*Анисимова Ксения, Романенко Максим, Хаертдинова Йолдуз*

**VOZNESENSKYLAB:POETICS. Предварительные исследования поэтики А.А. Вознесенского цифровыми методами**

**Аннотация**

В статье представлена попытка описать поэтическое наследие А. А. Вознесенского с помощью автоматических методов анализа естественного языка. Поэзия Вознесенского, практически не исследованная DH-специалистами, становится материалом для испытания методов тематического моделирования, метрик удобочитаемости (readability) и автоматического распознавания метра. Полученные результаты будут использованы для создания электронного сборника стихотворений нового типа – основанного на читательских сценариях.

<https://github.com/haertdinova/voznesenkylab_poetics>

**Введение**

VOZNESENSKYLAB:POETICS – проект, целью которого является создание нелинейного электронного издания стихотворений А.А. Вознесенского.

Нелинейность издания реализуется за счет читательских сценариев. Читательский сценарий подразумевает под собой: а) деление всего наследия на категории; б) чтение произведений в рамках выбранной категории; в) определенную для каждой категории внутреннюю навигацию; г) возможность в процессе чтения перейти в другой читательский сценарий.

Необходимость прибегать именно к комплексной концепции читательского сценария, которая помимо структурных составляющих включает в себя сам процесс использования издания, обусловлена предположением о том, что чтение полного массива наследия А.А. Вознесенского в хронологическом, алфавитном и других порядках, в сути которых не учтена семантика текстов, создает определенные сложности в формировании у читателя полноценного впечатления о творчестве поэта. В таком случае выбор стихотворений для чтения часто случайный, например, можно открыть случайную страницу издания. В случае с хронологией читатель, конечно, может осознанно выбрать интересующий его период, но это будет иметь смысл, если у него уже есть представление о творчестве этого поэта. Таким образом, основание структуры издания на читательских сценариях позволит широкому кругу читателей, в особенности тем, кто ещё не был знаком или мало знаком с наследием Вознесенского, сформировать полноценное впечатление о наследии поэта. При этом читатель не лишается субъектности в процессе чтения издания за счет возможности выбора сценария и реализованных между ними переходами.

Планируется реализовать следующие читательские сценарии:

1. “Метр”: деление на стихотворения, написанные в размерах, относящихся к силлабо-тонической системе и во всех остальных размерах.
2. “Сложность”: деление по удобочитаемости и лексической сложности;
3. “Тема”: деление издания по темам, которые прослеживаются в творчестве Вознесенского;
4. “Хронология”: деление наследия по датам написания;
5. “Песня”: набор стихотворений, которые легли в основу песен.
6. “Случайные”: случайный выбор стихотворения из издания.

Как мы видим, в издании будут отражены и “классические” способы организации текстов внутри издания – в сценариях “Хронология” и “Случайные”, и будут реализованы уже на этапе непосредственного создания электронного сборника. Остальные же сценарии предполагают деление на основе семантики текстов или их структуры (что в случае с поэтическими текстами представляет отдельный интерес). Последним и посвящена эта статья.

Реализовываться сценарии и деления на категории будут автоматическими, компьютерными методами. Например, деление на темы будет реализовано с помощью тематического моделирования. Цифровые методы ещё не так широко применяются к поэтическим текстам, поэтому дополнительной целью проекта становится разработка или совершенствование существующих методов автоматического анализа стихотворений и их апробация на материале творчества Вознесенского.

Важно отметить, что на данном этапе эта цель основная: настоящее исследование – «proof of concept», результатом которого должны стать выводы о применимости разработанных методов к анализу поэзии и к созданию задуманного издания.

**Данные**

Первый этап создания нелинейного цифрового издания, как, впрочем, любого другого, предполагает отбор произведений. Издание, созданное в рамках проекта VOZNESENSKYLAB:POETICS включает только стихотворные произведения, в том числе визуальную поэзию, и в частности видеомы, придуманные А.А. Вознесенским. Сам автор этот жанр определяет как “мир, увиденный через видео”[[1]](#footnote-0), под этим подразумевалось выражение поэтического через изобразительное: “Никакое другое искусство не находится столь близко к поэзии, как изобразительное, и чаще всего те образы и метафоры, которые не могут до конца выразиться словами, выражаются в этом”[[2]](#footnote-1).

Всего собрано 41 издание произведений Вознесенского и небольшое количество архивных материалов. Все издания предполагается обработать и построить на их основе полный размеченный стихотворный корпус наследия Вознесенского. Одна из особенностей этого корпуса состоит в том, что в него будут включены стихотворения 1985-2010 годов, которые ранее практически не были изучены исследователями[[3]](#footnote-2), этот корпус, таким образом, включая стихотворения и других периодов, окажется наиболее полным. На данный момент работа над этим корпусом не завершена. Размер корпуса, который предполагается сделать в рамках проекта VOZNESENSKYLAB:POETICS, составляет около 1700 произведений Вознесенского, включая визуальную поэзию и видеомы.

Для апробации методов автоматического распознавания метра (реализация сценария “Метр”) и автоматического определения сложности (сценарий “Сложность”) был создан семплированный корпус текстового наследия Вознесенского. Он включает примерно равное количество стихотворений (60-70 в нашем случае) для каждого десятилетия периода творческой активности А.А. Вознесенского: от 1950-х до 2000-х. При этом из корпуса исключены произведения периода 1940-х, т.к. не удалось найти достаточного количество стихотворений. Если бы найденные стихотворения этого периода всё же включили в корпус, он вышел бы неравномерным и непрезентативным. Всего произведений в корпусе чуть более 400, это около четверти от общего предполагаемого числа произведений, которые войдут в итоговое издание. Также в семплированный корпус не включены визуальные произведения Вознесенского: их предобработка для последующего анализа цифровыми методами – тема для отдельного исследования. Предполагается, что такого корпуса достаточно по крайней мере для апробации методов.

Семплированный корпус размечен также, как будет размечен полный. Пример разметки и сам корпус можно найти в github-репозитории проекта <https://github.com/haertdinova/voznesenkylab_poetics>.

Для тематического моделирования (сценарий “Тема”) использован набор данных, на основе которого написан материал ТАСС, посвященный 90-летию Вознесенского[[4]](#footnote-3). Он состоит примерно из 1300 стихотворений, для каждого из которых известно название и десятилетие написания. В данных от ТАСС нет видеом, но есть стихотворения “позднего” периода, 1985-2010гг., а самое раннее стихотворение относится к 1940-м. Датасет сформирован автоматически на основе издания 2020-го года “Андрей Вознесенский: полное собрание стихотворений в одном томе” [[5]](#footnote-4). Вручную добавлены несколько стихотворений, написанных в 2000-х годах. Датасет ТАСС состоит лишь из текстовых произведений, однако его можно назвать включающим основное наследие автора, однако данные в нем “грязные”: при парсинге издания буквы в некоторых словах разделились пробелами, во многих стихотворениях не сохранилось деление на строфы. Общее число “грязных” мест в текстах составляет около 20% от всех текстов корпуса. Это влияет на результаты обработки, но для этапа “proof of concept” такое состояние данных все равно позволило получить результат: для тематического моделирования сэмплированный корпус недостаточен, и полный, но неочищенный корпус оказался лучшим из доступных решений.

**Методы**

Для реализации читательских сценариев используются цифровые методы обработки текста: автоматическое распознавание метра, тематическое моделирование, подсчет различных индексов удобочитаемости для русского языка и метрика лексического разнообразия Type-Token Ratio. Рассмотрим детально каждый из них.

