Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт филологии и языковой коммуникации
Кафедра теории германских языков и межкультурной коммуникации

УТВЕРЖДАЮ
Заведующий кафедрой ТГЯиМКК
_____ О.В. Магировская
« ____» _____2021 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

# АТРИБУЦИЯ АВТОРСТВА НА ОСНОВЕ ОЦЕНКИ СТИЛЕМЕТРИЧЕСКИХ ПАРАМЕТРОВ ТЕКСТА (НА МАТЕРИАЛЕ АНГЛОЯЗЫЧНЫХ ТУРИСТИЧЕСКИХ БЛОГОВ)

45.04.02 Лингвистика
45.04.02.01 Межкультурная коммуникация и перевод

Магистрант _____ Л.А. Вдовина

Научный руководитель _____ д-р филол. наук, зав. каф. РЯиПЛ А.В. Колмогорова

Нормоконтролер _____ Я.М. Янченко

Красноярск 2021

## СОДЕРЖАНИЕ

# ВВЕДЕНИЕ

В процессе создания письменного текста каждый человек в силу своих индивидуальных особенностей использует различные языковые структуры. Обнаружение и описание таких структур позволяет с высокой точностью определять личность автора анонимного документа с помощью объективных научных методик.

Распространение Интернета привело к увеличению количества анонимных материалов, существенную долю которых составляют противоправные тексты, содержащие призывы к экстремизму, угрозы, оскорбления и пр. В связи с этим вопросы установления авторства представляют интерес не только для лингвистов, но и для экспертов-криминалистов в вопросах установления личности автора вредоносного кода, определения автора писем с угрозами, установления достоверности предсмертных записок, идентификации террористов и убийц. Также системы атрибуции текста используются юристами и журналистами, поскольку установление истинного автора текста может использоваться для обнаружения плагиата, привлечения к ответственности правонарушителей или оправдания невиновных.

Данная работа посвящена атрибуции авторства, основывающейся на количественной оценке стилеметрических параметров. Построение автоматической системы атрибуции и эффективность отдельных параметров рассматриваются на примере текстов англоязычных туристических блогов.

**Актуальность** исследования обусловлена тем, что, несмотря на существование достаточно точных систем атрибуции, исследователи до сих пор не пришли к согласию по поводу оптимального набора характеристик, описывающих авторский стиль. Многообразие возможных стилеметрических параметров требует дополнительных исследований в этой области.

**Научная новизна** исследования заключается в рассмотрении эффективности стилеметрических параметров для атрибуции текста среди

большого количества авторов-кандидатов. Большинство работ данной области рассматривает проблему выбора искомого автора из небольшого круга кандидатов, исследования обширных корпусов весьма немногочисленны.

**Целью** нашего исследования является сравнительный анализ стилеметрических параметров текста и оценка их эффективности для решения задачи атрибуции авторства.

Для достижения данной цели потребуется решение следующих **задач**:

1) провести обзор исследований в области стилеметрии: рассмотреть существующие подходы к анализу авторского стиля; подробно охарактеризовать основные параметры и современные методы атрибуции;

2) составить сбалансированный исследовательский корпус;

3) проанализировать показатели отдельных стилеметрических параметров для авторов нашего корпуса;

4) осуществить практическую реализацию автоматического извлечения характеристик и построить каноническую систему атрибуции;

5) провести анализ эффективности разработанного решения.

**Объектом** исследования является задача атрибуции текста; **предметом** служат стилеметрические параметры, количественная оценка которых позволяет охарактеризовать авторский стиль.

Корпус языкового **материала** был сформирован из записей англоязычных туристических блогов, размещенных на платформе WordPress. Всего корпус содержит 50 авторов, на каждого из которых приходится около 3000 слов. Суммарный объем проанализированного материала составляет 300 страниц формата А4.

**Теоретическая значимость работы** заключается в рассмотрении стилеметрических параметров на обширном корпусе, который позволяет эмпирически оценить их различающую способность и сопоставить с эффективностью в системе атрибуции.

**Практическая значимость** заключается в возможности применения разработанной системы атрибуции для решения реальных задач идентификации автора противоправного текста.

**Методологическую основу** данного исследования составляют труды ученых в области анализа индивидуального стиля автора (В.В. Виноградов, Ю.Н. Караулов, Г.Я. Мартыненко, М.Ю. Мухин), в области атрибуции текстов (З.И. Резанова, Д.В. Хмелев, А.С. Романов, Р.В. Мещеряков) и в области компьютерных методов обработки естественного языка и классификации текстов (М. Коппел, Дж. Шлер, Э. Стамататос, Т. Нил).

Основными **методами** исследования в работе являются метод сплошной и специальной выборки, метод экспертного лингвистического анализа, метод статистического анализа, метод контент-анализа.

Структура работы обусловлена ее содержанием.

Во введении изложено краткое содержание темы, ее актуальность, научная новизна, цель, задачи, объект и предмет исследования, материал и методы исследования, его практическая значимость и теоретическая база.

Первая глава «Теоретические основания атрибуции текста» посвящена описанию теоретических оснований исследования стиля и технологий решения задачи атрибуции текста. Глава состоит из трех параграфов. В параграфе 1.1. «Стилеметрия и атрибуция текста» описываются подходы к исследованию авторского стиля и типы задач, решаемых стилеметрией, а также дается формулировка задачи атрибуции авторства. Параграф 1.2. «История стилеметрических исследований» содержит краткий хронологический обзор научных работ, демонстрирующий становление и развитие стилеметрии. В параграфе 1.3. «Построение канонической системы атрибуции авторства» рассматриваются этапы построения системы автоматической атрибуции; типы стилеметрических характеристик и формальных методов классификации, применяемых для решения задачи, а также предоставляется теоретическое обоснование выбранной модели

классификации. Выводы по первой главе содержат основные выводы по всем параграфам.

Вторая глава «Анализ стилеметрических параметров текста и их эффективности в задаче атрибуции авторства» посвящена анализу основных стилеметрических характеристик, путем количественной оценки которых, проводится атрибуция авторства. Глава состоит из четырех параграфов. В параграфе 2.1. «Материалы исследования» представлено подробное описание корпуса исследования. Параграф 2.2. «Предварительная обработка корпуса» содержит описание техник обработки текста, используемых в нашей работе. В параграфе 2.3. «Анализ стилеметрических параметров текста» выделяются четыре подпункта, в которых рассматриваются лексические, символьные, синтаксические и морфологические группы параметров. В параграфе 2.4. «Результаты работы модели атрибуции авторства» представлена оценка точности разработанной системы атрибуции в зависимости от используемых стилеметрических характеристик. Выводы по второй главе содержат основные выводы по всем параграфам. В заключении представлены выводы по всему исследованию. Список использованной литературы содержит 115 источников.

# ГЛАВА 1. ТЕОРЕТИЧЕСКИЕ ОСНОВАНИЯ ИССЛЕДОВАНИЯ АТРИБУЦИИ АВТОРСТВА

## 1.1. Стилеметрия и атрибуция авторства

Интерес к языковым средствам, выражающим индивидуальность автора, возник с появлением в лингвистике антропоцентрической парадигмы. Язык стал рассматриваться как средство отражения внутреннего мира индивида, как зеркало души, а центральным объектом исследований стала языковая личность автора.

В монографии «Русский язык и языковая личность» Ю.Н. Караулов определяет *языковую личность* как «личность, выраженную в языке (текстах) и через язык, личность, реконструированную в основных своих чертах на базе языковых средств» [Караулов, 1987: 38]. Таким образом, в ходе создания текста каждый человек в силу своих индивидуальных особенностей использует различные языковые структуры, а также производит характерный для него выбор способов и средств выражения мыслей. Такой выбор отражает уникальный стиль автора.

Так, Г. Хердан (1966) рассматривает *стиль* как «общую характеристику индивидуального способа выражения личности в языке», как «подсознательный фактор, которому автор не может не подчиняться» [Herdan, 1966: 12]. Согласно более формальному определению стиля, предложенному В. Винтером, *стиль* представляет собой «систему периодически повторяющихся выборок из перечня произвольных черт языка» [Winter, 1969: 3]. При этом тип выборки может быть различным: исключение произвольных элементов, обязательное включение каких-либо характеристик и т. д.

В рамках изучения языковой личности при рассмотрении индивидуальных особенностей выражения мыслей автора наибольшее распространение получило понятие *идиостиль*. М.Н. Кожина определяет

*идиостиль* как «совокупность языковых, стилистико-текстовых особенностей, свойственных речи писателя, ученого публициста, а также отдельных носителей данного языка» [Кожина, 2006: 95]. Согласно Г.В. Напреенко, выбор из бесконечного множества средств и способов выражения коммуникативных намерений формирует индивидуальные предпочтения языковой личности – *индивидуальный стиль автора* или *идиостиль* [Напреенко, 2014: 17].

В данной работе будет использоваться определение *индивидуального стиля автора* как «системы периодически повторяющихся выборок из перечня произвольных черт языка» [Winter, 1969: 3], так как оно наиболее близко статистическим методам анализа, используемым в нашем исследовании.

Первые исследования, посвященные изучению индивидуального стиля автора, были проведены такими лингвистами, как М.М. Бахтин и Р.О. Якобсон. Одним из наиболее значимых первых трудов по вопросам определения индивидуального стиля является монография В.В. Виноградова «Проблема авторства и теория стилей» [Виноградов, 1980]. Вслед за этими авторами исследование идиостиля продолжили В.П. Григорьев и Г.О. Винокур, в результате чего была заложена основа изучения авторского стиля в отечественной лингвистике. Так, в начале двадцать первого века было опубликовано множество работ, изучающих идиостили отдельных авторов и в целом рассматривающих принципы статистического анализа стиля (например, Каменская, 2001; Горчакова, 2009; Мухин, 2009; Ахмедова, 2008).

На сегодняшний день одним из направлений исследований по изучению авторского стиля является *стилеметрия* – прикладная филологическая дисциплина, занимающаяся измерением стилевых характеристик с целью систематизации и упорядочения (типологии, атрибуции, датировки, диагностики, реконструкции и т.д.) текстов и их частей [Мартыненко, 1988].

Стилеметрия занимается исследованием и измерением стилевых характеристик текста с целью установления авторства или получения каких-либо сведений об авторе и условиях создания текстового документа. В отличие от других направлений, стилеметрия характеризуется формальным подходом к авторским характеристикам и, в первую очередь, занимается подсчетом и измерением стилистических явлений, а также их статистическим анализом с использованием лексико-статистических методов.

*Объектом стилеметрии* является текст, созданный конкретным автором, в конкретное время, в конкретной ситуации. *Предметом стилеметрического исследования* являются элементы стиля, которые понимаются как особенности периферии характеристики объекта. Стиль может быть описан через факультативные, поверхностные признаки текста, которые лишь неявным образом затрагивают его сущностные, глубинные характеристики [Базылев, 2007: 143].

Стилеметрия имеет дело с количественным классифицированием, а эта область классификационных занятий тесно соприкасается с несколькими научными направлениями: теорией группировок, теорией оценивания, распознаванием образов, теорией корреляции, количественной таксономией, методами психологического тестирования и др. Границы между этими направлениями стираются, и сегодня можно говорить о комплексе подходов и методов, занимающихся теми или иными видами количественной систематизации объектов произвольной природы [Базылев, 2007: 143].

Поскольку связь автора и используемых им языковых средств — проблема двухсторонняя, стилеметрия, занимающаяся анализом авторского стиля также условно делится на две области.

Так, Г.Я. Мартыненко (2019) замечает, что в процессе развития стилеметрия разделилась на два направления: традиционное атрибуционное направление, занимающееся вопросами установления авторства, и направление, уделяющее внимание классификационным статистико-

филологическими задачами — диагностикой, типологией, таксономией периодизацией.

Исходя из этого, все задачи стилеметрии в широком смысле можно разделить на два вида [Батура, 2012; Литвинова, Громова, 2020]:

1) идентификационные задачи, связанные с определением автора текста путем проведения раздельного и сравнительного анализа признаков, проявившихся в спорном тексте и текстах – образцах предполагаемого автора; при этом проблема идентификации автора также может рассматриваться в различных аспектах: когда нужно определить автора из относительно небольшого замкнутого круга лиц (проблема идентификации), либо же необходимо определить, является ли автором текста конкретное лицо (проблема верификации);

2) диагностические задачи, связанные с определением половозрастных, индивидуально-личностных характеристик, уровня коммуникативной компетенции, речевой культуры, сферы профессиональной деятельности автора текста.

Наиболее развернутую характеристику задач стилеметрии с точки зрения компьютерной лингвистики можно встретить в западной литературе. Так, согласно Э. Стамататосу (2009), все множество задач стилеметрии можно классифицировать следующим образом:

– *атрибуция авторства* или *идентификация автора*. Цель состоит в том, чтобы отнести текст неизвестного авторства к одному из кандидатов. Приводится набор кандидатов, для которых имеются тексты – образцы установленного авторства. Идея состоит в том, что автор оставляет в своих произведениях неповторимый паттерн, и анализ авторства может быть использован для выявления уникальных признаков, позволяющих классифицировать автора [Pillay and Solorio, 2010]. Если предположить, что для каждого рассматриваемого текста фактический автор находится в заданном наборе кандидатов, то с точки зрения машинного обучения эту

задачу можно считать стандартной задачей категоризации текста [Sebastiani, 2002].

– *Верификация авторства*. В данном случае эксперту предоставляют тексты – образцы одного автора и просят определить, был ли другой текст написан этим же автором. Как проблема категоризации верификация авторства значительно сложнее, поскольку для сравнения нет репрезентативных примеров текстов не автора, и необходимо установить относится ли разница значений анализируемых характеристик к нормальным колебаниям в пределах авторского стиля одного человека, или же является сигналом, отражающим стили разных авторов [Koppel, Schler, 2004].

– *Выявление плагиата*. Плагиат – это копирование чужой работы путем ее изменения, либо копирования оригинальных идей без надлежащей ссылки. В наши дни плагиат стало чрезвычайно трудно идентифицировать, поскольку для достижения надежного результата необходимо сравнить документ с миллиардами других документов, доступных в Интернете [Ramnial, Panchoo, Pudaruth, 2015]. Подобное сравнение требует существенных вычислительных и временных ресурсов, поэтому современные исследования направлены на обнаружение плагиата не путем сопоставления текстов, а методом обнаружения различий в авторском стиле. Таким образом, они частично следуют тем же шагам, что и другие задачи стилеметрии, а именно обнаруживают уникальные особенности автора и определяют, были ли фрагменты текста написаны одним и тем же человеком [Stamatatos, Koppel, 2011; Meyer zu Eissen, Stein, Kulig, 2007].

– *Портретирование* или *характеризация автора* занимается определением отдельных характеристик автора данного текста. Никакой набор кандидатов не предоставляется, а задача состоит в том, чтобы предоставить как можно больше информации об авторе. Например, является автор мужчиной или женщиной [Романов, Мещеряков, 2011; Степаненко, 2017; Koppel, Argamon, Shimoni, 2002]? Каков возраст автора [Schler, 2006]?

– *Выявление стилистических несоответствий*. Это направление пересекается с задачей по выявлению плагиата, поскольку плагиат обычно вызывает стилистические несоответствия. Тем не менее, здесь цель состоит не в том, чтобы определить, был ли текст написан несколькими авторами, а в том, чтобы найти стилистические несоответствия и помочь их устранить [Graham, Hirst, Marthi, 2005].

Данное исследование фокусируется в частности на задаче атрибуции авторства или идентификации автора (в данной работе термины «атрибуция авторства» и «идентификация автора» используются взаимозаменяемо) и не охватывает другие проблемы стилеметрии, хотя используемые функции и методы могут быть полезными и для других направлений.

На сегодняшний день существует несколько подходов к решению задачи атрибуции авторства, однако среди общего множества можно выделить три основных метода [Vašák, 1980]:

1. Метод документальный и фактический. Здесь для определения авторства используется информация двух видов: документы, созданные предполагаемым автором произведения, например, дневники, автографы, автобиография, письма, и т. п.; либо документы, полученные от иных лиц и учреждений, каким-либо образом соприкасавшихся с процессом создания текста. Тем не менее, такие документы должны рассматриваться с осторожностью и не могут считаться однозначным доказательством авторства.

2. Метод идейно-тематический. Атрибуция авторства базируется на противопоставлении идей, литературной школы, направления и тематики текста спорного авторства и других произведений предполагаемых авторов-кандидатов.

3. Метод языковой и стилистический. Данный метод исходит из понятия индивидуального стиля и заключается в сопоставлении лингвостилистических параметров спорного текста и документов предполагаемых авторов-кандидатов. Наибольшее внимание уделяется

характеристикам, которые автор генерирует неосознанно, основываясь на своем индивидуальном стиле.

В настоящем исследовании мы будем заниматься вопросом атрибуции авторства исходя исключительно из языкового и стилистического метода, поскольку такой лингвистико-статистический метод определения авторства включает в себя лишь сравнение языка и стиля при помощи объективных характеристик текста и в наименьшей степени зависит от субъективных факторов, которые могут присутствовать при анализе сторонних документов или сопоставлении идей.

## 1.2. История стилеметрических исследований

Можно сказать, что проблема установления авторства появилась вместе с первыми литературными произведениями, однако первым упоминаемым в отечественной литературе исследованием стиля текста, проведенным с целью его атрибуции, является опубликованный в XV в. трактат итальянского филолога Лоренцо Валла, в котором он на основе различных, в том числе стилистических критериев рассматривает проблему авторства дарственной грамоты Константина [Валла, 1963].

В зарубежной литературе в качестве одной из первых стилеметрических работ упоминается исследование Эдмонда Мэлоуна, в результате которого он, основываясь на нехарактерных для Шекспира чертах стиля (количество строк, заканчивающихся безударным слогом; с остановкой на конце; рифмующихся строк) пришел к выводу, что «Генрих VI» был написан другим автором. В 1812 году данное исследование было продолжено Генри Вебером, в результате чего, основываясь на частоте строк, заканчивающихся безударным слогом, он разделил авторство отдельных фрагментов трагикомедии «Два знатных родича» между Уильямом Шекспиром и Джоном Флетчером [Grieve, 2007].

Тем не менее, считается, что становление современной стилеметрии как статистической стилистики началось в 1851 г., когда английский математик Август де Морган предположил, что авторы могут быть детектированы посредством скрытых статистических черт и выдвинул гипотезу о том, что средняя длина слова в произведении автора может быть характерной чертой авторского стиля. Хотя сам де Морган не делал никаких вычислений, в 1887 г. Томас К. Менденхолл, заинтересованный его гипотезой, опубликовал статью «Характеристика состава», в которой попытался охарактеризовать стиль разных авторов через частотное распределение слов различной длины и предложил использовать данную характеристику для разрешения Шекспировского вопроса [Mendenhall, 1887].

Во второй половине XIX в. также существовала группа ученых, работавшая над так называемым методом «стилометрики». В качестве критериев они использовали изменение размера стихов и количество повторений отдельных слов. В результате своих исследований Дж. К. Ингрэм, Ф.Г. Флей и Ф. Фернивалл обнаружили и охарактеризовали постоянное изменение стиля Шекспира на протяжении всех лет его творчества [Tuldava, 2005].

Термин «стилеметрия» впервые был использован немецкий филологом Вильгельмом Диттенбергером, который в 1880 г. попытался решить проблему хронологии и атрибуции диалогов Платона. Стилеметрию ученый рассматривал как задачу решения спорного авторства. Он исследовал частоту употребления служебных слов, наличие которых не зависит от тематики текста. Позднее его исследования продолжили такие ученые, как Э. Целлер, Ф. Чада и К. Риттер [Мартыненко, 2019].

В конце XIX в. статистическая стилистика все еще не привлекла внимание большого количества ученых, поэтому можно назвать лишь небольшое количество работ того периода: так, Л.А. Шерман анализировал среднюю длину предложения в английской прозе (1888), В. Лутославский использовал методы статистики для решения вопроса о хронологии диалогов

Платона (1897), Л. Франк рассматривал частотность цветовых обозначений в работах Гете (1909), а П. Парцингер (1909) изучал эволюцию стиля Цицерона [Журавлева, 2012].

В России вероятностно-статистический метод в целях исследования стиля впервые применил А.А. Марков. В 1913 г. исследователь представил работу, в которой проанализировал 20000 символов из поэмы А.С. Пушкина «Евгений Онегин» и рассчитал вероятность появления гласных и согласных в определенных условиях. Вслед за этим в 1915 г. Н.А. Морозов опубликовал работу, посвященную проблеме отличия оригинальных работ известных авторов от плагиата. Для решения данной задачи ученый рассматривал частоту употребления служебных слов в отдельных текстах.

Следующий важный шаг в применении статистических методов при изучении стиля был сделан в 30-е гг. XX в. Джордж Ципф (1932), изучавший статистические закономерности, обнаружил, что одни слова используются намного чаще, чем другие и если распределить слова по популярности, то частота использования слова будет обратно пропорциональна его рангу в таблице. Помимо этого, были опубликованы работы Дж. В. Флетчера, который изучал развитие стиля Спенсера (1934), Дж. М. Боллинга, написавшего критическое эссе по статистическому исследованию стиля Гомера (1937) и Дж. Б. Кэрролла, рассматривающего проблему разнообразия словаря (1938).

В 1938 г. Дж. У. Юлл представил исследование распределения длины предложений как перманентную характеристику стиля, а в 1944 г. – опубликовал свою знаковую работу, в которой предлагал ввести критерий богатства словаря, основанный на частоте повторения одинаковых слов. Именно с этих работ начинается применение современных статистических методов в стилеметрии, а исследования авторского стиля начинают проводиться по всему миру.

В 1960-е гг. с появлением компьютеров интерес к статистической стилистике значительно возрос. Использование технических устройств

позволило хранить большие корпусы текстов, находить и считать слова, производить сложные вычисления и обрабатывать большие объемы информации. Первой попыткой атрибуции авторства с использованием компьютера считается исследование Мостеллера и Уоллеса (1964). Исследователи поставили перед собой цель определить авторство отдельных статей «Записок Федералиста», которые были написаны совместно А. Гамильтоном, Дж. Мэдисоном и Дж. Джеем, но публиковались под общим псевдонимом Публий. Для решения поставленной задачи ученые подсчитали Байесовскую вероятность принадлежности того или иного эссе к конкретному автору на основании частотности употребления служебных (функциональных) слов, таких как предлоги, союзы и артикли. Результаты исследования показали существенные различия в показателях, что позволило однозначно провести атрибуцию авторства и ознаменовало собой начало эпохи компьютерной стилеметрии – статистических методов, использующих измеримые характеристики текста, для анализа литературного стиля [Holmes, 1998].

Именно в этот период появляются и получают распространение разнообразные идеи анализа индивидуального стиля автора. Начинают публиковаться работы множества исследователей, среди которых можно упомянуть Г. Хетсо, Дж. Б. Кэрролл, Г. Хердана, Л.В. Милова, Ч. Мюллера, Дж. Мистрика, Л.Т. Милика, Б.Н. Головина, М.Н. Кожину, Д. Росса, М.Х.Т. Элфорда, Л.Лаббе и др.

Кроме того, стали выходить сборники статей, посвященных статистическим исследованиям стиля: «Математика и поэзия» (1965), «Компьютер и литературный стиль» (1966), «Статистика и стиль» (1969), «Вопросы статистической стилистики» (1974), «Компьютер в литературе и лингвистических исследованиях» (1976), а также журналы, например «Пражские исследования по математической лингвистике», «Изучение метрики и поэзии», «Глоттометрика», «Журнал квантитативной лингвистики» и множественные конференции по компьютерной лингвистике.

Последующие исследования в области стилеметрии также довольно многочисленны. Можно сказать, что в конце XX в. прогресс компьютерных технологий и повсеместное распространение Интернета, вызвавшие появление большого количества электронных текстов, изменили технологию атрибуции авторства. Были разработаны эффективные методы представления и классификации больших объемов текста, средства обработки естественного языка, способные эффективно анализировать текст, и мощные алгоритмы машинного обучения, способные обрабатывать тысячи стилеметрических признаков.

На сегодняшний день были проведены сотни исследований, освещающих различные аспекты стилеметрии в целом и атрибуции авторства в частности. Среди наиболее значимых современных отечественных лингвистов, занимавшихся стилеметрией необходимо отметить Г.Я. Мартыненко, написавшего в 1988 г. монографию «Основы стилеметрии», а также с 1971 по 2009 гг. опубликовавшего ряд статей, посвященных методам статистических исследований в языкознании и автоматической атрибуции авторства. Его поздние работы посвящены исследованию принципа «золотого сечения» в лингвистике и его применению на различных уровнях (фонемный, морфологический, синтаксический).

Другим современным отечественным лингвистом, занимающимся статистическими методами изучения авторского стиля, является М.Ю. Мухин (2011). Так, в своей работе Михаил Юрьевич проводит сопоставительный анализ произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова и составляет так называемые «концептуальные профили» их творчества. Сначала составляется корпус слов, наиболее характерных для творчества определенных авторов, но при этом не учитывается сверхчастотная лексика, используемая повсеместно, а затем проводится семантический анализ полученных результатов.

Также широкую известность получил М.А. Марусенко, разработавший метод распознавания образов – один из первых комплексных подходов к анализу индивидуального стиля автора, учитывающий не только лексический, но и синтаксический уровни. В основе данной методики лежит принцип многомерного статистического анализа, где авторский стиль представлен как набор параметров, характеризующих состав, способы объединения и стилистико-вероятностные закономерности речевых средств.

М.А. Марусенко разделяет процедуру атрибуции на три этапа [Марусенко, 1990]:

– формирование литературно-критической гипотезы,
– проверка литературно-критической атрибутивной гипотезы методами теории распознавания образов (проходит в несколько этапов, осуществляемых в строгой последовательности),
– интерпретация результатов проверки атрибутивной гипотезы.

Хотя данный метод атрибуции является весьма эффективным, не всегда представляется возможным применить его из-за определенных недостатков: большое число признаков необходимо определять вручную, рассмотрение большого количества авторов вызывает затруднения и др. [Суркова, 2014]

В зарубежной лингвистике статистическая лингвистика получила не меньшее освещение. Некоторые авторы предлагают новые стилеметрические характеристики [Koppel, Schler, 2004; Graham, Hirst, Marthi, 2005; Hedegaard, Simonsen, 2011], другие изучают производительность систем атрибуции с различными комбинациями набора признаков [Chakraborty, Bandyopadhyay, 2010; Canales et al., 2011; Pateriya et al., 2014] или фокусируются на частных задачах стилеметрии, таких как идентификация авторства [Koppel, Schler, Argamon, 2010; Tanguy et al., 2012], верификация авторства [Koppel, Schler, 2004; Fridman et al., 2013], составление портрета автора [Jankowska et al., 2013] или исследование более частных случаев, таких как изменение авторского стиля с течением времени [Tamboli, Prasad, 2019], категоризация текста с точки зрения жанра [Stamatatos, Fakotakis, Kokkinakis, 2001],

реализация различных методов обработки текста [Sarwar, Nutanong, 2016] и сравнение производительности различных алгоритмов машинного обучения [Jockers, Witten, 2010; Zheng et al., 2006].

Среди множества источников наиболее полно охватывают диапазон стилеметрических исследований обзор П. Джуола (2008), в котором рассматривается история атрибуции авторства, используемые стилеметрические параметры и математические модели; обзор современных методов атрибуции авторства, опубликованный Э. Стамататосом (2009), где рассматриваются как стилеметрические характеристики, так и методы атрибуции; а также обширное исследование эффективности верификации авторства в зависимости от алгоритмов, извлекаемых признаков, длины обучающих и тестовых фрагментов, проведенное Т. Нил и др. (2017). Всеобъемлющие обзоры последних зарубежных работ по стилеметрии и атрибуции авторства можно найти в работах С. Свейн и др. (2017), а также К. Лагутиной и др. (2019).

Как показывают исследования, к настоящему времени накоплен большой опыт в применении статистического метода для решения проблемы атрибуции текстов. Можно сказать, что этот метод стал хрестоматийным, причем не только в лингвистике, но и в криминалистике, медицинской диагностике, социальной психологии [Базылев, 2007]. Кроме того, с развитием компьютерных технологий и увеличением вычислительных мощностей проблема атрибуции авторства стала рассматриваться как задача автоматической классификации текста, решаемая с помощью алгоритмов обработки естественного языка и машинного обучения, на сегодняшний день не представляется возможным использование чисто экспертных методов, не использующих компьютер [Pillay and Solorio, 2010].

