**Задачи проекта**

***Электронное, интерактивное, нелинейное, многовариантное издание стихов А.А. Вознесенского***

Заявленное предполагает:

1. Собрание в рамках единого корпуса стихов Вознесенского разных периодов его творчества (***ранний***: *н. 1950-х – н. 1970-х*; ***поздний***: *н. 1970-х – сер. 1980-х.*)
2. Создание рекомендательной системы, которая позволит читателю выбрать траекторию знакомства с поэзией Вознесенского в соответствии с одним из следующих запросов:

* ***Поиск похожих стихотворений*** (схожесть с точки зрения различных предлагаемых на выбор читателю ***критериев***: *близость тематики*; *структурная сложность*; *удобочитаемость*; м.б. должна быть возможность поиска схожих *по совокупности всех признаков*)
* ***Поиск непохожих стихотворений*** (м.б. в соответствии с теми же характеристиками)
* ***Поиск наиболее / наименее известных стихотворений*** (можно представить с помощью тех же инструментов, что и предыдущие два пункта – но м.б. есть смысл вынести как отдельный пункт)
* ***Случайный поиск*** (м.б. случайный, но в пределах отдельных кластеров)

1. Таким образом, предполагаемый конечный продукт проекта можно метафорически представить в виде нелинейной ветвящейся многоуровневой структуры, каждый элемент которой (отдельный стих) связан с каждым другим. В рамках одного уровня этой структуры расстояние между элементами определяется исходя из выбранного читателем критерия.

Подробнее об этой структуре расскажет моя коллега.

**Сад расходящихся тропок**

… короткая метафора на основе рассказа

И чтобы в какой-то мере достичь этой красивой цели, следует произвести вполне конкретные действия (следующий слайд)

**Научные задачи**

Решение основной цели проекта: электронное интерактивное, нелинейное, многовариантное издание стихов А.А. Вознесенского – предполагает постановку следующих научных задач по аналитической обработке материала проекта:

1. Во-первых, необходимо определиться с пространством *переменных*, по которым будет проводиться количественное сравнение стихов Вознесенского (*признаковое пространство*)

* При этом кажется, что переменные могут быть как количественные (например, метрическая / лексическая сложность), так и качественные (напр., тематика)
* В дальнейшем, исходя из наблюдаемой корреляции между отдельным переменными, можно будет выделить более общие скрытые признаки – т.н. факторы, и свести все переменные к ним (*факторный анализ*, если я вообще правильно понимаю, что это)

1. Во-вторых, необходимо каждое стихотворение измерить (выразить в числовом виде) по единой шкале для каждой из выделенных переменных:

* Для этого могут быть использованы различные метрики: напр., ***индекс разнообразия Шеннона*** для измерения метрической сложности, ***метрики TTR*** и ***CTTR*** – для лексической; тематика может быть выражена списком из нолей и единиц (см. коммент 2)
* Таким образом, каждый стих редуцируется до некоторого набора из чисел, характеризующих его с разных сторон

1. Далее, чтобы определить близость / дальность стихов друг к другу по определенному, выбранному читателем критерию, каждый из которых для отдельного стиха, скорее всего, окажется представлен в виде **набора** чисел (напр., читательский критерий структурной сложности может складываться из измерений метрической и лексической сложностей), необходимо измерить расстояние между соответствующими ***наборами*** чисел – представить данные в виде *матрицы расстояний* и *матрицы сходств*:

* Метрики для измерения расстояния могут быть разные, соответственно, надо изучить этот вопрос, чтобы определиться с выбором метрики

1. Наконец, к полученным значениям можно применить метод ***кластеризации*** для объединения схожих стихов в сравнительно однородные группы:

* Это предполагает применение к данным одного из возможных методов статистического анализа, что также требует отдельного предварительного изучения

1. Итак, полученные в результате анализа и сравнения стихотворений Вознесенского матрицы расстояний и система кластеров необходимы для реализации каждого из охарактеризованных выше запросов читателя:

* ***Поиск похожих стихов*** реализуется на основании поиска кратчайших расстояний, не выходя при этом за пределы соответствующего кластера
* ***Поиск непохожих стихов*** – через предложение произведений из других кластеров
* ***Поиск по известности стихотворений*** можно осуществить с помощью кластеризации стихов по признаку известности (соответственно, степень известности можно представить в числовом эквиваленте и на этом основании измерить расстояние между стихами)
* ***Случайный поиск*** можно реализовать с помощью алгоритма, выбирающего случайные числа из матрицы расстояний для совокупности измеряемых признаков стихотворений; при этом можно настроить алгоритм так, чтобы, к примеру, каждый раз он обращался к иному кластеру в целях увеличения разнообразия выдаваемых читателю результатов

1. При этом остается еще ряд задач, не связанных напрямую с анализом самого материала проекта (стихов Вознесенского):

* Речь идет, например, о выборе предлагаемых читателю критериев для сравнения стихотворений (тематическая, структурная схожесть, удобочитаемость и т.д.), формируемых исходя из предполагаемых запросов читателя, которые сами по себе нуждаются в изучении
* Отдельную трудность представляет и формулировка функций, доступных пользователю финального продукта для выбора траектории знакомства с поэзией Вознесенского, которые также должны соответствовать его (пользователя) потребностям как читателя
* Нетривиальной задачей является и сбор самих стихов Вознесенского (а также их оцифровка и распознание, где необходимо) в единый корпус, подходящий для дальнейшего издания

Для более же детальной характеристики материалов нашего проекта передаю слово моей коллеге.

**Данные**

Собрать материалы для датасета/корпуса работ автора, которого считают классиком, казалось бы, достаточно простая задача. НО - *по-настоящему* полного собрания творчества Вознесенского не существует.

1. Собрание сочинений в 3 томах. — М.: Художественная литература, 1983—1984, 75 000 экз.
2. Собрание сочинений в 7 томах (8 книгах). — М.: Вагриус, 2000—2009
3. Полное собрание стихотворений и поэм в одном томе. — М.: Издательство АЛЬФА-КНИГА, 2012. — [ISBN 978-5-9922-1110-8](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D1%83%D0%B6%D0%B5%D0%B1%D0%BD%D0%B0%D1%8F:%D0%98%D1%81%D1%82%D0%BE%D1%87%D0%BD%D0%B8%D0%BA%D0%B8_%D0%BA%D0%BD%D0%B8%D0%B3/9785992211108). (2-е издание — 2020 г.)
4. Стихотворения и поэмы. В 2 томах. — Вступительная статья, составление, подготовка текста и примечания Г. И. Трубникова — СПб.: Издательство Пушкинского Дома; Вита Нова, 2015. — (серия «Новая Библиотека поэта»). — [ISBN 978-5-87781-038-9](https://ru.wikipedia.org/wiki/%D0%A1%D0%BB%D1%83%D0%B6%D0%B5%D0%B1%D0%BD%D0%B0%D1%8F:%D0%98%D1%81%D1%82%D0%BE%D1%87%D0%BD%D0%B8%D0%BA%D0%B8_%D0%BA%D0%BD%D0%B8%D0%B3/9785877810389). [Это первое академическое издание произведений поэта.]

Академическое издание содержит стихотворения и поэмы до 1985-го года. После период визуальной поэзии, перформансов. Видеомы.

Проблемы: разрозненность материалов, нет полного собрания сочинений в одной месте, в собраниях сочинений не содержится графика, некоторые стихотворения существуют в нескольких вариантах.

Основная проблема со сбором графики: возможно, какая-то часть графических работ содержится в архиве параллельной команды.

**Подготовка данных**

1. Формирование критериев отбора материалов (не только тексты, но и изображения).

Важный этап создания корпуса/датасета - отбор репрезентативных материалов. По каким критериям будет произведен этот отбор (скорее всего автоматический) - предстоит решить. Кроме этого следует очертить структуру данных. Пока понятно, что там будут данные и текстовые, и визуальные; и их обработка и кластеризация должна вестись схожим образом, т.е. автоматическими средствами придется доставать семантику графических работ. (ну чето такое)

1. Сбор материалов.
2. Подбор инструментов и технологий для разметки.
3. Разметка данных:

* акцентуация -> метр;
* схема рифмы;
* строфика;
* семантическая и синтаксическая разметка;
* и вполне возможно что-то ещё.

1. Обработка графики.