**Автоматическое распознавание метра**

Автоматическое распознавание метра давно волнует исследователей поэзии. Независимо друг от друга они пытаются выделить основные сложности этого направления исследования, предлагают вертикальный анализ текста (сравнение строк друг с другом)[[6]](#footnote-5), уделяют особое внимание сложным размерам (например, логаэд[[7]](#footnote-6)), пишут алгоритмы, которые проводят полный формальный анализ стихотворения и возвращают целый набор характеристик текста, касающихся не только метра[[8]](#footnote-7).

Похожая задача стоит и перед нами: деление стихотворений по метру – один из наиболее очевидных и нужных сценариев, и нам важно максимально автоматизировать его реализацию в связи с большим количеством текстов в корпусе.

При этом стоит признать, что наша задача всё же немного проще: ещё в начале работы над изданием было принято решение внутри этого сценария делить тексты на силлаботонику и всё остальное, не выделяя дольники, логаэды и т.д., чтобы любому читателю был интуитивно понятен принцип распределения стихотворений по группам. Действительно, даже не зная о метрах, практически каждый сможет понять, что стихотворения «Над темной молчаливою державой…» и «Верблюды пишут верлибры…» формально отличаются, не осознавая, что первый текст – силлаботоника, а второй – нет:

*Над темной молчаливою державой*

*какое одиночество парить!*

*Завидую тебе, орел двуглавый,*

*Ты можешь сам с собой поговорить.*

*Вознесенский А.А. Над темной молчаливою державой…[[9]](#footnote-8)*

*Верблюды пишут верлибры*

*Нимфы рифмуются с отражением*

*В полусвете воска вокзалов*

*вязнет моя судьба*

*Вознесенский А.А. Верблюды пишут верлибры…[[10]](#footnote-9)*

В связи с этим мы решили создать алгоритм распознавания метра, который определял бы метр стихотворений, написанных силлаботоникой, а остальным текстам присваивал бы категорию “иные” (other).

Работу алгоритма мы заранее разделили на два этапа:

1. Поиск строгих силлаботонических стихотворений без отклонений или с отклонениями в пределах нормы (редкие сверхсхемные ударения и любое количество пропусков ударений) и добавление их в сценарий;
2. В случаях, когда метр не определился, – расчёт уровня отклонений от общего метра стихотворения для последующего добавления в сценарий или исключения текста из выборки (если процент отклонений не превышает определённый нами допустимый уровень, то определяется метр, а текст добавляется в сценарий).

Таким образом, на первом этапе в сценарий мы принимаем только тексты, в которых нет отклонений от метра и сверхсхемных ударений, и тексты, в рамках одной строки которых количество сверхсхемных ударений не превышает количество схемных. Для реализации этого решения в алгоритме необходимо в первую очередь определить, что мы понимаем под сверхсхемным ударением и как оно образуется.

Сверхсхемное ударение – «ударение, стоящее на слабом слоге стиха в силлабо-тонических размерах»[[11]](#footnote-10), то есть не предусмотренное метрической схемой. В двусложных стихотворных размерах (хорей, ямб) нарушения в виде сверхсхемных ударений образуются за счёт односложных слов, а в трёхсложных (дактиль, амфибрахий, анапест) – за счёт односложных и двусложных слов. Связано это с запретом переакцентуации: «Сверхсхемное ударение и пропуск схемного ударения в русской силлабо-тонике не могут совмещаться в одном слове (“запрет переакцентуации”); поэтому сверхсхемные ударения могут приходиться лишь на слова, не превышающие объемом междуиктового интервала: на 1-сложные в ямбе и хорее, на 1- и 2-сложные в дактиле, амфибрахии и анапесте»[[12]](#footnote-11).

Стоит отметить, что запрет переакцентуации не абсолютен и может нарушаться, как и любой другой поэтический закон. Тем не менее «в русском классическом стихе его нарушение маловероятно»[[13]](#footnote-12). В связи с этим мы установили следующую границу отклонений от метра для нашего алгоритма: отступление внутри строки считается допустимым, если сверхсхемное ударение не сопряжено с пропуском схемного. При этом пропуски ударений никак не ограничиваются, так как являются нормой для русской силлаботоники.

Это позволяет нам определить несколько типов нарушений силлаботонического метра и условий, которые необходимо учесть в алгоритме, чтобы полноценно реализовать упомянутый ранее критерий отбора текстов в сценарий:

1. Если сверхсхемное ударение приходится на начало строки:
   1. в ямбе и амфибрахии оно создаёт последовательность из двух ударений подряд с интервалом в один шаг[[14]](#footnote-13), на первом и втором слоге, и приходится только на односложные слова;
   2. в анапесте – последовательность из двух ударений подряд, с интервалом в один шаг, на втором и третьем слоге, или последовательность из двух ударений с интервалом в два шага, на первом и третьем слоге, и приходится на односложные или двусложные слова в обоих случаях;
   3. в хорее и дактиле невозможно, потому что первое схемное ударение в этих метрах приходится на первый слог.
2. Если сверхсхемное ударение приходится на середину строки:
   1. в хорее и ямбе оно создаёт последовательность из трёх ударений подряд с интервалом в один шаг (например, на первом, втором и третьем слоге в хорее) и приходится на односложные слова;
   2. в дактиле, амфибрахии и анапесте – последовательность из трёх ударений с интервалом в два и один шаг (например, на третьем, пятом и шестом слоге в анапесте) или с интервалом в один и два шага (например, на втором, третьем и пятом слоге в амфибрахии), приходится на односложные или двусложные слова в обоих случаях.
3. На клаузуле, то есть в конце строки, сверхсхемное ударение считается невозможным.
4. Количество сверхсхемных ударений не превышает количество схемных в рамках одной строки.

Этот список формальных условий уже значительно упрощает процесс определения метра, но не является окончательным. Попробуем сократить список, вернувшись на шаг назад, к акцентуации текстов.

Для автоматической расстановки ударений использовалась библиотека Python «ru\_accent\_poet», созданная специально для работы с русской поэзией. Библиотека разрабатывалась с учётом недостатков существовавших до неё акцентуаторов, благодаря чему, авторам удалось достигнуть точности в 98%[[15]](#footnote-14). Особенность этой библиотеки заключается в том, что ударение никогда не ставится на односложные слова. Это помогает избегать некорректной переакцентуации, а также позволяет нам исключить из алгоритма условия 1.1 и 2.1 из вышеприведённого списка. Так, мы приходим к следующей схеме определения метра:

1. Компьютер получает текстовый файл со стихотворением, в котором предварительно были расставлены ударения.
2. Стихотворение делится на строки.
3. Для каждой строки выполняются следующие действия:
   1. Определяется порядковый номер каждого ударного слога:
   2. Рассчитывается интервал между ударными слогами.
      1. Если все интервалы в строке соответствуют одной из групп метров, то каждое ударение проверяется на предмет соответствия одному из метров. Определяется метр строки.
      2. Если хотя бы один интервал в строке отличается от ожидаемых (то есть не делится на 2 или 3), то строка проверяется на наличие допустимых сверхсхемных ударений. При выполнении условий, перечисленных выше, определяется метр строки.
      3. При нарушении хотя бы одного из условий (неверный интервал или недопустимые отклонения) метр строки определяется как «other» и не считается силлаботоническим.
4. После определения построчного метра строки сравниваются друг с другом, происходит вертикальная проверка текста. В случае совпадения метров всех строк определяется метр всего стихотворения. В противном случае метр стихотворения определяется как «other» и не считается силлаботоническим.
5. Если в стихотворении есть строки с одним ударением, то сперва определяется общий метр стихотворения. Если он является единым и силлаботоническим, происходит проверка соответствия строки с одним ударением общему метру. В случае несовпадения метр стихотворения определяется как «other» и не считается силлаботоническим.

