**МИНОБРНАУКИ РОССИИ**

Федеральное государственное бюджетное образовательное учреждение

высшего образования

**«РОССИЙСКИЙ ГОСУДАРСТВЕННЫЙ ГУМАНИТАРНЫЙ УНИВЕРСИТЕТ»**

**(РГГУ)**

ИНСТИТУТ ЛИНГВИСТИКИ

Учебно-научный центр

компьютерной лингвистики

**Балуева Дарья Владимировна**

Современный поэт и поэты серебряного века: количественное сравнение текстов

Курсовая работа обучающейся 3-го курса очной формы обучения

Направление подготовки 45.03.03 Фундаментальная и прикладная лингвистика

Научный руководитель канд. филологических наук А.Ч.Пиперски

2019

Оглавление

[Введение 3](#_Toc10156289)

[Глава 1. Сравнение корпусов 5](#_Toc10156290)

[Объект и материалы исследования 5](#_Toc10156291)

[Вычисление расстояний между корпусами 8](#_Toc10156292)

[Использованные меры 10](#_Toc10156293)

[Геометрические меры 10](#_Toc10156294)

[Статистические меры 11](#_Toc10156295)

[Сходство по ключевым словам 12](#_Toc10156296)

[Глава 2. Результаты компьютерных экспериментов 13](#_Toc10156297)

[Сравнение по геометрическим мерам 13](#_Toc10156298)

[Сравнение по мере χ2 16](#_Toc10156299)

[Ключевые слова: подбор параметров и интерпретация 17](#_Toc10156300)

[Итоги сравнения 22](#_Toc10156301)

[Заключение 25](#_Toc10156302)

[Библиография 28](#_Toc10156303)

[Приложение 30](#_Toc10156304)

# **Введение**

Сложно представить себе писателя или поэта, который никогда не читал произведения других. Авторы знакомятся с творчеством современников или предшественников, и это может отражаться в их собственных текстах. Изучение влияния авторов друг на друга помогает устанавливать и объяснять связи в литературе, а сравнение текстов само по себе важно для решения практических задач: определения плагиата, информационного поиска, автоматической классификации текстов и многих других.

Литературоведы проводят качественный анализ текстов вручную, опираясь на свои знания и наблюдения. Однако даже самому увлеченному человеку трудно прочитать, например, восемьдесят тысяч стихотворений, а затем сопоставить их и сделать содержательные выводы. Именно столько текстов включает в себя Поэтический подкорпус Национального корпуса русского языка, и сегодня он доступен любому исследователю. Электронные собрания текстов и работа с ними были бы невозможны без компьютеров и программирования, в 1960-х годах для изучения литературы начинают применяться и компьютерные технологии. [Schreibman et al. 2004] Точные методы, позволяющие находить интересные статистические закономерности в текстах, начали применяться еще раньше - в начале XX века: например, в работах Б.И.Ярхо [Ярхо 1924; Акимова и др. 2006], а затем и М.Л.Гаспарова [Гаспаров 1974].

На сегодняшний день цифровые исследования литературы (Literary Studies) развиваются как часть Цифровых гуманитарных наук, или Digital Humanities [Schreibman et al. 2004]. Они развиваются и привлекают все большее внимание: весной 2019 года вышла статья, в которой автор критикует цифровой подход к изучению литературы [Da 2019]. Эта статья вызвала широкий отклик, и в ходе спора были высказаны интересные мнения о значимости и достоверности подобных исследований[[1]](#footnote-1). С одной стороны, ученые согласны с тем, что полученные результаты, а чаще всего это числа, всегда требуют интерпретации, и очень осторожной. С другой стороны, цифровые литературоведы доказывали, что количественные методы не сводятся к простым подсчетам и могут дать совершенно новый взгляд на исследуемый материал. Данная курсовая работа тоже посвящена количественному сравнению текстов, а ее основной объект – стихотворения современного поэта.

Современный поэт Михаил Калинин привлек мое внимание тем, что он часто упоминает в интервью поэзию и поэтов серебряного века [Калинин 2016: 1-4:, 2017], а Игоря Северянина называет своим любимым автором. По его словам, именно стихи Северянина еще в школьные годы оказали на Калинина наибольшее впечатление [Калинин 2013, 2017: 57], и он до сих пор их перечитывает [Калинин 2018].

Цель нашего исследования – при помощи автоматических количественных методов сравнить тексты, написанные современным поэтом и тексты поэтов серебряного века, выявить, чем они похожи и чем отличаются. Наше предположение заключается в том, что увлечение поэзией Игоря Северянина отразилось на стиле Михаила Калинина, и их стихотворения будут похожи.

Чтобы проверить эту гипотезу, нужно, во-первых, собрать корпус сочинений Михаила Калинина, который будет состоять из стихотворений и текстов песен. Чтобы определить степень сходства Калинина и Северянина, мы сравним тексты Калинина с текстами еще восьми поэтов. Используя средства программирования, мы найдем расстояния между корпусами разных авторов. Сопоставлять корпуса мы будем на разных языковых уровнях, а расстояния вычислять по нескольким мерам. В конце мы сравним результаты, полученные разными способами. Мы проверим, верно ли, что больше всего тексты Михаила Калинина окажутся похожи именно на Северянина, то есть расстояние между их корпусами будет наименьшим. В таком случае мы будем считать, что творчество Северянина повлияло на Михаила Калинина, и Калинин действительно пишет похоже именно на этого автора.

# **Глава 1. Сравнение корпусов**

## **Объект и материалы исследования**

Расскажем об авторе, тексты которого были взяты как основной материал работы. Михаил Николаевич Калинин (псевдоним – Эм Калинин) родился 6 января 1988 года в городе Череповец Вологодской области. По образованию Михаил журналист [Калинин 2011], но посвятил свою жизнь не этой профессии, а поэзии и музыке. Калинин - создатель и участник шести музыкальных проектов [Калинин 2015], из которых в настоящее время действуют два, в том числе группа Аффинаж[[2]](#footnote-2), набирающая популярность на сцене российской альтернативной музыки. Стихи Калинин пишет с детства, два раза были изданы сборники его лирики [Калинин 2008, 2017], а недавно был издан сборник текстов песен группы Аффинаж [Калинин 2019]. В интервью Михаил Калинин часто рассказывает о своих взглядах на творчество и упоминает других поэтов и музыкантов. Наше внимание привлекло то, что, говоря о любимой поэзии, он неоднократно выделяет классику, а именно серебряный век [Калинин 2016:1-4, 2017]. С одной стороны, Калинин много раз называл любимым поэтом именно Игоря Северянина и упоминал других, говорил о своем следовании поэтическим традициям [Калинин 2016: 1-4]. Эта связь с классической поэзией стала его своеобразной визитной карточкой: «тексты в традиции русской поэзии Серебряного века»[[3]](#footnote-3), «отчетливые северянинские нотки в творчестве» [Черный и др. 2008], «баланс между русской поэтической традицией (Северянин, Сологуб, Есенин) и насыщенной, но легкой для восприятия музыкой»[[4]](#footnote-4) - эти и другие слова стали появляться в статьях журналистов и анонсах концертов. А с другой стороны, сам поэт говорил, что стремится выйти за рамки классических форм, подчеркивал свои отличия [Калинин 2016: 1-4], категорично говорил, что некоторые авторы ему не близки [Калинин 2017: 59]. Можно ли утверждать, что автор двадцать первого века пишет похоже на поэтов века двадцатого? Мы решили это проверить и провести исследование, а первым его этапом был сбор корпуса текстов Михаила Калинина.

