапшина Н. В., Романович И. К., Ярхо Б. И. Метрический справочник к стихотворениям А. С. Пушкина. — М.; Л.: Academia, 1934.
5. Плунгян В. А. Опыты неклассической метрики у Тютчева. Язык художественной литературы: традиционные и современные методы исследования. М.: Азбуковник, 2016, C. 393-406.
6. Русова Н. Ю. Терминологический словарь-тезаурус по литературоведению. От аллегории до ямба // М.: Флинта, наука. – 2004.
7. Гаспаров М. Л. Очерк истории русского стиха. – Фортуна Лимитед, 2002.
8. Тредьяковский В. К. Избранные произведения. – Советский писатель, 1963. – Т. 2.
9. Ломоносов М. Избранные произведения. – Strelbytskyy Multimedia Publishing, 2018.
10. Чернышевский Н. Г. <Критика>. Сочинения А. С. Пушкина. Изд. П. В. Анненкова // Чернышевский Н. Г. Полное собрание сочинений : В 15 т. М. : Гослитиздат, 1949. Т. 2. — С. 424–516.
11. Востоков А. Х. Опыт о русском стихосложении. – Directmedia, 2013.
12. Белый А. Символизм: Книга статей/Общ. ред. ВМ Пискунова //М.: Республика. – 2010.
13. Жирмунский В. М. Теория стиха. — Ленинград, 1975.
14. Томашевский Б. Русское стихосложение. Метрика. Пг.: Academia, 1923 //Вопросы поэтики. – №. 2.
15. Пушкин А. С. Евгений Онегин. М.: Художественная литература, 1967.
16. Холшевников В. Е. Сверхсхемное ударение // Энциклопедия К. Л. в 9-ти тт.–Т. 6. – 1962.
17. Гаспаров М. Л. Избранные труды, том III // О стихе. М.:«Языки русской культуры. – 1997.
18. Bailey J. Toward a statistical analysis of English verse: the iambic tetrameter of ten poets. – John Benjamins Publishing, 1975. – Т. 1.
19. Červenka M., Sgallová K. On a probabilistic model of the Czech Verse //Prague Studies in Mathematical Linguistics. – 1967. – Т. 2. – С. 105-120.
20. Тарановский К. Русские двусложные размеры. Статьи о стихе. – Litres, 2017.
21. Tarlinskaja M. English Verse //English Verse. – De Gruyter Mouton, 1976.
22. Колмогоров А. Н., Прохоров А. В. К основам русской классической метрики //Содружество наук и тайны творчества. – 1968. – С. 397-432.
23. Рогов, В.В. Что нужно знать переводчику о некоторых особенностях английского стихосложения Текст. /В.В. Рогов: в сборнике Keats J. Selected Poems. -M.: Прогресс, 1966.
24. Schumann A. K., Beierle C., Blößner N. Using finite-state machines to automatically scan Ancient Greek hexameter // Digital Scholarship in the Humanities. – 2022. – Т. 37. – №. 1. – С. 242-253.
25. De la Rosa J. et al. Transformers analyzing poetry: multilingual metrical pattern prediction with transfomer-based language models // Neural Computing and Applications. – 2021. – С. 1-6.
26. Estes A., Hench C. Supervised machine learning for hybrid meter // Proceedings of the fifth workshop on computational linguistics for literature. – 2016. – С. 1-8.
27. Perez Pozo A. et al. A bridge too far for artificial intelligence?: Automatic classification of stanzas in Spanish poetry // Journal of the Association for Information Science and Technology. – 2022.
28. Корчагин К. Русский стих: цезура. – Litres, 2021.
29. Rajan R., Raju A. A. Deep neural network based poetic meter classification using musical texture feature fusion // 2019 27th European Signal Processing Conference (EUSIPCO). – IEEE, 2019. – С. 1-5.