Если алгоритм не столкнулся ни с одним из исключений, то он определяет метр стихотворения. Этим завершается первый этап работы нашего алгоритма.

На втором этапе нам было необходимо задать алгоритму предельный процент строк, не подчиняющихся общему метру текста. Мы решили ввести этот шаг в реализацию сценария, потому что как и сам Вознесенский, так и весь период, на который приходится его творчество, экспериментальны в форме и в содержании. Ямб Вознесенского не будет и не может быть равен ямбу Пушкина, и чистая силлаботоника будет встречаться слишком редко для полноценной реализации сценария.

При этом нет общепринятого уровня отклонений, который мы можем считать допустимым, – в нашем алгоритме это 25%, то есть не более одной строки из четырёх может метрически не совпадать с остальными. Такое число позволяет нам учесть даже самые короткие стихотворения, но всё же оно строгое, поэтому свободные стихотворения и тексты, включающиеся в другие системы стихосложения, не попадут в сценарий.

Этот шаг, расчёт процента отклоняющихся от метра строк, завершает второй этап работы алгоритма и весь алгоритм в целом. Все стихотворения, не попавшие в выборку, не считаются силлаботоническими и в дальнейшем не рассматриваются.

Итогом становится таблица со столбцами, в которых отображается название текста, список ударений по строкам и метр стихотворения. Полный код опубликован в github-репозитории проекта – <https://github.com/haertdinova/voznesenkylab_poetics>.

**Тематическое моделирование**

Помимо добавления в корпус информации о метре для каждого стихотворения, естественным кажется выделить темы, к которым они относятся. Один из способов сделать это – построение тематической модели. Работа по тематическому моделированию корпуса, как правило, представляет собой последовательность из следующих этапов:

1. Первичная обработка текста, которая включает очищение текста от небуквенных символов, стоп-слов, а также может включать лемматизацию[[16]](#footnote-15) или отбор слов определенной части речи для учета алгоритмом (например, только существительных[[17]](#footnote-16)).
2. Выбор алгоритма и параметров для него (главный параметр – количество топиков) на основании оценки качества полученных тем.
3. Применение алгоритма и представление результатов в виде, доступном для интерпретации.
4. Интерпретация полученных тем, которая помимо реализации возможных исследовательских задач позволяет также представить результаты в понятном для пользователя виде.

Каждый из приведенных этапов тематического моделирования допускает различные способы его реализации, а потому специфика подхода, избранного авторами проекта, заслуживает отдельного обсуждения, чему будет посвящен соответствующий фрагмент следующего раздела статьи. Здесь же будут описаны лишь основные инструменты данной работы.

Первый этап работы состоял в очищении корпуса от небуквенных символов и стоп-слов, а также приведении лексем к начальным формам, для чего был использован автоматический анализатор Mystem.

Для второго этапа в качестве алгоритма тематического моделирования был выбран NMF (Non-negative matrix factorization), который, как считается, больше подходит для анализа поэтических текстов[[18]](#footnote-17). Этот алгоритм получает на вход терм-документную матрицу (наблюдения – документы, переменные – термы, выраженные у нас леммами), фиксирующую показатель встречаемости слов каждого документа. В качестве метрики использовалось значение tf-idf, формула которого следующая:

где – частотность слова в документе , – число всех документов, – число документов, в которых встречается слово .

Таким образом, показатель максимизируется при максимизации встречаемости слова в документе и минимизации количества документов с ним.

Далее, алгоритм проводит факторизацию матрицы, сокращая число переменных до некоторого определяемого вручную числа факторов с учетом распределения весов каждого предполагаемого фактора (топик) по переменным (термам). В итоге он производит разложение входной матрицы на произведение двух новых: W – документ-топик, и H – топик-терм. Таким образом, в матрице H содержится информация о ключевых словах каждого топика, а в матрице W – о главных топиках каждого документа (из нее же, к слову, можно извлечь информацию о главных документах, представляющих каждый топик, что не менее полезно).

Для выбора оптимального количества тем был использован метод измерения согласованности (coherence) ключевых слов топика – TC-W2V. Этот метод основан на использовании модели word2vec (архитектуры skipgramm) относительно всего нашего корпуса. Для построения модели word2vec использовалась библиотека Gensim[[19]](#footnote-18).

Модель word2vec является матрицей векторного представления слов, основанного на контекстуальной близости каждой лексемы корпуса с другими. Далее, показатель близости измеряется на основании этой матрицы для каждой пары ключевых слов топика из матрицы H, а среднее значение этого показателя для топика выражает меру его связности (coherence).

Наконец, чтобы определить оптимальный параметр числа тем (параметр k) для алгоритма NMF, выбираются границы градации этого числа и производится применение алгоритма тематического моделирования последовательно для каждого значения k в выбранном промежутке. Наконец, для каждой полученной модели измеряется показатель общей связности (overall coherence), равный средней связности для тем, и выбирается наиболее связная модель. Предполагается, что, чем выше значение общей связности модели, тем лучше она отражает дистрибутивные (т.е. семантические, согласно принципам дистрибутивной семантики) отношения между словами в корпусе.

**Автоматическое определение сложности**

Издание нацелено на широкий круг людей, в том числе не имеющих формального образования или просто юных читателей, наравне с людьми увлеченными поэзией и уже знакомыми с творчеством автора. Оценка сложности текстов позволит упростить подбор текстов для читателей с разным культурным бэкграундом.

Индексы удобочитаемости – это один из немногих формализованных методов, которыми можно измерить один из аспектов сложности текста: сложность восприятия. Хотя изначально предполагалось использовать эти метрики на учебных и научных текстах, они всё же показывают интерпретируемые результаты на художественных текстах[[20]](#footnote-19). В рамках данного исследования удобочитаемость измерялась индексами, адаптированными для русского языка: Flesh-Reading Ease (формула ), формула Оборневой[[21]](#footnote-20) (формула ) и Мацковского[[22]](#footnote-21) (формула ), все из которых были описаны и проанализированы в статье “Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics” В. Соловьева, В. Иванова и М. Солнышкиной. Формулы зависят от следующих количественных характеристик текста: ASL – среднее количество слов в предложениях, ASW – среднее количество слогов в словах, LW – количество слов, состоящих из трех или более слогов.

Для оценки лексической сложности использовалась метрика Type-Token Ratio, которая представляет собой отношение уникальных лемм (начальная, нормальная форма слова) к общему их числу. В литературе и названии метрики использовано слово “токен” т.е. разные словоформы одного слова (лексемы) – это разные токены. Русский язык можно отнести к синтетическим, поэтому имеет смысл здесь считать именно леммы.

Описанные выше цифровые методы по предположению авторов позволят реализовать читательские сценарии “Метр”, “Тема”, “Сложность” и расширить круг людей, которым было бы интересно наше издание.

**Эксперименты и обсуждение**

Избранные нами цифровые методы обработки текста – математические модели, которые в определенном смысле “дискретизируют”[[23]](#footnote-22) смыслы или структуры, содержащиеся в тексте. Это оказывается особенно заметным при работе с поэзией, где каждое слово и его место в тексте несет за собой определенное впечатление. Но как ни странно, в нашем случае достаточным оказывается такой грубый подход: исходить из того, что стихотворение можно однозначно отнести к одной из двух, трех, четырех категорий – это само по себе допущение такой “дискретизации”.

Этот раздел, одновременно с описанием результатов применения цифровых методов на корпусе стихотворений Вознесенского, посвящен подбору, настройке и интерпретациям результатов работы используемых нами математических моделей. Каждый этап использования цифрового метода сопряжен с некоторыми допущениями и выборами, описание которых позволит насколько это возможно очертить какие именно “точки смысла” мы пытаемся найти и описать.