Таким образом, в настоящей работе задачу атрибуции авторства мы будем понимать как прикладную задачу компьютерной лингвистики по разработке системы автоматической классификации, соотносящей спорный

текст с одним из авторов-кандидатов, на основании оценки отдельных стилеметрических параметров.

## 1.3. Построение канонической системы атрибуции авторства

В рамках компьютерной лингвистики законченный процесс атрибуции авторства состоит из сбора и предварительной обработки текстов; извлечения стилеметрических параметров; построения и обучения модели классификации, выполняющей фактическую работу по атрибуции на основе извлеченных параметров; и оценки эффективности модели [Tang, Liang, Liu, 2019; Neal et al., 2017].


Рисунок 1. Типовая система атрибуции авторства

Структура типовой системы атрибуции авторства показана на рисунке 1. В дальнейшем, основываясь на этой модели, мы рассмотрим каждый из этапов построения системы атрибуции в отдельности.

### 1.3.1. Предварительная обработка текста

С точки зрения вычислительных систем все данные делятся на две категории: структурированные и неструктурированные. Структурированные данные соответствуют табличному формату (например, таблица с именами, идентификаторами и адресами сотрудников). Этот тип данных идеально

подходит для компьютерной обработки и использования в машинном обучении, поскольку его элементы можно адресовать, обрабатывать и анализировать с относительной легкостью. Вся остальная информация, например, книги, аудио, изображения, журналы и т.п., относится к неструктурированным данным [Zhang et al., 2014]. Неструктурированные данные, напротив, непригодны для машинного обучения, поскольку их структура и параметры неизвестны, и компьютер не может их обработать [Nalini, Sheela, 2014].

Компьютерная обработка текстов сопряжена с рядом проблем: язык представляет собой гибкий и универсальный инструмент, в связи с чем текст трудно представить в структурированном формате данных. Для того, чтобы создать модель автоматической атрибуции авторства, необходимо сначала преобразовать корпус текстов в формат, подходящий для машинной обработки, другими словами, преобразовать неструктурированные данные в структурированные для эффективного индексирования и поиска в больших объемах текста [Nadkarni, Ohno-Machado, Chapman, 2011].

Процесс такой трансформации обычно называют «обработкой естественного языка» (Natural Language Processing, далее NLP). Наиболее эффективный и надежный комплекс средств обработки естественного языка можно найти в инструментарии естественного языка (Natural Language Toolkit, далее NLTK). NLTK представляет собой набор написанных на языке программирования Python библиотек и программ, предназначенных для символьной и статистической обработки естественного языка (в основном, английского) [Loper, Bird, 2002].

Таким образом, процесс обработки естественного языка, необходимый для приведения текста к структурированному формату, включает в себя два этапа количественного представления текстовых данных: предварительную обработку текста (нормализация и очистка данных) и извлечение информативных параметров.

Предварительная обработка текста является критическим этапом NLP, поскольку полученные в результате слова, символы и предложения являются фундаментальными единицами, передаваемыми на дальнейшие этапы работы системы [Kannan, Gurusamy, 2014], и, следовательно, влияют на результаты работы всей модели атрибуции текста. Нельзя не отметить важность предварительной обработки текста, поскольку ее отсутствие приводит к снижению точности и эффективности модели атрибуции авторства – машина не может должным образом обработать данные, не приведенные к нормализованному виду.

На сегодняшний день наиболее часто выделяют следующие этапы предварительной обработки текста [Kadhim, 2018; HaCohen-Kerner, Miller, Yigal, 2020]:

1. Удаление знаков препинания. Данный шаг предполагает удаление любых знаков препинания и позволяет исключить постороннюю информацию, которая может помешать обработке отдельных лексических единиц.

2. Перевод в нижний регистр. Все буквы преобразуются в строчные, чтобы, например «destination» и «DESTINATION», были восприняты компьютером как одно и то же слово, а не два разных. Этот шаг считается полезным для любого типа классификации текстов [HaCohen-Kerner, Miller, Yigal, 2020].

3. Лемматизация или стемминг. Одна из основных сложностей обработки естественного языка заключается в том, что текстовые данные содержат различные словоформы, которые интерпретируются машиной как совершенно разные слова. Так, для лингвиста-эксперта очевидно, что «visited», «visits» и «visit» представляют собой формы одного слова, а компьютерная система посчитает все три единицы абсолютно разными словами. Для решения данной проблемы необходимо провести лемматизацию – свести различные формы слова к начальной, например, заменить «visited» и «visits» леммой «visit».

Тем не менее, автоматическое проведение лемматизации не является тривиальной задачей. Компьютерные инструменты обработки естественного языка несовершенны, поскольку невозможно учесть каждую из множества языковых ситуаций, а текстовые данные содержат значительное количество неоднозначных случаев [Stamatatos, 2009]. Например, «meeting» в зависимости от контекста может быть как существительным, так и глаголом, а значит его леммы будет отличаться как «meeting» и «meet» соответственно.

В качестве альтернативы лемматизации иногда используют стемминг – отсечение от слова окончаний и суффиксов таким образом, чтобы оставшаяся часть была одинаковой для всех грамматических форм слова. Программная реализация такой процедуры значительно проще, однако анализатор стемминга допускает еще большее количество ошибок, поскольку не учитывает ряд морфологических изменений и может привести совершенно разные слова к одной форме. Например, слова «universal», «university» и «universe» алгоритм стемминга с высокой вероятностью сведет к одинаковой форме «univers».

Несмотря на определенную вероятность появления ошибочных данных, которые могут помешать анализу текста, лемматизация (иногда стемминг) является обязательным шагом предварительной обработки текста. Данный этап необходим для корректного извлечения стилеметрических параметров лексического уровня.

4. Токенизация. Текст разбивается на отдельные лексические единицы, чтобы система могла обращаться к каждому слову по отдельности. Этот процесс играет важную роль при выполнении многих задач, в частности, при подсчете общего количества слов или частоты встречаемости отдельных единиц [Kadhim, 2018].

5. Удаление стоп-слов и числительных. Стоп-слова – это наиболее распространенные слова, такие как артикли, предлоги и местоимения, которые не содержат в себе семантической информации и широко используются всеми авторами. Многие исследователи предлагают убрать

стоп-слова и числительные, чтобы сфокусироваться на лексике, определяющей смысл текста, и уменьшить количество нерелевантных характеристик [Kannan, Gurusamy, 2014].

6. Присвоение тегов части речи. Зачастую этот шаг объединяют с приведением слов к начальной форме: во время анализа лемматизатор в любом случае определяет часть речи, чтобы подобрать правильную лемму слова.

7. Обработка орфографических ошибок, аббревиатур и акронимов. Наиболее проблемным фактором анализа нехудожественных текстов, является наличие разговорных или официальных сокращений, а также присутствие орфографических ошибок [Eder, 2013]. Например, акроним «ASAP» с точки зрения компьютера не имеет ничего общего с «as soon as possible». Чтобы привести текст к каноническому виду, такие нестандартизированные слова должны быть преобразованы в их словарные формы, однако из-за количества возможных лингвистических вариаций эта задача далеко не тривиальна.

Данный этап предварительной обработки данных особенно важен при анализе текстов социальных сетей, твитов, форумов, SMS и других неформальных дискурсов, поскольку большая часть их текстового содержимого нестандартизирована [Clark, Araki, 2011].

В заключение необходимо отметить, что применение вышеупомянутых процедур предварительной обработки текста напрямую зависит от рассматриваемых стилеметрических параметров. Например, нельзя удалять стоп-слова, если целью исследования является анализ частотности местоимений, или нет смысла удалять знаки препинания, если их частотность рассматривается как часть авторского стиля.

### 1.3.2. Извлечение стилеметрических параметров текста

Как упоминалось ранее, для использования компьютерных систем автоматической атрибуции авторства текст сначала должен быть квантифицирован – представлен таким образом, чтобы компьютер мог «понять» его. Извлечение стилеметрических параметров позволяет представить текст в структурированном формате, в виде набора количественных характеристик, отображающих те или иные особенности стиля.

Выбор языковых характеристик, определяющих авторский стиль, является важнейшим этапом атрибуции авторства текста и одной из самых больших проблем стилеметрии. Исследователи выделяют около тысячи различных характеристик, используемых для представления авторского стиля на разных уровнях языка: лексическом, синтаксическом, семантическом и морфологическом. Такое многообразие характеристик отражает проблему сложности и многомерности текста, которая требует тщательной оценки параметров, отобранных для стилеметрического анализа, и их способности представить авторский стиль, для решения задачи идентификации автора [Juola, 2007; Neal et. al, 2017].

На сегодняшний день ученые не пришли к согласию по поводу оптимального набора стилеметрических параметров, наиболее полно и эффективно описывающих стиль автора. Их выбор в основном случаен и зачастую зависит от типа анализируемых текстов, применяемого классификатора и конкретной задачи.

Исследования, посвященные данной проблеме, делятся на два условных направления: поиск стилеметрических параметров, релевантных при решении конкретной исследовательской задачи, и поиск относительно универсального набора таких характеристик. В данном случае под универсальностью, как правило, понимается возможность эффективного

использования таких признаков для решения задач стилеметрии в пределах одного языка, но вне зависимости от жанра или стиля текста.

В данной работе мы рассмотрим классификацию стилеметрических параметров текста, предложенную З.И. Резановой и др. (2013). В отличие от большинства других классификаций, рассматривающих стилеметрические параметры со стороны информатики [Canales et. al, 2011; Stamatatos, 2009; Neal et. al, 2017], данная работа подходит к категоризации скорее с лингвистической точки зрения.

Полная классификация стилеметрических параметров представлена на рисунке 2. Используя данную схему в качестве логической основы, мы рассмотрим каждую из групп в отдельности для того, чтобы выделить потенциальные единицы анализа и возможные методы их статистической обработки.


Рисунок 2. Группы стилеметрических параметров текста, используемых для атрибуции авторства (по З.И. Резановой)

В общей сложности, все стилеметрические признаки можно разделить на две группы: признаки, анализирующие языковые элементы текстовой структуры, и собственно текстовые формальные характеристики, не использующие лингвистический анализ.

**Языковые элементы текстовой структуры**

1. *Элементы символьного уровня.*

Данная группа включает в себя стилеметрические признаки, основывающиеся на единицах лексического уровня, а также на отдельных знаковых элементах.

1.1. *Элементы плана выражения знака.* При извлечении признаков плана выражения знака текст рассматривается как формальная последовательность символов. К данной группе относятся такие признаки, как частота встречаемости отдельных символов, частота появления строчных и заглавных букв, цифр, знаков пунктуации и т.д.

Наиболее эффективным и всеобъемлющим признаком данной группы большинство исследователей считает символьные *n*-граммы, представляющие собой непрерывную последовательность из *n* символов [Stamatatos, 2009]. Значение *n* варьируется и обычно выбирается в зависимости от особенностей и свойств анализируемого языка.

Главным достоинством такого представления текста является его терпимость к орфографическим ошибкам и деформациям [Elberrichi, Aljohar, 2007]. Другими словами, в тех случаях, когда тексты содержат большое количество ошибок, что можно часто наблюдать в письмах электронной почты, текстах блогов, форумов и социальных сетей, статистические характеристики символьных n-грамм меняются не существенно. Так, при рассмотрении текста на уровне слов такие лексические единицы, как ‘acomodation’ и ‘accommodation’ будут восприняты машиной как абсолютно разные, что может привести к возникновению серьезной погрешности в текстах с большим количество ошибок. В то же время на уровне знаков ‘acomodation’ и ‘accommodation’ имеют большое количество совпадающих *n*-грамм, благодаря чему статистические показатели будут практически одинаковыми.

С другой стороны, повторяющиеся ошибки можно считать частью характерного стиля автора, что также дает подходу, позволяющему их анализировать, определенное преимущество. Так, с помощью символьных *n*-грамм можно различить американских и британских авторов по типичным

орфографическим различиям (например, ‘color’ – ‘colour’, ‘analyze’ – ‘analyse’). Для одного автора статистические параметры покажут высокую частоту триграмма ‘our’, а для другого будет характерно использование биграмма ‘or’. Лексический уровень не позволяет отразить данную информацию.

Многие исследователи отмечают, что данная группа признаков зачастую позволяет системе атрибуции достичь большей точности, чем признаки высоких уровней. Более того, данная закономерность наблюдается для многих языков, в числе которых русский [Litvinova T., Litvinova O., Panicheva, 2019], английский, китайский, греческий [Peng et. al, 2003] и др. Высокая релевантность и универсальность данных признаков обусловлена тем, что они относятся к низшему уровню языка и являются исключительно формальными. Таким образом, вероятность того, что автор будет сознательно контролировать их употребление в тексте, практически равна нулю. Благодаря этому проявление символьных признаков плана выражения знака в текстах разной жанровой и стилистической направленности одного автора можно считать инвариантным [Резанова и др., 2013].

Другим серьезным преимуществом данной группы признаков является простота анализа с точки зрения компьютерных алгоритмов. Так, для подсчета частоты встречаемости символа или символьных *n*-грамм не требуется предварительная обработка текста (например, лемматизация), а алгоритм позволяет достигнуть высокой точности даже для текстов, содержащих большое количество орфографических и пунктуационных ошибок, сленга или сокращений.

1.2. *Единицы лексического уровня.* Характерные особенности словаря автора одними из первых были использованы для установления авторства текста. Стилеметрические признаки данной категории направлены на отображение таких особенностей авторского стиля, как употребление автором определенных слов, богатство словарного запаса, анализ орфографических ошибок и длины используемых лексических единиц.

В лингвистических работах идиостиль автора чаще всего исследуется именно путем анализа используемых им лексических единиц, которые могут в той или иной мере отражать индивидуальность. В случае анализа текста экспертом-лингвистом характерные особенности лексики могут предоставить информацию об историческом периоде, когда был написан документ, о вероятном роде деятельности и образовании автора, а так же о географическом регионе, где автор вырос или жил.

В задаче установления авторства на лексическом уровне компьютерные системы, как правило, рассматривают текст как набор отдельных слов, игнорируя грамматические связи между ними, контекст и порядок слов. Такая модель представления текста получила название «мешок слов» или bag-of-words. Многие исследователи, использующие данную модель, достигают высокой точности систем атрибуции [Bozkurt, Baghoglu, Uyar, 2007; Abbasi, Chen, 2008]. Представление текста в виде мешка слов представляет собой словарь, содержащий все встреченные в текстах автора слова и меру их частотности. Для эффективного использования этой модели текст необходимо полностью нормализовать: преобразовать заглавные буквы в строчные, привести все словоформы к нормальной словарной лемме, удалить знаки препинания и стоп-слова. В результате преобразования текста в мешок слов можно получить информацию о наиболее частотных лексических единицах, используемых автором.

Необходимо отметить, что набор используемых слов может напрямую зависеть от обозреваемой темы, стиля и жанра документа. Так, когда тексты-образцы всех авторов посвящены одной теме, для каждого из них наиболее употребляемыми лексическими единицами будут контентно-специфические и служебные слова, которые нельзя отнести к признакам авторского стиля. Для того, чтобы решить данную проблему и исключить влияние общих контентно-специфических слов на процесс атрибуции совместно с составлением частотного словаря используется методика *tf-idf* взвешивания [Rajaraman, Ullman, 2011]. Для каждой лексической единицы составленного

словаря рассчитывается статистическая мера *tf-idf*, которая присваивает слову определенный вес на основании того, как часто оно встречается в документах одного автора и как редко используется другими авторами [Kim et. al, 2019]. Таким образом, слова, которые являются характерными для одного конкретного автора, но редко встречаются в текстах других авторов, будут иметь наибольший вес для алгоритма определения авторства.

Очевидным недостатком представленного выше подхода является отбрасывание контекста и связей между словами (например, фразовый глагол “give up” будет рассматривается как два отдельных слова “give” и “up”). Подобное допущение может внести определенную погрешность в систему атрибуции. Для решения данной проблемы вместо отдельных словарных единиц предлагается рассматривать *n*-граммы слов по аналогии с *n*-граммами символов [Stamatatos, 2009; Jockers, Witten, 2010; Tamboli, Prasad, 2019]. Кроме того, к *n*-граммам слов также можно применить *tf-idf* взвешивание для учета характерности анализируемых единиц. Такой подход учитывает контекст и в некоторых случаях может быть более точным, однако он обладает определенными недостатками. Полученное представление текста будет весьма разреженным, то есть многие триграммы или 4-граммы будут встречаться в тексте не более одного или двух раз, особенно если тексты короткие. Подобная статистика не предоставит исследователю никакой информации, но приведет к избыточности, что, в свою очередь, снизит скорость и эффективность работы модели атрибуции авторства.

*Измерение лексического разнообразия.* Люди с богатым словарным запасом, выражают свои мысли более емкими словами и фразами, их речь последовательна и выразительна, в то время как люди с небольшим словарным запасом зачастую повторяют одни и те же слова. В случае анализа текста лингвистом-экспертом такой критерий, как «лексическое разнообразие» речи автора чаще всего оценивается субъективно, однако в стилеметрии все характеристики текста имеют количественное представление. В системах атрибуции авторства для анализа словарного

запаса чаще всего используют «коэффициент лексического разнообразия» [Torruella, Capsada, 2013] и отношение слов, встретившихся в тексте один раз (*hapax legomena*) к размеру словаря автора или к общему количеству слов в тексте [Ali, Hussein, 2014].

Проблема таких мер заключается в том, что уровень лексического разнообразия в значительной степени зависит от длины текста: по мере увеличения длины текста, увеличивается и количество использованных уникальных слов. Для нормализации коэффициента лексического разнообразия по длине анализируемого текста учеными было предложены различные функции [Yule, 1944; Honore, 1979; Tweedie, Baayen, 1998], однако на сегодняшний день обоснованность таких мер остается под вопросом, в связи с чем использование критериев лексического разнообразия для решения задачи атрибуции авторства возможно только совместно с другими стилеметрическими признаками [Stamatatos, 2009].

Другая сложность анализа словаря заключается в том, что для флективных языков каждое из используемых автором слов необходимо привести к нормальной форме. Эта задача не является тривиальной, поскольку для автоматической лемматизации требуется разработка и применение соответствующих алгоритмов, учитывающих грамматические характеристики слова [Sanderson, Guenter, 2006].

В отдельную категорию характеристик, анализируемых на лексическом уровне, можно выделить *семантические признаки*. Такие признаки основываются на значение языковых единиц, фраз или предложений, однако автоматическое извлечение семантического значения слова представляет собой сложную задачу, требующую наличия комплексных инструментов обработки текста. Несовершенство таких инструментов приводит к возникновению погрешностей [Tamboli, Prasad, 2019].

На сегодняшний день существуют следующие методы извлечения семантической информации: анализ тональности, построение графов семантических зависимостей и латентно-семантический анализ. Наибольший

интерес представляют алгоритмы анализа тональности и модели векторного представления слов, основанные на дистрибутивной семантике.

*Анализ тональности* или *сентимент анализ* – класс методов, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки мнений авторов по отношению к объектам, речь о которых идет в тексте [Pang, Lee, 2008]. Анализ тональности не предполагает извлечения фактографической информации, он занимается только степенью эмоциональной окраски сообщений. Анализ тональности отслеживает не что говорят о каком-то человеке или явлении, а с какой эмоцией о нем говорят [Николаев и др., 2017]. Данные методы могут вычислять оценку позитивности или негативности высказываний, а также преобладающую в тексте эмоцию. Таким образом, склонность к положительной или отрицательной оценке событий и явлений или преобладающие эмоции будут являться частью психологического портрета автора и отражаться в его авторском стиле [Martins et. al, 2019]. Тем не менее, результаты сентимент анализа должны использоваться с определенной осторожностью и исключительно в совокупности с другими признаками, поскольку чаще всего их можно отнести к контентно-специфической информации, напрямую зависящей от тематики текста и описываемого в нем объекта [Patra et. al, 2013].

*Модели векторного представления слов (Word2vec)* основаны на дистрибутивной семантике, занимающейся вычислением семантической близости между лексическими единицами на основании их распределения в объемных текстовых корпусах [Islam, 2018]. Сначала с помощью нейронных сетей модель анализирует объемный корпус текстов, на основании которого рассчитывает семантическую близость слов и присваивает каждому слову такой вектор, что слова, использующиеся в одинаковых контекстах находятся рядом, а слова различных смыслов – далеко друг от друга. Затем полученные вектора могут быть использованы в задаче установления

авторства путем анализа синонимических конструкций и используемой лексики [Khatun et. al, 2019].

Изначально подход Word2vec был разработан для того, чтобы компенсировать недостатки представления текста в виде «мешка слов», а именно для того, чтобы учитывать контекст и порядок слов. Кроме того, этот подход позволяет машине одинаково воспринимать слова с орфографическими ошибками и их литературную норму: в таком представлении слова с ошибками имеют такие же векторы, как и «правильные», поскольку они обычно используются в одинаковом контексте.

Для реализации данного подхода, сначала необходимо предоставить модели обширный корпус текстов, который позволит ей получить информацию о нормальном использовании слов и контекстах, в которых они используются. После этого исследуемый текст можно будет преобразовать в векторы, и алгоритмы машинного обучения смогут анализировать данные. Существует два подхода к обучению модели Word2vec: предоставить тексты исследуемого корпуса (т.е. рассмотреть контекст употребления слов конкретно в анализируемых документах) или загрузить предварительно рассчитанные векторы слов (т.е. предположить, что в анализируемых текстах слова используются аналогично общему массиву документов, используемому для расчета). Как правило, расчет векторов из текстов анализируемого корпуса проводят, когда корпус достаточно обширен и содержит большое количество данных.

2. *Идиосинкразические признаки*. Идиосинкразические признаки: орфографические и грамматические ошибки, а также другие текстовые аномалии – выделяются в отдельную группу характеристик текста, поскольку они связаны с нарушением норм употребления единиц всех языковых уровней. Эксперты-криминалисты активно используют ошибки пунктуации, орфографии и построения предложений при проведении анализа авторства, поскольку такие отклонения от языковой нормы, как правило, являются

характерными признаками авторского стиля и зачастую совершаются бессознательно.

Тем не менее, в ряде случаев ошибки употребления языковых единиц нельзя рассматривать как надежную стилеметрическую характеристику. Например, когда есть основания полагать, что искомый автор имел намерение «обмануть» лингвистическую экспертизу и мог сознательно смоделировать данный аспект. Также данные приемы должны с осторожностью использоваться при анализе текстов цифрового пространства и различных жанров интернет-коммуникации, поскольку системы автоматической правки текста могут корректировать часть ошибок, влияя на репрезентативность данной характеристики, а в случае социальных сетей, блогов и т.д. орфографические и речевые ошибки могут быть результатом использования интернет-жаргона или осмысленного коверкания слов. Кроме того, исследование идиосинкразических признаков невозможно в случаях, когда текст подвергается редакторской и корректорской правке, а значит большая часть художественных текстов не может подлежать атрибуции, опирающейся на данные параметры [Резанова и др., 2013].

Автоматическое извлечение таких характеристик представляется затруднительным и возможно только с помощью средств проверки правописания, которые далеко не всегда могут обнаружить ошибки и еще реже предоставляют возможности их классификации [Koppel, Schler, 2003]. В связи с этим исследования, основанные на идиосинкразических признаках, немногочисленны. Так, М. Коппел и Дж. Шлер (2003) разработали модель атрибуции, основанную на идиосинкразических признаках, и достигли точности в 68% без использования каких-либо других стилеметрических признаков. Для выявления ошибок исследователи использовали стандартный инструмент проверки орфографии программы MS Word и авторские скрипты. Аналогичные результаты можно увидеть и в исследованиях [Chaski, 2001; Grant, Baker, 2001]. Достигнутые показатели точности считаются достаточно высокими и перспективным для атрибуции на основе одного признака,

однако из-за сложностей реализации такого анализа большинство разработанных систем атрибуции его не используют [Stamatatos, 2009].

В практической части данной работы мы также не будем рассматривать использование идиосинкразических признаков для решения задачи по определению авторства, поскольку извлечение таких признаков требует обширных навыков программирования и заслуживает рассмотрения в качестве отдельной темы исследования.

*3. Единицы грамматики текста.* Использование грамматических единиц текста для проведения стилеметрической экспертизы представляется достаточно многообещающим, поскольку данный уровень текста генерируется подсознательно и не контролируется автором направленно. Вследствие этого грамматические признаки проявляются инвариантно в текстах различных жанров и дискурсов и могут использоваться для отражения характерных особенностей авторского стиля. Грамматика текста представляет собой соотнесение морфологических признаков слов и их синтаксических позиций, определяемых структурой синтаксического целого – высказывания (предложения) [Резанова и др., 2013: 44]. В связи с этим на грамматическом уровне выделяют два типа признаков: морфологические и синтаксические.

К *морфологическим признакам*, чаще всего используемым в задаче атрибуции авторства, относятся распределение частей речи [Pillay, Solorio, 2010; Tanguy et. al, 2012; Tamboli, Prasad, 2019], полных грамматических классов и их сочетаний [Кукушкина, Поликарпов, Хмелев, 2001; Stamatatos, Fakotakis, Kokkinakis, 2001], а также n-грамм частей речи [Koppel, Argamon, Shimoni, 2002].

В отличие от элементов символьного уровня автоматическое выделение грамматических признаков представляет собой определенную сложность, поскольку грамматическая система разных языков может значительно различаться, а выделение самих признаков требует сложных программ анализа и обработки текста, учитывающих особенности

конкретного языка. Как правило, для большинства распространенных языков существуют морфологические анализаторы, автоматически определяющие морфологические характеристики слов с точностью до 99%. Благодаря этому представляется возможным использовать части речи и полный набор грамматических классов в качестве стилеметрических признаков.

Так, Кукушкина, Поликарпов и Хмелев (2001) представили исследование, в котором сравнили точность модели классификации, устанавливающей авторство на основании распределения биграмм символов (73%), частей речи (61%) и полных грамматических классов (4%). Интересно отметить, что точность, полученная благодаря биграммам частей речи, намного превышает значение, полученное на основании полных грамматических классов. По всей видимости, дополнительная информация о принадлежности единиц к конкретному семантико-грамматическому разряду является избыточной и не является инвариантной частью авторского стиля, в то время как распределение частей речи может быть использовано с достаточно высокой эффективностью.

*Синтаксические признаки* включают в себя характеристики словосочетаний и предложений, способов их образования и употребления. Считается, что авторы бессознательно используют схожие синтаксические паттерны, а значит – синтаксические признаки реже подвергаются намеренному моделированию и более надежны, чем элементы плана выражения знака [de Vel et. al, 2001; Tamboli, Prasad, 2013]. Однако, по аналогии с морфологическими признаками, для выделения синтаксических паттернов требуются комплексные анализаторы, способные достоверно определять границу высказывания и предложения. Такие анализаторы существуют, однако из-за присущей языковым единицам неоднозначности точность их работы остается под вопросом, а значит при выделении признаков, опирающихся на работу подобных анализаторов, может возникнуть существенная погрешность.

По этой причине в практике стилеметрического анализа чаще всего используют такие формальные признаки, как длина предложения [Argamon, Saric, Stein, 2003; Grieve, 2007; Bozkurt, Baghoglu, Uyar, 2007], знаки пунктуации, обусловливающие смысловое и функциональное членение предложения [Chaski, 2001; Grieve, 2007; Романов, Шелупанов, Мещеряков, 2011], а также частота использования отдельных служебных слов, отражающих грамматические отношения и связи [Zhao, Zobel, 2007; Graham, Hirst, Marthi, 2005; Hedegaard, Simonsen, 2011].

Несмотря на отсутствие семантической информации, *функциональные слова* являются неотъемлемой частью предложения, они выражают грамматические связи между словами, обычно используются бессознательно и не зависят от темы, а значит могут использоваться как стилеметрические характеристики [Segarra, Eisen, Ribeiro, 2015]. Кроме того, служебные слова встречаются в любых фрагментах текста намного чаще, чем характерная для автора лексика и являются более информативным признаком.

Впервые служебные слова были использованы для установления авторства «Записок Федералиста» Мостеллером и Уоллесом (1964). Исследование показало, что частота функциональных слов обладает высокой различительной способностью (например, “upon” встречается только 0.23 раза на 1000 слов в работах Мэдисона и примерно 3.24 раза в работах Гамильтона). Данная характеристика была признана эффективной для атрибуции авторства и стала активно использоваться в отечественных и зарубежных исследованиях, на сегодняшний день оставаясь одним из самых распространенных стилеметрических признаков [Stamatatos, 2009].