Все использованные тексты находятся в открытом доступе в интернете: публикуются автором на его официальной странице Вконтакте[[5]](#footnote-5) или на официальных страницах музыкальных групп[[6]](#footnote-6). В корпус вошли стихотворения (144 текста), две поэмы, тексты песен групп «Аффинаж» (73 текста), «Эм(а)СПИД» (70 текстов) и «никого не люблю» (116 текстов), кроме того - четыре перевода песен Мэрлина Мэнсона. В настоящий момент Калинин пишет новые песни только для группы Аффинаж, а также продолжает писать и публиковать новые стихотворения. Даже с момента завершения сбора корпуса (февраль 2019 года) было опубликовано больше десяти новых текстов. Итак, в корпус объемом почти 59000 слов вошло 406 текстов с 2000 по 2019 год. Корпус был собран отчасти вручную, отчасти тексты были скачаны и обработаны автоматически.

Тексты были собраны в корпус и подготовлены для автоматической обработки: корпус существует в текстовом формате (.txt), а так же в табличном (.csv). Была вручную добавлена метаразметка текстов: название, год создания, тип текста (стихотворение, поэма, текст песни, перевод). Тексты обрабатывались с помощью средств программирования: библиотек Python (pymorphy2, xlrd/xlwt, pandas, numpy, scipy) и R (quatenda). Так же в ходе работы использовался корпусный менеджер Sketch Engine [Kilgariff et al. 2014], где был создан корпус Михаила Калинина с морфологической разметкой.

Так же в Sketch Engine мы создали корпуса и восьми других поэтов, с которыми сравнивали Калинина: Игоря Северянина, Николая Гумилева, Анны Ахматовой, Александра Блока, Владимира Маяковского, Сергея Есенина, Федора Сологуба и Марины Цветаевой. Все их тексты были взяты из поэтического подкорпуса НКРЯ[[7]](#footnote-7). Почти все эти поэты были выбраны, потому что Калинин так или иначе упоминал их. Как мы уже сказали, Игоря Северянина он называет любимым поэтом и до сих пор перечитывает его стихи. Сологуба и Есенина он тоже упоминал, и потом о группе Аффинаж писали так: «среди своих ориентиров в лирике они выделяют известнейших поэтов «Серебряного века»: символиста Сологуба, футуриста-авангардиста Северянина и, конечно же, Есенина – причем эти влияния легко прослеживаются в произведениях группы»[[8]](#footnote-8). Мы постараемся количественно проверить, насколько «легко» прослеживаются эти влияния. Маяковского Михаил тоже называл автором, чье творчество ему близко. [Калинин 2015: ] Что касается Цветаевой и Ахматовой, этих поэтов он упоминал в контексте того, что он «совсем не принимает». [Калинин 2017: 59] Калинин говорил что творчество Ахматовой и Цветаевой ему не близко, «как и вся женская поэзия». Двух оставшихся авторов, Гумилева и Блока, Калинин не упоминал вообще, мы выбрали их как еще одних ярких представителей серебряного века русской поэзии. Таким образом, мы сравним корпус текстов Михаила Калинина с корпусами восьми авторов. Мы предполагаем, что корпус Калинина окажется больше похож на корпуса тех, чье творчество ему ближе. Наша основная гипотеза – Калинин пишет похоже именно на своего любимого поэта Игоря Северянина. В следующем разделе будут рассмотрены методы и параметры сравнения.

## **Вычисление расстояний между корпусами**

Расстояние между корпусами – это мера, позволяющая оценить, насколько похожи корпуса текстов между собой, наша курсовая работа посвящена нахождению этих расстояний. Теоретические основы подобных исследований были заложены еще в работах [Kilgariff 1997; Rayson et al. 2000]. Количественное сравнение текстов основывается их на частотных характеристиках, то есть мы находим расстояние не между текстами, а между частотными списками языковых единиц из них. Чаще всего сравниваются частотные списки слов – это могут быть списки лемм или лексем, а так же n-грамм слов: пар, троек и т.д. Корпуса могут сравниваться и на других языковых уровнях: например, на уровне буквенных сочетаний (частотные списки буквенных n-грамм), на уровне морфологии (списки частеречных тегов слов) и даже пунктуации.

В данной работе мы выбрали три уровня сравнения: леммы слов, скипграммы[[9]](#footnote-9) и символьные биграммы[[10]](#footnote-10). Списки лемм мы получили двумя способами: используя лемматизатор pymorphy2 [Korobov 2013]) и самого Sketch Engine. Использование тех или других списков на результат не влияло, по умолчанию мы использовали со списками Sketch Engine.

При сравнении мы учитывали некоторые свойства корпусов. Во-первых, корпуса разных авторов могут сильно различаться по размеру. Чтобы результат не зависел от размера корпусов, во всех экспериментах мы использовали не абсолютные частоты слов, а относительные. Частоты были выражены в *ipm (instance per million),* то есть они характеризуют число вхождений на миллион слов корпуса.

Во-вторых, слова могут быть распределены по корпусам неравномерно. Например, слово может встречаться десять раз в одном и том же тексте, а другое тоже десять раз, но в десяти разных текстах. Абсолютная частота у них одинаковая – 10, но первое слово скорее характеризует отдельный текст, а второе – весь корпус в целом. Чешские исследователи [Savický, Hlavácová 2002] поднимали в своей работе эту проблему и предложили несколько способов вычисления уточненных частот. Мы использовали один из них – ARF, average reduced frequency. Вычисляя эту частоту, мы делим корпус на равные части, и ARF учитывает, насколько равномерно распределены вхождения слова по этим частям. Чем ближе друг к другу соседние вхождения, тем ниже будет ARF. ARF самостоятельно, а воспользовались встроенной функцией Sketch Engine.

Возвращаясь к скипграмам и символьным биграмам, оба эти уровня используются в компьютерной лингвистике для автоматической классификации текстов и, в том числе, для классификации текстов по автору. [Stamatatos 2009; D’hondt et al 2012] Мы решили использовать их и для нашей задачи. Биграммы мы извлекали с помощью python, а для создания списков скипграм воспользовались функцией tokens\_skipgrams библиотеки quatenda языка R [Benoit et al. 2018]. Все частоты мы выражали в ipm.

Итак, мы сформировали частотные списки лемм, буквенных биграмм и скипграмм, а затем сравнивали их. Нашей задачей было определить – похожи ли тексты Михаила Калинина на тексты Игоря Северянина или тексты других авторов. Для этого мы по очереди сравнивали корпус Калинина с восемью другими корпусами, ранжировали авторов по расстоянию, и смотрели, для кого расстояние оказалось наименьшим. Перед тем как вычислять расстояния мы автоматически создавали таблицы сопряженности такого вида:

|  |  |  |
| --- | --- | --- |
|  | Калинин | Ахматова |
| иголка | 34,25244049 | 0 |
| послышаться | 0 | 35,59542243 |
| боб | 17,12622024 | 0 |
| излом | 34,25244049 | 17,79771121 |

В первом столбце был объединенный список слов двух корпусов, а напротив были записаны частоты каждого слова в обоих корпусах. Дальнейшие действия зависели от использованной меры.

### **Использованные меры**

Существуют разные меры вычисления расстояния между корпусами, в обзоре [Gomaa, Fahmy 2013] все они подразделяются на меры, основанные на символах (character-based, самая известная из них – расстояние Левинштейна) и меры, основанные на единицах более высокого уровня - словах, n-граммах и т.д. (term-based). В данной работе мы используем пять term-based мер, о которых расскажем в данном разделе.