**Автоматическое распознавание метра**

Как уже было сказано ранее, проверка выработанного алгоритма проводилась на семплированном корпусе стихотворений. Этот корпус содержит стихотворения, в которых сохранено правильное деление на строки и строфы, исправлены опечатки и текст которых приведён к своему образцу из опубликованного издания. Значит, вероятность корректного определения метра возрастает.

При этом даже вычитанный корпус потребовал дополнительной обработки до запуска алгоритма:

1. Из всех стихотворений были удалены ненужные данные: теги, дата написания, название.
2. «Лесенка»[[24]](#footnote-23) была объединена в одну строку. Ориентиром для объединения служили версии одного стихотворения в разных изданиях, графическое деление текста и рифма.
3. Латиница была транслитерирована с помощью библиотеки Python transliterate. Перед акцентуацией дополнительно была проведена проверка транслитерированного текста.
4. Числа были записаны словами.

После этих шагов была проведена акцентуация текста. Она также потребовала дополнительной проверки: были исправлены некорректные ударения (частые в том числе из-за окказионализмов Вознесенского и транслитерированной латиницы), логические речевые акценты, удалены лишние ударения (в основном на слове «или»). Ударения на словах не удалялись, если возможна различная интерпретация, потому что для точной акцентуации такого слова необходимо заранее знать метр стихотворения (это также объясняет необходимость учёта отклоняющихся от метра строк).

После всех описанных шагов тексты были обработаны алгоритмом автоматического распознавания метра. Результаты его работы удовлетворительны.

Всего нам удалось определить метр 156 стихотворений, что составляет около 38% семплированного корпуса. Когда алгоритм тестировался на полном неочищенном корпусе, мы получили выборку из 91 стихотворения или менее 7% полного корпуса.

Точность алгоритма также высока. Проверка по определившимся ударениям показывает, что написанный код точно распознаёт чистый метр, учитывает пропуски ударений и сверхсхемные ударения, а также не пропускает тексты с недопустимым процентом ударений. Не удаётся ему учесть тексты с отклонениями и в то же время с большим количеством строк, метр которых не может быть до конца определён из-за нехватки ударений или метр которых неоднозначен (например, строка с ударениями на втором и восьмом слогах может быть и ямбом, и амфибрахием). Сами результаты не менее интересны.

Больше 50% силлаботоники Вознесенского приходится на ямб. Это ожидаемый результат, потому что ямб уже в XIX веке стал традиционным для русской поэзии. Реже всего у него встречаются трехёсложные размеры (дактиль, амфибрахий, анапест), а среди них самый редкий – дактиль. Это тоже понятные результаты, вписывающие его в русскую традицию.

Что примечательно, много силлаботоники у Вознесенского встречается в начале творчества, в 1950-ые гг. – 47,5%. Это число резко снижается в 2000-ые гг. – до 33,8%. Однако нельзя сказать, что это закономерное снижение, поскольку линия сокращения числа силлаботонических стихотворений не прослеживается[[25]](#footnote-24). Вероятно, в 2000-ые гг. Вознесенский начинает экспериментировать с поэзией и усложнять её, выводя из силлаботоники. Однако это лишь «заметки на полях». Основной итог в нашем случае – формирование сценария.

Полученные результаты показывают, что сценарий «Метр» может быть успешно реализован по существующему алгоритму и имеющимся данным, а тексты могут быть разделены на категории «Традиционное» (силлаботоника) и «Нетрадиционное»[[26]](#footnote-25). Теперь он лишь требует проверки читательской интуицией, после которой (в случае успеха) он будет расширен до сценария «Размер».

**Тематическое моделирование**

Каждый из этапов работы по тематическому моделированию допускает множество различных способов его реализации. Чтобы охарактеризовать специфику нашего подхода, обосновать его выбор, а также обеспечить воспроизводимость результатов, подробнее остановимся на каждом из этапов проведенной нами работы.

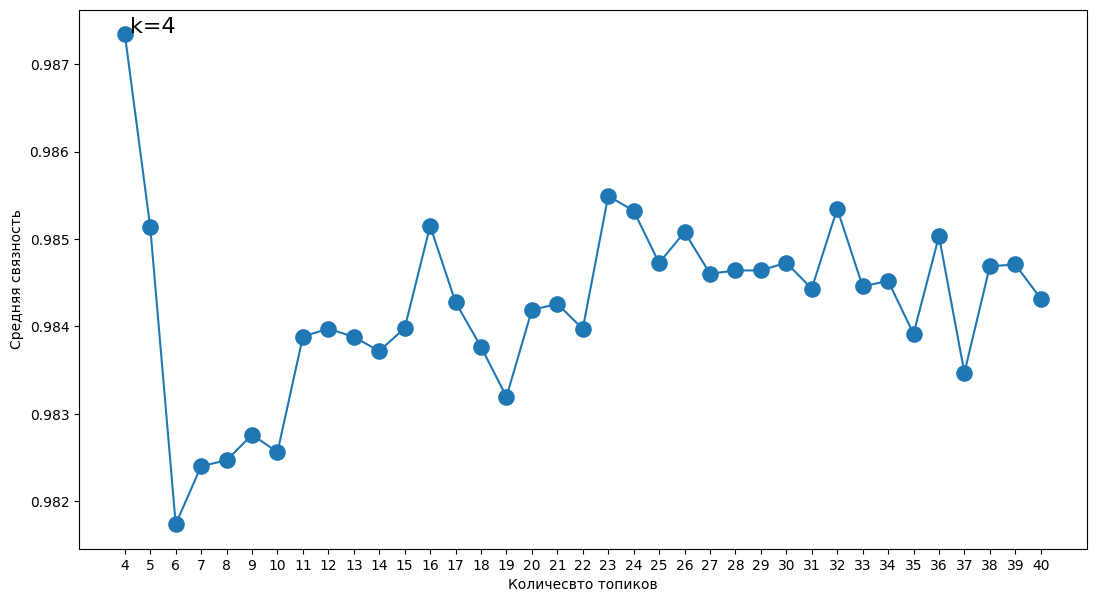
Первый этап: выбор стратегии первичной обработки текста – важный компонент работы, в значительной степени определяющий вид и интерпретируемость получаемых тем. Так, в зависимости от того, будет ли произведена лемматизация текста или нет, полученные топики могут иметь преимущественно семантический или же смешанный характер (к примеру, при отсутствии лемматизации текста в результате могут быть выделены группы лексем, представленные в тексте в виде рифмующихся словоформ, не образующих семантическое единство). Чтобы не усложнять устройство модели, было решено ограничиться только одними видом тем, что предполагало проведение лемматизации.

Что касается выбора лексем конкретной части речи в качестве единиц анализа, то для задачи проекта это видится нецелесообразным. Так как исключение каких-либо категорий слов на первом этапе сохранит дистрибутивное отношение между словами оставшейся категории (исходя из формулы tf-idf, приведенной в предыдущем разделе статьи), то облик получившихся топиков изменится только за счет отсутствия ключевых слов исключенных категорий, что способно лишь ограничить интерпретируемость этих топиков.

Второй этап: выбор параметра количества топиков (параметр k) для алгоритма. Напомним, что для избрания оптимального значения k проводилась оценка связности полученной модели (overall coherence) для каждого из значений k алгоритма в интервале. Надо отметить, что в случае нашего проекта, ориентирующегося помимо связности тем на их понятность для читателя, был выбран небольшой интервал, ограниченный следующими показателями: min k = 4, max k = 40.