При работе с набором служебных слов необходимо учитывать различие грамматических структур языков. В аналитических языках большинство синтаксических связей передается порядком слов в предложении, а также с помощью предлогов и других служебных слов, в то время как во флективных языках служебные слова имеют меньшую роль в выражении синтаксических связей. Выбор конкретных служебных слов, частотность которых будет

использоваться в качестве стилеметрических признаков, обычно основывается на произвольных критериях, но требует определенных языковых компетенций [Резанова и др., 2013].

**Собственно текстовые признаки.**

1. *Структура и графическое оформление текста.*

Структурные и графические признаки в целом можно отнести к контентно-специфической группе, поскольку они в большей степени зависят от типа и канала коммуникации. Наиболее распространенными структурными признаками являются количество пробелов, предшествующих знакам пунктуации, отступы, позиции табуляции, разделение текста на фрагменты, наличие у фрагментов заголовков и их стиль. Однако большинство структурных признаков релевантно для текстов ограниченного набора дискурсов [Резанова и др., 2013]. Так, стилеметрическая экспертиза научных текстов в качестве признака может рассматривать цитирование автором других источников и способ указания ссылок. При анализе исходного кода компьютерных программ, учитываются особенности оформления комментариев, условных конструкций и циклов, количество пробелов или знаков табуляции в отступах при переходе к следующему уровню вложенного кода и т.д. [Oman, Cook, 1989].

Использование структурных и графических признаков наиболее широко распространено в экспертизе писем электронной почты и сообщений онлайн-форумов. Так, де Вель и др. (2001) при анализе писем электронной почты используют такие формальные характеристики, описывающие структуру документа, как количество строк, количество пустых строк, средняя длина строки (в символах) и количество символов табуляции. Аббаси и Чэнь (2005) к данному набору характеристик добавляют выбор параметров шрифта (тип, размер и цвет), наличие в тексте ссылок или встроенных изображений.

Чжэнь, Ли, Хуан (2006) в качестве структурных характеристик предлагают использовать общее количество предложений, строк и абзацев;

среднее количество предложений, символов и слов в одном абзаце; наличие приветствия, разделителей между абзацами и абзацных отступов; наличие и местоположение цитирования предыдущих писем (перед или после текста ответного письма); наличие телефона, адреса электронной почты или ссылки в подписи. Результаты данного исследования показывают, что добавление структурных признаков к другим наборам стилеметрических параметров позволило увеличить точность системы атрибуции авторства на 4–13% (в зависимости от языка и алгоритма классификации).

Структурные признаки приобретают особую актуальность в очень коротких текстах, где стилистические свойства не могут быть адекватно представлены с помощью языковых элементов. Здесь использование структурных признаков может предоставить дополнительную информацию об авторском стиле и существенно увеличить точность атрибуции [Zheng et al., 2006].

Необходимым условием использования признаков данной категории является наличие текстов-образцов, принадлежащих к тому же дискурсу, что и текст спорного авторства. Такое требование возникает из необходимости извлечь структурные характеристики текста, присущие непосредственно анализируемому каналу коммуникации. Выполнение данного условия в реальных задачах не всегда представляется возможным, кроме того, большинство исследователей фокусируется на разработке универсальных стилеметрических признаков, позволяющих описать авторский стиль вне зависимости от жанра, тематики или канала коммуникации [Stamatatos, 2009]. Кроме того, большинство современных программ и интернет-сервисов поддерживают автоматическое форматирование текста, в результате чего исходная разметка документа может быть изменена сервисом, его редакторами и модераторами, что лишает текст достоверных структурных признаков. Стоит также отметить, что автоматическое извлечение признаков, отражающих структуру текста, не является тривиальной задачей и требует точных компьютерных анализаторов [Резанова, 2013].

2. *Метаданные документа.*

В данном случае под метаданными понимаются дополнительные служебные данные, не относящиеся непосредственно к содержанию текста [Резанова и др., 2013]. Такие данные автоматически генерируются текстовыми редакторами и сервисами при создании документа. Так, редактор MS Word сохраняет служебную информацию о расположении программы и системном имени пользователя, дате создания документа, имени автора, истории изменений, изначальном имени файла, имени автора правок и др.

Необходимо отметить, что полагаться исключительно на метаданные не стоит, поскольку их, как и любые явно выделяющиеся характеристики, можно намеренно скрыть, изменить или смоделировать [Juola, 2008]. Характеристики данного типа могут сыграть ключевую роль при определении авторства, одна они выходят за рамки чисто лингвистической экспертизы текста и не будут рассматриваться в нашей работе.

Можно заметить, что стилеметрические параметры используемые для определения авторства текста, разноплановы и весьма разнообразны: от формального подхода к тексту и электронному документу до признаков, анализирующих высшие уровни языковой системы. На сегодняшний день не существует единого мнения по поводу эталонного набора стилеметрических параметров, поскольку в каждом частном случае их выбор зависит от материалов исследования: учитываются особенности языка, анализируемого дискурса, электронного представления текстов и ряд других факторов. Тем не менее, анализ современных работ показывает, что при построении систем автоматической атрибуции авторства наблюдается тенденция к упрощению. Большинство исследователей считает, что признаки низших уровней и формальное представление текста дают лучшие результаты, чем рассмотрение синтаксического или семантического уровней [Abbasi, Chen, 2008; Juola, 2008; Stamatatos, 2009; Резанова и др., 2013; Neal et al., 2017 и др.]. Вероятнее всего, такая закономерность вызвана особенностями компьютерной обработки естественного языка: извлечение признаков

высшего уровня требует сложных анализаторов и может вносить существенную погрешность, в то время как посимвольный анализ текста позволяет учесть все имеющиеся данные и не требует существенной предобработки.

При выборе стилеметрических параметров также важно учитывать используемую модель классификации: так, часть методов наиболее эффективно работает с обширным количеством параметров (сотни тысяч слов, символов и *n*-грамм), другие модели имеют жесткие ограничения по количеству анализируемых параметров, а подходы, основанные на сжатии текста, и вовсе предлагают пропустить этап классического извлечения стилеметрических характеристик.

В следующем параграфе мы рассмотрим используемые в автоматических системах атрибуции методы установления авторства текстов, особенности их применения, достоинства и недостатки. Анализ существующих методов позволит нам подобрать оптимальную модель классификации с учетом материалов и цели нашего исследования.

### 1.3.3. Формальные методы определения авторства текстов

В каждой задаче автоматической идентификации авторства существует набор авторов-кандидатов, охватывающий всех возможных авторов, набор образцов текста с достоверно установленным авторством (учебный корпус) и набор образцов текста неизвестного авторства (тестовый корпус); для решения задачи каждый из текстов тестового корпуса должен быть приписан одному из авторов-кандидатов.

В данной работе, мы вслед за Э. Стамататосом (2009) будем рассматривать классификацию методов атрибуции в зависимости от того, как они обрабатывают учебные тексты отдельных авторов – индивидуально или объединяя все доступные образцы в один файл. Ключевое различие заключается в том, что в некоторых подходах все доступные для каждого

автора учебные тексты объединяются в один большой файл, откуда уже извлекают совокупное представление стиля этого автора (так называемый «профиль»). Иными словами, различия между отдельными текстами, написанными одним и тем же автором, игнорируются. Такие подходы принято называть «основанными на составлении профиля».

Другая группа подходов для разработки точной модели атрибуции требует наличия нескольких обучающих образцов текста на одного автора, то есть каждый учебный текст рассматривается как отдельный экземпляр авторского стиля. Такие подходы называют «основанными на отдельных экземплярах класса».

**Подходы, основанные на составлении профиля (профильно-ориентированные подходы)**

Такой вариант работы с доступными учебных текстами предполагает, что все доступные для автора тексты будут объединены в один файл, который в последствии будет использоваться для извлечения свойств авторского стиля в соответствующий вектор. Затем анализируемый текст неизвестного авторства сравнивается с каждым из получившихся объединенных текстов, и наиболее вероятный автор оценивается на основе функции оценки расстояния между векторами извлеченных характеристик. Следует отметить, что отдельных представлений каждого из образцов текста нет, а хранится только один большой файл для каждого автора.

*1. Вероятностные модели*.

Одним из самых ранних подходов к идентификации автора является использование вероятностных моделей. Несмотря на раннее появление [Mosteller, Wallace, 1964], такие подходы до сих пор используется во многих современных исследованиях. Общий принцип работы вероятностных моделей можно описать следующим образом: модель атрибуции ищет автора, для которого значение метрики подобия, учитывающей логарифм вероятности авторства определенного кандидата по отношению к остальным, будет максимальным. Наиболее известными вероятностными

классификаторами являются Наивный Байесовский классификатор [Sebastiani, 2002] (в т.ч. модифицированный с помощью статистических моделей языка [Peng et al., 2004]); методики, основанные на цепях Маркова [Хмелев, 2000]; метод распознавания образов [Хозяинов, 2017].

*2. Модели сжатия данных.*

Наиболее эффективные из алгоритмов, использующих сжатие данных, относятся к подходам, основанным на составлении профиля [Хмелев, 2003; Кукушкина, Поликарпов, Хмелев, 2001; Marton, Wu, Hellerstein, 2005]. Такие модели не требуют векторного представления характеристик авторского профиля. Изначально все доступные для автора тексты объединяют в один файл $X_a$, затем используют алгоритм сжатия данных, в результате работы которого получают сжатый файл $C(X_a)$. После чего в текстовый файл добавляют текст неизвестного авторства $x$ и снова используют алгоритм сжатия, получая файл $C(X_a + x)$. Разница в битовом размере сжатых файлов $d(x, x_a) = C(x_a + x) - C(x_a)$ показывает степень сходства авторских стилей анализируемого и неизвестного авторов [Stamatatos, 2009: 547]. Даная разница определяет значение перекрестной энтропии двух текстов, и чем меньше значение, тем вероятнее, что текущий анализируемый автор является искомым. В рамках данного подхода было протестировано несколько распространенных алгоритмов сжатия, включая RAR, LZW, GZIP, BZIP2, 7ZIP и д. В большинстве случаев наиболее эффективный результат был достигнут при использовании алгоритма RAR [Хмелев, 2003; Marton, Wu, Hellerstein, 2005; Pavelec et al., 2009; Oliveira, Justino, Oliveira, 2013].

Подходы, основанные на сжатии данных, позволяют достигнуть высокой точности атрибуции авторства и достаточно легки в использовании из-за отсутствия необходимости подготавливать текст для обработки. Тем не менее, их существенным недостатком является длительное время работы алгоритма при больших объемах корпуса.

3. *Модели совпадающих n-грамм*.

Впервые метод совпадающих *n*-грамм был описан В. Кешелем и др. (2003). В данной методике профиль автора представляется в виде $L$ наиболее частотных *n*-грамм встречающихся в его текстах. Затем полученный профиль сравнивается с профилем текста спорного авторства и вычисляется относительная разница между совпадающими в профилях *n*-граммами. Каждый *n*-грамм не являющийся общим для обоих профилей, увеличивает значение дистанции между ними. Здесь важно отметить, что эффективность работы метода напрямую зависит от параметров $L$ – количество *n*-грамм в авторском профиле и $n$ – длина самих *n*-грамм. В работе В. Кешеля и др. (2003) наибольшая точность атрибуции была достигнута при значениях $1000 \leq L \leq 5000$и $3 \leq n \leq 5$.

Рассматривая модели совпадающих *n*-грамм, также необходимо упомянуть об одной из ключевых проблем атрибуции авторства – *дисбалансе классов*. Данная проблема возникает, когда тексты корпуса распределены между авторами-кандидатами неравномерно, то есть для одних авторов имеется большое количество лингвистического материала, в то время как другие авторы представлены лишь небольшими текстовыми отрывками. В реальной жизни, например, при решении задачи идентификации автора анонимного текста в криминалистике, практически невозможно получить равное количество текстовых образцов для каждого из подозреваемых. В то же время большинство научных исследований основано на сбалансированных наборах данных, где предоставляется примерно одинаковое количество текстового материала для каждого автора-кандидата. На основании результатов таких исследований едва ли представляется возможным предсказать точность методов атрибуции при решении реальных задач [Stamatatos, 2009].

Метод совпадающих *n*-грамм позволяет достигнуть высокой точности, когда учебный корпус относительно сбалансирован, однако в несбалансированных случаях, когда профиль хотя бы одного из авторов короче $L$, показатели резко ухудшаются [Stamatatos, 2007]. Например при $L =$

$3000$ и $n = 4$, в условиях малого количества материала для определенного автора, количество *4*-грамм в его профиле может быть меньше 3000. В такой ситуации модель атрибуции в большинстве случаев будет приписывать тексты данному автору, поскольку значение результирующей функций, оценивающей количество несовпадающих *n*-грамм будет ниже, а значит – профили авторов будут считаться схожими.

Для решения данной проблемы было предложено несколько модификаций функции, определяющей схожесть профилей. Дж. Франческу и др. (2006) предложили метод расчета, названный «упрощенное пересечение профилей». В таком подходе предлагается учитывать только количество совпадающих *n*-грамм, игнорируя остальные показатели. Интересно, что такой упрощенный вариант функции показал большую эффективность, чем оригинальный подход за исключением случаев, когда для всех авторов-кандидатов кроме одного представлены короткие тексты-образцы. В таком случае результирующая функция будет отдавать предпочтение автору длинного обучающего текста.

Последующая модификация функции расчета дистанции между профилями авторов была предложена Э. Стамататосом (2007). Здесь в формулу было введено понятие корпусной нормы, представляющей собой тексты всех авторов, и частоты встречаемости отдельных *n*-грамм в такой корпусной норме. В результате для измененной функции в значение дистанции между профилями стали вносить вклад исключительно *n*-граммы текста неизвестного авторства. Подобная модификация помогла решить проблему дисбаланса классов, однако в сбалансированных случаях оригинальная методика более эффективна [Stamatatos, 2007].

В отечественной лингвистике эксперименты по определению авторства на основе метода совпадающих *n*-грамм с использованием различных метрик можно найти в работе Леоновой А.В. и Леоновой И.В. (2018).

**Подходы, основанные на отдельных экземплярах класса**

Большинство современных систем идентификации авторства рассматривает каждый образец обучающего текста как отдельную единицу, которая вносит свой вклад в модель атрибуции. Каждый из образцов текста преобразуют в вектор атрибутов, выбранных исследователем, а алгоритм классификации обучается на данных векторах и строит модель атрибуции, способную идентифицировать неизвестного автора.

Для создания надежной модели атрибуции таким алгоритмам требуется несколько текстов-образцов для каждого автора-кандидата. В случаях, когда имеется только один, но довольно длинный текст-образец стиля конкретного автора-кандидата (например, целая книга), его необходимо искусственно разделить на несколько предположительно равных частей. С другой стороны, при наличии нескольких образцов обучающего текста разной длины требуется нормализация, для чего обучающие тексты каждого автора сегментируются на выборки одинакового размера [Sanderson, Guenter, 2006].

Важно отметить, что индивидуальные образцы должны быть достаточно длинными, чтобы лингвостатистические параметры представления текста могли адекватно отобразить авторский стиль. Различными исследователями был проведен ряд экспериментов по установлению зависимости между точностью атрибуции и длиной отдельного текстового примера [Sanderson, Guenter, 2006; Koppel et al., 2007; Hirst, Feiguina, 2007] Так, Г. Херст и О. Фейгина (2007) провели эксперименты с текстовыми блоками различной длины (200, 500 и 1000 слов) и сообщили о значительном снижении точности по мере уменьшения длины текстового блока. Таким образом, выбор обучающего экземпляра текстовой выборки не является тривиальным и напрямую влияет на производительность модели атрибуции [Stamatatos, 2009].

*1. Модели векторных пространств.*

Учитывая, что в классической модели атрибуции обучающие тексты представлены в виде вектора стилеметрических характеристик, каждый текст

можно рассматривать как вектор в многомерном пространстве. Для построения классификационной модели в таком случае могут быть использованы различные статистические алгоритмы или алгоритмы машинного обучения: включая дискриминантный анализ [Ермолаева, 2009; Bagavandas, Manimannan, 2008; Chaski, 2005], метод опорных векторов [Li, Zheng, Chen, 2006; Sanderson, Guenter, 2006; Романов, Мещеряков, 2009; Awad, Khanna, 2015], деревья принятия решений [Uzuner, Katz, 2005; Zhao, Zobel, 2007], нейронные сети [Khosmood, Levinson, 2006; Zheng et al., 2006] и др.

Несомненным достоинством алгоритмов данного класса является способность эффективно обрабатывать многомерные, разрозненные данные, содержащие в себе определенное количество помех, что позволяет более полно представлять тексты в виде векторов. Например, метод опорных векторов не сталкивается с проблемой переобучения даже при использовании нескольких тысяч функций и считается одним из лучших решений в современной практике атрибуции авторства [Li, Zheng, Chen, 2006; Sebastiani, 2002]. Тем не менее, эффективность данного метода уменьшается при нарастании классового дисбаланса.

*2. Модели, основанные на мере сходства.*

Основная идея таких моделей заключается в вычислении попарных мер сходства между текстом неизвестного авторства и каждым из текстов-образцов, а затем наиболее вероятный автор оценивается с помощью алгоритма поиска ближайшего соседа.

Одна из наиболее известных методик, основанных на мере сходства, была предложена Джоном Барроузом (2002). Исследователь назвал свой метод «Дельта». Данный метод подразумевает вычисление относительных частот встречаемости набора служебных слов (обычно 150 самых частотных) в каждом из текстов. Затем для каждого текста вычисляется отклонение частоты каждого слова от нормы с помощью *z*-оценки, которая указывает, используется ли в данном тексте слово больше (положительная *z*-оценка) или

меньше (отрицательная *z*-оценка) раз, чем в среднем. Затем проводится собственно дельта-измерение, отображающее разницу между набором обучающих текстов, написанных одним и тем же автором, и неизвестным текстом. Значение дельта вычисляется как среднее модулей разности между z-оценками набора функциональных слов в обучающих текстах и соответствующими *z*-оценками неизвестного текста. Чем меньше показатель дельта, тем выше вероятность совпадающего авторства у сравниваемых текстов [Будаев, 2017: 45].

Дельта-расстояние Барроуза оказалось довольно эффективным при решении проблем авторской атрибуции и позволило использовать комбинированные различия в сотнях и тысячах слов одновременно. Метод, основываясь на частотах первых сотен самых распространенных служебных слов, при атрибуции авторства демонстрирует очень высокую точность (90% и выше) [Шеля, Плехач, Зеленков, 2020; Hoover, 2004]. Однако, следует заметить, что точность метода зависит от размера выборок — чтобы различия в частотах слов начали проявляться и складываться в устойчивые закономерности, нужен достаточно объемный текст (приблизительно в диапазоне между 2000 и 5000 слов [Eder 2015]).

Среди других подходов, основывающихся на мере сходства, можно отметить предложенную Дарио Бенедетто и др. (2002) методику вычисления схожести текстов. Исследователь предлагает использовать сжатие данных в подходах, основанных на отдельных экземплярах класса, с помощью вычисления меры сходства. Таким образом, сначала необходимо сжать отдельные образцы текстов с помощью стандартного алгоритма сжатия (GZIP), а затем сжать каждый из образцов текста совместно с текстом неизвестного авторства и вычислить разницу в их битовых размерах. Итоговое решение о наиболее вероятном авторе принимается алгоритмом поиска первого ближайшего соседа.

Данный подход подвергся широкой критике со стороны исследователей [Khmelev, Teahan, 2003], поскольку из-за необходимости

дважды использовать алгоритм сжатия для каждого из текстов образцов, программа атрибуции работает чрезвычайно медленно. Кроме того, стандартные модели сжатия, придерживающиеся профильно-ориентированного подхода, позволяют достигнуть более высокой точности при гораздо меньшем времени работы [Marton, Wu, Hellerstein, 2005].

Альтернативный способ вычисления меры сходства для комбинации с подходами сжатия данных предложили Пол Витаньи и Руди Цилибрази (2005). Основываясь на понятии Колмогоровской сложности, они ввели нормированное расстояние сжатия и испытали свою методику для классификации текстов русскоязычных авторов, достигнув выдающихся показателей точности системы.

Таким образом, каждый из методов имеет свои особенности и с разной степенью успешности используется для атрибуции авторства. С лингвистической точки зрения подходы, основанные на сжатии данных, представляют собой наименьший интерес, поскольку они не предусматривают этапа извлечения стилеметрических характеристик и не позволяют проанализировать их эффективность. Модели совпадающих *n*-грамм и метод дельта-расстояния рассматривают стилеметрические характеристики лишь частично.

Для рассмотрения обширного набора стилеметрических параметров наиболее оптимальным выбором представляется модель классификации, использующая машинное обучение. В таком случае алгоритм классификации может носить как вероятностный характер, так и геометрический (модели векторных пространств). Теоретическое сравнение данных алгоритмов требует определенных математических навыков и ввиду различной природы алгоритмов не может однозначно определить наиболее эффективный алгоритм классификации. Большинство экспертов соглашается с тем, что эффективность алгоритма классификации зависит от количества и типа анализируемых характеристик, а также от вида задачи: так, считается, что Наивный Байесовский классификатор лучше справляется с задачей

категоризации текстов по тематике, в то время как метод опорных векторов (SVM) работает эффективнее в классификации текста, в частности, при атрибуции авторства [Zheng et al., 2006; Stamatatos, 2009; Романов, Мещеряков, 2010; Neal et al., 2017].

В качестве классификатора, проводящего атрибуцию, в нашем исследовании будет использоваться модель машинного обучения, основанная на методе опорных векторов (SVM). В следующем параграфе мы рассмотрим теоретическое обоснование выбранного решения, а также особенности его работы в задаче установления авторства.

### 1.3.4. Машинное обучение и метод опорных векторов

В общем виде задача машинного обучения выглядит следующим образом: имеется некоторое множество, называемое множеством объектов. Каждому из объектов по какой-то системе приписывается признак из множества, именуемого множеством ответов. Систему, по которой объекту приписывается ответ, называют целевой функцией. В некоторых задачах такая функция представляет собой «черный ящик» – для каждого конкретного объекта можно сказать, какой именно ответ дает целевая функция, но сам принцип описать либо очень трудно, либо вовсе невозможно. Функция, имитирующая целевую, подбирается из некоторого ограниченного множества. Выбор этой функций и подбор ее параметров и осуществляются одним из алгоритмов машинного обучения [Николаев и др., 2017].

Подходы, основанные на использовании машинного обучения делятся на машинное обучение с учителем (*supervised*) и машинное обучение без учителя (*unsupervised*).

*Алгоритмы обучения без учителя* или неконтролируемого обучения – это еще один вид алгоритмов, в которых известны только объекты, а ответов нет. Данные алгоритмы, как правило выполняют задачу кластеризации –

выборка объектов на усмотрение классификатора разбивается на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались. Хотя есть много успешных сфер применения этих методов, их, как правило, труднее интерпретировать и оценить. Машинное обучение без учителя зачастую используется для решения таких задач, как определение тем в наборе записей или сегментирование клиентов на группы с различными предпочтениями [Jockers, Witten, 2010].

*Алгоритмы обучения с учителем* или контролируемое обучение – это вид алгоритмов, в которых пользователь предоставляет пары объект-ответ для всех возможных ответов, а алгоритм анализирует их по заданным параметрам и находит некоторые зависимости. Концепция контролируемого машинного обучения представлена на рисунке 3.

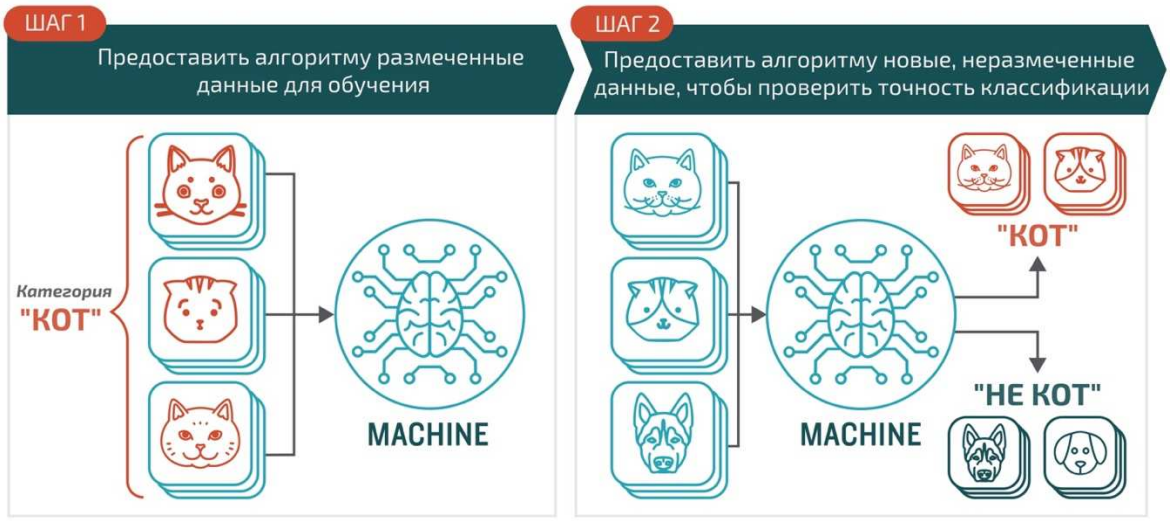


Рисунок 3. Концепция машинного обучения с учителем. Источник: boozallen.com

Исследования, посвященные атрибуции авторства, как правило используют машинное обучение с учителем, поскольку в таких задачах имеется заранее размеченный корпус примеров установленного авторства для всех возможных авторов кандидатов.

В общем случае можно сказать, что алгоритм выводит функцию $y = f(x)$, которая сопоставляет входные данные $x$ с выходной меткой $y$ [Awad, Khanna, 2015]. Такие алгоритмы называют контролируемыми, потому что они подбирают функцию, наблюдая за заданным набором обучающих данных. Поскольку правильные ответы известны, алгоритм итеративно делает предположения о принадлежности объекта к тому или иному классу, и в случае ошибки корректируется «учителем». К концу обучения алгоритм имеет некоторую функцию, определяющую связь между входными данными и выходными метками. В результате алгоритм способен без стороннего вмешательства выдать ответ для объекта, который он никогда раньше не встречал.

Несмотря на то, что создание набора с объектами и ответами зачастую требует существенных усилий и временных затрат, алгоритмы обучения с учителем более интерпретируемы и качество их работы легко оценить. При использовании такого подхода задача идентификации автора сводится к задаче классификации текстов, которая может быть решена путем обучения классификатора на размеченной выборке [Мюллер, Гвидо, 2017].

Для наглядной демонстрации работы алгоритма рассмотрим упрощенный пример. Предположим, у нас есть обучающие данные, содержащие частоту встречаемости цифр в тексте и имена авторов этих текстов (таблица 1). Часть таких данных будет использоваться для обучения (тренировочный набор), часть – для последующей оценки точности (тестовый набор).

Итеративный процесс обучения может проходить следующим образом:

(1) $x = 0.17$, метка «Дерек» – Хорошо, тогда для всех $x = 0.17$ в будущем метка будет «Дерек».

(2) $x = 0.08$, метка «Дерек»? – Нет, «Эшли». – Хорошо, тогда возьмем среднее значение: $(0.17 + 0.08)/2 = 0.125$, для всех $x \leq 0{,}125$ метка будет «Эшли», в противном случае – «Дерек».

(3) $x = 0.163$, метка «Дерек»? – Да. – Хорошо, существующая функция работает корректно, никаких изменений.

(4) $x = 0.181$, метка «Дерек»? – Да. – Хорошо, существующая функция работает корректно, никаких изменений.

(5) $x = 0.13$, метка «Дерек»? – Нет. – Хорошо, тогда возьмем среднее значение: $(0.13 + 0.163)/2 = 0.1465$, для всех $X \leq 0.1465$ метка будет «Эшли», в противном случае – «Дерек».

Все данные тренировочного набора рассмотрены, результирующая функция: «Для всех $X \leq 0.1465$ метка «Эшли», в противном случае – «Дерек».