#### **Геометрические меры**

Чтобы применить геометрическую меру для сравнения корпусов, необходимо сначала представить корпуса в виде векторов частот слов. [Salton et al 1975] Мы использовали созданные таблицы и создавали векторы такого вида: *Калинин = (34; 34; …; 0),* *Ахматова = (35; 0; …; 17)*. Затем мы находили расстояния между ними так же, как в геометрии находятся расстояния между точками многомерного пространства. Мы использовали три геометрические меры – Манхэттенское, Евклидово и Косинусное расстояния.

Во всех приведенных далее формулах *a,b* – это векторы частот слов двух корпусов, *d(a,b)* – расстояние между ними, *n* – число слов в объединенном списке,  – частоты *i*-го слова из списка в обоих корпусах.

Так, **Евклидово** **расстояние[[11]](#footnote-11)** между точками равно квадратному корню из суммы квадратов разностей их координат:

Для нас «точка» – это слово из объединенного списка, а ее координаты – частоты слова в обоих корпусах.

**Манхэттенское** **расстояние[[12]](#footnote-12)** равно сумме модулей разностей соответвующих частот:

Чтобы найти **Косинусное расстояние[[13]](#footnote-13)** между векторами, нужно сначала найти косинус угла между векторами, это значение – степень сходства корпусов: ( – скалярное произведение векторов; – длины векторов)

Автоматически эти расстояния можно вычислить, используя функции библиотеки scipy [Jones et al 2001] Python - euclidean(a,b), cityblock[[14]](#footnote-14) (a, b), cosine(a,b).

#### **Статистические меры**

Статистические меры расстояния основаны на соответствующих статистических тестах – χ2 и расстояние Спирмена. Килгарифф доказывал эффективность этих мер для сравнения корпусов [Kilgariff 2001].

Мы использовали только χ2. Именно эта мера оказалась самой эффективной, когда Килгарифф сравнивал разные меры расстояния. Из ранее созданных таблиц сопряженности мы делали таблицы такого вида:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Корпус A | Корпус B | Суммы строк |
| Слово |  |  | + |
| Слово |  |  | + |
| … | … | … | … |
| Слово |  |  | + |
| Суммы столбцов |  |  |  |

Затем мы вычисляли хи-квадрат по обычной формуле, где i,j – номера всех строк и столбцов таблицы, – наблюдаемые значения частот, – предполагаемые значения[[15]](#footnote-15):

#### **Сходство по ключевым словам**

Еще один способ сравнения, который мы будем использовать в данной работе – сравнение корпусов по ключевым словам. Этот метод был введен Адамом Килгариффом [Kilgarriff 2009] и используется в Sketch Engine. Под ключевыми словами подразумеваются слова, характерные для одного корпуса по сравнению с другим. Вычислить, насколько слово характерно для корпуса можно по следующей формуле:

A – фокусный[[16]](#footnote-16) (focus) корпус, B – референсный[[17]](#footnote-17) (reference) корпус, – частоты слова w в корпусах A,B. Если слово w из фокусного корпуса отсутствует в референсном, = 0, а на ноль делить нельзя, поэтому n никогда не равно нулю. Разные параметры дают разное ранжирование списка ключевых слов. Расстояние между сравниваемыми корпусами будет равно среднему арифметическому полученных значений *k*.

Списки ключевых слов с вычисленным *k* мы получим с помощью Sketch Engine. Мы будем использовать разные параметры n, выразим частоты в *ipm*, ARF, а затем сравним результаты, полученные разными способами. Наконец мы рассмотрим, какие именно слова выделяются как ключевые для корпуса Калинина, а какие как нет, и постараемся проинтерпретировать результаты.

В следующей главе мы расскажем о результатах нашего исследования.

# **Глава 2. Результаты компьютерных экспериментов**

Мы провели серию компьютерных экспериментов: сравнили корпус текстов Михаила Калинина с корпусами восьми поэтов серебряного века. Эксперименты проводились на трех языковых уровнях и с помощью пяти мер расстояния. Подбирая параметры и уточняя частоты, всего мы проделали 24 опыта. Результатом работы программ был список расстояний от корпуса Калинина до каждого из восьми корпусов. Затем мы сортировали списки по возрастанию и присваивали каждому поэту ранг от 1 до 8: например, если расстояние между текстами Калинина и Ахматовой было наименьшим из всех, ей присваивался ранг 1. В конце мы суммировали ранги каждого поэта по всем экспериментам. Если сумма рангов какого-то поэта была меньше, чем у других, мы считали, что тексты Калинина больше всего похожи именно на тексты этого поэта.

В данной главе мы сначала отдельно расскажем о расстояниях, вычисленным с помощью разных мер, а затем постараемся обобщить и проинтерпретировать результаты.

## **Сравнение по геометрическим мерам**

На первом этапе мы вычисляли Манхэттенское, Евклидово и Косинусное расстояния, используя частотные списки, сгенерированные с помощью Sketch Engine, python и R. Все использованные частоты относительные, а именно - число вхождений на миллион. Sketch Engine так же позволил нам получить списки лемм с частотами ARF. В этой серии экспериментов применение ARF на результат не повлияло. Мы предполагали, что больше всего тексты Калинина окажутся похожи на тексты тех поэтов, кого он сам называл любимыми или близкими по творчеству авторами. В первую очередь это Северянин, а также Есенин, Маяковский и Сологуб. Соответственно, мы считали, что расстояние до текстов Цветаевой, Ахматовой, Гумилева и Блока будет больше. Вычисление расстояний между корпусами не всегда давало результаты, соответствующие ожиданиям.

Рассмотрим сначала, с кем из авторов у Калинина оказалось наибольшее сходство по лексике, то есть по леммам и скипграммам. Для этого выделим разными цветами тех авторов, которых мы ожидали видеть вверху таблицы, и посмотрим, как меняются их позиции:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Евклид | | | Косинус | | | |
| Леммы | Леммы, ARF | Скипграммы | Леммы | Леммы, ARF | | Скипграммы |
| Есенин | Есенин | Цветаева | Есенин | | Есенин | Цветаева |
| Ахматова | Ахматова | Ахматова | Ахматова | | Ахматова | Ахматова |
| Сологуб | Сологуб | Блок | Сологуб | | Сологуб | Есенин |
| Северянин | Северянин | Северянин | Северянин | | Северянин | Блок |
| Блок | Цветаева | Гумилев | Цветаева | | Цветаева | Северянин |
| Цветаева | Блок | Есенин | Блок | | Блок | Маяковский |
| Гумилев | Гумилев | Маяковский | Гумилев | | Гумилев | Гумилев |
| Маяковский | Маяковский | Сологуб | Маяковский | | Маяковский | Сологуб |

Таблица 1: Ранжирование авторов по степени сходства их текстов с текстами Калинина, Евклидово и Косинусное расстояния

Когда сравнивались частотные списки лемм, результаты по двум мерам оказались практически идентичными. Интересно, что из тех, кого мы предполагали видеть вверху списков, там оказался только Есенин. Например, Маяковский - наоборот оказался очень низко. Сологуб и Северянин занимали средние позиции. Когда мы перешли к скипграммам, результат изменился и стал еще более неожиданным: Цветаева стала лидировать, а Сологуб оказался ниже всех. В то же время позиция Ахматовой тоже всегда была высокой, и, к тому же, самой стабильной.