Оценка связности ключевых слов топиков каждой модели проводилось с помощью матрицы word2vec для всего корпуса. При этом важно уточнить, что в качестве датасета для word2vec был использован наш стихотворный корпус, а не общий для языка. Это решение обусловлено предполагаемой уникальностью словоупотребления внутри поэтического контекста и в особенности в рамках поэзии Вознесенского.

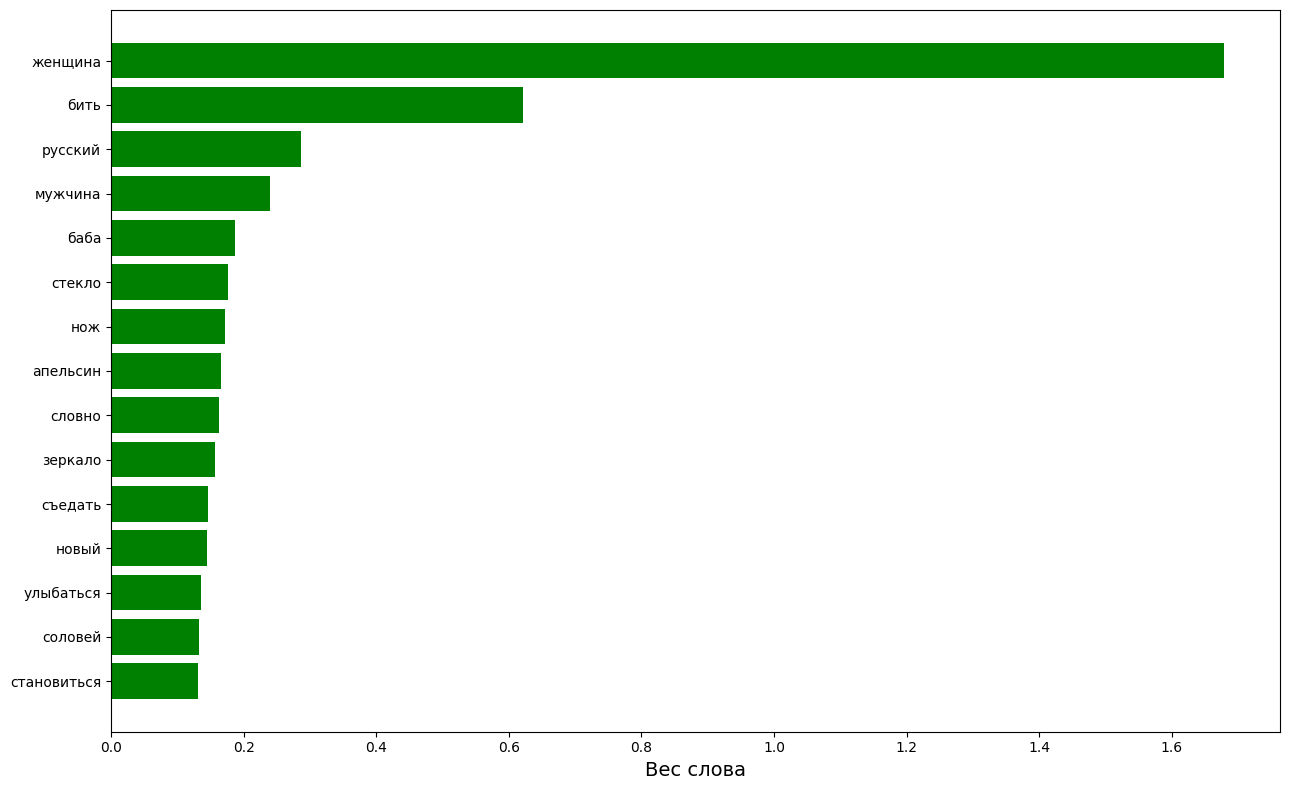
Для того, чтобы выбрать оптимальное значение параметра k, результаты процедуры оценки связности были отражены на графике (рис. 1). Несмотря на то, что максимальный показатель связности обнаруживается для 4 и 5 тем, было решено не останавливаться на столь малом количестве, что могло повредить интерпретируемости. На промежутке примерно между k = 25 и k = 31 на графике заметно относительное плато без резких колебаний, на котором наблюдается переход тренда от постепенного восхождения связности к его постепенному снижению. При этом, в отличие от последующих, этот фрагмент графика остается относительно стабильным при экспериментировании со значением параметра random\_state (для задействующих случайность алгоритмов, если задан отличным от “None”, этот параметр обеспечивает воспроизводимость результата). В конечном счете, был выбран k=25, учитывая, что колебания связности в целом невелики, а увеличение числа топиков может привести к появлению «шумовых» с точки зрения интерпретируемости тем.

*Рисунок 1. Изменение общей связности модели в зависимости от количества выделяемых топиков.*

Третий этап: применение модели и представление результатов. Как было сказано, в качестве результатов алгоритм приводит матрицы W (документ-топик) и H (топик-терм). Из последней были выведены 25 топиков (тем), которые представлены в таблице (табл. 1) в виде наборов из 10 ключевых слов, упорядоченных по убыванию веса. Как видно из таблицы, темам присвоены имена (в скобках рядом с номером), которые являются результатом процесса интерпретации, о котором будет сказано далее. Пока стоит отметить, что тема, обозначенная как “Failed”, образовалась из-за неправильно распознанного текста в датасете от ТАСС, используемого для тематического моделирования. С помощью вычитки текстов стихов эта тема будет переопределена алгоритмом. Так или иначе очевидно, что в таком виде она не может быть представлена читателю.

На диаграмме ниже (рис. 2) изображено распределение веса по ключевым словам в одном из топиков. В следующей же таблице (табл. 2) представлены самые характерные стихи для этой же темы, отсортированные по убыванию показателя представленности темы в стихах (из матрицы W).

| Номер темы | Ключевые слова для темы |
| --- | --- |
| Тема 0 (Домашняя) | стена, дом, окно, труба, тень, чужой, жить, звук, плечо, сквозь |
| Тема 1 (Failed) | ст, ен, ет, го, ер, сь, ся, ск, зн, ва |
| Тема 2 (Мотив белого) | белый, надевать, роза, темный, словно, выходить, бить, джинсы, свет, ходить |
| Тема 3 (Мотив неба) | небо, синий, крест, написать, лес, ребенок, подымать, голубой, цвет, облако |
| Тема 4 (Морская) | море, берег, пляж, палец, чайка, волна, плыть, напоминать, выходить, начинать |
| Тема 5 (Музыкальная) | музыка, бывать, скрипка, играть, окно, слово, скрывать, жить, звезда, сыграть |
| Тема 6 (Любовная) | любить, говорить, москва, глядеть, спасать, ждать, счастливый, русски, голова, хороший |
| Тема 7 (Мотив души) | душа, хороший, господь, ангел, птица, глаз, народ, жить, лес, место |
| Тема 8 (Молитва) | давать, господи, свет, дерьмо, боже, спасать, речь, господь, россия, называть |
| Тема 9 (Мотив устремления) | идти, новый, ребенок, мать, след, тянуть, встречать, дыхание, дверь, господь |
| Тема 10 (Мотив снега) | снег, красный, прийти, метель, река, хороший, снежок, слепой, солнце, пахнуть |
| Тема 11 (Мотив своего) | свой, пить, жить, ребенок, становиться, чужой, друг, отец, красивый, подруга |
| Тема 12 (Фигура поэта) | поэт, понимать, убивать, свет, звезда, последний, земной, россия, меж, кричать |
| Тема 13 (Фигура женщины) | женщина, бить, русский, мужчина, баба, стекло, нож, апельсин, словно, зеркало |
| Тема 14 (Время) | век, новый, принимать, счастливый, стол, русский, конец, школа, вещь, замечать |
| Тема 15 (Мотив прошедшего) | оставаться, умирать, родина, пушкин, читать, самый, жить, бог, пусть, думать |
| Тема 16 (Мотив чувства) | сердце, слеза, боль, дух, понимать, дождь, видеть, точно, пойти, мука |
| Тема 17 (Мотив черного) | черный, голубой, свеча, квартира, ствол, нестись, зеленый, новый, роща, дыра |
| Тема 18 (Территориальная) | страна, странный, становиться, рим, спасать, спина, блин, боль, страдание, стол |
| Тема 19 (Мотив сна) | спать, сон, глаз, сниться, ангел, ночь, час, париж, дух, огонь |
| Тема 20 (Мотив покидания, сохранения) | земля, оставлять, уходить, любовь, берег, голова, вода, выходить, увидеть, ландыш |
| Тема 21 (Мотив полета) | лететь, помнить, красный, красивый, упасть, шар, песенка, стоять, гореть, кровь |
| Тема 22 (Мотив прощения) | прощать, боже, лицо, щека, золотой, слеза, взгляд, плакать, вода, слава |
| Тема 23 (Мотив поля) | поле, русский, сон, ах, туз, звезда, улыбка, россия, андрей, вонзаться |
| Тема 24 (Мотив руки (часто Божьей)) | рука, бежать, христос, знать, глаз, мир, протягивать, сторона, нога, успевать |