Таблица 1. Пример входных данных алгоритма машинного обучения

| | Частотность цифр, % (*x*) | Автор (*y*) |
|---|---|---|
| Тренировочный набор | | |
| Текст 1 | 0.17 | Дерек |
| Текст 2 | 0.08 | Эшли |
| Текст 3 | 0.163 | Дерек |
| Текст 4 | 0.181 | Дерек |
| Текст 5 | 0.13 | Эшли |
| Тестовый набор | | |
| Текст 1 | 0.09 | Эшли |
| Текст 2 | 0.14 | Дерек |
| Текст 3 | 0.19 | Дерек |

В дальнейшем для оценки эффективности работы системы имитируется реальная задача классификации, то есть алгоритм получает набор тестовых данных, которые он раньше не встречал, и на основании выведенной функции присваивает ответ каждому из образцов. После этого ответы классификатора сравниваются с эталонными и оценивается результирующая точность [Sebastiani, 2002].

В рассматриваемом примере алгоритм классифицирует данные тестового набора следующим образом:

(1) $x = 0.09$ – «Эшли»

(2) $x = 0.14$ – «Эшли»

(3) $x = 0.19$ – «Дерек»

Предсказания (1) и (3) верны, в то время как (2) ложно, следовательно точность рассматриваемой модели составляет $2/3 \approx 67\%$, а значит, правильно будет классифицировано примерно 67% текстов.

Описанный выше пример, безусловно, представляет собой крайне упрощенную версию работы алгоритма, однако позволяет получить общее представление о процессе машинного обучения. Число итераций зависит от количества пар объект-ответ в наборе обучающих данных, и чем обширнее тренировочный набор, тем точнее результирующая функция.

В рассмотренном примере итоговое правило представляет собой простейшую линейную функцию, которую легко можно вывести в результате обычного наблюдения, однако в реальных практических задачах авторство не может быть установлено на основании одного признака. Количество характеристик, составляющих авторский стиль, ошеломляет, и вручную вывести правило, описывающее связь между сотнями стилеметрических параметров и каждым из авторов, не представляется возможным. Таким образом, большое число параметров и существенные объемы обрабатываемых данных делают машинное обучение важным инструментом классификации.

**Метод опорных векторов**

Ключевым моментом в построении модели машинного обучения является выбор непосредственно алгоритма, который определяет способ обучения системы. Существует широкий спектр алгоритмов контролируемого обучения, каждый из которых имеет свои достоинства и недостатки. Не существует единого алгоритма, обеспечивающего

наибольшую эффективность для любых задач [Jockers, Witten, 2010]. Выбор алгоритма в значительной степени зависит от набора данных.

Согласно Т. Иоахимсу (1998), при выборе метода атрибуции для классификации текстов следует учитывать следующие характеристики текста:

– *высокая размерность анализируемого пространства*. Точное представление текста обычно требует тысяч стилеметрических параметров, и текстовый классификатор должен быть способен обработать такой объем данных. Также стоит учитывать, что благодаря высокой размерности, большинство задач классификации текста *линейно разделимы,* то есть в многомерном пространстве можно найти такую гиперплоскость, для которой все точки одного класса будут расположены с одной стороны, а другого — с другой.

– *малое количество нерелевантных параметров*. Для снижения размерности данных обычно предполагается, что большинство параметров является нерелевантным и их можно исключить из анализа. Однако, в вопросах атрибуции текста большая часть параметров несет в себе информацию об авторском стиле, и выбор ограниченного количества важных и релевантных характеристик может привести к потере информации и, следовательно, ухудшить эффективность системы.

– *разреженное векторное представление.* Как правило, большая часть координат векторного представления документа равна нулю (например, частотность вхождения редких слов, встретившихся в других документах). Это создает определенные сложности в процессе обработки и приводит к чрезмерному использованию памяти. Не все алгоритмы классификации способны обрабатывать такой формат данных.

В нашем исследовании мы используем *машинный классификатор с линейным алгоритмом опорных векторов (SVM)*.

Метод опорных векторов – это алгоритм контролируемого машинного обучения, в котором каждый элемент данных отображается в виде точки в *n*-

мерном пространстве (*n* – общее количество параметров), причем значение каждого из параметров является значением соответствующей координаты. Классификация выполняется путем нахождения гиперплоскости, которая наилучшим образом разделяет набор данных на классы [Шеля, Плехач, Зеленков, 2020]. В нашей работе не рассматриваются математические принципы SVM, формулы, лежащие в основе процесса нахождения гиперплоскости, можно найти в Авад, Ханна (2015).

SVM является оптимальным решением для классификации текстов, поскольку он учитывает все вышеперечисленные особенности текстовых данных: SVM имеет защиту от перенасыщения, благодаря чему он может обрабатывать объекты высокой размерности; существуют как теоретические, так и эмпирические доказательства того, что метод SVM подходит для обработки разреженных векторов; и, наконец, сама концепция SVM заключается в поиске линейного разделителя, который обычно хорошо работает с высокоразмерным представлением текста. Благодаря этим особенностям многие авторы признают метод опорных векторов одним из наиболее эффективных и универсальных решений для атрибуции текстов [Joachims, 1998; Sebastiani, 2002; Juola, 2008; Stamatatos, 2009; Jockers, Witten, 2010 и др.]

На рисунке 4 представлена графическая демонстрация тренировочных данных и результирующая функция SVM для рассмотренного выше примера. В этом случае у нас есть только один параметр – частотность цифр, поэтому разделитель представляет собой линию, проведенную на одной координате.

На рисунке 5 продемонстрирован несколько более сложный случай. В трехмерном пространстве отображены 3 параметра: частотность цифр, запятых и восклицательных знаков. На рисунке отчетливо видно, что точки, соответствующие текстам Дерека, в большинстве случаев ближе друг к другу, а не смешаны с текстами Эшли. Кроме того, заметно, что эти две выборки можно линейно разделить, построив плоскость в трехмерном пространстве.


Рисунок 4. Диаграмма процесса обучения SVM для рассматриваемого примера

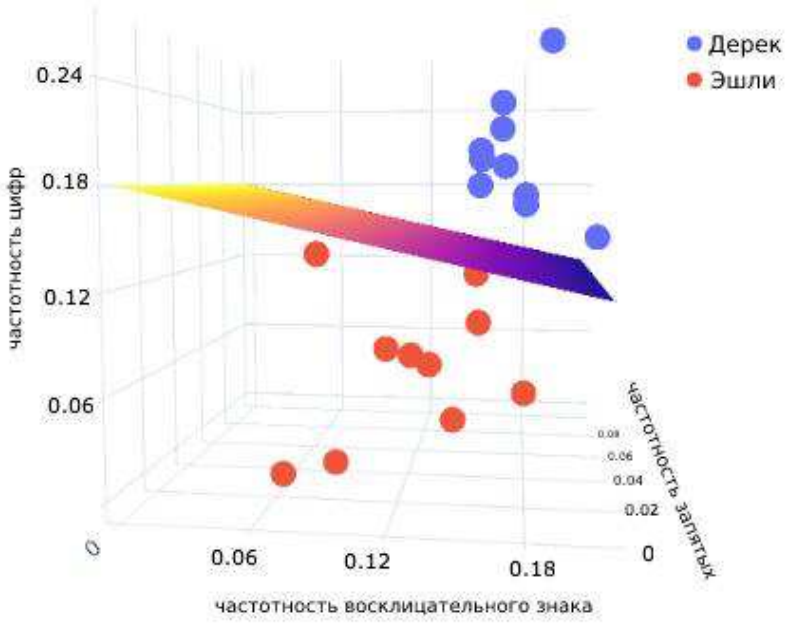


Рисунок 5. Диаграмма работы SVM для случая трехмерного пространства

Для нахождения оптимального разделителя гиперплоскость строится при помощи опорных векторов, что и обуславливает название метода. В процессе фактической атрибуции авторства тексты присваиваются авторам в зависимости от их положения относительно гиперплоскости. Аналогичным

образом алгоритм работает и для более высоких размерностей недостижимых человеческому представлению, находя разделяющую гиперплоскость для заданного числа анализируемых параметров.

**Оценка качества атрибуции**

Существует два подхода к оценке точности алгоритмов контролируемого машинного обучения: разделение заранее размеченных данных на тренировочный и тестовый набор либо подход перекрестной проверки по *k* блокам.

*Разделение на тренировочный и тестовый набор*. Исходный корпус данных делится на два набора: обучающий и тестовый. Классификатор итерационно анализирует параметры текстов обучающего набора (учится), а затем тестовый набор используется для оценки того, насколько хорошо модель классифицирует неизвестные данные. Как правило, тексты исходного корпуса делят в соотношении 70-80% текстов – для тренировочного набора и оставшиеся 20-30% данных – для тестирования [Sebastiani, 2002].

Такой подход к валидации результатов был проиллюстрирован в вышеприведенном примере: мы разделили исходные данные на тренировочный набор, содержащий 5 образцов, и тестовый набор, содержащий 3 текста. Алгоритм прошел обучение на тренировочных данных, а итоговая оценка была рассчитана на основании точности классификации для 3 текстов тестового набора. В результате точность классификатора была оценена в 67%, но очевидно, что мы могли получить совершенно другой показатель, если бы разделили исходные данные иным образом. Для того, чтобы избежать подобных неточностей, была разработана перекрестная проверка по *k* блокам.

*Перекрестная проверка по k блокам.* Идея перекрестной валидации заключается в повторении разделения исходных данных на тренировочный и тестовый наборы *k* раз. Исходный корпус случайным образом разбивается на *k* одинаковых по размеру блоков. Один из блоков используется в качестве

тестового набора, а остальные – для обучения. Этот процесс повторяется $k$ раз, пока каждый из блоков не будет использован в качестве тестового набора. Итоговая оценка точности рассчитывается как среднее значение всех показателей [Arlot, Celisse, 2010]. Рисунок 6 иллюстрирует схему 4-кратной перекрестная проверки данных.

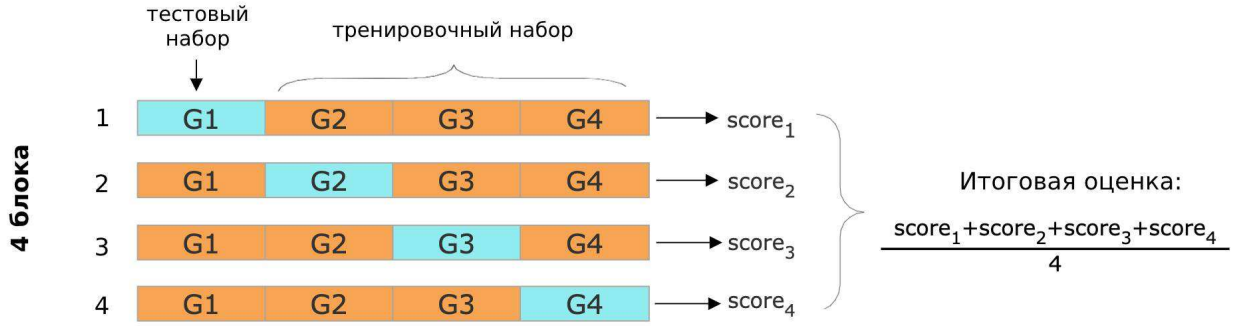


Рисунок 6. Схема 4-кратной перекрестной проверки

Перекрестная проверка считается более точным методом, поскольку она дает возможность обучать классификатор на разных множествах обучающих наборов и предоставляет информацию о том, как классификатор будет работать на различных тестовых данных. С другой стороны, перекрестная проверка требует большей вычислительной мощности и выполняется в $k$ раз дольше, так как классификация выполняется $k$ раз. Такая временная прогрессия может быть серьезным ограничением для больших наборов данных [Berrar, 2019].

В нашем исследовании в качествен меры оценки точности работы классификатора мы будем использовать перекрестную проверку с $k = 10$, поскольку такой подход дает лучшее понимание того, насколько хорошо работает классификатор, а вычислительное время и мощность для нашего корпуса являются приемлемыми.

## ВЫВОДЫ ПО ГЛАВЕ 1

Рассмотрев различные подходы к определению понятия «стиль», можно сделать вывод о том, что используемые автором языковые средства в большей степени являются подсознательным фактором и представляют собой систему периодически повторяющихся выборов. Таким образом, автор оставляет в своих произведениях неповторимый паттерн, и статистические методы анализа стиля могут быть использованы для установления авторства текста.

Проблема атрибуции авторства текстов давно привлекает внимание исследователей. До появления методик компьютерной обработки естественного языка процедура идентификации автора спорного текста, как правило, сводилась к тому, что филолог-литературовед, хорошо знакомый с творчеством писателей рассматриваемого периода, пытался найти какие-либо доказательства, позволяющие приписать анализируемый текст конкретному автору. При таком подходе субъективные факторы зачастую играли решающую роль, а определить авторство анонимных сообщений или работ, написанных людьми, не имеющими непосредственного отношения к литературе, не представлялось возможным. На сегодняшний день наиболее оптимальным подходом к решению задачи атрибуции авторства является стилеметрия или лингвистико-статистический метод, поскольку сравнение языка и стиля производится исключительно при помощи объективных характеристик текста и не зависит от субъективных факторов.

В рамках стилеметрии задачу идентификации автора можно рассматривать как задачу автоматической атрибуции текста с использованием стилеметрических параметров и алгоритма машинного обучения. Для реализации такой системы необходимо провести предварительную обработку корпуса текстов, выбрать и рассчитать стилеметрические параметры, а также настроить непосредственный классификатор. На этапе предварительной обработки текста необходимо

учесть особенности исследуемого текстового материала и планируемые к анализу стилеметрические параметры.

Выбор параметров, отображающих языковые особенности авторского стиля, является важнейшим этапом установления авторства текста, так как именно они повлияют на конечный результат и принятие решения о личности автора. Стилеметрические параметры могут быть получены путем анализа различных уровней языка: фонемного, лексического, синтаксического и морфологического, однако большинство исследований сходятся во мнении, что анализ признаков низшего уровня является более эффективным. В связи с этим в нашей работе наибольшее внимание будет уделено элементам символьного уровня.

В результате анализа формальных методов атрибуции авторства в соответствии с целью нашего исследования и с учетом основных характеристик представления текста и особенностей решаемой задачи в качестве алгоритма классификации был выбран метод опорных векторов, способный быстро и эффективно обрабатывать большое количество параметров. Для оценки результатов работы модели будет использоваться 10-кратная перекрестная проверка как более точный показатель по сравнению с единичным разделением.

# ГЛАВА 2. ОЦЕНКА ЭФФЕКТИВНОСТИ ИСПОЛЬЗОВАНИЯ СТИЛЕМЕТРИЧЕСКИХ ПАРАМЕТРОВ ТЕКСТА ДЛЯ РЕШЕНИЯ ЗАДАЧИ АТРИБУЦИИ АВТОРСТВА

## 2.1. Материалы исследования

Большинство исследований рассматривает задачу атрибуции для малого набора авторов-кандидатов: от двух до десяти человек. Безусловно, малое количество претендентов значительно облегчает процесс установления авторства, однако одной из задач нашего исследования было проведение нетривиального эксперимента, когда число возможных авторов относительно велико. Вместе с тем, было принято решение ограничить тематику корпуса для того, чтобы исключить влияние темы на процесс атрибуции авторства.

Принимая во внимание вышеизложенные факторы, в качестве материала исследования были выбраны посты популярных англоязычных блогов, посвященных путешествиям и размещенных на платформе WordPress. Выбор формата блогов и туристической тематики обусловлен относительной доступностью и обширностью эмпирического материала. В результате для анализа нами был собран корпус из текстов 50 авторов, порядка 3000 слов на каждого (рисунок 7). Полную информацию о корпусе (имена авторов, источники и общее количество слов) можно найти в приложении А.

Отдельно следует отметить, что в процессе сбора материала мы постарались дополнительно сузить тематическое разнообразие. Были выбраны записи, в основном посвященные общим вопросам путешествий, таким как «советы» и «лайфхаки», чтобы избежать специфической лексики (например, топонимов или имен собственных) и уменьшить влияние факторов, связанных с темой, на окончательную модель. Тем не менее, для некоторых авторов было невозможно подобрать тексты заданной узкой

тематики, в связи с чем в корпусе также присутствуют тексты, посвященные отдельным направлениям путешествий, событиям, местной кухне и т.д.

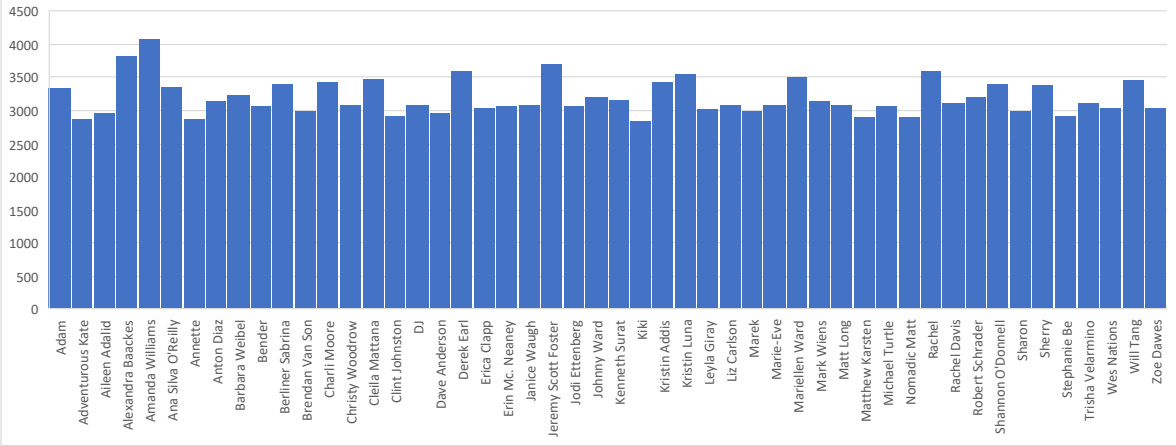


Рисунок 7. Распределение количества слов по авторам анализируемого корпуса

Для того чтобы исследуемый корпус был сбалансированным и общее количество слов на одного автора сохранялось примерно равным, текстовые данные отдельного автора могут состоять как из нескольких коротких постов, объединенных в один документ, так и из одной полной записи в блоге или даже ее части необходимой длины (в зависимости от доступных материалов). Так, если конкретный автор размещает исключительно короткие посты, состоящие из 200 – 300 слов, результирующий документ будет содержать 10 – 15 объединенных записей, но если автор публикует только длинные статьи из 5000 слов, его файл будет содержать отрывок одной из записей длинной в 3000 слов. Членение фрагментов проводилось по границе предложения. Таким образом, в нашей работе предполагается, что все тексты одного автора имеют одинаковый авторский стиль, и любые возможные различия между отдельными текстовыми записями игнорируются. Иначе говоря, оригинальный корпус представляет собой 50 txt-файлов с названиями, соответствующими их авторам. Каждый файл содержит все текстовые записи автора без явного указания границ между образцами исходного текста.

## 2.2. Предварительная обработка корпуса

Как было отмечено ранее, предварительная обработка текста является критическим этапом в построении систем атрибуции. Для того чтобы модель могла взаимодействовать с анализируемыми документами и корректно извлекать стилеметрические параметры, текст должен быть представлен определенным образом. В данном параграфе мы поэтапно рассмотрим процесс приведения исследуемых текстов к структурированному формату и обозначим получившиеся в результате наборы данных, используемые для анализа.

Техники предобработки текста, использованные в нашей работе, будут продемонстрированы на примере одного из отрывков. Так, фрагмент оригинального текста Дэйва Андерсона выглядит следующим образом:

*Looking for the best Australia Travel Tips for traveling down the east coast? Well, look no further, because you've come to the right place!*

1. Удаление знаков препинания. При извлечении параметров, ориентированных исключительно на словарную составляющую, знаки препинания представляют собой избыточную информацию и могут помешать анализу характеристик.

*Looking for the best Australia Travel Tips for traveling down the east coast Well look no further because you ve come to the right place*

2. Перевод в нижний регистр. Все заглавные алфавитные символы были преобразованы в строчные.

*looking for the best australia travel tips for traveling down the east coast well look no further because you ve come to the right place*

3. Лемматизация. В качестве алгоритма приведения различных словоформ к единой мы выбрали лемматизацию, поскольку для английского языка существуют относительно эффективные морфологические анализаторы, а стемминг приводит к большей погрешности. В результате обработки слова были приведены к следующим формам:

*look for the good australia travel tip for travel down the east coast well look no further because you ve come to the right place*

В нашем исследовании использовался лемматизатор spaCy, который анализирует зависимости между словами и может довольно эффективно обрабатывать двусмысленные грамматические ситуации. В данном примере было проведено четыре корректных трансформации: «looking» – «look», «best» – «good», «tips» – «tip», «traveling» – «travel», однако сравнительная форма прилагательного «far» осталась без изменений, что указывает на несовершенство анализатора.

4. Токенизация. Текстовая строка преобразуется в набор токенов, один токен – одно слово.

[*look, for, the, good, australia, travel, tip, for, travel, down, the, east, coast, well, look, no, further, because, you, ve, come, to, the, right, place*]

5. Удаление стоп-слов и числительных. Для анализа контентно-специфических характеристик, например лексического разнообразия, не имеет смысла рассматривать широко употребляемые слова, в том числе и служебные. Для исключения погрешности, такие слова и числительные убирают из набора токенов.

В нашей работе используется базовый список стоп-слов английского языка, предоставленный библиотекой NLTK, дополненный словарными формами количественных и порядковых числительных от одного до миллиона. Большую часть списка составляют функциональные слова, так, в нем можно встретить следующие единицы: «i», «her», «a», «an», «the», «and», «but», «if», «or», «because», «as», «didn't», «wasn't». Полный список исключаемой лексики содержит 237 слов и находится в приложении Б.

[*look, good, australia, travel, tip, travel, east, coast, well, look, further, come, right, place*]

Как видно из приведенного выше примера, количество токенов сократилось почти в два раза, и в наборе остались только значимые слова.

6. Присвоение тегов части речи. Набор токенов преобразуется в набор полученных на этапе лемматизации пар: слово – часть речи.

[*look, VERB*], [*good, ADJ*], [*australia, PROPN*], [*travel, NOUN*], [*tip, NOUN*], [*travel, VERB*], [*east, PROPN*], [*coast, PROPN*], [*well, INTJ*], [*look, VERB*], [*come, VERB*], [*right, ADJ*], [*place, NOUN*]

Используемые сокращения частей речи расшифровываются следующим образом: INJ – interjection – междометие, PROPN – proper noun – имя собственное, ADJ – adjective – прилагательное. На этом шаге мы получаем дополнительные данные о токенах (т.е. словах), которые могут быть проанализированы различными способами, например, путем сравнения распределения частей речи в текстах разных авторов.

7. Обработка орфографических ошибок, аббревиатур и акронимов. Количество отклонений от речевой нормы в нашем корпусе незначительно, в основном анализируемые тексты грамматически и орфографически корректны. Реализация исправления орфографических ошибок или расшифровки аббревиатур не окажет существенного влияния на производительность модели. По этой причине мы пропускаем данный этап предварительной обработки, однако следует отметить, что морфологический анализатор способен корректно обрабатывать наиболее распространенные сокращения.

Таким образом, в процессе предварительной обработки текст претерпел существенные изменения. Наше исследование подразумевает всесторонний обзор стилеметрических параметров, следовательно, мы не можем использовать лишь одно «итоговое» представление текста, поскольку оно лишено части характеристик, составляющих авторский стиль. В связи с этим на этапе анализа для каждого типа параметров мы будем указывать использованное в данном случае представление текста (необработанный текст, текст в форме токенов, текст без знаков пунктуации и т.д.)

## 2.3. Анализ стилеметрических параметров текста

В данном параграфе будут рассмотрены стилеметрические параметры, используемые нами для анализа авторского стиля и последующей атрибуции авторства. Этот этап работы представляет наибольший интерес, поскольку мы проанализируем показатели формальных характеристик авторского стиля, на основании которых алгоритм машинного обучения будет принимать решение об установлении авторства.

Группы параметров будут рассматриваться по возрастающей сложности: от простых к более комплексным, в каждой из групп будут представлены частные характеристики и их показатели для отдельных авторов. Текстовые примеры, используемые для демонстрации конкретных стилеметрических черт, будут выбраны случайным образом, поскольку эмпирическое рассмотрение параметров каждого из 50 авторов с приведением и анализом лингвистических примеров требует существенных временных затрат и не представляется нам целесообразным.

### 2.3.1. Единицы лексического уровня

Простой и естественный способ восприятия текста – это представление его в виде последовательности токенов, сгруппированных в предложения, где каждый токен соответствует слову, числу или знаку препинания. Следует отметить, что хотя единицы лексического уровня более сложны, чем элементы плана выражения знака, мы начинаем анализ с них, отдавая дань традициям стилеметрии.

**Модель bag-of-words**

При анализе единиц лексического уровня классическим подходом является представления текста в виде «мешка слов», то есть в форме вектора, содержащего все единицы словаря и количество их вхождений в определенном тексте (рисунок 8).

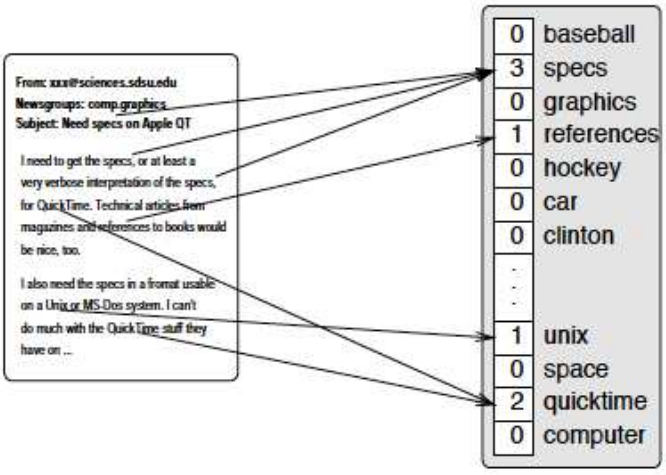


Рисунок 8. Представление текста в виде словарного вектора [Joachims, 1998]

В результате атрибуция авторства проводится на основании сравнения используемых авторами лексических единиц и их частотности. В нашей работе такая репрезентация основана на полностью обработанном тексте, то есть не содержит стоп-слов и пунктуации, а различные словоформы приведены к начальной. Рисунки 9 и 10 иллюстрируют результаты работы модели *bag-of-words* для двух авторов нашего корпуса. Визуализация при помощи облака слов предполагает, что частота встречаемости слова в тексте и кегль, используемый при его написании на изображении, находятся в прямой зависимости.


Рисунок 9. Облако слов Sherry


Рисунок 10. Облако слов Matt Long

Таким образом было выявлено, что наиболее частотные слова, такие как *travel*, *world*, *life* и *people*, совпадают для обоих авторов. Конечно это обусловлено единой тематикой текстов, однако при рассмотрении менее распространенных лексических единиц различия становятся видны.

Таблица 2. 20 наиболее частотных лемм для Александры Бакс, Чарли Мур, Сабрины Берлинер и Дерека Эрла

| | Alexandra Baackes | Charli Moore | Berliner Sabrina | Derek Earl |
|---|---|---|---|---|
| 1 | flight | travel | job | travel |
| 2 | travel | flight | travel | tourist |
| 3 | trip | retirement | money | crowd |
| 4 | time | cheap | sell | price |
| 5 | destination | budget | online | destination |
| 6 | check | find | time | time |
| 7 | book | consider | people | away |
| 8 | new | plan | good | local |
| 9 | way | like | way | visit |
| 10 | around | spend | video | rule |
| 11 | mile | accommodation | work | experience |
| 12 | point | study | website | place |
| 13 | fly | cost | help | walk |
| 14 | seat | london | pay | people |
| 15 | use | need | write | country |
| 16 | want | money | need | around |
| 17 | search | offer | social | traveler |
| 18 | friend | destination | product | world |
| 19 | day | free | create | much |
| 20 | good | life | love | city |

В таблице 2 приведены 20 наиболее часто встречающихся слов для четырех случайно выбранных авторов. Здесь так же есть слова, употребляемые каждым из авторов, например, *travel, destination, time*. В то же время у отдельных авторов выделяются лексические единицы связанные с

конкретными темами: например, *flight, book, seat, fly* для Александры Бакс; *cheap, budget, plan, cost* для Чарли Мур; *job, money, online, work* для Сабрины Берлинер.