Применение Манхэттенской меры дало немного другой результат:

|  |  |  |  |
| --- | --- | --- | --- |
| Ранг | Леммы | Леммы, ARF | Скипграммы |
| 1 | Ахматова | Ахматова | Ахматова |
| 2 | Есенин | Есенин | Есенин |
| 3 | Цветаева | Цветаева | Цветаева |
| 4 | Маяковский | Маяковский | Северянин |
| 5 | Северянин | Северянин | Блок |
| 6 | Гумилев | Гумилев | Гумилев |
| 7 | Блок | Блок | Маяковский |
| 8 | Сологуб | Сологуб | Сологуб |

Таблица 2: Ранжирование авторов по степени сходства с текстами Калинина, Манхэттенское расстояние

Переход с лемм на скипграммы немного изменил ранжирование, однако лидеры всегда оставались неизменными. Цветаева и Ахматова по-прежнему оставались там, где мы не ожидали их увидеть – вверху списка.

И наконец, перейдем к буквенным биграммам:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Ранг | Манхэттен | Евклид | Косинус | Сумма рангов |  |
| 1 | Блок | Ахматова | Есенин | Есенин | 7 |
| 2 | Маяковский | Есенин | Северянин | Блок | 12 |
| 3 | Гумилев | Северянин | Цветаева | Северянин | 12 |
| 4 | Есенин | Сологуб | Ахматова | Ахматова | 13 |
| 5 | Сологуб | Блок | Маяковский | Маяковский | 15 |
| 6 | Цветаева | Гумилев | Блок | Гумилев | 16 |
| 7 | Северянин | Цветаева | Гумилев | Цветаева | 16 |
| 8 | Ахматова | Маяковский | Сологуб | Сологуб | 17 |

Таблица 3: Ранжирование авторов по степени сходства с текстами Калинина на уровне буквенных биграмм

Разные меры давали непохожие результаты и внесли разный вклад в общую сумму рангов. Оказывается, что сравнение по биграммам показало наибольшее сходство текстов Калинина и Есенина, эти результаты похожи на то, что мы видели в предыдущих опытах.

Подведем итоги всех сравнений по геометрическим мерам:

1. Ахматова - 15
2. Есенин - 19
3. Цветаева - 32
4. Северянин - 39
5. Блок - 49
6. Сологуб - 52
7. Гумилев - 58
8. Маяковский – 60

По результатам всех опытов чаще всего тексты Михаила Калинина оказывались больше похожи на тексты Анны Ахматовой или Сергея Есенина. Результаты сравнений с другими авторами показывают гораздо меньшее сходство – в том числе и с текстами Игоря Северянина. Уже на данном этапе можно сказать, что наша гипотеза о наибольшем сходстве Калинина и Северянина не подтверждается. Более того, интуитивное предположение о том, что тексты Калинина окажутся не похожи на тексты поэтов, которые ему «не близки», тоже не оправдалось: как и Ахматова, Цветаева тоже оказывается вверху «рейтинга» по результатам сравнений.

В следующем разделе перейдем вычислению расстояния по методу хи-квадрат.

## **Сравнение по мере χ2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Ранг | Леммы | Леммы, ARF | Скипграммы | Биграммы | Сумма |  |
| 1 | Цветаева | Цветаева | Цветаева | Цветаева | Цветаева | 4 |
| 2 | Маяковский | Маяковский | Маяковский | Северянин | Маяковский | 11 |
| 3 | Блок | Блок | Блок | Есенин | Блок | 15 |
| 4 | Северянин | Северянин | Гумилев | Ахматова | Северянин | 16 |
| 5 | Ахматова | Есенин | Есенин | Маяковский | Есенин | 19 |
| 6 | Есенин | Ахматова | Северянин | Блок | Ахматова | 22 |
| 7 | Гумилев | Гумилев | Ахматова | Гумилев | Гумилев | 25 |
| 8 | Сологуб | Сологуб | Сологуб | Сологуб | Сологуб | 32 |

Таблица 4: Ранжирование авторов по степени сходства с текстами Калинина, мера **χ2**

Результаты сравнения по леммам и скипграммам получились практически одинаковыми, переход на биграммы, как было и ранее, немного изменил распределение мест. Отличие от предыдущих экспериментов в том, что Ахматова и Есенин уже не оказываются так высоко. Зато наверху остается Цветаева: сравнение по мере хи-квадрат показывает наибольшее сходство текстов Калинина именно с ее текстами. Так же сильно меняет позиции Маяковский: в результате сравнения по геометрическим мерам он от Калинина был дальше всего. Северянин по прежнему не оказывается выше всех.

Обобщим результаты, полученные по всем четырем мерам:

1. Есенин - 45
2. Ахматова - 50
3. Цветаева - 52
4. Северянин - 67
5. Блок - 76
6. Маяковский - 81
7. Гумилев - 99
8. Сологуб - 101

На данном этапе можно сказать, что наши предположения о том, на чьи тексты больше всего окажутся похожи тексты Калинина, подтвердились только в отношении Есенина. Сходство с текстами Ахматовой и Цветаевой оказалось скорее неожиданным, как и то, что все меры показывали очень низкое сходство текстов Калинина и Сологуба.

В следующем разделе мы сравним корпуса поэтов на основе выделенных ключевых слов и обратимся к самим ключевым словам: мы посмотрим на общие слова Калинина и поэтов, с кем его сходство окажется наибольшим.

## **Ключевые слова: подбор параметров и интерпретация**

Перейдем к выделению ключевых слов и сравнению их списков. Слова, которые употребляются в двух корпусах наиболее различно, будут иметь более высокий коэффициент «ключевости», расстояние между корпусами – среднее арифметическое коэффициентов всех слов. Таким образом, более похожими оказываются корпуса, слова в которых употребляются наиболее похоже. С помощью Sketch Engine мы получили 8 списков ключевых слов каждого автора, меняя параметры (1, 10, 100, 1000) и нормализуя частоты (ipm и ARF/mill[[18]](#footnote-18)).

Более низкие параметры (1,10) позволяют выделить более редкие слова: это слова, которые встречаются в фокусном корпусе и вообще не встречаются в референтном или встречаются в референтном корпусе очень редко. [Kilgariff 2009] Такие слова особенно заметны, если сравнить корпус Калинина с объединенным корпусом всех восьми поэтов. За те почти сто лет, что разделяют современного поэта и поэтов серебряного века, русский язык изменился. Это закономерно, и так же закономерно меняется поэзия: отсюда слова, которые встречаются у Калинина и вообще не встречаются у его предшественников. На первый взгляд такие ключевые слова можно разделить на несколько групп. В основном, это существительные, и во-первых, это слова, обозначающие современные предметы или явления[[19]](#footnote-19) (*фотка, спортзал, ниндзя, рэп, помада, телевизор, монитор, клип, сэлфи* и т.д.). Во-вторых, это нецензурная лексика. Подобные слова встречаются даже в стихотворениях некоторых авторов серебряного века, но в среднем современный автор употреблял их чаще и разнообразнее – поэтому они и выделились как ключевые. Так же при маленьких n как ключевые выделяются имена собственные (*Сидней, Олег, Кэтрин, Мища, Киркоров* и т.д.) и современные просторечия и сокращения (*чо, пацан, т.к., фишка, ава*[[20]](#footnote-20) и т.д.) Посмотрим теперь на результаты сравнения корпуса Калинина с корпусами других поэтов. Заметим, что в этой серии из четырех опытов использование ARF стало влиять на результат уже при n=10. Так как более высокая ARF была у слов, которые характерны для конкретного корпуса целиком (а не для отдельного стихотворения, например), эти слова больше подходят на роль ключевых. При маленьких значениях n наибольшее сходство у Калинина оказалось с Маяковским:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ранг | n=1 | n=1,АRF | n=10 | n=10, ARF |
| 1 | Маяковский | Маяковский | Маяковский | Маяковский |
| 2 | Цветаева | Цветаева | Цветаева | Блок |
| 3 | Блок | Блок | Северянин | Гумилев |
| 4 | Северянин | Северянин | Блок | Северянин |
| 5 | Гумилев | Гумилев | Гумилев | Цветаева |
| 6 | Есенин | Есенин | Есенин | Сологуб |
| 7 | Ахматова | Ахматова | Ахматова | Ахматова |
| 8 | Сологуб | Сологуб | Сологуб | Есенин |