*Таблица 1. Ключевые слова по каждой из выделенных алгоритмом топиков.*

*Рисунок 2. Распределение весов ключевых слов в Теме 13.*

| *Номер стиха* | *Название стиха* |
| --- | --- |
| *736* | *ВЫПИСКА ИЗ КНИГИ «ЧАРОДЕЙСТВО,* |
| *51* | *БЬЮТ ЖЕНЩИНУ* |
| *142* | *БЫТЬ ЖЕНЩИНОЙ* |
| *121* | *БЬЕТ ЖЕНЩИНА* |
| *1129* | *Нынче время — крупных лаж,* |
| *404* | *БОГОМАТЕРЬ-37* |
| *721* | *ДОНЖУАНИЗМ* |
| *761* | *ЖЕНСКОЕ ПЛАМЯ* |
| *926* | *ДАЧА* |
| *1006* | *ХАКЕР* |

*Таблица 2. Десять наиболее характерных стихов для Темы 13.*

Четвертый этап: интерпретация. Основная трудность на этом этапе связана с отсутствием в академической литературе попыток вручную выделить набор тем, наиболее свойственных поэзии Вознесенского, что усложняет процесс интерпретации, а также делает невозможным проведение компаративного исследования. В связи с этим выдвижение общезначимой гипотезы о тематическом разнообразии наследия поэта, как и ее проверка посредством применения обозначенного метода изначально представлялись невозможными. Поэтому процедуру интерпретации полученного результата пришлось ограничить двумя операциями, применяемыми к каждой из получившихся тем. Во-первых, из рейтинга ключевых слов каждого топика было выбрано (в том числе, на основании распределения весов) характеризующее большинство из них слово или составлено описание темы на их основании. Во-вторых, мы обращались к перечню наиболее характерных для темы стихов и при помощи “close-reading” некоторых из них пытались определить меру и качество, в которых выведенная нами характеристика представлена в текстах. Это позволило, с одной стороны, уточнить формулировку темы, а с другой – понять, описывает ли топик тематическую направленность стиха, или же отдельные художественные приемы внутри его текста.

Надо сказать, что не все топики имеют одинаковое распределение весов ключевых слов. Так, если опираться на показатель веса самого тяжелого слова для темы, то все топики можно условно разделить на три группы. Первая группа – топики, вес самого выраженного ключевого слова для которых находится в пределах интервала от 2 до 2.5. В нее входят топики, наличие которых в стихе наиболее сильно коррелирует с одним или несколькими словами. Топики первой группы следующие: Темы 0 (самые весомые слова: “стена”, “дом”, “окно”), 5 (самое весомое слово: “музыка”, вес которого наибольший для всех тем: более 2.5), 6 (“любовь”, “говорить”) и 10 (“снег”).

Вторая группа – характеризуется интервалом от 1.5 до 2. В эту группу входит большинство топиков, в том числе и Тема 13 на Рис. 2.

Наконец, третья группа – интервал от 0.8 до 1.5. В нее входит 8 топиков: Темы 12 (“поэт”, “понимать”), 15 (“оставаться”, “умирать”, “родина”), 18 (“страна”, “странный”), 19 (“спать”), 20 (“земля”, “оставлять”, “уходить”, “любовь” - у темы минимальный показатель веса ключевых слов, а именно, ни одно не поднимается выше 0.8), 22 (“прощать”, “боже”), 23 (“поле”, “русский”) и 24 (“рука”, “бежать”, “христос”). Малый вес ключевых слов в этих темах может означать, помимо прочего, что наиболее выраженного слова для топика нет, и он представляет собой, таким образом, слияние нескольких тем, коррелирующих с различными словами (возможно, не очень распространенными во всем корпусе) в разных текстах.

Это можно продемонстрировать на примере Темы 20, которая в таких стихах, как “Где заснеженная Россия?”, “Ушла душа. Земле до лампочки” или “Пропорции”, представлена, главным образом, словом “Земля”, а в стихах “Пасата”, “Переезд” или “Говорит мама” – словом “оставлять”. Редкое сочетание ключевых слов в одном стихе (как в тексте “Где заснеженная Россия?”, в котором также присутствует лемма “уходить”) при этом не способно компенсировать явное семантическое различие слитых топиков.

Еще одно препятствие на пути однозначной интерпретации топиков можно проиллюстрировать также на примере Темы 20. Оно связано с различием в значениях одной лексемы, которое может скрадываться при изъятии слов из контекста. Так, слово “оставлять” употребляется в значении “покидать” в стихе “Пасата” и “Переезд”, а в “Говорит мама” – в значении “сохранять”. Та же шероховатость метода проявляет себя и в Теме 22, в основном характеризуемой лексемой “прощать”. Так, если в “Тобольском романсе” речь идет о прощении, то в знаменитой “Осени в Сигулде”, также отнесенной алгоритмом к этому топику, лирический герой прощается с тем, что проносится мимо его вагона, используя слова, превращенные лемматизатором в ту же самую форму “прощать”. Очевидная разница смыслов здесь утрачивается в ходе предобработки текстов, тогда как в случае с Темой 20 изъян допущен самим алгоритмом тематического моделирования, как правило, к слову, чувствительным к разнице контекстов.

Последнее замечание, которое необходимо сделать относительно процедуры интерпретации тем, заключается в упомянутой выше разнице между топиками, характеризующими тематическую направленность стиха, и теми, что обозначают отдельные художественные приемы внутри его текста. На это последнее особое значение топиков для художественных текстов указывал Наварро-Колорадо в своей статье по тематическому моделированию для испанской поэзии Золотого Века, где он, говоря о топике как о поэтическом мотиве, определял последний как набор повторяющихся слов и образов, используемых поэтом в разных стихах для раскрытия разных тем[[27]](#footnote-26). Так, поэт может, используя мотив моря или музыки, говорить, например, о любви.

В случаях, где при внимательном прочтении наиболее характерных для топика стихов Вознесенского можно было с некоторой степенью уверенности определить его в разряд стилистического мотива (т.е. набора метафор и образов, используемых в том или ином сюжете), в название топика добавлялось слово “Мотив”. Остальные названия указывают на категории людей, объекты или явления, часто выступающие центральными предметами внимания Вознесенского в его творчестве. Разумеется, подобная классификация довольно условна и нуждается в уточнении, возможно, с привлечением независимых аннотаторов.