Тем не менее, нельзя сказать, что анализ частоты встречаемости слов является оптимальным подходом для атрибуции авторства, поскольку количество вхождений в первую очередь зависит от объема и тематики текста и уже в меньшей степени от индивидуальных особенностей авторского стиля.

На сегодняшний день существует два подхода к усовершенствованию модели «мешка слов».

1. В процессе обработки заменить базовый список стоп-слов процедурой удаления лексических единиц, которые встречаются в текстах определенной доли авторов (например, более 25%). Так, в нашем корпусе при использовании данной техники будут удалены следующие контентно-специфические слова: *travel* – встречается в текстах 41 автора из 50, *time* – 39/50, *like* – 23/50, *world* – 20/50, *good* – 20/50, *way* – 17/50, *find* – 14/50, *day* – 17/50, *people* – 16/50, *life* – 16/50 и *want* – 13/50. Поскольку эти единицы присутствуют в документах стольких авторов, они вряд ли могут быть использованы для определения авторства текста и могут затруднить работу алгоритма классификации.

2. Выполнить *tf-idf* взвешивание, учитывающее (1) как часто термин встречается в документах конкретного автора и (2) насколько он специфичен (в текстах какого количества других авторов он встречается). В результате такого взвешивания наибольший вес для определения авторства получают не просто самые частотные слова, а лексические единицы, которые часто встречаются в текстах одного конкретного автора, но редко – в текстах других авторов.

Результаты *tf-idf* взвешивания векторов нашего корпуса продемонстрированы в таблице 3. Для большей наглядности в приведенной таблица рассматриваются те же авторы, что и в предыдущей.

Сравнивая список наиболее частотных слов со словами, имеющими наибольший вес, мы видим определенные изменения. Например, появляются новые лексемы, которых не было в списке частотных слов: *tsa, kayak, aisle, airline* – для Александры Бакс; *luxurious, pet, airfare, make* – для Чарли Мур; *make, shirt, affiliate, seo, skill, medium* – для Сабрины Берлинер; *shopkeeper, bargaining, touristy, interaction, lisbon* – для Дерека Эрла. В то же время некоторые часто употребляемые слова с учетом их веса ранжируются ниже: например, *spend* с 10-го ранга по частотности упало до 20-го ранга по его специфике для Чарли Мур, *find* – с 6 до 17, *people* – с 7 до 19, *visit* – с 9 до 14 и т.д.

Таблица 3. 20 лемм, обладающих наибольшим *tf-idf* весом

| | Alexandra Baackes | Charli Moore | Berliner Sabrina | Derek Earl |
|---|---|---|---|---|
| 1 | flight | retirement | job | crowd |
| 2 | mile | travel | sell | travel |
| 3 | travel | budget | online | shopkeeper |
| 4 | trip | cheap | money | tourist |
| 5 | tsa | flight | travel | price |
| 6 | check | london | make | bargaining |
| 7 | seat | study | website | rule |
| 8 | destination | accommodation | video | destination |
| 9 | kayak | consider | product | romania |
| 10 | get | retire | shirt | touristy |
| 11 | time | make | affiliate | away |
| 12 | flyer | cost | seo | walk |
| 13 | book | luxurious | social | learning |
| 14 | entry | pet | skill | visit |
| 15 | aisle | house | medium | traveler |
| 16 | fly | plan | time | local |
| 17 | pre | find | write | interaction |
| 18 | airline | offer | get | bargain |
| 19 | frequent | airfare | people | lisbon |
| 20 | point | spend | pay | get |

Демонстрируемые примеры наглядно показывают особенности подхода: *bag-of-words* захватывает скорее контентно-специфическую информацию, чем черты авторского стиля и может быть совершенно неэффективным, когда обучающие тексты разных авторов относятся к строго заданной тематике, а спорный текст – к другой, поскольку именно тема текста определяет наиболее часто встречающиеся слова. В задаче атрибуции авторства имеет смысл использовать данную модель на текстах, содержащих служебные слова, поскольку они являются одним из наиболее эффективных параметров и корректно обрабатываются в «мешке слов». В нашем исследовании служебные слова рассматриваются отдельно.

В целом такой подход наиболее эффективен в категоризации текстов по тематике или задачах сентимент-анализа. Тем не менее, он также используется при атрибуции авторства, и результаты работы данного представления текста будут рассмотрены нами в соответствующем параграфе.

**_N_-граммы слов**

Одним из недостатков *bag-of-words* является полное игнорирование контекста и отношений между словами. В качестве методики, компенсирующей данный недостаток, рассмотрим представление текста в виде вектора, содержащего частотность словарных *n*-грамм. В отличие от предыдущего представления текста здесь мы не будем удалять функциональные слова, потому что они зачастую выполняют роль грамматической связки и являются частью фразовых глаголов. Таким образом, анализируемый текст обработан лишь частично: удалена пунктуация и все символы переведены в нижний регистр.

В таблице 4 представлены 15 наиболее часто встречающихся биграмм, 3-грамм и 4-грамм слов для Александры Бакс и Чарли Мур. Сокращенные формы глаголов, вводимые апострофом, например, *ve – have, m – am, s – is* также считаются отдельными словами.

Таблица 4. 15 наиболее частотных 2-, 3-, 4-грамм слов для Александры Бакс и Чарли Мур

| № | Alexandra Baackes | | | Charli Moore | | |
|---|---|---|---|---|---|---|
| | 2 | 3 | 4 | 2 | 3 | 4 |
| 1 | i m | where to go | i built a trip | you can | if you re | a lot of people |
| 2 | i ve | in general i | built a trip to | if you | a lot of | make the most of |
| 3 | you can | i m not | when it comes to | you re | you need to | make a budget holiday |
| 4 | of the | and you can | ticket out of the | in the | do not have | a budget holiday feel |
| 5 | in the | frequent flyer miles | over the years i | when you | house and pet | budget holiday feel like |
| 6 | and i | in order to | the years i ve | you ll | your post retirement | holiday feel like a |
| 7 | to go | tsa pre check | i ve yet to | to the | not have to | feel like a luxury |
| 8 | it s | ve yet to | summer i built a | need to | you do not | like a luxury break |
| 9 | for a | i want to | right now i m | to be | emergency exit seats | do not have to |
| 10 | you re | i built a | i ve always wanted | for a | one of the | how can you make |
| 11 | have a | built a trip | ve always wanted to | the most | lot of people | you do not have |
| 12 | to the | a trip to | but in general i | a lot | the cost of | we all know that |
| 13 | i have | always wanted to | sign up for a | of the | cost of living | example if you re |
| 14 | if you | when it comes | if you want to | you are | make the most | the cost of living |
| 15 | to book | it comes to | i d say it | i m | the most of | cost of living is |

Легко заметить, что теперь, когда мы не убирали стоп-слова, первые строки частотных биграмм занимают их комбинации: *if you, you re, i m, i ve, to be, for a, of the, in the*, и т. д. Триграммы дают представление о контексте и встречаются в текстах не слишком малое количество раз, в то время как 4-граммы выглядят наиболее информативными, однако полностью совпадающие комбинации слов встречаются достаточно редко и, вероятно, не будут играть значительной роли при классификации. *N*-граммы слов

также можно нормализовать с помощью *tf-idf* взвешивания для учета их специфичности.

**Измерение лексического разнообразия**

В качестве показателей лексического разнообразия автора в нашей работе будут использоваться коэффициент лексического разнообразия (доля уникальных единиц в общем количестве слов, с англ. *type-token ratio*) и отношение количества *hapax legomena* к размеру словаря автора и общему количеству слов в тексте. Продемонстрируем расчет данных параметров на примере.

Исходный текст, состоит из 13 слов: *I love rainy days, but I hate sunny days. Sunny days are better.* Среди них уникальными являются 10 лексем: *I, love, rainy, day, but, I, hate, sunny, be, good.* Единожды в тексте встречаются 6 лексем: *love, rainy, but, hate, be, good.* Показатели лексического разнообразия, где $V$ – количество уникальных лексем, $N$ – общее количество слов, а $HL$ – количество лексем, встречающихся в тексте один раз, будут рассчитываться следующим образом:

$$V = 10, N = 13, HL = 6$$

$$TTR = \frac{V}{N} = 10/13 \approx 0.769$$

$$HL_V = \frac{HL}{V} = 6/10 = 0.6 \qquad HL_N = \frac{HL}{N} = 6/13 \approx 0.46$$

В нашей работе расчет соответствующих параметров проводился на полностью нормализованном тексте, содержащем леммы. Диаграмма значений коэффициента лексического разнообразия для всех авторов анализируемого корпуса представлена на рисунке 11.

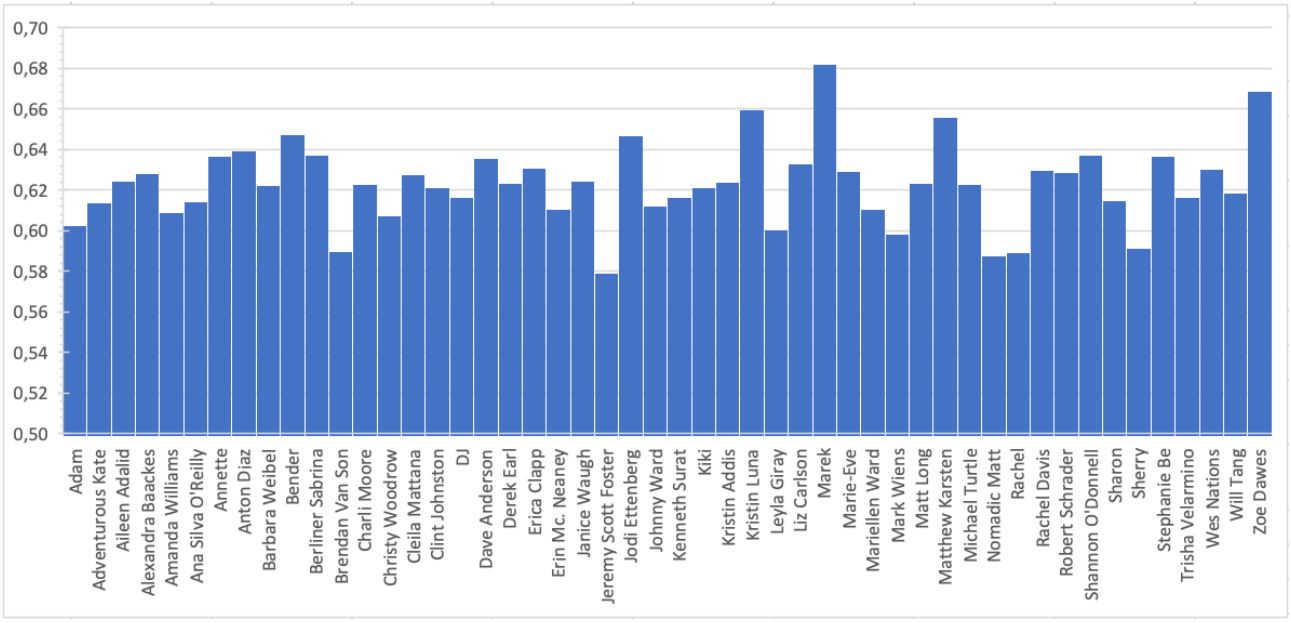


Рисунок 11. Коэффициент лексического разнообразия ($TTR$)

Рисунки 12 и 13 демонстрируют распределение коэффициентов *hapax legomena*. Как можно заметить, распределения не сильно отличаются между собой, придерживаясь общей тенденции, так как делимое одинаковое, а соотношение делителей не слишком различается для отдельных авторов (в соответствии с *type-token ratio*).

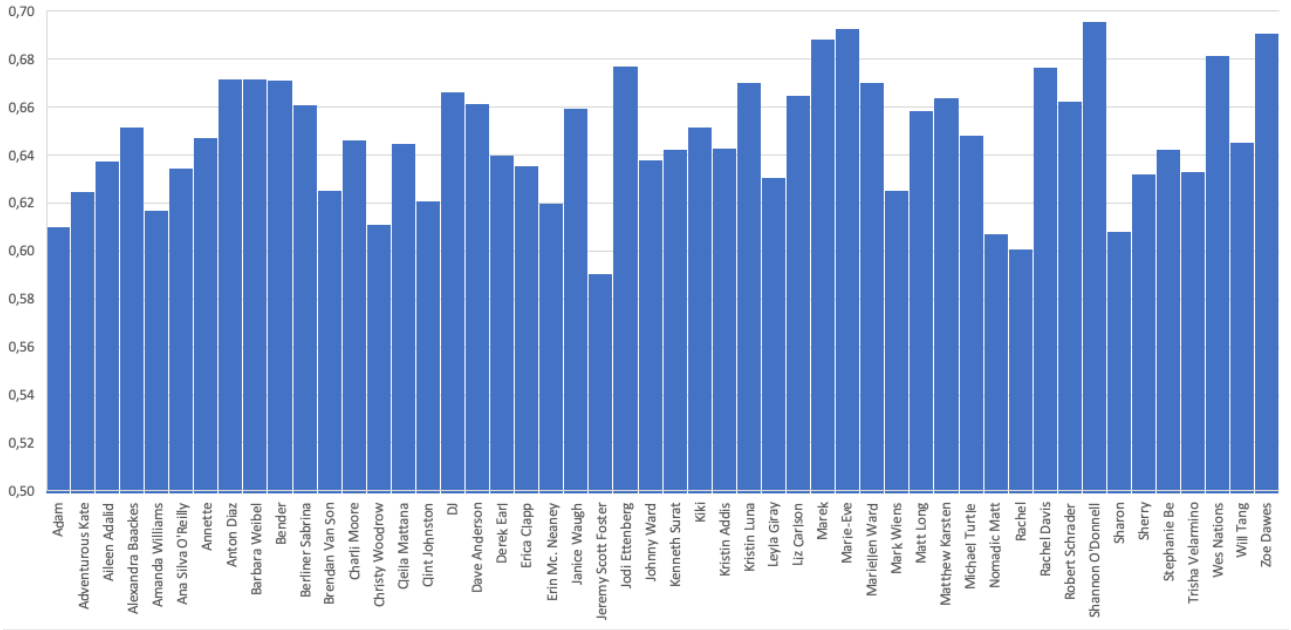


Рисунок 12. Отношение *hapax legomena* к количеству уникальных лексем ($HL_V$)

Распределение отличается для авторов с наиболее выделяющимися значениями коэффициента лексического разнообразия: так, показатель $HL_V$ у Шери намного выше, чем у Шэрон, но коэффициенты $HL_N$ аналогичных авторов практически равны. Причиной является относительно небольшое словарное разнообразие текстов Шэрон (количество уникальных лексем значительно меньше, чем в среднем по авторам).

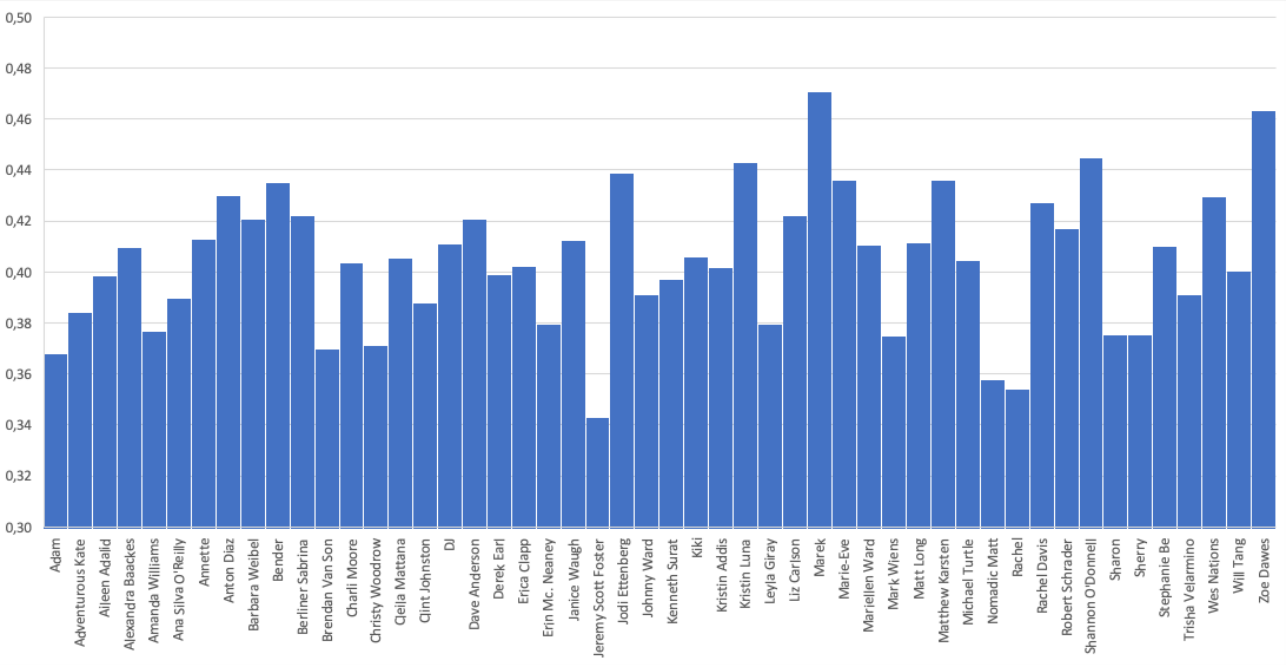


Рисунок 13. Отношение *hapax legomena* к общему количеству слов ($HL_N$)

Поскольку последние два параметра практически полностью пропорциональны, нет необходимости использовать их оба. В нашей системе атрибуции будет использоваться коэффициент лексического разнообразия *TTR* и отношение *hapax legomena* к количеству уникальных слов. В данном случае $HL_V$ представляется нам более обусловленным стилем автора, поскольку количество уникальных слов различается для каждого автора, а при обучении системы общее количество слов в одном тексте будет всегда примерно равно 300 словам.

**Средняя длина слова**

Значение средней длины слова рассчитывалось из оригинального текста без учета цифр и знаков препинания. На рисунке 14 представлено распределение средней длины слова для авторов исследуемого корпуса. И хотя все авторы писали на одном языке, диаграмма демонстрирует различия между средней длиной слова отдельных авторов. Самое низкое значение зарегистрировано для Уэс Нэйшнс – 3.84 буквы на слово, а самое высокое для Шеннон О'Доннел – 4.66 буквы на слово.

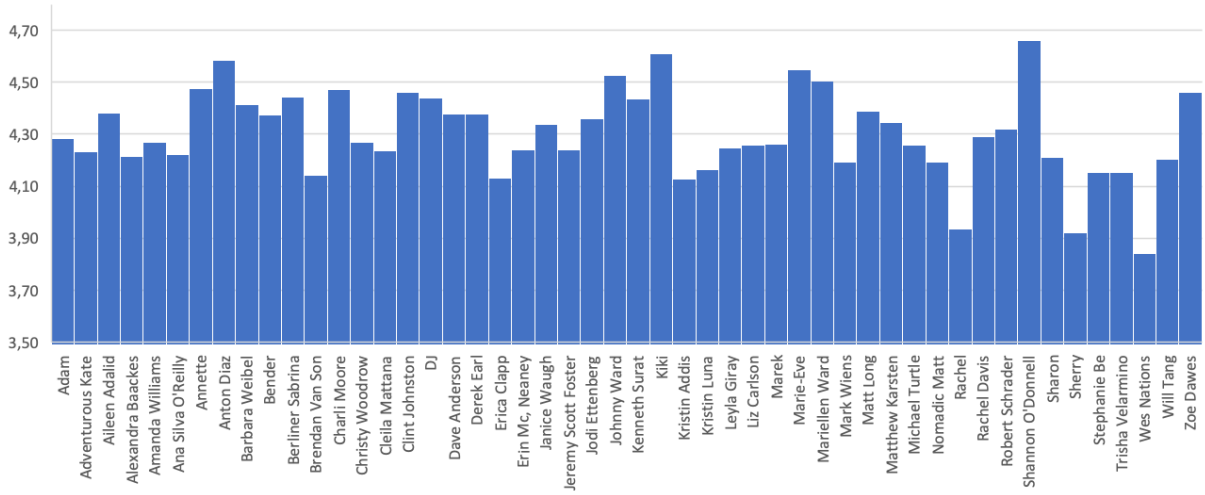


Рисунок 14. Распределение средней длины слова

**Оценка тональности**

Для применения технологий сентимент-анализа в нашем исследовании мы используем модель VADER (Valence Aware Dictionary and sEntiment Reasoner). Данный инструмент позволяет получить количественную меру тональности предложения с помощью комбинации правил оценки и тональности встречаемой лексики. Используемый словарь тональности содержит набор общепринятых для выражения оценки слов, количественно размеченных экспертами как положительные, отрицательные или нейтральные. В свою очередь обобщенные правила работы учитывают грамматические и синтаксические паттерны, которые люди используют при выражении и усилении оценки. По словам разработчиков, VADER проводит

анализ тональности на уровне сравнимом с работой экспертов, а иногда и превосходит их. [Hutto, Gilbert, 2015].

При проведении оценки тональности VADER учитывает следующие факторы:

1. Предшествующий оценочному слову триграмм. В 90% случаев слово, изменяющее полярность или интенсивность оценки, входит в число трех предшествующих оценочной единице.

2. Союзы. Такие союзы, как «*однако»* или «*но»*, информируют об изменении полярности оценки, при этом идущая после союза часть высказывания считается ключевой. Например, предложение «*The room was very spacious, but dirty»* будет оценено как в определенной степени отрицательное, поскольку вторая часть является доминирующей и содержит негативное прилагательное *dirty*.

3. Использование верхнего регистра. Оценка, выраженная прописными буквами, считается более интенсивной. Например, высказывание «*The trip was GREAT»* будет считаться более позитивным, чем «*The trip was great»*.

4. Модификаторы тональности – слова, которые влияют на интенсивность оценки, увеличивая или уменьшая ее. Например, выражение «*London is absolutely stunning»* будет оценено позитивнее, чем «*London is stunning*», в то время как «*London is slightly better than Paris*» считается менее позитивным, чем «*London is better than Paris*».

5. Пунктуация. Восклицательный знак увеличивает интенсивность выражения оценки, поэтому фраза «*I love Paris*» будет оценена менее позитивно, чем «I love Paris!!!».

Анализатор тональности возвращает 4 значения: *compound* – общая оценка тональности, *neg* – оценка негативности высказывания, *neu* – нейтральности и *pos* – позитивности. Предложение считается положительным, если $compound >= 0.05$, нейтральным, если $-0.05 < compound < 0.05$, и отрицательным, если $compound <= -0{,}05$. Для

иллюстрации работы инструмента, проиллюстрируем результаты оценки отдельных высказываний нашего корпуса.

*London is very beautiful, lively, and interesting city.*

compound: 0.8771, neg: 0.0, neu: 0.332, pos: 0.668.

*I actually don't like France...*

Оценка тональности: compound: -0.2755, neg: 0.413, neu: 0.587, pos: 0..

*You will ruin your vacation, if you go to Africa.*

Оценка тональности: compound: -0.5859, neg: 0.297, neu: 0.703, pos: 0.0,

*Cadiz is the best destination ever!!!*

compound: 0.7249, neg: 0.0, neu: 0.496, pos: 0.504,

*Madrid is abcolutely PERFECT!*

compound: 0.6932, neg: 0.0, neu: 0.388, pos: 0.612,

*I would never return to Washington.*

compound: 0.0, neg: 0.0, neu: 1.0, pos: 0.0,

*The trip was very long, but interesting.*

compound: 0.6054, neg: 0.0, neu: 0.603, pos: 0.397,

В нашей модели атрибуции мы использовали общую оценку тональности для расчета позитивности целого текста. Результаты сентимент-анализа нашего корпуса продемонстрированы на рисунке 15

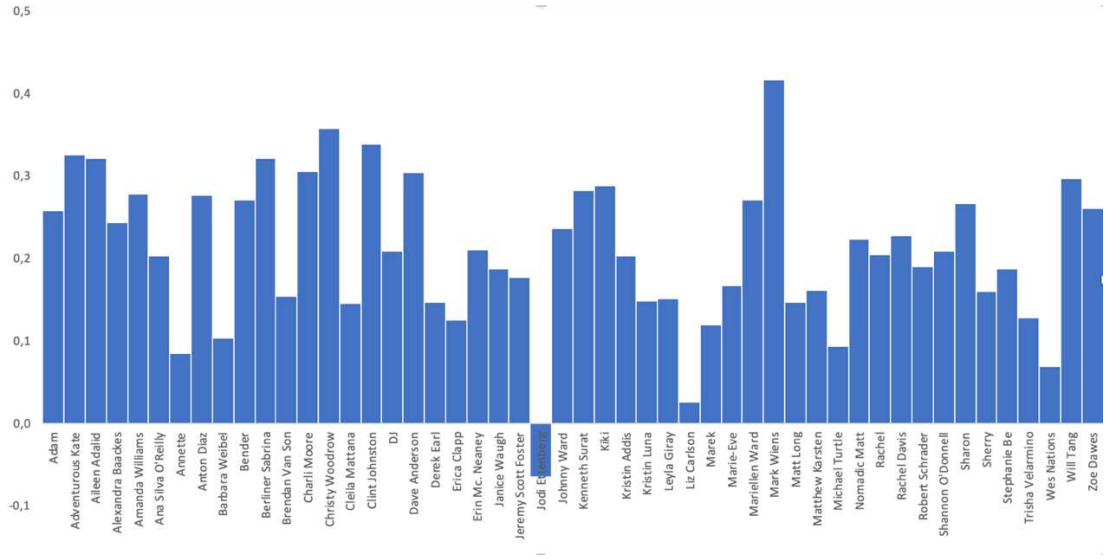


Рисунок 15. Общая оценка тональности

Единственный автор, чей текст признан скорее негативным, чем позитивным, – это Джоди Эттенберг. Единственный нейтральный автор – Лиз Карлсон. Тексты большинства авторов были оценены как позитивные. Вероятнее всего такой результат был получен из-за того, что VADER предназначен для анализа коротких отзывов и комментариев, где мнение выражается напрямую, а записи блогов не всегда содержат какую-либо эксплицитную оценку, и авторы, как правило, придерживаются позитивной линии повествования. Тем не менее, VADER предоставляет нам числовую оценку тональности текста, которая будет использована в нашем эксперименте в качестве стилеметрического параметра.

**Модель векторного представления слов**

В нашем исследовании для представления текста в виде векторов, учитывающих взаимосвязи между словами, мы будем использовать предварительно рассчитанные векторы слов *GloVe*. Данный набор векторов основывается на обширном массиве текстовых данных (6 миллиардов английских слов), состоящем из всех статей Википедии, написанных в 2014 году, и 5-го издания корпуса *Gigaword*, составленного Консорциумом лингвистических данных. Мы выбрали этот набор из-за того, что анализируемый корпус довольно мал, и модель не сможет не получить достаточно информации о вхождении каждого из слов. В нашем случае заранее подготовленное векторное представление слов позволит получить более точное результирующее векторное представление текстов.

Ниже рассмотрим пример работы и возможности модели *Word2vec*.

1. Функция поиска *n* схожих слов (слова, которые обычно используются в схожих контекстах и имеют аналогичные характеристики).

Так, при запросе слов, похожих на название испанского города Кадис, модель выдает имена других городов Испании: «Malaga», «Algeciras», «Tenerife», «Valladolid», «Salamanca», «Tarragona». В свою очередь словами, аналогичными прилагательному «stunning», считаются «spectacular», «breathtaking», «impressive», «astonishing», «dazzling», «remarkable».

2. Сложение и вычитание слов на основании семантических отношений между ними. Так, «poetry» + «prose» = «poems», а «woman» + «king» – «man» = «queen».

3. Поиск лишнего слова. Например, среди слов «Paris», «Madrid» и «Barcelona» лишним модель Word2vec считает «Paris». Если задать ряд слов с более тонкими семантическими различиями, например, «Shakespeare», «Cervantes», «Pushkin», «Obama», лишней будет выбрана фамилия «Obama».

4. Показатель схожести слов. Данная оценка рассчитывается на основании расстояния между векторами слов и варьируется от 1 – одинаковые слова, до 0 – совершенно различные. Так, показатель схожести слов «travel» и «money» равен 0.5293254, что указывает на частое использование слов в схожих контекстах; схожесть «travel» и «cry» близка к 0 и равняется 0.0988148, поскольку данные слова редко пересекаются; «go» и «walk» являются близкими по значению словами, их показатель схожести ближе к 1 и равен 0.72357756.