Таблица 5: Ранжирование авторов по степени сходства с текстами Калинина, ключевые слова (n=1,10)

Обратим теперь внимание на более высокие параметры: 100 и 1000. Они позволяли выделять более общие слова: все они так или иначе встретились в обоих корпусах, но более характерны для фокусного корпуса. В отличие от слов, которые выделялись при маленьких n, их сложнее разделить на группы. Мы снова сравнили корпус Калинина с объединенным корпусом серебряного века, уже при n=100 (ARF), и заметили много интересного. Можно считать, что эти слова характеризуют тематику, настроение стихотворений, а значит дают больше содержательной информации о творчестве поэта. Стоит еще раз уточнить, что лемматизация слов происходила в Sketch Engine, а части речи мы определяли с помощью морфологического анализатора pymorphy2.

Во-первых, в числе ста самых частотных ключевых слов есть личные, возвратные и притяжательные местоимения (*ты, твой, я, себя*), местоимение *мой* тоже было в числе ключевых, но на 229 месте. О том, что слово *ты* встречается чуть ли не в каждом стихотворении Калинина, говорят цифры: частотность местоимений я и ты – около 20 и 10 тысяч слов на миллион. Сам Калинин говорил про творчество группы Аффинаж: «Главная тема — это человек. Мир вокруг него и внутри него». [Калинин 2018] И об этом человеке автор не говорит отстраненно в третьем лице: лирический герой чаще всего либо *я*, либо *ты*. Тексты очень «личные», и, обращаясь на ты, поэт вовлекает в них и читателя. Интересно, что он так же говорил, что учится «ощущать не через бесконечное «я», а через «ты»», пока он «на этапе «я и ты», «мы» сказать не могу, не могу взять на себя такую ответственность». [Калинин 2017: 64-66] Такое использование местоимений стало отличительный чертой поэта.

Что касается прилагательных, интересно, что большинство из них имеет положительный оттенок: *тёплый, счастливый, красивый, нужный, идеальный, добрый, главный, важный*. Полный список 100 самых частотных ключевых слов в корпусе стихотворений Михаила Калинина приводится в Приложении. [Приложение, Таблица 1]

Посмотрим теперь на то, на тексты каких авторов тексты Калинина оказались наиболее похожи по ключевым словам:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Ранг | 100 | 100А | 1000 | 1000А |
| 1 | Цветаева | Ахматова | Ахматова | Ахматова |
| 2 | Ахматова | Гумилев | Гумилев | Есенин |
| 3 | Маяковский | Северянин | Есенин | Гумилев |
| 4 | Гумилев | Есенин | Блок | Северянин |
| 5 | Есенин | Маяковский | Цветаева | Блок |
| 6 | Северянин | Блок | Северянин | Сологуб |
| 7 | Блок | Цветаева | Сологуб | Цветаева |
| 8 | Сологуб | Сологуб | Маяковский | Маяковский |

Таблица 5: Ранжирование авторов по степени сходства с текстами Калинина, ключевые слова (n=100,1000)

Ранжирование менялось при разных n и при переходе на ARF. При n=100 оказалось, что ключевые слова Калинина больше всего совпадают с ключевыми словами Цветаевой, в следующих тестах всегда побеждала Ахматова. Насчет Северянина наши предположения не оправдались и в данной серии опытов: он не лидировал ни при каких параметрах. Что касается Сологуба, как и по геометрическим мерам и мере хи-квадрат, так и по ключевым словам сходство текстов Калинина с его текстами оказывается невысоким.

Мы решили внимательнее посмотреть на результаты сравнений с несколькими поэтами. Чаще всего мы сравнивали при таких параметрах, при которых Калинин оказывался больше похож именно на этих авторов. Для Цветаевой это n=100, для Есенина – n=1000(ARF), для Маяковского – n=10(ARF), а в случае Ахматовой и Северянина мы остановились на n=100(ARF). Мы снова смотрели на сто самых частотных ключевых слов, только теперь не самых разных, а наоборот – самых похожих (k=1).

С самого начала мы не ожидали сходства текстов Калинина с текстами Ахматовой и Цветаевой. В Приложении [Приложение, Таблица 2] приводим списки слов, совпавших у Калинина с ними, но сложно выделить среди них отдельные группы и проинтерпретировать сходство.

В отличие от поэтов-женщин, Есенина и Северянина Калинин называл своими любимыми поэтами. При этом он говорил, что Северянин для него более «минорный», «трогательный» и «тонкий», чем Есенин. [Калинин 2017: 57-59] Посмотрим и сравним, какие ключевые слова у Калинина совпали с Есениным, а какие с Северяниным.

На первый взгляд, то что совпадает у Северянина и Калинина и чего нет у Калинина и Есенина – слова на тему семьи, дома: *мать, дочь, дом, близкие*. Есть у них так же общие глаголы, связанные с отношениями между людьми: *расставаться, покидать, пожелать, лгать, встретиться, обмануть*. С Северяниным у Калинина совпадает больше прилагательных. С одной стороны: *чистый, уютный, любящий, смелый, воздушный, блестящий, а с другой: темный, суровый, ничтожный, едкий, острый*. У Калинина и Есенина меньше общих прилагательных, среди них можно выделить: *чужой, страшный, злой, глупый*. Полные списки совпавших слов мы приводим в Приложении. [Приложение, Таблица 3]

Наравне с Есениным более «мажорным» автором Калинин называл Маяковского [Калинин 2017:58]. Так же Михаил говорил, что Маяковский близок ему как «дионисический» автор. [Калинин 2015: 2-3, 2017: 58] Противопоставление дионисического и аполонического ввел философ Шеллинг, под первым он подразумевал бурный, разрушительный творческий порыв, а под вторым – форму и порядок.[[21]](#footnote-21) Мы посмотрели на 100 самых частотных общих ключевых слов Маяковского и Калинина при n=10 (ARF). На наш взгляд, показательнее всего общие прилагательные (*торжественный, праздничный, интересный, золотой, багровый*) и глаголы: *трясти, стараться, сделаться, сверкнуть, ринуться, рассыпать, разлиться, потухнуть, победить, мелькнуть, запеть, гоняться, вертеть*… Кажется, что стихотворения Калинина и Маяковского объединяет торжественность, яркость, динамичность.

В следующем разделе мы подведем итоги сравнения по всем мерам и по ключевым словам.