30. Yousef W. A. et al. Learning meters of Arabic and English poems with Recurrent Neural Networks: a step forward for language understanding and synthesis //arXiv preprint arXiv:1905.05700. – 2019.
31. Al-shaibani M. S., Alyafeai Z., Ahmad I. Meter classification of arabic poems using deep bidirectional recurrent neural networks // Pattern Recognition Letters. – 2020. – Т. 136. – С. 1-7.
32. Lamar A., Chambers A. Generating homeric poetry with deep neural networks // 2019 First International Conference on​ Transdisciplinary AI (TransAI). – IEEE, 2019. – С. 68-75.
33. Gopidi A., Alam A. Computational analysis of the historical changes in poetry and prose //Proceedings of the 1st International Workshop on Computational Approaches to Historical Language Change. – 2019. – С. 14-22.
34. Барахнин В. Б. и др. Автоматизация комплексного анализа русского поэтического текста: модели и алгоритмы // Вестник Новосибирского государственного университета. Серия: Информационные технологии. – 2015. – Т. 13. – №. 3. – С. 5-18.
35. Трошина Н.Н. Исследования по лингвистической поэтике // Социальные и гуманитарные науки. Отечественная и зарубежная литература. Сер. 6, Языкознание: Реферативный журнал. 2018. №4. URL: https://cyberleninka.ru/article/n/issledovaniya-po-lingvisticheskoy-poetike (дата обращения: 15.01.2022).
36. Пильщиков И.А., Старостин А.С. Автоматическое распознавание метра: проблемы и решения // Славянский стих. М.: Рукописные памятники Древней Руси, 2012. С. 492–498.
37. Аншаков О. М. Проблема автоматического определения метра и ритма русского стиха: математические модели и алгоритмы // Института русского языка им. ВВ Виноградова. – 2017. – С. 179.
38. Бойков В. Н., Каряева М. С., Пильщиков И. А. Формально-языковая модель стиха и регулятивов его метра для автоматизированной метрической идентификации // Компьютерная лингвистика и вычислительные онтологии. – 2019. – №. 3. – С. 124-141.
39. Козьмин А. В. Автоматический анализ стиха в системе Starling // Компьютерная лингвистика и интеллектуальные технологии: Тр. междунар. конф. «Диалог 2006» (Бекасово, 31 мая – 4 июня 2006 г.). М., 2006. С. 265–268.
40. Зализняк А. А. Грамматический словарь русского языка. – Русские словари, 2003.
41. Викисловарь. URL: <https://ru.wiktionary.org> (дата обращения: 16.01.2022).
42. Wikitextparser. URL: <https://pypi.org/project/wikitextparser/> (дата обращения: 15.01.2022).
43. Sberbank AI. Classic AI. Phonetic Baseline. URL: <https://github.com/sberbank-ai/classic-ai/tree/master/examples/phonetic-baseline> (дата обращения: 15.01.2022).
44. Национальный корпус русского языка. 2003—2022. URL: <https://ruscorpora.ru/> (дата обращения 23.01.2022).
45. Scikit-learn. URL: <https://scikit-learn.org/> (дата обращения: 01.03.2022).
46. Towards Data Science. URL: <https://towardsdatascience.com/> (дата обращения: 02.03.2022).
47. Russtress. URL: <https://pypi.org/project/russtress/> (дата обращения: 23.01.2022).
48. Ponomareva M. et al. Automated Word Stress Detection in Russian //Proceedings of the First Workshop on Subword and Character Level Models in NLP. – 2017. – С. 31-35.
49. Вычисление стихотворного размера / Хабр. URL: <https://habr.com/ru/post/651395/> (дата обращения: 28.12.2021).