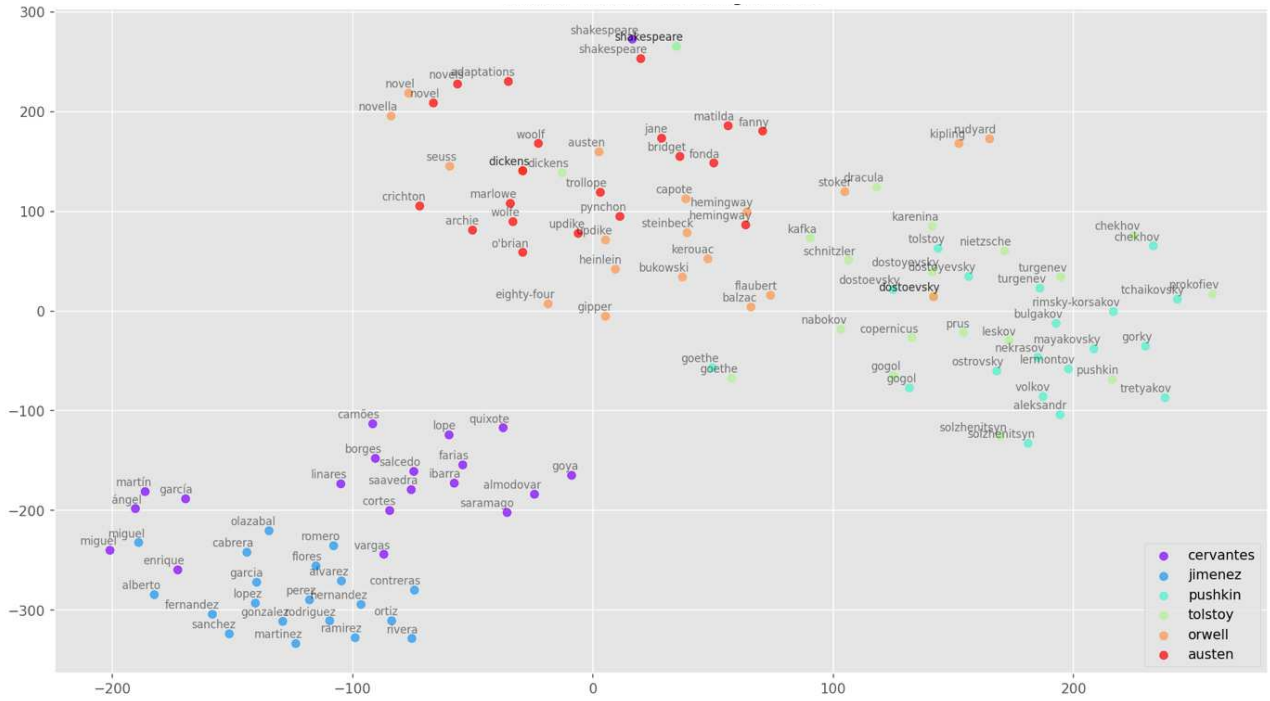


Рисунок 16. Наборы слов, ближайших к именам известных писателей, в модели Word2vec

Приведенные выше примеры показывают, что модель Word2vec улавливает семантическое значение слов, и векторное представление слов может быть полезным для атрибуции авторства, а также для классификации текстов по тематике и тональности.

На рисунке 16 представлена визуализация векторного представления слов, наиболее близких к именам известных писателей, таких как М. Сервантес, Х.Р. Хименес, А.С. Пушкин, Л.Н. Толстой, Дж. Остин и Дж. Оруэлл, на основе предварительно обученных векторов *GloVe*. Этот график представляет собой интерес в нескольких аспектах. Во-первых, слова, которые модель считает наиболее похожими на имена авторов (таблица 5). Некоторые из них вполне ожидаемы, в то время как другие, на первый взгляд, кажутся ошибочными, но их присутствие в списке означает, что они часто используются в тех же контекстах, что и имена авторов. Во-вторых, распределение групп слов по координатам: слова, относящиеся к авторам одной страны (например, А.С. Пушкин и Л.Н. Толстой), имеют близкие координаты, в то время как каждая из «страны» расположена отдельно.

Таблица 5. Топ 10 слов, ближайших к фамилиям Cervantes, Jimenez, Pushkin, Tolstoy, Orwell, Austen

| Cervantes | Jimenez | Pushkin | Tolstoy | Orwell | Austen |
|---|---|---|---|---|---|
| quixote | garcia | lermontov | dostoyevsky | updike | dickens |
| borges | hernandez | dostoyevsky | turgenev | dickens | jane |
| goya | ortiz | tolstoy | dostoevsky | kipling | wolfe |
| salcedo | martinez | solzhenitsyn | kafka | eighty-four | woolf |
| ibarra | romero | goethe | chekhov | steinbeck | matilda |
| miguel | lopez | aleksandr | pushkin | stoker | archie |
| almodovar | perez | shakespeare | karenina | kerouac | crichton |
| vargas | cabrera | mayakovsky | nietzsche | bukowski | novels |
| ángel | miguel | gogol | dracula | capote | bridget |
| camões | rodriguez | tchaikovsky | dickens | hemingway | hemingway |

Таким образом, векторное представление слов учитывает семантические значения и контекстуальное употребление единиц и может

быть использовано в качестве параметров для модели классификации текста. В параграфе «Результаты работы модели атрибуции авторства» нами будет рассмотрена точность атрибуции для векторного представления слов.

### 2.3.2. Элементы плана выражения знака

Наиболее распространенным подходом к извлечению характеристик в автоматических моделях атрибуции авторства является посимвольный анализ текста. Такой метод прост в реализации и зачастую показывает наилучшие результаты. Все характеристики данной группы мы будем извлекать из необработанного текста, поскольку посимвольный анализ не требует никакой предварительной обработки.

**Символьные *n*-граммы**

В первую очередь рассмотрим наиболее комплексный признак плана выражения знака: *n*-граммы символов. В сущности, представление текста как набора символьных последовательностей представляют собой скорее всеобъемлющую репрезентацию текста, чем отдельный частный признак, поскольку в виде набора *n*-грамм представляется весь текст, а не его отдельные черты. Продемонстрируем разбиение текста на *n*-граммы, для удобства заменив символ пробела нижним подчеркиванием.

Исходный текст:

*My_acomodation_is_clean.*

Текст в виде символьных 4-грамм:

*‘My_a’, ‘y_ac’, ‘_aco’, ‘acom’, ‘como’, ‘omod’, ‘moda’, ‘odat’, ‘dati’, ‘atio’, ‘tion’, ‘ion_’, ‘on_i’, ‘n_is’, ‘_is_’, ‘is_c’, ‘s_cl’, ‘_cle’, ‘clea’, ‘lean’, ‘ean.’*.

Очевидно, что такое представление текста учитывает разноплановые особенности авторского стиля и позволяет одновременно анализировать

лексический уровень (*'is_c'*), часть контекстуальной информации (*'My_a'*), использование заглавных букв (*'My_a'*), пунктуацию (*'ean.'*) и пр.

В процессе анализа, каждый из текстов преобразуется в вектор, содержащий количество вхождений самых частотных *n*-грамм (размер вектора определяется вычислительной мощностью оборудования и особенностями используемой модели атрибуции). Затем на основании векторных представлений модель сравнивает спорный текст с текстами, предоставленными ей на этапе обучения и определяет личность автора.

Так, в таблице 6 представлены 20 наиболее частотных биграмм, триграмм и 4-грамм для Лиз Карлсон и Мэтта Лонга. Практически все позиции заняты служебными словами, на первых строках находятся *n*-граммы, составляющие артикль «*the*» и биграмма «*e_*», которую можно интерпретировать как любое слово, оканчивающееся на «*e*». Неудивительно, что эти последовательности символов были признаны самыми частотными не только для Лиз Карлсон и Мэтта Лонга, но и для всех других авторов корпуса.

Однако, есть и различия: так, на 16 и 17 позициях биграмм Лиз Карлсон находятся знаки препинания, в то время как у Мэтта Лонга их нет, из чего можно сделать вывод, что Лиз значительно чаще использует грамматическое членение мыслей, разделяя текст на предложения – «._» и составные части – «,_» . Также у Лиз превалирует количество отрицаний – «*n't*» и местоимений «*_I_*» и «*_my_*». Кроме того, среди 4-грамм Мэтта можно заметить особенности используемой им лексики: частое употребление слов, начинающихся на «*trav*» и существительных, оканчивающихся на «*ion*».

В целом можно заметить, что *n*-граммы символов действительно охватывают разноплановые особенности авторского стиля и вероятно продемонстрируют высокие результаты при установлении авторства. В итоговой системе атрибуции для разбиения текста мы будем рассматривать последовательности символов длиной $n = 4$ и $n = 5$. В ходе эксперимента

нами было установлено, что при больших или меньших значениях *n* точность атрибуции резко падает, следовательно, для английского языка такая длина является оптимальной. При больших значениях параметра точность системы не увеличивается, а вычислительное время значительно возрастает, в то время как при меньших значениях точность падает.

Таблица 6. 20 наиболее частотных 2-, 3-, 4-грамм для отдельных авторов

| | Liz Carlson | | | Matt Long | | |
|---|---|---|---|---|---|---|
| | 2-граммы | 3-граммы | 4-граммы | 2-граммы | 3-граммы | 4-граммы |
| 1 | e_ | _th | _the | e_ | _th | _the |
| 2 | _t | the | the_ | _t | the | the_ |
| 3 | t_ | ing | ing_ | s_ | he_ | and_ |
| 4 | in | he_ | _to_ | _a | nd_ | _and |
| 5 | s_ | nd_ | and_ | t_ | and | _to_ |
| 6 | _a | _I_ | _and | th | _an | ing_ |
| 7 | th | ng_ | _in_ | in | _to | _of_ |
| 8 | he | _to | _my_ | an | ing | _tha |
| 9 | d_ | _an | _of_ | he | ng_ | hat_ |
| 10 | n_ | to_ | hat_ | _i | to_ | _in_ |
| 11 | _w | and | _tha | d_ | _in | that |
| 12 | _m | _in | that | n_ | _of | rave |
| 13 | y_ | in_ | n_th | _o | of_ | avel |
| 14 | _i | _my | n’t_ | er | hat | _tra |
| 15 | er | _of | _you | _w | ave | trav |
| 16 | ._ | _a_ | ith_ | nd | _a_ | vel_ |
| 17 | ,_ | ed_ | _wit | ha | at_ | _we_ |
| 18 | _o | en_ | with | ve | re_: | _is_ |
| 19 | an | my_ | _for | re | e_t | t’s_ |
| 20 | _I | of_ | _is_ | y_ | s_a | tion |

По аналогии с лексическими единицами можно преобразовать количество вхождений символьных *n*-грамм, используя схему взвешивания *tf-idf*. Такой подход позволит учесть специфичность единиц для текстов отдельного автора и, как правило, является более эффективным.

Эффективность работы системы, использующей представление текста в виде *n*-грамм символов и их *tf-idf* весов будет представлена в параграфе «Результаты работы модели атрибуции авторства».

**Частотность алфавитных символов**

Данная группа параметров отражает частоту встречаемости отдельных буквенных символов для каждого автора и основывается на предположении, что авторы подсознательно склонны использовать слова с определенными буквами. Тем не менее, частота используемых алфавитных символов также зависит от особенностей языка. Как правило, данные параметры не используются отдельно, а сочетаются с частотностью цифр и знаков препинания, таким образом квантифицируя весь текст. Частотность символов любого типа рассчитывается по исходному тексту без какой-либо предварительной обработки. На рисунке 17 показана частотность наиболее распространенных буквенных символов для трех авторов нашего корпуса. Можно заметить, что в зависимости от автора для всех букв присутствует колебание значения частоты в пределах 1%.

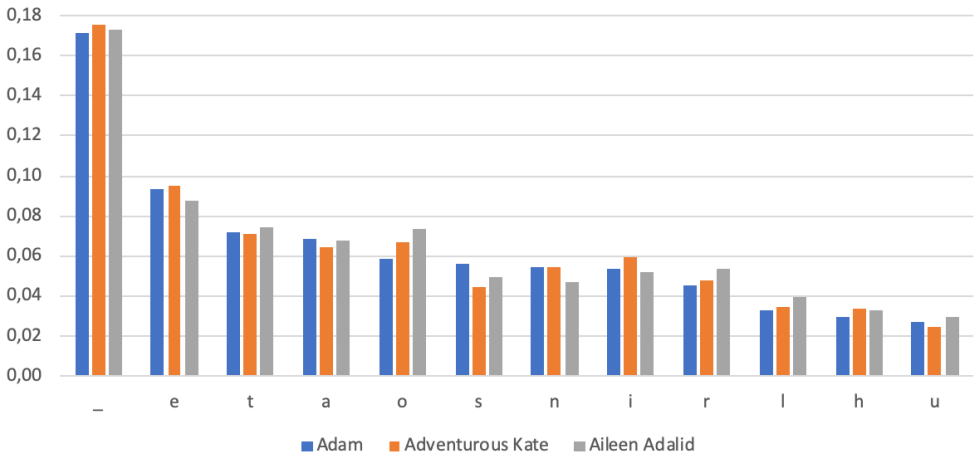


Рисунок 17. Частотность алфавитных символов

**Частотность прописных букв**

Частота использования заглавных букв отражает количество заголовков в тексте, склонность автора к нестандартному написанию слов, например, использованию «верблюжьего регистра» и написанию обычных слов заглавными буквами, а также количество имен собственных и аббревиатур.

Рисунок 18 демонстрирует распределение частоты прописных символов в нашем корпусе. Показатель варьируется в зависимости от автора, при этом на диаграмме выделяются два пиковых значения – для Клейлы Маттана и Марка Винса. Причина в том, что эти два автора выделяют все пункты и основные идеи заглавными буквами.

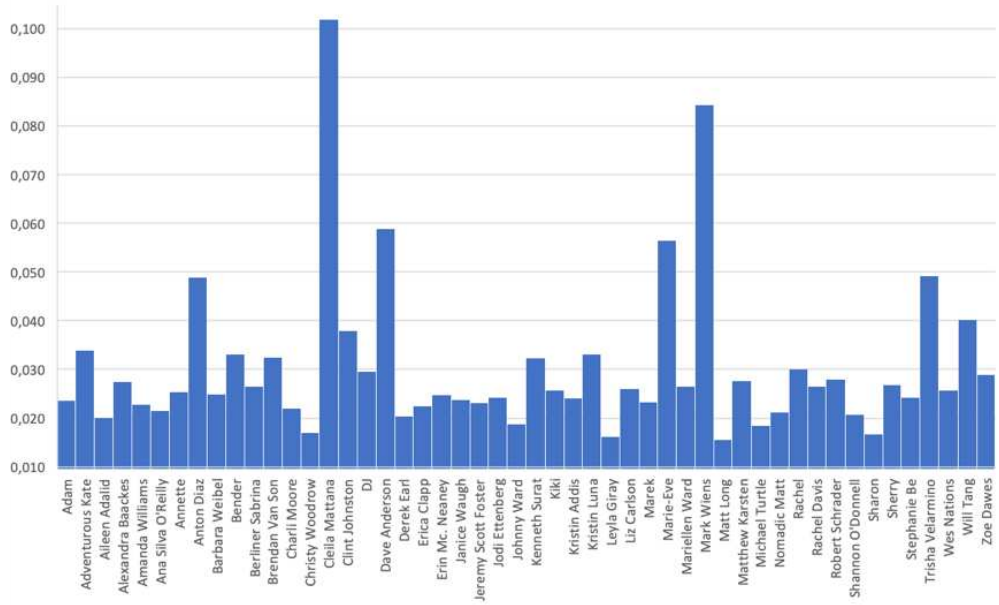


Рисунок 18. Частотность прописных букв

**Частотность цифр**

Несмотря на то, что появление цифр в тексте может зависеть от темы, в равных условиях одни авторы используют цифры чаще, чем другие, например, если они предпочитают нумерованные списки алфавитным. Этот параметр будет использоваться в сочетании с другими для дополнения общего набора. На рис. 16 показано распределение частоты цифр в нашем

корпусе. Легко заметить, что распределение неравномерно и зависит от каждого автора.

Наибольшее количество цифр встречается в корпусах Антона Диаза и Марека, однако путем эмпирического анализа материала мы установили, что в данном случае значения обусловлены тематикой текста. Исследуемые посты Антона Диаза посвящены различным фестивалям, для каждого из которых он указывает обширные числовые данные: дата проведения, график работы, адрес и номер телефона. Пост Марека содержит состоящий из 91 пункта пронумерованный список «*странных и прекрасных уроков, которые я извлек, путешествуя по миру в течение 2 лет*». В обоих случаях тексты содержат большое количество цифр, но нельзя сказать, что они являются специфической характеристикой авторского стиля.

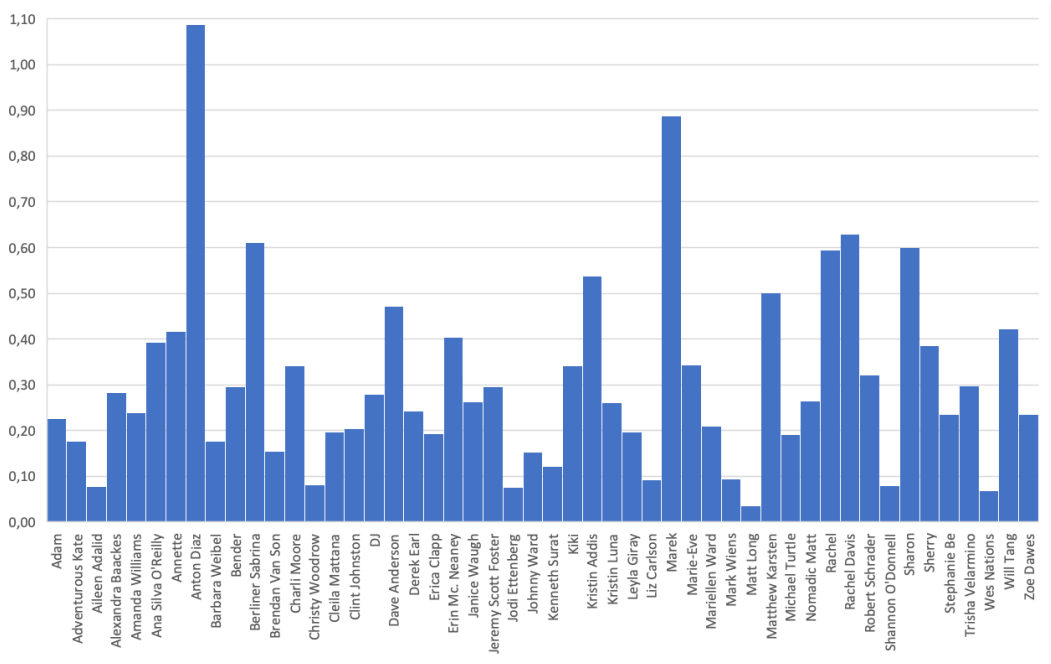


Рисунок 19. Частота встречаемости цифр в тексте, проценты

Тем не менее, у большинства авторов не было таких явных причин использовать большое количество цифр, при этом величина параметра частотности значительно различается и, по-видимому, может использоваться для идентификации авторского стиля.

**Частотность специальных символов**

В нашей работе к специальным символам мы относим следующий набор: ~ , @, #, $, %, ^, &, *, -, _, = ,+, >, <, [, ], {, }, /, |. Частотность этих символов будет учитываться в модели атрибуции авторства наравне с другими параметрами плана выражения знака.

### 2.3.3. Единицы синтаксического уровня

**Средняя длина предложения**

Средняя длина предложения является одним из наиболее примитивных способов анализа синтаксических паттернов. Длина высказывания является универсальной характеристикой всех человеческих языков и в большей степени связана с особенностями авторского стиля, чем с контекстом. В нашей работе длина предложения и все другие параметры синтаксического уровня будут рассчитываться на необработанных текстах, содержащих пунктуацию и все лексические единицы.

Рисунок 20 показывает распределение средней длины предложения среди авторов нашего корпуса. Параметр средней длины предложения минимален для Мэтью Карстена – 13 слов и максимален для Эйлин Адалид – 23.3 слова и Роберта Шнайдера – 23.2 слова.

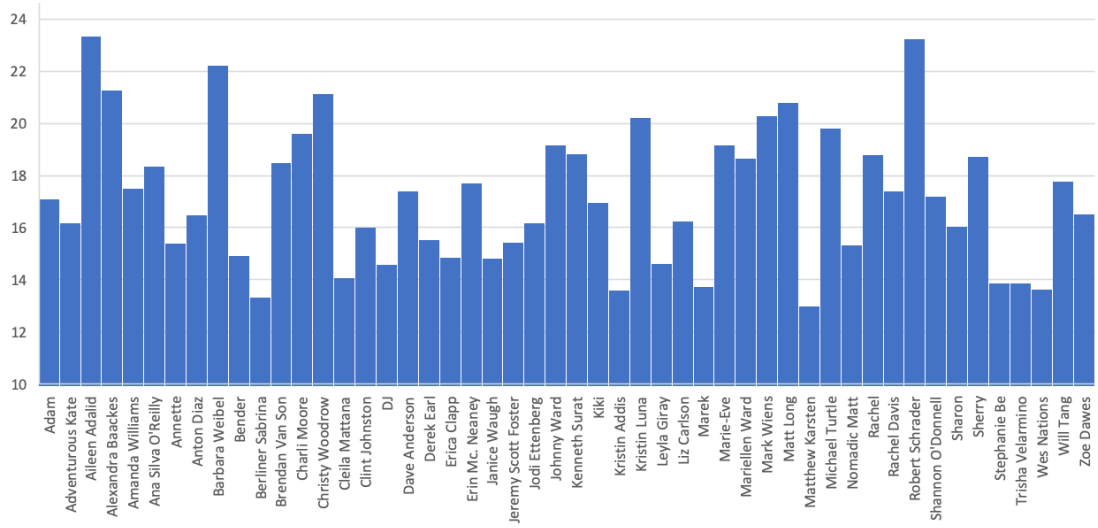


Рисунок 20. Средняя длина предложения

Действительно, при анализе текста данных авторов видны различия. Так, Мэтью Карстен большинство идей выражает простыми короткими предложениями.

*ATMs out of money? Great! Take an unplanned road trip over to the next town and explore. Sometimes freakouts happen regardless. Just take a deep breath and remind yourself that it could be worse.*

В свою очередь большая часть высказываний Эйлин Адалид представлена в виде составных предложений, содержащих обособленные дополнения, причастные обороты и пр.

*Travel to low-cost-of-living countries and/or visa-free countries first: this advice applies best especially if you're someone like me who holds a third world passport. By doing this, you will avoid costly visa fees and you will also spend less (since such countries ordinarily have lower cost of living).*

Аналогичную тенденцию можно заметить и в текстах Роберта Шрейдера, содержащих составные предложения с большим количеством однородных членов.

*Let's start out our round-up of 2 week vacation ideas with one you might not have considered: Europe's Balkan region, which is the area wedged between the Alps, the Black Sea and Greece. For the purposes of my popular two weeks in the Balkans post this includes Croatia, Bosnia, Montenegro, Serbia and maybe Kosovo, but you can expand this to countries like Albania, Bulgaria and Macedonia (or substitute them), depending upon your preferences.*

Таким образом, средняя длина предложения действительно отражает особенности выражения мыслей автора и может использоваться для описания стиля.

**Частотность знаков пунктуации**

Данная характеристика также направлена на выявление характерных синтаксических паттернов, которые множо установить путем анализа знаков

препинания: например, обилие запятых может говорить о частом использовании обособляемых оборотов или однородных членов, а вопросительные знаки – о склонности к риторическим вопросам. Для анализа этого параметра мы будем рассматривать частотность следующих знаков пунктуации: точка (.), запятая (,), вопросительный знак (?), восклицательный знак (!), двоеточие (:) и точка с запятой (;).

На рисунке 21 изображено распределение частотности анализируемых знаков для восьми авторов нашего корпуса. Из диаграммы можно заметить, что частота знаков препинания коррелирует с автором, поскольку для каждого из них показатели различаются. Например, Барбара Вайбель не использует восклицательный знак ни в одном из своих постов, в то время как Аннет использует его даже чаще, чем вопросительный знак, двоеточие или точку с запятой. Аманда Уильямс использует запятую и точку почти с одинаковой частотой, в то время как другие авторы демонстрируют преобладание одного из этих знаков препинания.

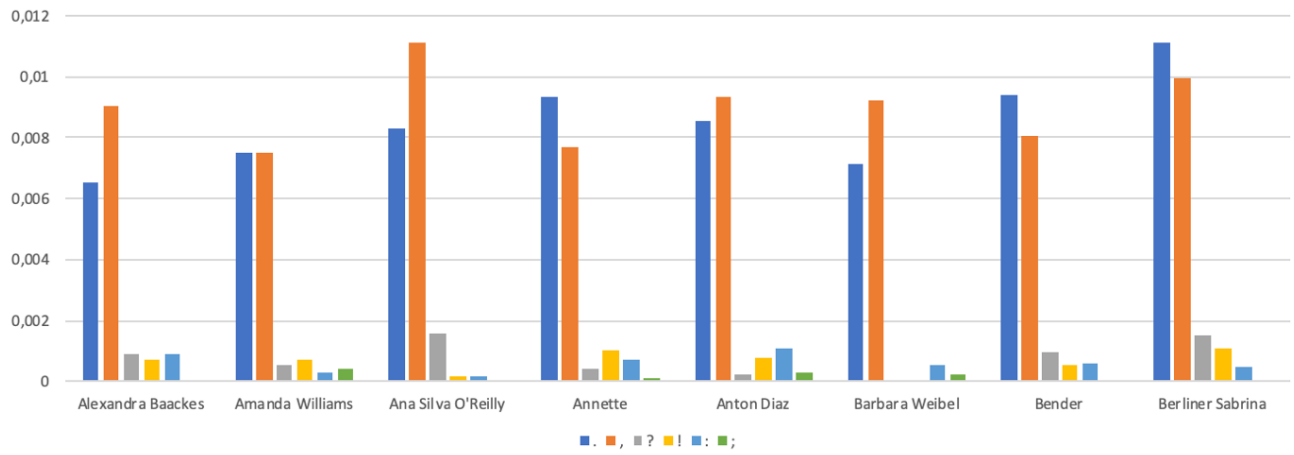


Рисунок 21. Частотность знаков пунктуации

Очевидно, что использование знаков препинания зависит от способа выражения мыслей и является индивидуальным выбором автора, особенно в условиях личного блога, где нет необходимости придерживаться строгих правил пунктуации.

**Частотность функциональных слов**

В нашей работе используется 265 функциональных слов английского языка, предложенные О'Ши, Бандаром и Крокеттом (2012). Полный список слов можно найти в приложении Б.

Вслед за каноническим исследованием Мостеллера и Уоллеса для наибольшего удобства и наглядности в качестве меры частотности мы будем использовать не относительную частоту функциональных слов в корпусе, а частоту анализируемого слова на каждую 1000 слов. В результате эксперимента нами были получены следующие результаты: для 30 из 50 авторов корпуса наиболее частым словом является *the*, для 10 – *to*, для 4 – *you*, для 3 – *I*, для 2 – *a* и для 1 – *and*. На рисунках 22 – 24 показано распределение функциональных слов *the*, *to* и *for* соответственно.