## **Итоги сравнения**

Итак, проведя все эксперименты, мы суммировали ранги всех поэтов, которые они получали по степени сходства текстов Калинина с их текстами. Ниже приведены суммарные результаты по всем 24 экспериментам, а также подсчет того, сколько раз тот или иной поэт побеждал в отдельных сравнениях:

|  |  |  |  |
| --- | --- | --- | --- |
| Сумма рангов во всех экспериментах | | Количество побед | |
| Ахматова | 83 | Ахматова | 7 |
| Цветаева | 83 | Цветаева | 7 |
| Есенин | 85 | Есенин | 5 |
| Северянин | 101 | Маяковский | 4 |
| Блок | 110 | Блок | 1 |
| Маяковский | 114 | Гумилев | 0 |
| Гумилев | 128 | Северянин | 0 |
| Сологуб | 160 | Сологуб | 0 |

Таблица 6: Результаты всех экспериментов – суммы рангов и число «побед» для каждого автора

Ахматова и Цветаева, разделившие первое место, напомнили нам о спорах, которые ведутся вокруг цифровых гуманитарных исследований: результаты получились неожиданными, но плохо интерпретируемыми. Больше всего тексты Калинина оказались похожи именно на тексты тех, кого мы предполагали видеть внизу списка: Калинин говорил, что «женская поэзия ему не близка». Делать выводы о подлинном сходстве текстов Калинина с текстами других авторов необходимо очень осторожно. Совсем недавно нам удалось еще раз, но уже лично, спросить Калинина, каких поэтов серебряного века он предпочитает. Ответ был таким: «Северянин, Маяковский и Есенин». [Калинин 2019] Можно сказать, что его предпочтения не изменились: это те же авторы, кого он и раньше упоминал в интервью. Проведя серию количественных экспериментов, мы не можем утверждать, что больше всего тексты Калинина похожи на тексты тех авторов, которым он отдает предпочтение. Оказалось, что:

* Цветаева и Ахматова побеждали по 7 раз (из них Цветаева – 5 раз по мере хи-квадрат, Ахматова – 4 раза по ключевым словам и 3 раза по геометрическим мерам)
* Есенин (по геометрическим мерам) и Маяковский (по ключевым словам) – 5 и 4 раза соответственно. Несмотря на победы, общая сумма рангов Маяковского по всем опытам оказалась меньше даже суммы Блока.
* Сумма рангов Есенина почти равна суммам Ахматовой и Цветаевой (85 и 83). Можно считать, что больше всего Калинин оказался похож именно на этих трех авторов.

Наше главное предположение о том, что увлечение лирикой Северянина повлияло на стиль Калинина и он пишет похоже именно на этого автора, не оправдалось. Расстояние между их текстами не оказывалось наименьшим ни в одном из опытов, и общая сумма рангов Северянина не столь велика.

Даже на официальной странице группы можно прочитать, что «Аффинаж – поэзия в духе Сологуба…». В нашей работе оказалось, что расстояние между текстами Сологуба и Калинина наибольшее. Первые упоминания этого автора в контексте группы встречаются в публикациях 2013 года, а в 2014 году вышел первый альбом группы – «Я и Мёбиус едем в Шампань»[[22]](#footnote-22). В 2015 году выходит интервью Калинина и еще одного участника группы, басиста Сергея Шиляева. В этом интервью Шиляев, сооснователь группы, говорит о задумке альбома, которая появилась еще в 2012 году: «сделать что-то безумное, в духе Сологуба». [Шиляев 2015: 7 м.] В этом же году выходит второй альбом группы – «Русские песни», и в одной из рецензий[[23]](#footnote-23) лирика Аффинажа снова сравнивается с лирикой Сологуба. И наконец, в 2017 году Захар Прилепин в эфире музыкальной телепрограммы «Соль» спрашивает Калинина о трех его любимых поэтах «Северянине, Есенине и Сологубе», но Калинин Сологуба уже не упоминает [Калинин 2017: 56]. Судя по всему, если Калинин и увлекался творчеством Сологуба (тексты песен из альбома 2013 года могли быть написаны и раньше), то это влияние может прослеживаться только в небольшом количестве текстов.

# **Заключение**

Мы провели количественное сравнение текстов современного поэта Михаила Калинина и восьми поэтов серебряного века. Почти все из них упоминались самим Калининым или в контексте группы Аффинаж. Мы предполагали, что больше всего тексты Калинина окажутся похожи на тексты Игоря Северянина: Калинин неоднократно называл его своим любимым автором, рассказывал, что творчество Северянина оказало на него сильное впечатление еще в школьном возрасте. Также мы ожидали, что тексты Калинина скорее окажутся похожи на тексты Владимира Маяковского и Сергея Есенина, которым он сам отдавал предпочтение, или на тексты Федора Сологуба. Выбирая для сравнения Анну Ахматову, Марину Цветаеву (их творчество Калинин упоминал в контексте того, что ему не нравится), Александра Блока, Николая Гумилева (их современный автор не упоминал вообще), мы думали, что сходство Калинина с ними будет меньше. Однако результаты получились не такими предсказуемыми.

Средства программирования (Python, R) и корпусной менеджер Sketch Engine позволили нам охватить собрания текстов объемом от 50 до 100 тысяч слов. Корпус Калинина мы собрали вручную, а тексты классиков взяли из НКРЯ. Мы сравнивали не сами тексты, а их частотные списки на трех языковых уровнях - лексем, скипграмм из двух слов и буквенных биграмм. Мы вычисляли расстояния между текстами, используя три геометрические меры (Манхэттенское расстояние, Косинусное расстояние, расстояние Евклида), статистическую меру (хи-квадрат), а также сравнили ключевые слова разных частей речи. Использовали относительные частоты – ipm, ARF/mill, меняли параметры в сравнении ключевых слов.

Результатом каждого сравнения был список авторов, отсортированный по возрастанию расстояния между их корпусами и корпусом Калинина. В конце серии экспериментов суммировали ранги каждого поэта в каждом опыте. Если сумма рангов какого-то поэта была меньше, чем у других, мы считали, что сходство его текстов и текстов Калинина выше. Разные меры давали разные результаты, также различались результаты на разных языковых уровнях. Нет одного автора, который бы побеждал всегда или всегда бы занимал высокие места. Всего было 24 опыта, соответственно, минимальная сумма рангов – 24 (24 первых места), а максимальная – 8\*24=192. Наименьшие суммы оказались у Ахматовой, Цветаевой (83) и Есенина (85), а максимальная – у Сологуба (160). Северянин на четвертом месте, его сумма – 101. Маяковский оказался даже ниже Блока (109) с суммой 114.

Таким образом, наше предположение о том, тексты Калинина больше всего окажутся похожи на тексты Северянина, не оправдалось.

Мы не только узнали, на кого из этих восьми авторов Калинин похож больше, но и обратили внимание на разные группы слов, которые оказались ключевыми для его корпуса:

* Слова, отличающие его от всех восьми поэтов серебряного века
* Слова, в равной степени ключевые и для него, и для других поэтов (мы рассмотрели его общие ключевые слова с Маяковским, Цветаевой, Ахматовой, Северяниным и Есениным).

В некоторых случаях мы попытались дать интерпретацию выделенным ключевым словам, ссылаясь на самого поэта, а в других случаях мы этого делать не стали. В частности, трудно проинтерпретировать сходство текстов Калинина с текстами Цветаевой и Ахматовой.

В данной работе мы исследовали художественные тексты с помощью точных методов, а значит, у исследования есть как минимум три пути развития:

1) Лингвистический: Подбор новых языковых параметров для сравнения, например, пунктуация, морфология, синтаксис. Более подробное сравнение результатов, уже полученных по леммам, биграммам и скипграммам; дополнительный морфологический анализ уже полученных частотных списков.

2) «Точный»: Например, оценка результатов, полученных по каждой мере. Мы не установили, почему разные меры показывали разную степень сходства текстов.

3) Литературоведческий: Например, более подробная интерпретация результатов – изучение совпавших ключевых слов, сходства с одними авторами и непохожести на других; продолжение описания индивидуального стиля современного поэта.

Как и в данной работе, оба пути могут снова пересечься, с использованием новых методов.

