НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

МАТЕРИАЛЫ К ГОСУДАРСТВЕННОМУ ЭКЗАМЕНУ

НАУЧНЫЙ ДОКЛАД

по результатам подготовленной научно-квалификационной работы

(диссертации)

«Лингвистическое моделирование как инструмент атрибуции текста»

ФИО Хоменко Анна Юрьевна

Направление: Языкознание и литературоведение

Профиль (направленность): 10.02.19 Теория языка

Аспирантская школа по филологическим наукам

Аспирант _____/А. Ю.Хоменко

подпись

Научный руководитель _____/Т.В.Романова

подпись

Директор Аспирантской школы _____/ А.С.Варенкова

подпись

Нижний Новгород, 2020 г.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

В современной лингвистике справедливо главенствует междисциплинарный подход к исследованиям языка, ученые пытаются совместить знания нескольких областей при анализе языковых реалий. Сходная тенденция имеет место и в ветви лингвистики, занимающейся вопросами атрибуции текстового материала. Данная отрасль в отечественной науке является достаточно консервативной. В судебной лингвистике в подавляющем большинстве случаев используются исключительно методы качественного лингвистического анализа [Вул 1973, 1977, 1982, 1992, 2007; Галяшина 2005а,б, Рубцова, Ермолаева, Безрукова и др. 2007]. В сфере внесудебных, научных атрибуционных исследований методы качественного анализа на современном этапе развития науки практически не разрабатываются, в то время как количественные методики осваиваются очень активно [Захаров 2000, Labbe, Labbe, 2001, Koppel, Schler, 2003, Coulthard 2004, Juola, Sofko, Brennan 2006, Марусенко 2003, Родионова 2008а, б, Романов 2010, Медведева 2010, Батура 2012, Мартыненко 2015, Korobov 2015, Wright 2016, Litvinova, Sboev, Panicheva 2018, Custódio, Paraboni 2018, Murauer, Tschuggnall, Specht 2018, Panicheva, Mirzagitova, Ledovaya 2018, Bacciu, Morgia, La 2019, Muttenthaler, Lucas, Amann 2019, Gomzin, Laguta, Stroev 2018, и пр.]. Эта тенденция сформировалась в англоязычной школе, где автороведение традиционно связано с количественными, стилометрическими методами анализа [Campbell L. 1867, Mendenhall 1887, Lutoslawski 1897, Jule 1944, Mosteller, Wallace 1964, Somers 1972, Foster 1989, 2001; Merriam 1989, 2003, Holmes 1994, Holmes, Forsyth 1995, Baayen, van Halteren, Tweedie, 1996].

На современном этапе существует острая необходимость интеграции качественного и количественного анализа в атрибуции. Становится все более очевидным, что построение моделей авторских идиостилей лишь на основании

традиционных стилостатистических данных не может в полной мере удовлетворить атрибуционную лингвистику. Попытки интеграции разнородных подходов впервые появились в начале века в России и на Западе и имеют место до настоящего момента [Баранов 2001: 43–52; Koppel, Schler 2003, Хоменко 2013; 2014а,б,в, 2018, 2019а,б,в, г, 2020; Pimonova, Durandin, Malafeev 2020].

В настоящей работе речь идёт именно об интегративной методике атрибуционного анализа текста, основанной на соединение результатов интерпретативного исследования материала и объективации этих результатов посредством математической статистики. Методика реализуется в следующем наборе шагов: 1) автоматическое извлечение из текста параметров, описывающих идиостиль с точки зрения прагматикона, тезауруса и лексикона автора; 2) поиск традиционных стилеметрических текстовых данных; 3) присвоение веса каждому параметру; 4) построение математических моделей сравниваемых текстов; 5) сравнение математических моделей с целью выявления уровня их корреляции между собой.