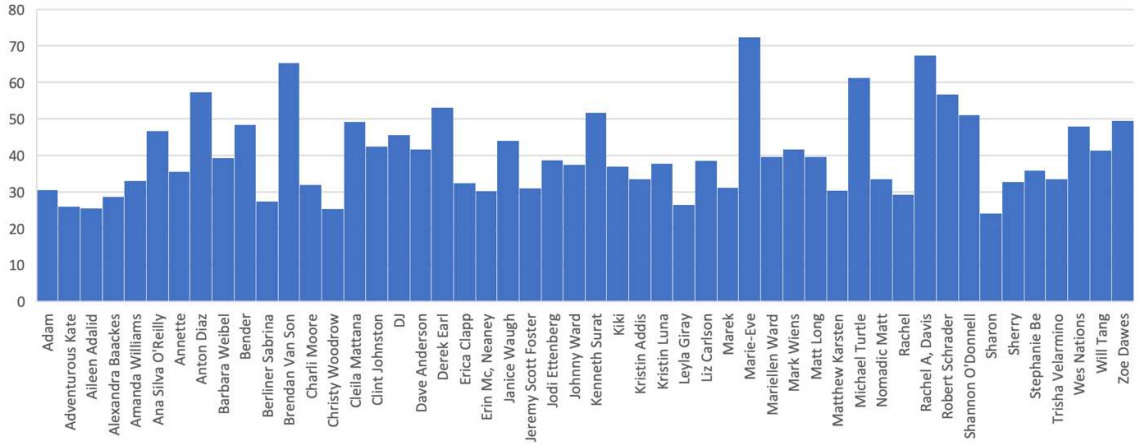


Рисунок 22. Частотность *the* для 50 авторов корпуса

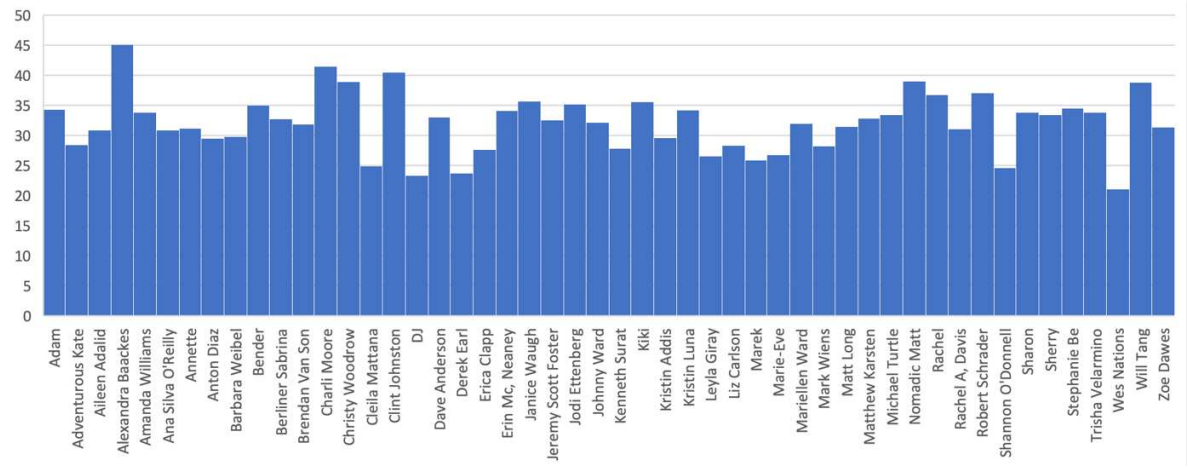


Рисунок 23. Частотность *to* для 50 авторов корпуса

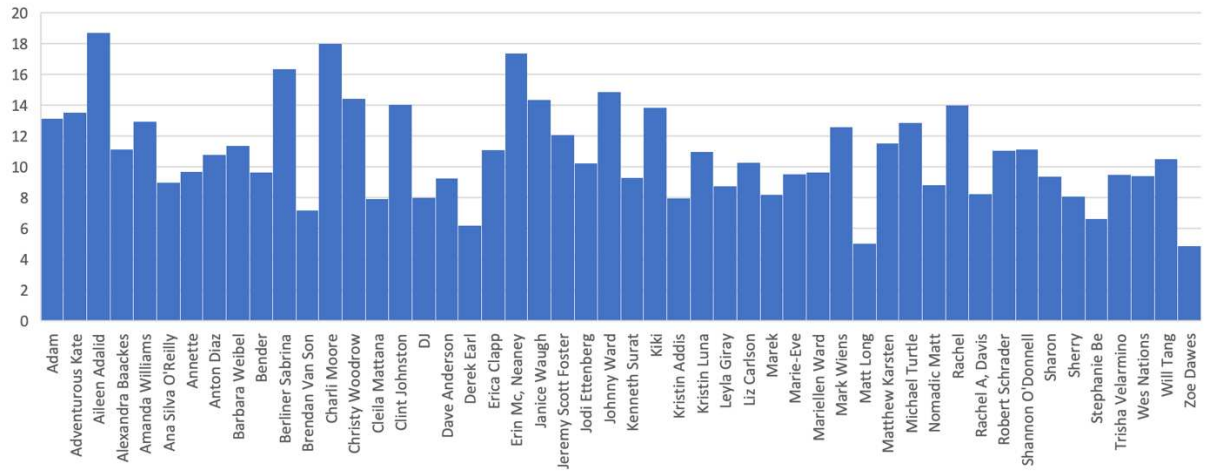


Рисунок 24. Частотность *for* для 50 авторов корпуса

Графики подтверждают, что значения частоты функциональных слов значительно различаются в зависимости от автора и могут рассматриваться как различающий признак. Кроме того, большое количество функциональных слов в совокупности может достаточно полно охватить особенности авторского стиля и внести существенный вклад в решение задачи атрибуции.

### 2.3.4. Единицы морфологического уровня

**Распределение частей речи**

Как упоминалось ранее, во время предварительной обработки морфологический анализатор spaCy присваивает словам грамматические теги: *PRON, ADJ, NOUN, ADV, INTJ, AUX, CCONJ, PROPN, NUM, SCONJ, VERB, ADP, DET*.

*PRON* – местоимение (включает в себя личные, вопросительные, относительные, неопределенные и отрицательные местоимения); *ADJ* – прилагательное, *VERB* – глагол; *NOUN* – существительное, *ADV* – наречие, INTJ – междометие; *AUX* – вспомогательный глагол, данный тег присваивается словам, сопровождающим основной глагол и отображающим грамматические характеристики (например, время, залог, модальность и

глагольные связки); *CCONJ* – сочинительный союз и *SCONJ* – подчинительный союз; ADP – адлог (предлоги и послелоги); *DET* – детерминатив и *NUM* – число.

Для проведения анализа мы подсчитали частотность отдельных частей речи для каждого автора. На рисунке 25 можно увидеть сравнительное распределение наиболее частотных частей речи для 8 авторов нашего корпуса. Для демонстрации мы использовали глагол, прилагательное, существительное, наречие, числительное и союз (без разделения на сочинительные и подчинительные союзы).

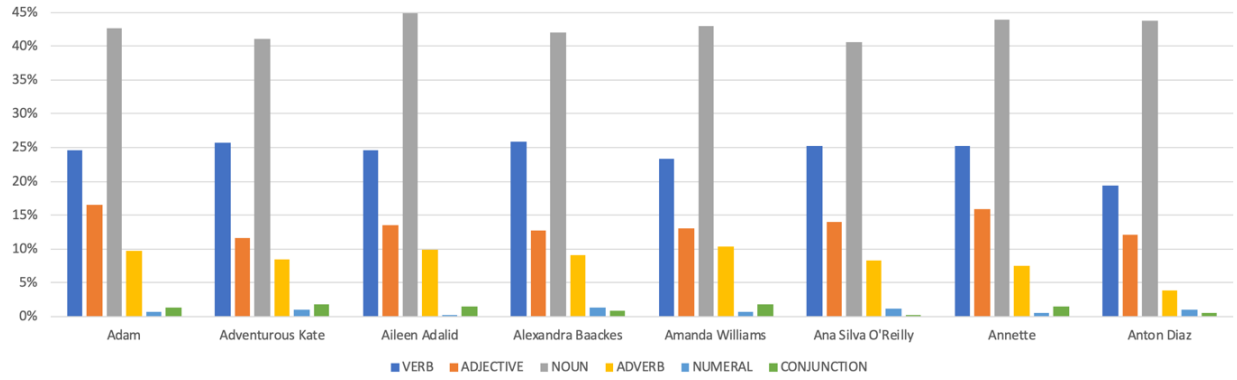


Рисунок 25. Частотность наиболее распространённых частей речи

Хотя конкретные значения различны для всех авторов, диаграмма показывает, что общая тенденция распределения одинакова для глаголов, прилагательных, существительных и наречий, которые являются наиболее распространенными частями речи. Таким образом, существительное является наиболее распространенной частью речи и составляет от 40 до 45 процентов от общего количества, за ним следует глагол с частотностью от 20 до 25 процентов, затем прилагательное и наречие, другие части речи встречаются значительно реже – в пределах 2% от общего числа. Тем не менее, можно заметить, что значения встречаемости для различных авторов колеблются достаточно существенно и могут отражать специфические характеристики идиостиля.

В таблице 7 приведен список наиболее часто встречающихся прилагательных, существительных и глаголов для 5 авторов корпуса. Можно отметить, что существительные между авторами варьируются сильнее, чем глаголы или прилагательные, однако все слова скорее отражают контентно-специфическую составляющую.

В итоговой системе атрибуции авторства в качестве морфологических параметров мы будем использовать распределение таких частей речи, как существительное, глагол, прилагательное и наречие. Другие грамматические категории составляют лишь незначительный процент от общего числа, кроме того, в результате дополнительных экспериментов было установлено, что их добавление в набор параметров не влияет на итоговую точность модели.

Таблица 7. Наиболее распространенные существительные, прилагательные и глаголы

| | Adam | Adventurous Kate | Aileen Adalid | Alexandra Baackes | Amanda Williams |
|---|---|---|---|---|---|
| СУЩЕСТВ. | account | friend | travel | flight | cruise |
| | saving | people | tip | trip | work |
| | money | hostel | price | time | river |
| | thing | tour | flight | travel | time |
| | summer | activity | budget | destination | day |
| ГЛАГОЛЫ | make | meet | travel | check | work |
| | use | travel | help | travel | want |
| | save | make | would | fly | travel |
| | travel | find | know | use | set |
| | get | look | give | book | make |
| ПРИЛАГАТ. | easy | social | free | new | active |
| | good | many | good | good | remote |
| | important | new | cheap | big | good |
| | sure | local | local | free | new |
| | many | easy | short | international | many |

## 2.4. Результаты работы модели атрибуции авторства

В нашей работе были подробно рассмотрены наиболее распространенные стилеметрические параметры, используемые для

атрибуции авторства: *bag-of-words*, *n*-граммы слов, *tf-idf* взвешивание, меры лексического разнообразия, средняя длина слова, оценка тональности, векторное представление слов, *n*-граммы символов, средняя длина предложения, частотность символов различного типа, знаков пунктуации, функциональных слов и частей речи.

Следует отметить, что эти параметры различаются по своей природе: некоторые из них представляют весь текст с помощью одного числа, например, средняя длина слова или коэффициент лексического разнообразия, другие возвращают набор значений, например, частотность алфавитных символов описывает авторский стиль 26 числовыми величинами. На данном этапе мы также выделили параметры, представляющие собой полноценное представление текста, в отдельную категорию, поскольку их зачастую используют отдельно от всех других характеристик. Всеобъемлющая числовая репрезентация текста, как правило, содержит большое количество частных параметров, например, при подходе «мешок слов» количество результирующих значений равно размеру словаря.

Отождествляя количество параметров с количеством числовых значений, которые они предоставляют, в общей сложности наша модель анализирует 362 параметра и 9 текстовых представлений. Заключительным шагом нашего исследования является непосредственно проведение эксперимента по атрибуции авторства при помощи разработанной нами модели статистической обработки текстов. В результате эксперимента мы оценили эффективность всех рассмотренных выше стилеметрических параметров для решения задачи установления авторства и выделили параметры, позволяющие наиболее точно атрибутировать документы нашего корпуса. Результаты сравнения эффективности параметров приведены в таблице 8.

Рассматривая полученные значения, также необходимо учитывать размер корпуса исследования: низкие абсолютные показатели обусловлены большим количеством авторов-кандидатов. Увеличение количества

возможных кандидатов существенно уменьшает точность, а уменьшение – увеличивает. Так, Джек Грив (2007) в своем исследовании показывает, что при увеличении количества возможных авторов с 2 до 40 точность оценки, основанной на отдельных параметрах, в среднем падает в 10 раз: с 70% до 7%.

Таблица 8. Результаты работы модели атрибуции в зависимости от анализируемых параметров

| Стилеметрические параметры | Точность атрибуции, % |
|---|---|
| Числовые репрезентации текста | |
| 1. Мешок слов (8604 элемента) | 68.1 |
| 2. Мешок слов + *tf-idf* взвешивание | 77.3 |
| 3. Биграммы слов (62585 элементов) | 31.7 |
| 4. Биграммы слов + *tf-idf* взвешивание | 53.8 |
| 5. Символьные 4-граммы (41394 элемента) | 72.5 |
| 6. Символьные 5-граммы (105044 элемента) | 73.3 |
| 7. Символьные 4-граммы + *tf-idf* взвешивание | 82.5 |
| 8. Символьные 5-граммы + *tf-idf* взвешивание | 81.9 |
| 9. Векторное представление слов (вектора *GloVe*) | 55 |
| Лексические параметры | 8.6 |
| 1. Средняя длина слова | 5 |
| 2. Коэффициент лексического разнообразия | 4.8 |
| 3. Отношение *hapax legomena* к количеству уникальных лексем | 3.8 |
| 4. Оценка тональности (*VADER*) | 2.3 |
| Параметры плана выражения знака | 26.3 |
| 1-26. Частотность прописных букв | 16.3 |
| 27-36. Частотность цифр | 4.8 |
| 37-62. Частотность алфавитных символов | 19.8 |
| 63-82. Частотность специальных символов (~ , @, #, $, %, ^, &, *, -, _, = ,+, >, <, [, ], {, }, /, \|) | 4.8 |
| Синтаксические параметры | 36.7 |
| 1-6. Частотность знаков пунктуации (“,”, “.”, “?”,“!”, “:”, “;”) | 13.5 |
| 7-271. Частотность функциональных слов (265 слов) | 35.3 |
| 272. Средняя длина предложения | 5 |
| Морфологические параметры | 6.1 |
| 1-4. Частотность отдельных частей речи (существительное, глагол, прилагательное, наречие) | 6.1 |

При реализации репрезентаций текста для подхода *bag-of-words* мы рассматриваем только отдельные слова и биграммы слов, а для *n*-грамм символов используются только 4-х и 5-символьные последовательности. Оптимальные значения *n* были получены из дополнительного эксперимента, показавшего, что при больших или меньших значениях точность резко падает.

Можно заметить, что в целом представления текста, базирующиеся на элементах плана выражения знака, превосходят представления, анализирующие текст на уровне слов. Одной из причин этого может быть то, что символьные представления текста получены из исходных документов, не проходивших предварительную обработку, т.е. в ходе анализа учитывается каждый авторский выбор: прописные и строчные буквы, знаки препинания, функциональные слова. Другая возможная причина заключается в том, что, помимо особенностей текста, игнорируемых *bag-of-words,* 4-х и 5-символьные *n*-граммы захватывают 56% и 66% всех слов соответственно, учитывая статистическое распределение средней длины слова в английском языке [Norvig, 2013], а также абсолютное большинство обычно коротких функциональных слов.

Еще одно интересное наблюдение заключается в том, что *tf-idf* взвешивание повышает точность атрибуции для всех текстовых представлений. Такой результат был ожидаем, поскольку по своей природе *tf-idf* оценка предназначена для улучшения работы модели и рассмотрения того, насколько специфичен термин для конкретного типа документов (в нашем случае для конкретного автора), а значит она соответствует основной цели извлечения стилеметрических параметров – найти признаки, которые могут наиболее полно и характерно описать авторский стиль, т.е. те, которые типичны для одного автора и редко встречаются у других.

Лучшее значение точности в 82.5% было достигнуто для представления текста в виде набора 4-грамм символов, взвешенных с помощью *tf-idf*. В то время как подход «мешок слов» продемонстрировал меньшую

эффективность, а использование векторного представления слов по сравнению с другими текстовыми репрезентациями не показало впечатляющих результатов, достигнув точности всего в 55%. Это, по-видимому, говорит о том, что использование предварительно обученных векторов слов может быть не так эффективно, как мы предполагали, и полученная точность, вероятно, может быть улучшена с помощью векторов слов, обученных на анализируемом наборе данных.

Что касается отдельных групп параметров, мы заметили, что значение точности выше для объемных наборов, содержащих большое количество характеристик. Так, морфологическая группа включают в себя исключительно параметр частотности частей речи и соотносит тексты с низкой точностью в 6.1%. Лексические параметры анализируют 4 различных характеристики и, объединяя их все, достигает точности 8.6%. Параметры плана выражения знака возвращают 82 значения, и их результирующая точность составляет 26.3%. Синтаксические параметры содержат 272 характеристики и показывают наиболее эффективный результат, корректно определяя автора в 36.7% случаев.

Среди отдельных параметров наиболее эффективными являются классическая частотность функциональных слов с точностью 35.3%, частотность алфавитных символов – 19.8% и частотность прописных букв – 16.3%. Наименьшую эффективность продемонстрировали параметры, возвращающие одно значение: например, оценка тональности текста – 2.3% и отношение *hapax legomena* к количеству уникальных лексем – 3.8%

## ВЫВОДЫ ПО ГЛАВЕ 2

Одним из важнейших аспектов проведения достоверного эксперимента по атрибуции авторства является сбор соответствующего корпуса текстовых материалов. Характеристики анализируемого корпуса могут серьезно повлиять на итоговую точность системы: так, контентно-специфическая информация может серьезно упростить задачу классификации, в результате чего исследователи получат ложно высокие показатели точности. Такие значения не будут отражать реальной эффективности системы и не могут быть получены для текстов другого корпуса. В нашем исследовании мы постарались с помощью узкого ограничения тематики уменьшить влияние контентно-специфической информации и подобрать схожие тексты для каждого из авторов. Корпус также был искусственно сбалансирован по количеству материала. Кроме того, мы провели эксперимент по работе системы атрибуции в относительно неблагоприятных условиях большого количества авторов-кандидатов, что позволило рассмотреть точность параметров для широкого набора возможных авторов.

Техники предобработки текста также оказывают существенное влияние на процесс атрибуции, поскольку они напрямую преобразовывают текст. Чтобы избежать стирания или искажения отдельных черт авторского стиля, такое преобразование должно проводиться со всей осторожностью и последующим учетом извлекаемых стилеметрических параметров. Поскольку первостепенной целью нашего исследования являлся сравнительный анализ и оценка стилеметрических параметров текста, для каждой группы параметров мы использовали различные комбинации техник предобработки.

Процесс атрибуции авторства строится на анализе параметров текста, предположительно отражающих уникальный стиль автора. В нашей работе мы проанализировали четыре группы параметров: лексические, символьные, синтаксические и морфологические. Зачастую принадлежность какого-либо

параметра к определенной группе достаточно условна и зависит от подхода исследователей: так, функциональные слова можно отнести как к единицам лексического уровня, поскольку анализ проходит на уровне слов, так и к элементам синтаксиса, отражающим грамматические отношения. Поскольку данное исследование проводилось в рамках лингвистики, параметры распределены по группам в первую очередь в соответствии с их функциональными ролями, а не планом воплощения.

Проанализировав отдельные характеристики лексического уровня, мы пришли к выводу о том, что они могут использоваться для атрибуции авторства только в совокупности с другими наборами значений. Все параметры данной группы стремятся описать текст одним числом, что не может дать всеобъемлющего представления сложной природы текста, однако добавление этих параметров в систему атрибуции увеличивает ее точность. Отдельная оценка лексических параметров нашего корпуса позволяет атрибутировать тексты с точностью 8.6%. Наиболее эффективная лексическая репрезентация в виде *tf-idf* взвешенного «мешка слов» позволяет достигнуть значения в 77.3%.

Рассмотрение элементов плана выражения знака по своей природе является наиболее формальным и далеким от лингвистики. Параметры базируются на подсчете частотности отдельных типов символов без возможности корректной лингвистической интерпретации и обоснования полученных значений. Символьные характеристики базируются на предположении о том, что любой текст состоит из символов и, может быть представлен в виде количества вхождений каждого из уникальных знаков. Причем такое представление будет косвенно отражать все присутствующие в тексте речевые структуры и выборы, поскольку любое письменное выражение мысли обличено в знаки. Результаты анализа показывают, что рассмотрение отдельных символов действительно является наиболее эффективным для описания авторского стиля. Полноценная символьная репрезентация текста позволяет среди 50 кандидатов установить автора с

точностью более 82%. Набор отдельных символьных параметров позволяет корректно идентифицировать автора примерно одной четверти текстов.

Синтаксические параметры являются одними из наиболее лингвистически обоснованных, поскольку они стремятся отразить характерные для индивида структуры и способы выражения мысли. По сравнению с другими наборами характеристик, значения синтаксических параметров позволяют достигнуть наибольшего значения точности атрибуции – 36.7%. Морфологические параметры, анализируемые в нашем исследовании, представлены лишь частотностью наиболее распространённых частей речи и являются перспективными при комбинации с характеристиками других типов, однако при отдельном использовании правильно соотносится всего 6% текстов.

В общей сложности, наиболее эффективными оказались всеобъемлющие репрезентации текста, учитывающие не отдельные параметры, а стремящиеся представить весь текст целиком. Такие результаты говорят о том, что индивидуальный стиль – это сложный феномен, который отражается в каждом символе, слове и синтаксическом паттерне, а значит – он не может быть адекватно представлен небольшим набором характеристик.

# ЗАКЛЮЧЕНИЕ

Развитие цифровых технологий и сети Интернет значительным образом повлияло на человеческую коммуникацию: на сегодняшний день объемы ежедневно генерируемой текстовой информации увеличиваются в геометрической прогрессии и, как следствие, увеличивается потребность в установлении авторства анонимных текстов. Большинство членов современного общества ежедневно генерирует цифровые текстовые документы, в результате чего формируется объемный корпус авторских текстов, хранящихся на серверах цифровых систем коммуникации (социальные сети, мессенджеры, сотовые операторы). Таким образом, в случае необходимости правоохранительные органы могут получить доступ к достаточному для описания авторского стиля количеству текстового материала и использовать его для атрибуции противоправных текстов. Развитие компьютерных технологий, в свою очередь, позволило существенно усовершенствовать процесс атрибуции авторства и перейти от субъективных методов к более объективным методикам, основанным на количественной оценке параметров текста.

В теоретической части работы мы рассмотрели основные подходы к определению понятия «стиль», историю и этапы становления стилеметрии, современную формулировку проблемы атрибуции авторства и методы ее решения.

При анализе языковых элементов текстовой структуры используются параметры различных уровней: лексического, символьного и грамматического. На сегодняшний день различными исследователями было предложено более 1000 параметров, характеризующих авторский стиль, но универсальный набор характеристик, позволяющий со 100% вероятностью отличить одного автора от другого до сих пор не найден. В каждом частном случае выбор параметров зависит от материалов исследования: учитываются особенности языка, анализируемого дискурса, электронного представления

текстов и используемая для атрибуции модель классификации. Одним из наиболее популярных подходов к классификации текста является машинное обучение, основанное на методе опорных векторов (*SVM*). Такая модель позволяет обрабатывать многомерные массивы параметров и с их помощью разграничивать тексты разного авторства.

Практическая часть исследования посвящена анализу стилеметрических параметров и развертыванию полноценной системы атрибуции авторства. В нашем исследовании рассматривались исключительно языковые черты текста без учета структурных особенностей и метаданных документа. В соответствии с классификацией параметров, представленной в теоретической главе, мы проанализировали единицы лексического уровня, плана выражения знака, синтаксического и морфологического уровней, всего 362 параметра и 9 текстовых представлений. Анализ наиболее распространенных характеристик показал, что все они обладают различающей способностью и могут быть использованы для атрибуции авторства с различной степенью эффективности. Результаты анализа для большей наглядности представлены в виде сравнительных диаграмм.

Практическая реализация системы атрибуции показала, что наиболее эффективными для установления личности автора оказались всеобъемлющие текстовые репрезентации и группы параметров, возвращающие большое количество значений: чем больше значений содержит набор параметров, тем точнее определяется личность автора. Для эксперимента на корпусе 50 авторов максимальная точность атрибуции была получена с помощью представления текста в виде *tf-idf* взвешенных символьных 4-грамм и составила 82.5%. Для достижения такой точности тексты были представлены в виде 41394 значений частотности соответствующих 4-грамм. Среди наборов частных характеристик максимальная точность установления личности автора составила 36.7%. для синтаксических параметров. Полученные значения считаются весьма удовлетворительными,

разработанная система может использоваться для решения реальных практических задач.

Текст представляет собой сложную систему, в которой выражаются индивидуальные особенности и предпочтения языковой личности и, конечно, он не может быть всеобъемлюще описан единичными параметрами. Для построения эффективной системы атрибуции авторства необходимо достичь такого количественного представления текста, которое будет учитывать как можно большее число лингвистических явлений. Нельзя однозначно оценить результативность отдельных параметров, потому что только в совокупности с общим множеством характеристик они представляют собой текст, а по-отдельности содержат лишь частичную информацию, которая не всегда позволяет однозначно отнести документ к тому или иному автору. В случае необходимости ограничить количество стилеметрических параметров, на наш взгляд, предпочтение стоит отдавать более объемным характеристикам, которые часто встречаются в тексте и охватывают наибольший диапазон специфических черт идиостиля.

В перспективе дальнейшего исследования нам представляется актуальным улучшение представленной системы с помощью настройки отдельных параметров и масштабирования значений характеристик; рассмотрение эффективности различных методов атрибуции при использовании одного и того же набора стилеметрических параметров; проведение эксперимента, отражающего зависимость точности атрибуции от количества авторов кандидатов.

## СПИСОК ИСПОЛЬЗОВАННОЙ ЛИТЕРАТУРЫ


1. Ахмедова Ю.А. Идиостиль сонетов И. Северянина из цикла «Медальоны»: дис. … канд. филол. наук : 10.02.01. Челябинск, 2008. 189 с.

2. Базылев В.Н. Общее языкознание: учебное пособие. М. : Гардарики, 2007. 285 с.

3. Будаев Э.В. Зарубежная юридическая лингвистика: становление и проблематика // Эволюция лингвистической экспертизы: методы и приемы : монография. Екатеринбург : Уральский государственный педагогический университет, 2017. С. 7 – 61.

4. Горчакова И.А. Фразеография идиостиля И. Нолль в романе «Мертвый петух» // Известия Уральского государственного университета. Сер. 2, Гуманитарные науки. 2009. № 3 (65). С. 109 – 115.

5. Ермолаева Ю.Е. Классификация стихотворных текстов методом дискриминантного анализа // Вестник ТГУ. 2009. № 7. С. 292 – 296.

6. Журавлева Н.Н. Применение количественных методов при анализе стиля автора и решении проблем атрибуции // Вестник Тюменского государственного университета. 2012. № 1. С. 150 – 155.

7. Каменская Ю.В. Ирония как компонент идиостиля А.П. Чехова : дис. … канд. филол. наук : 10.02.01. Саратов, 2001. 173 с.

8. Караулов Ю.Н. Русский язык и языковая личность. М. : Наука, 1987. 262 с.

9. Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Проблемы передачи информации. 2001. Т. 37, вып.2. С. 96 – 109.

10. Леонова А.В., Леонова И.В. Определение авторства текстов на основе подхода n-грамм // Научное обозрение. Технические науки. 2018. № 6. С. 37 – 40.

11. Литвинова Т.А., Громова А.В. Компьютерные технологии в судебной автороведческой экспертизе: проблемы и перспективы

использования // Вестник Волгоградского государственного университета. Серия 2, Языкознание. 2020. Т. 19, № 1. С. 77 – 88.

12. Валла Л. Рассуждения о подложности так называемой дарственной грамоты Константина // Итальянские гуманисты XV века о церкви и религии. М. : АН СССР, 1963.

13. Мартыненко Г.Я. Методы математической лингвистики в стилистических исследованиях. СПб. : Нестор-История, 2019. 296 с.

14. Мартыненко Г.Я. Основы стилеметрии. Л. : ЛГУ, 1988. 173 с.

15. Марусенко М.А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов.Л. : ЛГУ, 1990. 164 с.

16. Мухин М.Ю. Лексическая статистика и идиостиль автора // Вестник Южно-Уральского государственного университета. Серия: Лингвистика. 2009. № 2(135). С. 51 – 55.

17. Мухин М.Ю. Лексическая статистика и идиостиль автора: корпусное идеографическое исследование (на материале произведений М. Булгакова, В. Набокова, А. Платонова и М. Шолохова) : автореф. дис. … доктора филол. наук : 10.02.19. Екатеринбург, 2011. 43 с.