В заключение хотелось бы привести цитату из рецензии, которую мы уже упоминали:

*… эти влияния легко прослеживаются в произведениях группы. Первый [Сологуб] одарил «Аффинаж» правильным пессимизмом, второй [Северянин] помог пристальней смотреть внутрь себя, а Сергей Александрович [Есенин] научил высоко ценить простоту природы и традиции «босоногой России».*

Оценка и интерпретация лирики может быть субъективной, можно высказывать предположения о связи авторов, влиянии их друг на друга. Однако в большинстве случаев любое предположение необходимо доказать. Наша работа показала, что гипотезу о сходстве авторов можно проверить с помощью автоматических количественных методов. А результаты точных исследований всегда требуют проверки и содержательной интерпретации, если эта интерпретация возможна.

# **Библиография**

1. Бурцева А. Компьютерный анализ помог объяснить разницу между Сенекой и его подражателями // N+1 Интернет-издание, 2017. // URL: <https://clck.ru/FxwbX>
2. Гаспаров М.Л. Современный русский стих. Метрика и ритмика. / М., Наука, 1974.
3. Давыдова В. Изменение интервала сравниваемых слов для измерения расстояния между корпусами / Дипломная работа, М., РГГУ, 2018.
4. Калинин М. «Аффинаж»: герои нуар-шансона взрывают фестивальные сцены. Интервью для портала Московский Комсомолец, 2018 // URL: https://clck.ru/GGGSz
5. Калинин М. «Досье», 2011 / URL: <https://clck.ru/Fxwkn>
6. Калинин М., Шиляев С. Аффинаж. Фото и тексты песен. Лучше всех / М., Эксмо, 2019.
7. Калинин М. Выпуск телепрограммы «Соль» на канале НТВ с участием группы «Аффинаж», ведущий Захар Прилепин. 1 октября 2017 // [Запись телепрограммы] URL: https://clck.ru/Fxwbr
8. Калинин М. Интервью группы «Аффинаж» для II Всероссийского фестиваля современной поэзии в г. Вологда, 2016. // [Видеозапись интервью] URL: https://clck.ru/FxwcD
9. Калинин М. Интервью группы «Аффинаж» для портала «Spirit Of Rock» / Страница «Spirit Of Rock» Вконтакте, 2013. // URL: https://clck.ru/FxwcL
10. Калинин М. Люди не хотят думать и выходить за пределы своего клишированного счастья / Интервью, Интернет-портал «Еженедельник «Вечерний Саранск», 2018. // URL: https://clck.ru/Fxwbi
11. Калинин М. Ответ на вопрос во время презентации книги «Аффинаж. Фото и тексты песен. Лучше всех» в книжном магазине Республика, М. 24 мая 2019 / [Запись отсутствует]
12. Калинин М. Ответы на вопросы зрителей после концерта, "Китайский летчик Джао Да", М., 2015 / [Видеозапись] URL: https://clck.ru/GD362
13. Калинин М. Собрание сочинений, 2015 // URL: https://clck.ru/Fxwn5
14. Калинин М. Эм Калинин. Музыка высшей пробы. / Интервью для портала «Live Show Голос Рока», 22 ноября 2018 // URL: <https://clck.ru/FxwcU>
15. Калинин М., Шиляев С. Интервью для портала DVA CORABLYA, 2015 / [Видеозапись интервью] URL: <https://clck.ru/GGHro>
16. Калинин Эм. Стихотворения / Центр поддержки образования, науки и искусства, Санкт-Петербург, 2017.
17. Пиперски А.Ч. Измерение расстояний между корпусами. Tampere Summer School in Multilingual Corpora. Tampere, Finland, 28 August–1 September 2017.
18. Черный А. Михаил Калинин. Плюсовая поэзия. Молодые вологодские стихотворцы начала XXI столетия. / под общ. ред. А. Черного. Вологда: Союз Российских Писателей, 2008. – 150 с. // URL: <https://is.gd/r3KKOr>
19. Ярхо Б. И., Методология точного литературоведения: Избранные труды по теории литературы / Акимова М., Пильщиков И., Шапир М., М.: Языки славянских культур, 2006.
20. Benoit K, Watanabe K, Wang H, Nulty P, Obeng A, Müller S, Matsuo A (2018). “quanteda: An R package for the quantitative analysis of textual data.” Journal of Open Source Software, 3(30), 774. doi: 10.21105/joss.00774, https://quanteda.io.
21. D’hondt E., Verberne S., Weber N., Koster K., Boves. L. Using skipgrams and PoS-based feature selection for patent classification. Computational Linguistics in the Netherlands Journal 2: 52–70, 2012.
22. Gerard Salton, A. Wong, and C. S. Yang. A vector space model for information retrieval. Communications of the ACM, 18(11):613–620, 1975.
23. Gomaa W. H., Fahmy A. A. A Survey of Text Similarity Approaches International / Journal of Computer Applications, 68(13), April, pp. 13–18, 2013.
24. Jones E, Oliphant E, Peterson P, et al. SciPy: Open Source Scientific Tools for Python, 2001.// URL: http://www.scipy.org
25. Kilgarriff A. Simple maths for keywords / Proceedings of Corpus Linguistics Conference CL2009, University of Liverpool, UK, July 2009.
26. Kilgarriff A., Baisa V., Bušta J., Jakubíček M., Kovář V., Michelfeit J., Rychlý P., Suchomel V. The Sketch Engine: ten years on / Lexicography, 1: 7-36, 2014. // URL: https://www.sketchengine.eu
27. Kilgarriff, A. Comparing corpora / International Journal of Corpus Linguistics 6(1). 97–133, 2001.
28. Korobov M. pymorphy2 Documentation / Выпуск v0.1, 2013.
29. Literary Studies / A Companion to Digital Humanities, ed. Susan Schreibman, Ray Siemens, John Unsworth. Oxford: Blackwell, 2004.
30. Rayson, P., & Garside, R. Comparing corpora using frequency profiling. / In Proceedings of the Comparing Corpora Workshop at ACL 2000. Hong Kong, 2000.
31. Savický P., Hlavácová J. Measures of Word Commonness, Journal of Quantitative Linguistics, 9:3, 215-231, 2002.
32. Scott M. PC analysis of key words — and key key words / System 25(2),pp. 233–245, 1997.
33. Stamatatos, E. On the robustness of authorship attribution based on character n-gramfeatures. / Journal of Law and Policy, 21(2), 421–440, 2013.

# **Приложение**

1) Таблица 1: 100 самых частотных ключевых слов в корпусе текстов Михаила Калинина относительно объединенного корпуса 8 поэтов серебряного века (*n*=100, *ARF*)

|  |  |
| --- | --- |
| **Часть речи** | **Ключевые слова** |
| **Существительное** | экран, чувство, фраза, фото, талант, сигарета, пустота, подъезд, планет, пара, папа, мысль, момент, мама, кожа,  квартира, карман, дурак, доктор, дождь, дерьмо, врач, вопрос, вещь |
| **Глагол** | хотеться, хотеть, улыбаться, украсть,  считать, становиться, смочь, смеяться,  сделать, пытаться, прочесть, простить,  просить, потерять, поверить, писать, оставаться, нибыть, написать, купить, желать, думать, бояться, болеть |
| **Прилагательное** | хороший, тёплый, твой, такой,  счастливый, скорый, самый, похожий, плохой, нужный, лишний, красивый, который, идеальный, добрый, главный, весь, важный, большой |
| **Наречие** | уже, тоже, сложно, сейчас, потому, потом, постоянно, пока, очень, наверное, легко, вряд, вообще, возле |
| **Местоимение** | я, ты, себя, ничто |
| **Союз** | чтобы, что, хотя, то, или, будто |
| **Междометие** | спасибо |
| **Частица** | это, просто, ну, даже, бы, нужно, можно, внутри |