Отдельная важная и непрезентированная в изданиях часть - визуальные работы Вознесенского. Стоит подумать о том, как произвести разметку графических данных аналогично разметке текстовых, чтобы потом обработать их как единый корпус.

**Подготовка данных (первый этап): к проблемам и примерам**

Уже на начальном этапе подготовки данных возникает несколько вопросов-проблем, связанных с отбором материала. Они в первую очередь касаются, как уже упомянула моя коллега, разнообразия творчества Вознесенского (и тексты, и графика), а также наличия нескольких вариантов одного и того же текста. Два этих аспекта создают разные сложности в процессе формирования корпуса.

**Визуальная поэзия Вознесенского**

Как классифицировать тексты, подобные представленным на экране? Как соотносить их с “чистыми” текстами? Можно ли впоследствии разделять их по тем же категориям, что и остальные стихотворения (сложность, схожесть (и в чём их схожесть, кроме “жанра”, – в игре слов?)?

**А если и текст, и графика одновременно?..**

Как быть (как размечать и анализировать), если стих сочетает в себе две стороны творчества Вознесенского? → ещё острее вопрос об одинаковой обработке любых текстов Вознесенского (если тексты и графику не объединять в единый корпус, то что делать с подобными текстами? разделять нецелесообразно, это единое целое)

**Видеомы**

Нужно ли включать видеомы в корпус?

Под видеомами мы понимаем те стихотворения, в которых визуальная часть преобладает над текстовой. Если же текстовая часть преобладает (или по крайней мере имеет достаточно большой вес), то мы считаем (по велению Вознесенского) стих кругомётом. Взять кругомёты, не брать видеомы? (но видеомы – тоже стихи → Осы Осипа)

Как размечать видеомы и как их анализировать, если они войдут в корпус? Как их категоризировать?

**Несколько вариантов одного текста**

Текстологическая проблема: часто встречаются варианты и разночтения. На что мы должны ориентироваться, когда выбираем тот или иной вариант?

Важно заметить, что это не просто попытка угодить читательскому вкусу; разночтения очень легко меняют формальное устройство стиха, что-то ломается, например, стих становится ещё сложнее метрически.

Кроме того, встречаются варианты стихотворений, в которые добавляется целая строфа (или даже несколько) → меняется метрический строй, добавляется лексика, добавляется новый слой к тематике.

**Предварительные задачи**

Из описанного моими коллегами и мной вытекают следующие предварительные задачи:

В первую очередь нам нужно приступить непосредственно к созданию корпуса текстов Вознесенского. Для этого нам предстоит понять масштаб проблемы с разрозненностью текстов и их реальным количеством. Так мы сможем определить источники, из которых мы будем добывать “тексты”, объём нашего издания, а также критерии отбора материала (исходя из необходимого нам количества текстов и из возможности получить те или иные тексты для включения их в проект).

Конечно, придётся лучше познакомиться с цифровыми методами анализа поэзии. Во-первых, это то с чем мы будем постоянно работать, и хотя коллеги рассказали о методах работы с текстом, которые могут нам пригодится, нам ещё предстоит изучить их и всё попробовать, чтобы понять, насколько те или иные методы действительно нам подходят.

Во-вторых, это также поможет нам сузить выборку текстов (если мы поймём, что не сможем проанализировать какие-то из них по тем или иным причинам).

Отсюда вытекает 4 пункт.

Когда у нас сформируются представления о корпусе текстов Вознесенского и о методах, с которыми мы будем работать, мы сможем определить зоны ответственности каждого и распределить роли в проекте.

**Планы на ближайшее время**

Как я уже сказала, главное для нас сейчас – начать сбор корпуса текстов Вознесенского, чтобы получить подробное представление о материале, с которым мы будем работать.

Кроме того, мы бы хотели найти и изучить похожие проекты. Интересно посмотреть, как такие проекты реализуются, что уже сделано и успешно существует и что нового мы можем предложить.

Научный бэкграунд проекта → работы по Вознесенскому и по цифровому стиховедению (?)

Изучив всё упомянутое ранее, мы наконец-то сможем прийти к какой-то более конкретной идее и понять, в каком виде мы хотим представить итог нашего проекта. Эта более конкретная идея поможет нам сформулировать ТЗ и распределить/перераспределить роли.