18. Мюллер А. Введение в машинное обучение с помощью Python. Руководство для специалистов по работе с данными / А. Мюллер, С. Гвидо. Санкт-Петербург : ООО «Альфа-книга, 2017. 480 с.

19. Напреенко Г.В. Идентификация текста по его авторской принадлежности на лексическом уровне (формально-количественная модель) // Вестник Томского государственного университета. 2014. № 379. С. 17 – 23

20. Прикладная и компьютерная лингвистика / под ред. И.С. Николаева, О.В. Митрениной, Т.М. Ландо. изд.2-е. М. : ЛЕНАНД, 2017. 320 с.

21. Резанова З.И., Романов А.С., Мещеряков Р.В. О выборе признаков текста, релевантных в автороведческой экспертной деятельности //

Вестник Томского государственного университета. Филология. 2013. № 6(26). С. 38 – 52.

22. Романов А.С., Мещеряков Р.В. Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. междунар. конф. «Диалог–2009» (Бекасово, 27–31 мая 2009 г.). М. : РГГУ, 2009. Вып. 8 (15). С. 432 – 437.

23. Романов А.С., Мещеряков Р.В. Идентификация авторства коротких текстов методами машинного обучения // Компьютерная лингвистика и интеллектуальные технологии: по матер. ежегод. междунар. конф. «Диалог» (Бекасово, 26–30 мая 2010 г.). М. : Изд-во РГГУ, 2010. Вып. 9 (16). С. 407 – 413

24. Романов А.С., Мещеряков Р.В. Определение пола автора короткого электронного сообщения // Компьютерная лингвистика и интеллектуальные технологии: матер. ежегод. Междунар. конф. «Диалог» (Бекасово, 25 – 29 мая 2011 г.). М.: Изд-во РГГУ, 2011. Вып. 10 (17). С. 620 – 626

25. Романов А.С., Шелупанов А.А., Мещеряков Р.В. Разработка и исследование математических моделей, методик и программных средств информационных процессов при идентификации автора текста. Томск : В-Спектр, 2011. 188 с.

26. Степаненко А.А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений // Вестник Томского государственного университета. 2017. №415. С. 17 – 25.

27. Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. 2-е изд., испр. и доп. М. : Флинта, 2006. 696 с.

28. Суркова А.С. Идентификация авторства текстов на основе информационных портретов // Вестник Нижегородского университета им. Н.И. Лобачевского. 2014. № 3(1). С. 145 – 149.

29. Хетсо Г. Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах «Время» и «Эпоха». Осло : Solum Forlag, 1986. 82 с.

30. Хмелев Д.В. Классификация и разметка текстов с использованием методов сжатия данных [Электронный ресурс] // Всё о сжатии данных, изображений и видео. 2003. URL: http://compression.ru/download/articles/classif/intro.html (дата обращения: 09.01.2021).

31. Хмелев Д.В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестн. МГУ. Сер. 9: Филология. 2000. № 2. С. 115 – 126.

32. Хозяинов С.А. Классификация текстов методами распознавания образов // Вестник Сыктывкарского университета. Серия 1. Математика. Механика. Информатика. 2017. №22. С. 3 – 20.

33. Шевелев О.Г. Методы автоматической классификации текстов на естественном языке : учеб. пособие. Томск : ТМЛ-Пресс, 2007. 144 с.

34. Шеля А., Плехач П., Зеленков Ю. Феномен Батенькова и проблема верификации авторства: многомерный статистический подход к нерешенному вопросу // Acta Slavica Estonica XII. 2020. Вып. 2. С. 131 – 165.

35. Abbasi A., Chen H. Identification and comparison of extremist-group Web forum messages using authorship analysis // IEEE Intelligent Systems. 2005. Vol. 20, № 5. P. 67 – 75.

36. Abbasi A., Chen H. Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace // ACM Transactions on Information Systems. 2008. № 26(2). P. 1 – 29.

37. Ali S., Hussein K. The Comparative Power of Type/Token and Hapax legomena/Type Ratios: a Corpus-based Study of Authorial Differentiation // Advances in Language and Literary Studies. 2014. № 5(3). P. 112 – 119.

38. Argamon S., Saric M., Stein S.S. Style mining of electronic messages for multiple authorship discrimination: first results // Proceedings of the 9th ACM

SIGKDD International Conference on Knowledge Discovery and Data Mining. NY : ACM, 2003. P. 475 – 480.

39. Arlot S., Celisse A. A survey of cross-validation procedures for model selection // Statistics surveys. 2010. № 4. P. 40 – 79.

40. Awad M., Khanna R. Support vector machines for classification // Efficient Learning Machines. 2015. P. 39 – 66.

41. Bagavandas M., Manimannan G. Style consistency and authorship attribution: A statistical investigation // Journal of Quantitative Linguistics. 2008. Vol. 15 (1). P. 100 – 110.

42. Benedetto D., Caglioti E., Loreto V. Language trees and zipping // Physical Review Letters. 2002. № 88(4).

43. Berrar D. Cross-validation // Encyclopedia of bioinformatics and computational biology. 2019. № 1. P. 542 – 545

44. Bozkurt I., Baghoglu O., Uyar E. Authorship attribution // 2007 22nd international symposium on computer and information sciences. 2007. P. 1 – 5.

45. Canales O. A stylometry system for authenticating students taking online tests / O. Canales, V. Monaco, T. Murphy, E. Zych, J. Stewart, C. Tappert, A. Castro, O. Sotoye, L. Torres, G. Truly // Proceedings of Student-Faculty Research Day, CSIS. 2011. P. B4.1 – B4.6

46. Chaski C.E. Empirical evaluations of language-based author identification // Forensic Linguistics. 2001. Vol. 8, № 1. P. 1 – 65.

47. Chaski C.E. Who's at the keyboard? // Authorship attribution in digital evidence investigations. International Journal of Digital Evidence. 2005. Vol. 4(1). P. 1 – 13.

48. Cilibrasi R., Vitanyi P.M.B. Clustering by compression // IEEE Transactions on Information Theory. 2005. № 51(4). P. 1523 – 1545.

49. Clark E., Araki K. Text Normalization in Social Media: Progress, Problems and Applications for a Pre-Processing System of Casual English // Social and Behavioral Sciences. 2011. № 27. P. 2 – 11.

50. De Vel O. Mining e-mail content for author identification forensics / O. de Vel, A. Anderson, M. Corney, G. Mohay // ACM SIGMOD Record. 2001. № 30(4). P. 55 – 64.

51. Eder M. Does size matter? Authorship attribution, small samples, big problem // Digital Scholarship in the Humanities. 2015. №. 30(2). P. 167 – 182.

52. Eder M. Mind your corpus: systematic errors in authorship attribution // Literary and Linguistic Computing. 2013. № 28(4). P. 603 – 614.

53. Elberrichi Z., Aljohar B. N-grams in texts categorization // Scientific Journal of King Faisal University (Basic and Applied Sciences). 2007. № 8(2):1428H.

54. Frantzeskou G., Stamatatos E., Gritzalis S., Katsikas S. Effective identification of source code authors using byte-level information // Proceedings of the 28th International Conference on Software Engineering. NewYork : ACM Press, 2006. P. 893 – 896.

55. Graham N., Hirst G., Marthi B. Segmenting documents by stylistic character // Natural Language Engineering. 2005. № 11(4). P. 397 – 415.

56. Grant T., Baker K Identifying reliable, valid markers of authorship: A reponse to Chaski // Forensic Linguistics. 2001. Vol. 8, № 1. P. 66 – 79.

57. Grieve J. Quantitative Authorship Attribution: An Evaluation of Techniques // Literary and Linguistic Computing. 2007. № 22(3). P. 251 – 270.

58. HaCohen-Kerner Y., Miller D., Yigal Y. The influence of preprocessing on text classification using a bag-of-words representation // PLOS ONE. 2020. № 15(5):e0232525.

59. Hedegaard S., Simonsen J. Lost in translation: authorship attribution using frame semantics // Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. 2011. P. 65 – 70.

60. Herdan G. The advanced theory of language as choice and chance / G. Herdan. Berlin : Springer-Verlag, 1966. 365 p.

61. Hirst G., Feiguina O. Bigrams of syntactic labels for authorship discrimination of short texts // Literary and Linguistic Computing. 2007. Vol. 22(4). P. 405 – 417.

62. Honore A. Some simple measures of richness of vocabulary // Association for Literary and Linguistic Computing Bulletin. 1979. № 7(2). P. 172 – 177.

63. Hoover D. Delta prime? // Literary and Linguistic Computing. 2004. № 19(4). P. 477 – 495.

64. Hutto C., Gilbert E. VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text // Proceedings of the 8th International Conference on Weblogs and Social Media. 2015.

65. Islam S. A Comparative Analysis of Word Embedding Representations in Authorship Attribution of Bengali Literature // 21st International Conference of Computer and Information Technology (ICCIT). 2018. P. 1 – 6.

66. Joachims T. Text categorization with support vector machines: Learning with many relevant features // European conference on machine learning. 1998. P. 137 – 142.

67. Jockers M., Witten D. A comparative study of machine learning methods for authorship attribution // Literary and Linguistic Computing. 2010. № 25(2). P. 215 – 223.

68. Juola P. Future trends in authorship attribution // Advances in Digital Forensics III. 2007. P. 119 – 132.

69. Kadhim A. An Evaluation of Preprocessing Techniques for Text Classification // International Journal of Computer Science and Information Security. 2018. № 16(6). P. 22 – 32.

70. Kannan S., Gurusamy V. Preprocessing Techniques for Text Mining // International Journal of Computer Science and Communication Networks. 2014. № 5(1). P. 7 – 16.

71. Keselj V., Peng F., Cercone N., Thomas C. N-gram-based author profiles for authorship attribution // Proceedings of the Pacific Association for Computational Linguistics. 2003. P. 255 – 264.

72. Khatun A., Rahman A., Islam M.S., Marium-E-Jannat. Authorship Attribution in Bangla literature using Character-level CNN / A. Khatun, A. Rahman, M.S. Islam, Marium-E-Jannat // 22nd International Conference on Computer and Information Technology (ICCIT). 2019. P. 1 – 5.

73. Khmelev D.V., Teahan W.J. A repetition based measure for verification of text collections and for text categorization. // Proceedings of the 26th ACM SIGIR. – NewYork : ACM Press, 2003. P. 104 – 110.

74. Kim D. Multi-co-training for document classification using various document representations: TF–IDF, LDA, and Doc2Vec / D. Kim, D. Seo, S. Cho, P. Kang // Information Sciences. 2019. № 477. P. 15 – 29.

75. Koppel M., Argamon S., Shimoni A. Automatically Categorizing Written Texts by Author Gender // Literary and Linguistic Computing. 2002. № 17(4). P. 401 – 412.

76. Koppel M., Schler J. (2004) Authorship verification as a one-class classification problem // Proceedings of the 21st International Conference on Machine learning. 2004. P. 62 – 68.

77. Koppel M., Schler J. Exploiting stylistic idiosyncrasies for authorship attribution // IJCAI03 Workshop on Computational Approaches to Style Analysis and Synthesis. 2003. P. 69 – 72.

78. Koppel M., Schler J., Argamon S. Authorship attribution in the wild // Language Resources and Evaluation. 2010. № 45(1). P. 83 – 94.

79. Lagutina K. A survey on stylometric text features / K. Lagutina, N. Lagutina, E. Boychuk, I. Vorontsova, E. Shliakhtina, O. Belyaeva, I. Paramonov, P. Demidov // 25th Conference of Open Innovations Association (FRUCT). 2019. P. 184 – 195.

80. Li J., Zheng R., Chen H. From fingerprint to writeprint // Communications of the ACM. 2006. № 49(4). P. 76 – 82.

81. Litvinova T., Litvinova O., Panicheva P. Authorship Attribution of Russian Forum Posts with Different Types of N-gram Features // Proceedings of the 2019 3rd International Conference on Natural Language Processing and Information Retrieval. 2019. P. 9 – 14.

82. Loper E., Bird S. NLTK: The Natural Language Toolkit // Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics. 2002. P. 63 – 70.

83. Martins R. A sentiment analysis approach to increase authorship identification / R. Martins, J.J. Almeida, P. Henriques, P. Novais // Expert Systems. 2019. P. 1 –12.

84. Marton Y., Wu N., Hellerstein L. On compression-based text classification // Proceedings of the European Conference on Information Retrieval. Berlin, Germany : Springer, 2005. P. 300 – 314.

85. Mendenhall T.C. The characteristic curves of composition // Science, IX. 1887. P. 237 – 249.

86. Mosteller F., Wallace D.L. Inference and Disputed Authorship: The Federalist. Reading, MA : Addison-Wesley, 1964. 287 p.

87. Nadkarni P., Ohno-Machado L., Chapman W. Natural language processing: an introduction // Journal of the American Medical Informatics Association. 2011. № 18(5). P. 544 – 551.

88. Nalini K., Sheela L.J. Survey on Text Classification // International Journal of Innovative Research in Advanced Engineering (IJIRAE). 2014. № 1(6). P. 412 – 417.

89. Neal T. Surveying stylometry techniques and applications / T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard // ACM Computing Surveys. 2017. № 50(6). P. 1 – 36.

90. Norvig P. English Letter Frequency Counts: Mayzner Revisited or ETAOIN SRHLDCU [Электронный ресурс] : технические статьи, обзоры,

отчеты и другие материалы Питера Норвига. 2013. URL: http://norvig.com/mayzner.html (дата обращения: 14.03.2021).

91. O'Shea J., Bandar Z., Crockett K. A Multi-classifier Approach to Dialogue Act Classification Using Function Words // Transactions on Computational Collective Intelligence VII. 2012. P. 119 – 143.

92. Oliveira W., Justino E. and Oliveira L. Comparing compression models for authorship attribution // Forensic Science International. 2013. Vol. 228(1–3). P. 100 – 104.

93. Oman W.P., Cook R.C. Programming style authorship analysis // Proceedings of the 17th Annual ACM Computer Science Conference. 1989. P. 320 – 326.

94. Pang B., Lee L. Opinion Mining and Sentiment Analysis // Foundations and Trends in Information Retrieval. 2008. №2(1-2). P. 1 – 135.

95. Patra B.G. Feeling may separate Two Authors: Incorporating Sentiment in Authorship Identification Task / B.G. Patra, S. Banerjee, D. Das, S. Bandyopadhyay // 10th International Conference on Natural Language Processing (ICON 2013). 2013. P. 121 – 126.

96. Pavelec D., Oliveira L., Justino E., Nobre Neto F., Batista L. Compression and stylometry for author identification // 2009 International Joint Conference on Neural Networks. 2009. P. 936 – 940.

97. Peng F. Language independent authorship attribution using character level language models / F. Peng, D. Shuurmans, V. Keselj, S. Wang // Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics. 2003. P. 267 – 274.

98. Pillay S.R., Solorio T. Authorship attribution of web forum posts // 2010 eCrime Researchers Summit, Dallas, TX. 2010. P. 1 – 7.

99. Rajaraman A.. Mining of Massive Datasets / A. Rajaraman, J.D Ullman. New York : Cambridge University, 2011. 513 p.

100. Ramnial H., Panchoo S., Pudaruth S. Authorship Attribution Using Stylometry and Machine Learning Techniques // Advances in Intelligent Systems and Computing. 2015. P. 113 – 125.

101. Sanderson C., Guenter S. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: An investigation // Proceedings of the International Conference on Empirical Methods in Natural Language Engineering. – Morristown, NJ : Association for Computational Linguistics, 2006. P. 482 – 491.

102. Sebastiani F. Machine learning in automated text categorization // ACM Computing Surveys (CSUR). 2002. № 34(1). P. 1 – 47.

103. Segarra S., Eisen M., Ribeiro A. Authorship Attribution through Function Word Adjacency Networks. // IEEE Transactions on Signal Processing. 2015. №. 63(20). P. 5464 – 5478.

104. Stamatatos E., Fakotakis N, Kokkinakis G. Computer-based authorship attribution without lexical measures // Computers and the Humanities. 2001. Vol. 35, № 2. P. 193 – 214.

105. Tamboli M.S., Prasad R.S. A robust authorship attribution on big period // International Journal of Electrical and Computer Engineering (IJECE). 2019. № 9(4). P. 3167 – 3174.

106. Tang X., Liang S., Liu Z. Authorship Attribution of The Golden Lotus Based on Text Classification Methods // Proceedings of the 2019 3rd International Conference on Innovation in Artificial Intelligence. 2019. P. 69 – 72.

107. Tanguy L. Authorship Attribution: Using Rich Linguistic Features when Training Data is Scarce / L. Tanguy, F. Sajous, B. Calderone, N. Hathout // PAN Lab at CLEF. 2012.

108. Torruella J., Capsada R. Lexical Statistics and Tipological Structures: A Measure of Lexical Richness // Social and Behavioral Sciences. 2013. № 95. P. 447 – 454.

109. Tuldava J. Stylistics, author identification // Quantitative linguistics: an international handbook. 2005. P. 368 – 387.

110. Tweedie F., Baayen R. How variable may a constant be? Measures of lexical richness in perspective // Computers and the Humanities. 1998. № 32(5). P. 323 – 352.

111. Uzuner O., Katz B. A comparative study of language models for book and author recognition // Proceedings of the 2nd International Joint Conference on Natural Language Processing. 2005. P. 969 – 980.

112. Vašák P. Metody určování autorství / P. Vašák. – Praha, 1980. – 233 p.

113. Winter W. Styles as dialects // Statistics and style / Edited by L. Doležel, R.W. Bailey. New York : American Elsevier Publishing Company, 1969. P. 3 – 9.

114. Zhao Y., Zobel J. Searching with style: Authorship attribution in classic literature // Proceedings of the 30th Australasian Computer Science Conference. 2007. P. 59 – 68.

115. Zheng R., Li J., Chen H., Huang Z. A framework for authorship identification of online messages: Writing style features and classification techniques // Journal of the American Society of Information Science and Technology. 2006. № 57(3). P. 378 – 393.

## ПРИЛОЖЕНИЕ А

Таблица А.1 – Описание корпуса исследования

| № п/п | Имя автора | Источник | Общее кол-во слов |
|---|---|---|---|
| 1 | Adam | https://travelsofadam.com/ | 3349 |
| 2 | Adventurous Kate | https://www.adventurouskate.com/ | 2873 |
| 3 | Aileen Adalid | https://iamaileen.com/ | 2979 |
| 4 | Alexandra Baackes | https://www.alexinwanderland.com/ | 3836 |
| 5 | Amanda Williams | https://www.dangerous-business.com/ | 4104 |
| 6 | Ana Silva O'Reilly | https://mrsoaroundtheworld.com/ | 3354 |
| 7 | Annette | https://bucketlistjourney.net/ | 2866 |
| 8 | Anton Diaz | https://www.ourawesomeplanet.com/ | 3136 |
| 9 | Barbara Weibel | http://holeinthedonut.com/ | 3248 |
| 10 | Bender | https://travelwithbender.com/ | 3081 |
| 11 | Berliner Sabrina | https://www.justonewayticket.com/ | 3402 |
| 12 | Brendan Van Son | https://www.brendansadventures.com/ | 2989 |
| 13 | Charli Moore | https://wanderlusters.com/ | 3432 |
| 14 | Christy Woodrow | https://ordinarytraveler.com/ | 3094 |
| 15 | Cleila Mattana | https://www.keepcalmandtravel.com/ | 3478 |
| 16 | Clint Johnston | https://triphackr.com/ | 2937 |
| 17 | DJ | http://www.dreameurotrip.com/ | 3100 |
| 18 | Dave Anderson | https://www.jonesaroundtheworld.com/ | 2986 |
| 19 | Derek Earl | https://www.wanderingearl.com/ | 3611 |
| 20 | Erica Clapp | https://youngadventuress.com/author/erica | 3060 |
| 21 | Erin Mc. Neaney | https://www.neverendingvoyage.com/ | 3062 |
| 22 | Janice Waugh | https://solotravelerworld.com/ | 3094 |
| 23 | Jeremy Scott Foster | https://travelfreak.com/ | 3697 |
| 24 | Jodi Ettenberg | https://www.legalnomads.com/ | 3072 |
| 25 | Johnny Ward | https://onestep4ward.com/ | 3214 |
| 26 | Kenneth Surat | http://kennethsurat.com/ | 3176 |
| 27 | Kiki | https://theblondeabroad.com/travel-blog/ | 2864 |
| 28 | Kristin Addis | https://www.bemytravelmuse.com/ | 3446 |
| 29 | Kristin Luna | https://www.camelsandchocolate.com/ | 3562 |

Окончание таблицы А.1

| № п/п | Имя автора | Источник | Общее кол-во слов |
|---|---|---|---|
| 30 | Leyla Giray | www.women-on-the-road.com | 3017 |
| 31 | Liz Carlson | https://youngadventuress.com/author/elizabeth | 3095 |
| 32 | Marek | https://www.indietraveller.co/ | 3001 |
| 33 | Marie-Eve | https://www.toeuropeandbeyond.com/ | 3091 |
| 34 | Mariellen Ward | https://breathedreamgo.com/ | 3512 |
| 35 | Mark Wiens | https://migrationology.com/blog/ | 3141 |
| 36 | Matt Long | https://landlopers.com/ | 3099 |
| 37 | Matthew Karsten | https://expertvagabond.com/ | 2905 |
| 38 | Michael Turtle | https://www.timetravelturtle.com/ | 3070 |
| 39 | Nomadic Matt | https://www.nomadicmatt.com/ | 2907 |
| 40 | Rachel | https://hippie-inheels.com/ | 3598 |
| 41 | Rachel A. Davis | http://vagabondbaker.com/ | 3117 |
| 42 | Robert Schrader | https://leaveyourdailyhell.com/ | 3230 |
| 43 | Shannon O'Donnell | https://alittleadrift.com/ | 3425 |
| 44 | Sharon | https://www.wheressharon.com/ | 3010 |
| 45 | Sherry | https://www.ottsworld.com/ | 3391 |
| 46 | Stephanie Be | https://www.travel-break.net/ | 2929 |
| 47 | Trisha Velarmino | https://www.psimonmyway.com/ | 3128 |
| 48 | Wes Nations | http://johnnyvagabond.com/ | 3044 |
| 49 | Will Tang | https://goingawesomeplaces.com/ | 3461 |
| 50 | Zoe Dawes | https://www.thequirkytraveller.com/ | 3054 |

## ПРИЛОЖЕНИЕ Б

Полный список стоп-слов, используемых для предварительной обработки текста (265 единиц): *'i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', "you're", "you've", "you'll", "you'd", 'your', 'yours', 'yourself', 'yourselves', 'he', 'him', 'his', 'himself', 'she', "she's", 'her', 'hers', 'herself', 'it', "it's", 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', "that'll", 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'and', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'below', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'further', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', "don't", 'should', "should've", 'now', 'd', 'll', 'm', 'o', 're', 've', 'y', 'ain', 'aren', "aren't", 'couldn', "couldn't", 'didn', "didn't", 'doesn', "doesn't", 'hadn', "hadn't", 'hasn', "hasn't", 'haven', "haven't", 'isn', "isn't", 'ma', 'mightn', "mightn't", 'mustn', "mustn't", 'needn', "needn't", 'shan', "shan't", 'shouldn', "shouldn't", 'wasn', "wasn't", 'weren', "weren't", 'won', "won't", 'wouldn', "wouldn't", 'one', 'two', 'three', 'four', 'five', 'six', 'seven', 'eight', 'nine', 'ten', 'eleven', 'twelve', 'thirteen', 'fourteen', 'fifteen', 'sixteen', 'seventeen', 'eighteen', 'nineteen', 'twenty', 'thirty', 'forty', 'fifty', 'sixty', 'seventy', 'eighty', 'ninety', 'hundred', 'thousand', 'first', 'second', 'third', 'fourth', 'fifth', 'sixth', 'seventh', 'eighth', 'ninth', 'tenth', 'eleventh', 'twelfth', 'thirteenth', 'fourteenth', 'fifteenth', 'sixteenth', 'seventeenth', 'eighteenth', 'nineteenth', 'twentieth', 'thirtieth', 'fortieth', 'fiftieth', 'sixtieth', 'seventieth', 'eightieth', 'ninetieth', 'hundredth', 'thousandth'*.

**ПРИЛОЖЕНИЕ В**

Полный список функциональных слов (265 единиц): *'a', 'about', 'above', 'across', 'after', 'afterwards', 'again', 'against', 'all', 'almost', 'alone', 'along', 'already', 'also', 'although', 'always', 'am', 'among', 'amongst', 'amoungst', 'an', 'and', 'another', 'any', 'anyhow', 'anyone', 'anything', 'anyway', 'anywhere', 'are', 'around', 'as', 'at', 'be', 'became', 'because', 'been', 'before', 'beforehand', 'behind', 'being', 'below', 'beside', 'besides', 'between', 'beyond', 'both', 'but', 'by', 'can', 'cannot', 'could', 'despite', 'did', 'do', 'does', 'done', 'down', 'during', 'each', 'eg', 'either', 'else', 'elsewhere', 'enough', 'etc', 'even', 'ever', 'every', 'everyone', 'everything', 'everywhere', 'except', 'few', 'first', 'for', 'former', 'formerly', 'from', 'further', 'had', 'has', 'have', 'he', 'hence', 'her', 'here', 'hereafter', 'hereby', 'herein', 'hereupon', 'hers', 'herself', 'him', 'himself', 'his', 'how', 'however', 'i', 'ie', 'if', 'in', 'indeed', 'inside', 'instead', 'into', 'is', 'it', 'its', 'itself', 'last', 'latter', 'latterly', 'least', 'less', 'lot', 'lots', 'many', 'may', 'me', 'meanwhile', 'might', 'mine', 'more', 'moreover', 'most', 'mostly', 'much', 'must', 'my', 'myself', 'namely', 'near', 'need', 'neither', 'never', 'nevertheless', 'next', 'no', 'nobody', 'none', 'noone', 'nor', 'not', 'nothing', 'now', 'nowhere', 'of', 'off', 'often', 'on', 'once', 'one', 'only', 'onto', 'or', 'other', 'others', 'otherwise', 'ought', 'our', 'ours', 'ourselves', 'out', 'outside', 'over', 'per', 'perhaps', 'rather', 're', 'same', 'second', 'several', 'shall', 'she', 'should', 'since', 'so', 'some', 'somehow', 'someone', 'something', 'sometime', 'sometimes', 'somewhere', 'still', 'such', 't', 'than', 'that', 'the', 'their', 'them', 'themselves', 'then', 'thence', 'there', 'thereafter', 'thereby', 'therefore', 'therein', 'thereupon', 'these', 'they', 'third', 'this', 'those', 'though', 'through', 'throughout', 'thru', 'thus', 'to', 'together', 'too', 'top', 'toward', 'towards', 'under', 'until', 'up', 'upon', 'us', 'used', 'very', 'via', 'was', 'we', 'well', 'were', 'what', 'whatever', 'when', 'whence', 'whenever', 'where', 'whereafter', 'whereas', 'whereby', 'wherein', 'whereupon', 'wherever', 'whether', 'which', 'while', 'whither', 'who', 'whoever', 'whole', 'whom', 'whose', 'why', 'whyever', 'will', 'with', 'within', 'without', 'would', 'yes', 'yet', 'you', 'your', 'yours', 'yourself', 'yourselves'*.

Федеральное государственное автономное
образовательное учреждение
высшего образования
«СИБИРСКИЙ ФЕДЕРАЛЬНЫЙ УНИВЕРСИТЕТ»
Институт филологии и языковой коммуникации
Кафедра теории германских языков и межкультурной коммуникации

УТВЕРЖДАЮ
Заведующий кафедрой ТГЯиМКК
_____ О.В. Магировская
« 18 » июня 2021 г.

МАГИСТЕРСКАЯ ДИССЕРТАЦИЯ

# АТРИБУЦИЯ АВТОРСТВА НА ОСНОВЕ ОЦЕНКИ СТИЛЕМЕТРИЧЕСКИХ ПАРАМЕТРОВ ТЕКСТА (НА МАТЕРИАЛЕ АНГЛОЯЗЫЧНЫХ ТУРИСТИЧЕСКИХ БЛОГОВ)

45.04.02 Лингвистика
45.04.02.01 Межкультурная коммуникация и перевод

Магистрант _____ Л.А. Вдовина

Научный руководитель _____ д-р филол. наук, зав. каф. РЯиПЛ А.В. Колмогорова

Нормоконтролер _____ Я.М. Янченко

Красноярск 2021