2) Таблица 2: общие ключевые существительные, глаголы и прилагательные Михаила Калинина с Анной Ахматовой и Мариной Цветаевой (из числа 100 самых частотных)

|  |  |  |
| --- | --- | --- |
| **Часть речи** | **Ахматова, n=100 ARF** | **Цветаева, n=100** |
| **Существительные** | часть, ужин, трубка, танец, сок, речка, почка, подушка, письмо, осколок, мишка, рай, раз, одиночество, обман, мишка, малый, лопата, ласка, лапа, кривая, красота, краса, кот, колодец, ковёр, картина, карта, кабак, девчонка, дама, гнев, воин, воздух, вершина,  будущее, боль, бой, бес | чернила, чердак, хвоя,  троп, тоска, ток, ткань,  табор, сэр, судьба, стопка, сто, ставка, солнышко, служанка,  cлог, слезинка, слеза,  синева, сеть, свидание,  рим, ремень, раз, пророк, прок, подушка,  плита, первое, нога, небо, молоко, молитва,  мир, малость, лето, крыса, искра, злодей, зверь, дно, дама, граф,  голос, голова, вкус, вино, бык, бес, башка, ада |
| **Глаголы** | увидеть, убить, тонуть, спеть, спасать, скучать, сеять, сесть, рождать, прочитать, проходить,  приехать, почувствовать, потушить, построить, ползти, показывать, поймать, отпустить,  открыться, открыть, обманывать, найти, меняться, махать, лелеять, кончаться, кончать, истратить, замечать,завидовать,  достаться, говорить,  гнать, выносить, вынести,  везти, быть | шить, шептать, целоваться,  упасть, уводить, таить,  сыскать, струиться, спасать,  смущать, следить, светить,  садиться, рассыпать, работать, проснуться, пропасть, покидать, пойти, переть, начаться, назвать, крикнуть, делить, встречать, вести, бросить |
| **Прилагательные** | чей, холодный, французский, тяжкий,  трупный, сытый, столовый, спорый, скрипучий, пресный | явственный, явный, южный,  шумный, шампанский, хилый, утренний, упругий,  удушливый, тройной, торжественный, стёртый, столовый, сомнительный, собственный, смутный |

3) Таблица 3: общие ключевые существительные, глаголы и прилагательные Михаила Калинина с Игорем Северяниным и Сергеем Есениным (из числа 100 самых частотных)

|  |  |  |
| --- | --- | --- |
| **Часть речи** | **Северянин, n=100 ARF** | **Есенин, n=1000 ARF** |
| **Существительные** | черта, часть, церковь, урна, ужас, танец, талия, стул, стая, стадо,  росток, раз, птица, поступок, половина, оркестр, окова, окно, наряд, мёртвый, молоко, мода, могила, мать, малый, ложа, лада, кружка, краска, испуг, идеал, желание, дура, дочь, дом, долг, демон,  грудь, вера, будущее, близкие, бинокль | шея, шаг, улыбка, труп,  течь, счастие, сон, смерть,  рота, рай, простой, привет,  палец, обида, нож,  надежда, мёртвый, мор,  мир, листва, конец,  книга, капля, имя, звук,  дождь, деньга, грязь,  вера, боль |
| **Глаголы** | уехать, сквозить, сидеть, расставаться, раздевать, пускать,  послушать, покидать,  пожелать, подниматься,  плакать, падать, опускать,  одевать, обмануть, нибыть, назвать, манить,  ломаться, литься, лгать,  иссякнуть, изображать ,вырастить, встретиться, впадать, виться, вериться | чувствовать, читать, улыбнуться, стучать, смотреть, сидеть, рваться,  растить, рассказать, пропасть, послушать, подумать, молчать,  ловить, иметь, значить, здравствовать, дрожать,  держать, выходить, влюбить, висеть, вести,  вернуться, бывать, бросить |
| **Прилагательные** | чистый уютный тёмный точный суровый сплошной смелый случайный скрытый  острый один ничтожный любящий какой едкий возможный воздушный болотный блестящий | чужой, страшный, последний, полный, мокрый, лёгкий,  злой, глупый,  глубокий, высокий, верный, ваш, богатый |

1. Май, труд, DH / Центр цифровых гуманитарных наук НИУ ВШЭ // Вестник Digital Humanities, № 11, 2019. [↑](#footnote-ref-1)
2. Официальная страница группы Аффинаж Вконтакте // URL: https://vk.com/affinage\_group [↑](#footnote-ref-2)
3. Анонс концерта группы Аффинаж/ Страница «Клуб-ресторан "Маяковский. Жёлтая Кофта"», 2016 // https://vk.com/affinage\_group?w=wall-43343704\_16676 [↑](#footnote-ref-3)
4. Анонс концертов группы Аффинаж / Официальная страница группы, 2013 // URL: https://vk.com/affinage\_group?w=wall-43343704\_128 [↑](#footnote-ref-4)
5. Официальная страница Михаила Калинина Вконтакте // URL: https://vk.com/acnug [↑](#footnote-ref-5)
6. Официальная страница проектов «никого не люблю» и «Эм(а)СПИД» Вконтакте // URL: https://vk.com/affinage\_acnu9 [↑](#footnote-ref-6)
7. Поэтический подкорпус НКРЯ // URL: http://www.ruscorpora.ru/search-poetic.html [↑](#footnote-ref-7)
8. Аффинаж - Русские песни / Рецензия музыкального портала DA Vision, 2015 // URL: https://vk.com/wall-50845529\_2783 [↑](#footnote-ref-8)
9. В данной работе под скипграмами подразумеваются пары слов, расположенных в тексте на расстоянии 2 друг от друга (не заходя за границы предложений) [↑](#footnote-ref-9)
10. При выделении биграмм учитывались пробелы. [↑](#footnote-ref-10)
11. Euclidean distance. Wikipedia, The Free Encyclopedia, 1 April 2019. // URL: https://clck.ru/FxrtA [↑](#footnote-ref-11)
12. Taxicab geometry. Wikipedia, The Free Encyclopedia, 29 April 2019. // URL: https://clck.ru/FxruE [↑](#footnote-ref-12)
13. Cosine similarity. Wikipedia, The Free Encyclopedia, 3 March 2019. // URL: https://clck.ru/Fxrte [↑](#footnote-ref-13)
14. *City Block distance* или *Расстояние городских кварталов* – одно из названий Манхэттенской меры [↑](#footnote-ref-14)
15. Предполагаемое (Expected) значение в клетке таблицы – произведение маргинальных сумм, поделенное на сумму всех значений. Нулевая гипотеза – частоты слов распределены в обоих корпусах равномерно. [↑](#footnote-ref-15)
16. Корпус, из которого извлекаются ключевые слова [↑](#footnote-ref-16)
17. Корпус, с которым сравнивается фокусный корпус [↑](#footnote-ref-17)
18. ARF на миллион словоформ корпуса [↑](#footnote-ref-18)
19. Или явления, которые хоть и могли существовать в 20 веке, но в 21 веке стали гораздо более распространенными [↑](#footnote-ref-19)
20. Сокращение от аватар – фотография профиля в социальной сети [↑](#footnote-ref-20)
21. Дионисическое – Аполоническое / Философский словарь, 3000 статей, 2012-2019 // [Электронный словарь] URL: http://www.harc.ru/slovar/124.html [↑](#footnote-ref-21)
22. Я и Мёбиус едем в Шампань / Релиз альбома на официальной странице группы Аффинаж «Вконтакте», 21 февраля 2014 [↑](#footnote-ref-22)
23. Аффинаж - Русские песни / Рецензия музыкального портала DA Vision, 2015 // URL: https://vk.com/wall-50845529\_2783 [↑](#footnote-ref-23)