# Приложения

**Приложение 1 - Список стихотворений**

Хорей (1008 строк):

- Белый, Посиди со мной

- Анненский, Солнца в высях нету

- Брюсов, Белый цвет магнолий

- Сологуб, Вывески цветные

- Вяземский, Сколько слёз я пролил

- Майков, Думал я, что небо

- Бальмонт, Если зимний день тягучий

- Сологуб, Лишь спустится мрак ночной

- Анненский, Недоспетым поле сжато

- Блок, Новый блеск излило небо

- Брюсов, Как завидна в час уныний

- Горький, На страницы вашего альбома

- Бунин, Не угас ещё вдали закат

- Бунин, Горячо сухой песок сверкает

- Бунин, На распутье в диком древнем поле

- Фет, Как красив на утренней заре

- Некрасов, Вы в своей душе благословенной

- Тютчев, Вот бреду я вдоль большой дороги

- Гиппиус, Ходит, дышит, вьётся между нами

- Брюсов, Золото, убранство тайного ковчега

- Жуковский, Царскосельский лебедь

- Бальмонт, Придорожные травы

- Дельвиг, Одинок месяц плыл

- Ломан, Ах, взгляните, закрывает тучкой боязливо

- Лохвицкая, В час полуденный

- Маковский, Зноен день, но с гор прохлада

- Цетлин, Театр войны

- Кондратьев, Маковея

- Бальмонт, Косогор

- Кузмин, Ожидание

- Клычков, Манул и Харикли

- Баратынский, Наслаждайтесь, всё проходит

- Вяземский, Хандра

- Лермонтов, Песнь Ингелота

- Одоевский, По дороге столбовой

- Языков, Вино

- Тургенев, Дай мне руку - и пойдём мы в поле

- Некрасов, Школьник

- Майков, Ночная гроза

- Фет, Прекрасная ночь

- Мережковский, Ювенал о Древнем Риме

Ямб (1002 строки):

- Сологуб, Вот осень наступила

- Брюсов, Случайность и намеренность

- Блок, Где ты паришь теперь

- Бальмонт, Я помню, ночь кончалась

- Жуковский, Уж думал я, что я забыт

- Пушкин, Дубравы, где в тиши свободы

- Пушкин, Христос воскрес, питомец неба

- Иванов, Монастырь в Субнако

- Брюсов, Флоренция Декамерона

- Бунин, К прибрежью моря длинная аллея

- Фет, Всё дождь и дождь, и солнце лик свой прячет

- Фёдоров, Сумерки

- Брюсов, Последние слова

- Мережковский, Весь этот жалкий мир отчаянья и муки

- Надсон, Пугая мысль мою томящей тишиной

- Вяземский, Дорогою

- Лохвицкая, Синий дьявол

- Лермонтов, Стояла серая скала на берегу морском

- Капнист, Жил царь

- Брюсов, Презрение

- Блок, Что было год назад, всё то же

- Блок, Когда отдамся чувствам страстным

- Тютчев, Сияет солнце, воды блещут

- Некрасов, Так это шутка? Милая моя

- Полонский, Затворница

- Щербина, Два образа предстали пред тобою

- Григорьев, Две судьбы

- Мей, Беги её, чего ты ждёшь

- Кюхельбекер, Совет

- Огарев, А вы меня забыли

- Лермонтов, Приветствую тебя я, злое море

- Ростопчина, Виктору Гюго

- Губер, Могила моя

- Бенедиктов, Предчувствие

- Лермонтов, День гас, в наряде голубом

- Грибоедов, Душа

- Пушкин, О муза пламенной сатиры

- Дельвиг, Вдохновение

- Баратынский, Везде бранит поэт Клеон

Дактиль (1019 строк):