Итак, проведенный анализ показывает, что тематическое моделирование может быть использовано для реализации сценария “Тема”, который будет представлять собой модель из 24 тем (за исключением темы “Failed”), в рамках каждой из которых читателю будет предложен набор соответствующих ей стихов Вознесенского. При этом тот факт, что каждый стих, согласно данной модели, может быть атрибутирован сразу несколькими топиками, делает неизбежным наличие пересечений между темами. Перемещаясь среди текстов одной из тем, читатель всегда будет иметь возможность либо остаться в рамках актуального раздела, либо переместиться к стихам, другой тематико-стилистической специфики.

**Автоматическое определение сложности**

В рамках реализации сценария “Сложность” для стихотворений семплированного корпуса были подсчитаны метрики удобочитаемости: Flesh Reading Ease (FRE), формула Оборневой, формула Мацковского; для подсчета лексической сложности выбран Type-Token Ratio (TTR).

Избранные индексы удобочитаемости – уточненные для русского языка метрики Flesh-Kincaid. Формула Мацковского является аналогом метрики Flesh для русского языка и показывает количество необходимых лет образования для отсутствия сложностей понимания текста, формула Обориной описывается аналогично и является доработанной версией формулы Мацковского[[28]](#footnote-27).

Результаты подсчетов FRE и формулы Обориной примерно совпадают: для понимания самого простого стихотворения из семплированного корпуса может вовсе не понадобиться образования (см. значение Max для FRE и Min для формулы Оборневой в таблице 3), а для понимания самого сложного стихотворения корпуса может понадобиться высшее образование[[29]](#footnote-28). В среднем же читателю Вознесенского, исходя из этих подсчетов, лучше закончить начальную школу.

Таблица 3. *Значения и интерпретация измерений метрик удобочитаемости.*

|  | **Min** | **Max** | **Mean** | **Единицы измерения** |
| --- | --- | --- | --- | --- |
| **FRE** | 41.32 | 126.75 | 91.51 | Оценка в среднем от 100 до 0 (чем меньше, тем сложнее) |
| **Оборнева** | -3.7 | 16.73 | 2.64 | Годы образования |
| **Мацковский** | 2.66 | 26.4 | 6.92 | Годы образования |
| **TTR** | 23 | 100 | 79 | Проценты |

Формула Мацковского показывает завышенные примерно на половину показатели, при этом сохраняя очень похожую динамику[[30]](#footnote-29) с остальными двумя формулами при подсчете среднего значения по всем стихотворениям из корпуса для определенного десятилетия (см. рис. 3-5). Мы видим самые высокие показатели (т.е. самые сложные тексты) в 60-х для Мацковского и 70-х для FRE и Оборневой, и самые низкие (т.е. самые простые тексты) в 90-х для Мацковского и Оборневой и в 50-х для FRE и Оборневой.

| *Рисунки 3-5. Среднее значение метрик удобочитаемости стихотворного наследия Вознесенского (по десятилетиям)* | | |
| --- | --- | --- |
|  |  |  |
| *Рисунок 3. Flesh Reading Ease (для русского языка)* | *Рисунок 4. Формула Мацковского* | *Рисунок 5. Формула Оборневой* |

Было показано, что стихотворения Вознесенского для этих данных в среднем имеют TTR = 0,79 (см. таб. 3), что может говорить о значительном лексическом разнообразии стихотворений. Стоит также отметить, что показатели TTR для других поэтов, приведенные в статье Б.В. Орехова “Метрическое и лексическое разнообразие в стихах А. А. Вознесенского” также выделяют Вознесенского как более лексически сложного поэта по сравнению с другими своими современниками-поэтами и некоторыми другими.

Результаты измерения сложности текстов А.А.Вознесенского, включенных в семплированный корпус, позволяют реализовать читательский сценарий “Сложность”, так как метрики показывают интерпретируемый результат в относительных значениях, т.е. в соотношении друг с другом. В категорию “Простые” попадут те стихотворения, которые показывают одинаково небольшое (меньше среднего) необходимое для понимания количество лет образования и показатель лексической сложности ниже среднего, например: “Гарь”, “Шахты”, в группу “Сложные”, соответственно, попадут стихотворения с показателями выше среднего, например: “Демонстрация языка”, “Сан-Франциско - Коломенское…”.

**Заключение**

Выбранные цифровые методы действительно позволяют реализовать сложные читательские сценарии “Метр”, “Сложность”, “Тема”.

При измерении сложности было показано, что стихотворения Вознесенского отличает высокое лексическое разнообразие, тематическое моделирование показало, что творчество поэта представляет собой не только предметное, но и стилистическое разнообразие, которое характеризуется широким набором используемых им образов и метафор.

Результаты автоматического распознавания метра также демонстрируют разнообразие, однако его степень не так высока, но это доказывает, что творчество Вознесенского в целом метрически вписывается в русскую традицию (при этом значимые отклонения и новаторство поэта, конечно, отрицать нельзя).

Следующим этапом в создании электронного издания станет завершение полного размеченного корпуса наследия Вознесенского, включающего визуальную поэзию автора. Отдельно следует выработать способы описания визуальной поэзии, на которых возможно было бы сделать выводы, используя описанные в этой статье цифровые методы. Использование этих методов на полном и чистом размеченном корпусе может прояснить и некоторые особенности поэтики Вознесенского.

**Список литературы**

1. *Аншаков О.М.* Проблема автоматического определения метра и ритма русского стиха: математические модели и алгоритмы // Труды Института русского языка им. В.В. Виноградова, 2017. № 1. С. 179-188.
2. *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.
3. *Вознесенский А.А.* Аксиома самоиска. М: СП ИКПА, 1990 г.
4. *Вознесенский А.А.* Андрей Вознесенский: полное собрание стихотворений в одном томе. М.: Альфа-книга, 2020.
5. *Вознесенский А.А.* Полной собрание стихотворений и поэм в одном томе. М.: Альфа-книга, 2012.
6. *Гаспаров М.Л.* Современный русский стих: метрика и ритмика. М.: Наука, 1974.
7. *Короткова О.Ю.* Комбинированный словарно-нейросетевой акцентуатор для разметки русского поэтического текста // Труды Института русского языка им. В.В. Виноградова, 2022. № 3. С. 181-190.
8. *Мацковский М.С.* Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения в условиях массовой коммуникации. М., 1976. С. 126–142.
9. *Оборнева И.В.* Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. канд. пед. наук. М., 2006. 165 с.
10. *Околита А. (ТАСС)*. "Слишком западник и слишком патриот". К 90-летию Андрея Вознесенского, 2023.
11. *Орехов Б. В.* Башкирский стих XX века. Корпусное исследование, 2019.
12. *Орехов Б. В*. Метрическое и лексическое разнообразие в стихах А. А. Вознесенского, 2022.
13. *Пильщиков И.А., Старостин А.С.* Автоматическое распознавание метра: проблемы и решения // Славянский стих. Т. 9. М.: Рукописные памятники Древней Руси, 2012. С. 492-498.
14. *Пильщиков И.А.* Заседание московского лингвистического кружка 1 июня 1919 г. и зарождение стиховедческих концепций О. Брика, Б. Томашевского и Р. Якобсона // Revue des études slaves, LXXXVIII 1-2, 2017.
15. *Плунгян В. А., et al.* Поэтический корпус в рамках Национального корпуса русского языка: общая структура и перспективы использования*,* 2009.
16. *Холшевников В.Е*. Сверхсхемное ударение // Краткая литературная энциклопедия / Гл. ред. А. А. Сурков. М.: Советская энциклопедия, 1971. Т. 6: Присказка – «Советская Россия». С. 699.
17. Цифровые гуманитарные исследования. Монография / А.Б. Антопольский, А.А. Бонч-Осмоловская, Л.И. Бородкин [и др.]. — Красноярск : Сиб. федер. ун-т, 2023. — 272 с.
18. *Jockers, M. L., Mimno, D.* Significant themes in 19th-century literature // Poetics. 2013. Vol. 41. P. 750–769. doi: 10.1016/j.poetic.2013.08.005
19. *Jones, E., & Nulty, P*. “Quantitative measures of lexical complexity in modern prose fiction”, 2019
20. *Kincaid, J.P., Fishburne, R.P., Rogers, R.L., & Chissom, B.S*. Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis, 1975
21. *Navarro-Colorado, B.* On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry // Front. Digit. Humanit., 2018. Vol. 5. P. 1–12. doi: 10.3389/fdigh.2018.00015
22. *Rehurek, R., Sojka, P.* Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta: ELRA). 2010. P. 45–50.
23. *Senter, R.J.; Smith, E.A*. (November 1967). "Automated Readability Index" (PDF). *Amrl-Tr. Aerospace Medical Research Laboratories (U.s.)*. Wright-Patterson Air Force Base: 1–14.
24. *Solnyshkina, Marina & Ivanov, Vladimir & Solovyev, Valery.* (2018). Readability Formula for Russian Texts: A Modified Version: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II. 10.1007/978-3-030-04497-8\_11.
25. *Solovyev, Valery & Ivanov, Vladimir & Solnyshkina, Marina.* (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. Journal of Intelligent & Fuzzy Systems. 34. 1-10. 10.3233/JIFS-169489.
26. *Vilém Zouhar* “Poetry, Songs, Literature, Legalese and Translationese: Automated Sentence Complexity Perspective”, 2022.