Актуальность исследования определяется тем, что в современном обществе особое значение имеет аутентификация письменного материала. Так, методики атрибуции становятся необходимыми в филологических экспертизах при определении авторства известных художественных произведений (статьи Ф.М.Достоевского, работы М.А.Шолохова), в судебных автороведческих экспертизах при решении диагностических и идентификационных задач, при анализе контента сети Интернет на предмет автоматического поиска содержания деликтной направленности (текстов, связанных с террористической, экстремистской деятельностью, педофилией и пр.), при решении научных задач, связанных с идентификацией лиц по письменной речи и определением характеристик пола, возраста, социального статуса этих лиц. Каждое из перечисленных научных и практических полей деятельности нуждается в использовании полных и всесторонних методик атрибуции, дающих

объективные результаты. Использование количественных, автоматизированных атрибуционных моделей несет в себе опасность, связанную с неполнотой описания объекта моделирования, оригинала. Количественные модели обычно используют несколько легко вычленимых атоматически текстовых параметров (длины слов и предложений, распределение текстового материала по частям речи и пр.), делая конечную модель неполной, значит, недостаточно объективной. Квалификативные методы анализа могут конструировать достаточно полную лингвистическую модель текста, но эта модель всегда интерпретативна, соответственно, тоже не наделена достаточным уровнем объективности. Интеграция же двух этих подходов позволит сделать модель достаточно полной, всесторонне имитирующей оригинал и одновременно объективной.

В основе исследования лежит **гипотеза** о том, что интегративная атрибуционная модель, получаемая в результате ряда итераций (а) построение лингвистических моделей языковых личностей авторов сравниваемых текстов; б) объективации моделей посредством методов математической статистики; в) сравнение моделей с помощью разных статистических метрик) способна успешно решать идентификационную задачу атрибуционной лингвистики на текстах любого объема.

Создание интегративной методики текстовой атрибуции, основанной на сочетании квантитативного и квалификативного подходов к анализу текста в рамках модельной лингвистики, и разработка программного обеспечения на ее основе является **целью** настоящего исследования.

Достижение поставленной цели предполагает решение следующих **задач**:

– на основе анализа массива исследований в области прикладной (в том числе компьютерной) лингвистики, теоретической лингвистики, лингвистики моделей сформировать теоретическую базу исследования;

– разработать проблему наиболее используемых квантитативных и квалификативных методов и методик анализа; адаптировать их для применения

в прототипе автоматизированном программном комплексе по атрибуции текста;

– сформулировать рабочее понятие модели, создать дизайн ее архитектуры; применить их при создании автоматизированного программного комплекса по атрибуции текста.

Объектом исследования являются методы и методики текстовой атрибуции, модели текстовой атрибуции, имеющие квантитативную, квалификативную и интегративную основы.

Предметом исследования становится определение уровня работоспособности различных атрибуционных параметров, методов, методик атрибуции и атрибуционных моделей на различном текстовом материале.

Материалом для анализа стали отечественные и зарубежные методы и методики, алгоритмы и программные комплексы, предназначенные для решения атрибуционных задач, а также атрибуционные параметры различных уровней, используемые в разных атрибуционных подходах. Автором была оценена их работоспособность, описаны положительные и отрицательные компоненты их применения (глава 1). Также автор после вычленения наиболее удачных компонентов и атрибуционных параметров указанных методик и программных комплексов апробировал их на аутентичном разнородном текстовом материале (тексты официально-делового стиля, корпоративная переписка, короткие текстовые сообщения, публицистика и пр.) с целью поиска наилучшего сочетания параметров идентификации автора письменного текста (глава 3). На основе полученных результатов и разработки проблемы лингвистического моделирования (глава 2) автор с командой программистов-разработчиков создал прототип автоматизированного программного продукта, основанный на принципах модельной лингвистики и сочетающий в модели атрибуции как качественные, так и количественные ее параметры (глава 4).

Методико-методологическая база исследования представляет собой сочетание принципов когнитивной лингвистики, психолингвистики и

структурной лингвистики в совокупности с исследованиями в области компьютерных наук. Методы когнитивной лингвистики и психолингвистики, как то: анализ языковой личности автора письменного текста по методике Ю.Н. Караулова, анализ компетенций языковой личности по методике С.М. Вула и Е.С. Горошко – помогают установить глубинную природу речевых структур, что позволяет объективно оценить сущностность получаемой атрибуционной модели. Традиционный структуралистский подход к языку как уровневой системе (с применением морфологического, словообразовательного, лексического, синтаксического, семантического анализов) позволяет упростить сложные когнитивные структуры для их машинной формализации. Компьютерные методы анализа речевых структур (преимущественно автоматическая обработка текстов с целью вычленения идентификационных параметров) дают возможность преобразовать лингвистические модели в математические (с помощью методов математической статистики и теории вероятности) и сравнить их (посредством корреляционного анализа).