- Брюсов, Волны волос упадали

- Анненский, Сила господняя с нами

- Блок, В полночь глухую рождённая

- Вяземский, Чёрные очи

- Некрасов, Где твоё личико смуглое

- Некрасов, Демону

- Блок, Отдых напрасен

- Фофанов, Грешница

- Бальмонт, Майя

- Трефолев, Странник

- Жуковский, Рыцарь Роллон

- Случевский, Снежною степью лежала душа одинокая

- Бальмонт, Жизнь - отражение лунного лика

- Якубович, Помнишь ли, друг мой, дорогу

- Колосова, Бор мой

- Рождественский, Нет, не напрасно я в звёздном лесу

- Сологуб, Стыд

- Ширяевец, Мил мне смиренник пырей

- Блюменфельд, Бледные зори на небе

- Платонов, Солнце уж в силе

- Парнок, Зеркало держит Эрот

- Иванов, Лес опрокинут в реке

- Арский, Бойцы

- Цветаева, Волосы я или воздух целую

- Цветаева, На смех и на зло

- Ходасевич, Швея

- Дрожин, К песне

- Клюев, Молитва солнцу

- Набоков, Карлик безрукий во фраке

- Садовской, Душный туман заплели

- Бунин, Спутница

- Март, В курильне

- Белый, Асе

- Пастернак, Прощанье

- Фиолетов, Переменность

- Иванов, В небе над дымными долами

- Зенкевич, Сумрачный бог

- Цветаева, Правда

- Ходасевич, Ангелы

- Случевский, Женщина и дети, Философ

- Фофанов, Что ты сказала, я не расслышал

- Льдов, Город полуночной грезой объят

- Апухтин, Еду я ночью, темно и угрюмо

- Некрасов, Свадьба

- Никитин, Буря

Амфибрахий (1007 строк):

- Белый, Последнее слово

- Сологуб, Лунный свет

- Анненский, Я думал, что сердце из камня

- Случевский, Гуляя в сиянье заката

- Майков, Поэзии гений крылатый

- Сологуб, Был ветер так нежен

- Гиппиус, Там

- Языков, Голицыной

- Мей, В то время, когда по пучинам

- Кюхельбекер, Родство со стихиями

- Баратынский, Мадонна

- Брюсов, Старый викинг

- Бальмонт, Затон

- Толстой, Сижу да гляжу я

- Фет, К Цирцее

- Некрасов, И скучно, и грустно, и некого

- Полонский, в гостиной

- Лермонтов, Листок

- Гнедич, Гифтак

- Рождественский, Тряхнула б ты, Память, кошёлкою

- Злобин, Свиданье

- Андреев, Холод пространств

- Блок, Бред

- Брюсов, Из Ада изведенные

- Бальмонт, От капли росы, что трепещет, играя

- Бунин, И снилося мне, что осенней порой

- Гиппиус, Никогда

- Мережковский, Не надо желаний

- Соловьев, Эпиграмма на Победоносцева

- Гиппиус, Баллада

- Жуковский, Свисток

- Карамзин, К Лиле

- Волков, Зимою к друзьям

- Державин, Четыре возраста

- Востоков, На смерть воробья

- Баратынский, Лазурные очи

- Сатин, Природа с своей красотой беспредельной

- Бенедиктов, Реки

Анапест (1022 строки):

- Пастернак, Все наденут сегодня пальто

- Мандельштам, Воздух пасмурный

- Белый, Леопардовая лапа

- Пастернак, Зима

- Набоков, Беженцы

- Бунин, Гор сиреневых тучи встают

- Клюев, Ель мне подала лапу

- Блок, Я увидел во сне

- Шершневич, Обращайтесь с поэтами как

- Эллис, Если бедное сердце незримо рыдает

- Гумилёв, Однажды вечером

- Северянин, Воздушная яхта

- Есенин, Подражанье песне

- Мандельштам, О красавица Сайма

- Гнедич, Скиллодим

- Мей, Видение

- Рубцов, Чудный месяц плывет над рекою

- Бродский, Время года - зима

- Ахмадулина, Зимняя замкнутость

- Самойлов, Соловьи не прельщают мотивом

- Кирсанов, Временный дом

- Вознесенский, Здесь гостям наливают

- Заболоцкий, Петухи поют

- Пастернак, Бабье лето

- Парнах, Я всего натерпелся, поверь

- Тарковский, Дождь в Тбилиси

- Перелешин, Мы плетём кружева

- Набоков, К России

- Петров, Елка

- Паркау, Злая осень

- Твардовский, Не стареет твоя красота

- Колосова, Лебединые перья

- Чегринцева, Вальс

- Северянин, перевал через Ловчен

- Штейнберг, Волчья облава

- Михайлов, Сон невольника

- Случевский, Из природы. Снега

- Богданов, Проезжим