1. *Вознесенский* *А.А.* Видеомы. Стихи, визуальные образы, проза, Рик ‘Культура’, 1992 [↑](#footnote-ref-0)
2. *Там же.* [↑](#footnote-ref-1)
3. *Орехов* *Б. В.* Метрическое и лексическое разнообразие в стихах А. А. Вознесенского, 2022 [↑](#footnote-ref-2)
4. *Околита А. (ТАСС)*. "Слишком западник и слишком патриот". К 90-летию Андрея Вознесенского, 2023 [↑](#footnote-ref-3)
5. *Вознесенский А.А.* Андрей Вознесенский: полное собрание стихотворений в одном томе. М.: Альфа-книга, 2020. [↑](#footnote-ref-4)
6. *Пильщиков И.А., Старостин А.С.* Автоматическое распознавание метра: проблемы и решения // Славянский стих. Т. 9. М.: Рукописные памятники Древней Руси, 2012. С. 492-498. [↑](#footnote-ref-5)
7. *Аншаков О.М.* Проблема автоматического определения метра и ритма русского стиха: математические модели и алгоритмы // Труды Института русского языка им. В.В. Виноградова, 2017. № 1. С. 179-188. [↑](#footnote-ref-6)
8. *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. [↑](#footnote-ref-7)
9. *Вознесенский А.А.* Над темной молчаливою державой… // А.А. Вознесенский. Аксиома самоиска. М: СП ИКПА, 1990 г. С. 237. [↑](#footnote-ref-8)
10. *Вознесенский* *А.А.* Верблюды пишу верлибры… // А.А. Вознесенский. Полной собрание стихотворений и поэм в одном томе. М.: Альфа-книга, 2012. С. 990. [↑](#footnote-ref-9)
11. *Холшевников В.Е.* Сверхсхемное ударение // Краткая литературная энциклопедия / Гл. ред. А. А. Сурков. М.: Советская энциклопедия, 1971. Т. 6: Присказка – «Советская Россия». С. 699. [↑](#footnote-ref-10)
12. *Гаспаров М.Л.* Современный русский стих: метрика и ритмика. М.: Наука, 1974. С. 14. [↑](#footnote-ref-11)
13. *Пильщиков И.А.* Заседание московского лингвистического кружка 1 июня 1919 г. и зарождение стиховедческих концепций О. Брика, Б. Томашевского и Р. Якобсона // Revue des études slaves, LXXXVIII 1-2, 2017. С. 159. [↑](#footnote-ref-12)
14. Интервал рассчитывается путём вычитания большего слога в последовательности из меньшего. [↑](#footnote-ref-13)
15. *Короткова О.Ю.* Комбинированный словарно-нейросетевой акцентуатор для разметки русского поэтического текста // Труды Института русского языка им. В.В. Виноградова, 2022. № 3. С. 181-190. [↑](#footnote-ref-14)
16. О возможных издержках лемматизации художественного текста для интерпретируемости полученного результата см.: *Navarro-Colorado, B.* On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry // Front. Digit. Humanit., 2018. Vol. 5. pp. 1–12. doi: 10.3389/fdigh.2018.00015 [↑](#footnote-ref-15)
17. См., например: *Jockers, M. L., Mimno, D.* Significant themes in 19th-century literature // Poetics. 2013. Vol. 41. pp. 750–769. doi: 10.1016/j.poetic.2013.08.005 [↑](#footnote-ref-16)
18. Цифровые гуманитарные исследования. Монография. Красноярск, 2023. С. 141. [↑](#footnote-ref-17)
19. См.: *Rehurek, R., Sojka, P.* Software framework for topic modelling with large corpora // Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks (Valletta: ELRA). 2010. P. 45–50. [↑](#footnote-ref-18)
20. *Jones, E., & Nulty, P*. “Quantitative measures of lexical complexity in modern prose fiction”, 2019 [↑](#footnote-ref-19)
21. *Оборнева И.В.* Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. канд. пед. наук. М., 2006. 165 с. [↑](#footnote-ref-20)
22. *Мацковский М.С.* Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения в условиях массовой коммуникации. М., 1976. С. 126–142. [↑](#footnote-ref-21)
23. Использование этого математического термина, который примерно означает описание непрерывной математической структуры совокупностью её точек – красивая идея нашего коллеги Евгения Дуненкова. [↑](#footnote-ref-22)
24. Под «лесенкой» мы подразумеваем целую стихотворную строку, разделённую на несколько печатных строк. Зачастую она имеет вид лестницы, например, как в стихотворении 2009 г. «Свежесть чувств»:

    *свежие чувства*

    *в себе*

    *сохраняйте* [↑](#footnote-ref-23)
25. У Вознесенского наблюдается чередование силлаботоники по десятилетиям: большой процент силлаботоники – маленький процент силлаботоники – большой процент силлаботоники. При этом до 2000-ых он никогда не опускается ниже 35, а после 1950-ых – не поднимается выше 40. [↑](#footnote-ref-24)
26. Названия категорий рабочие, будут модифицироваться и тестироваться на обычном читателе. [↑](#footnote-ref-25)
27. *Navarro-Colorado B.* Op. cit. p. 10. [↑](#footnote-ref-26)
28. *Solovyev, Valery & Ivanov, Vladimir & Solnyshkina, Marina.* (2018). Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics. Journal of Intelligent & Fuzzy Systems. 34. 1-10. 10.3233/JIFS-169489. [↑](#footnote-ref-27)
29. Для интерпретации оценки FRE см. статью: *Flesch, Rudolf.* "How to Write Plain English". *University of Canterbury*. Archived from the original on July 12, 2016. Retrieved July 12, 2016. [↑](#footnote-ref-28)
30. NB: для сопоставления результатов FRE и остальных формул график FRE следует зеркально отразить по оси х. [↑](#footnote-ref-29)