Основным **инструментом** исследования стал язык программирования

Степень разработанности проблемы. Атрибуционная лингвистика со времен Л. Кэмпбелла [Campbell 1867] и В. Лютославского [Lutoslawski 1897] на западе и Н.А. Морозова [Морозов 1916] в России всегда шла двумя параллельными путями: путем стилеметрии [Mendenhall 1887, Mosteller, Wallace 1964, Захаров 2000, , Merriam 2003, Labbe, Labbe, 2001, Juola, Sofko, Brennan 2006, Мартыненко 2015, Litvinova, Seredin, Litvinova, etc., 2017, Wright 2017, Karlgren, Esposito, Gratton, etc. 2018, и пр.] и путем качественного анализа текста [Вул 1973, 2007; Горошко 2003, Комиссаров 2001, McMenamin 2002, Галяшина 2003, Coulthard 2004 и пр.]. На современном этапе развития исследовательского поля координация двух указанных ветвей происходит посредством объяснения стиметрических данных с точки зрения традиционной квалификативной

лингвистики: объяснение длины предложения как отражения уровня компетенций автора в письменной речи [Степаненко 2017:19-20], объяснение n-грамм как косвенной экспликации грамматических текстовых реалий [Захаров, Хохлова 2008: 41-42]. Разработкой вопроса формализации глубинных синтаксических структур занимается санкт-петербургская школа прикладной лингвистики [Марусенко 1990; Родионова 2008 и пр.]. Интегративный подход, основанный на сочетании анализа традиционных стилостатистических параметров (длин слов и предложений, наиболее частотных n-грамм, служебных слов и POS-tags) и анализа авторских идиосинкразем, одними из первых предложили М. Коппел и Дж. Шлер [Koppel, Schler 2003]. В России он разрабатывается на базе разных методик [Баранов 2001: 43–52; Хоменко 2013; 2014а,б,в, 2018, 2019а,б,в, г, 2020; Pimonova, Durandin, Malafeev 2020].

В настоящее время существует ряд значимых научных мероприятий, на которых обсуждаются результаты современных исследований в области автоматической обработки текста с целью его атрибуции. В России к таким мероприятиям относится ежегодная международная конференция «Диалог» (URL: http://www.dialog-21.ru/), международная конференция «AINL: Artificial Intelligence and Natural Language Conference» (URL: https://ainlconf.ru/), международная конференция «International Conference on Analysis of Images, Social Networks and Texts» (URL: https://aistconf.org/). За рубежом – серия мероприятий «PAN» в рамках «Conference and Labs of the Evaluation Forum, или Cross-Language Evaluation Forum», (URL: https://pan.webis.de/). В рамках данных мероприятий формируется тенденция к полностью автоматизированным системам, использующим разные модели, алгоритмы и метрики. Большой популярностью до сих пор пользуются различные модели и алгоритмы, основанные на исследовании n-граммного состава речи индивида [Bacciu, Morgia, La 2019], [Litvinova, Sboev, Panicheva 2018], [Custódio, Paraboni 2018], [Murauer, Tschuggnall, Specht 2018], [Muttenthaler, Lucas, Amann 2019], частеречной

отнесенности единиц [Litvinova, Sboev, Panicheva 2018], различного рода длин [Custódio, Paraboni 2018] с использованием кластеризационного подхода [Panicheva, Mirzagitova, Ledovaya 2018], традиционных [Gomzin, Laguta, Stroev 2018] и модифицированных [Korobov 2015] библиотек Python, алгоритмов векторных преобразований [Bacciu, Morgia, La 2019] и пр. Существуют также удачные попытки использования собственно лингвистических моделей (с применением векторного подхода к анализу текста) для определения авторства текста: [Pimonova, Durandin, Malafeev 2020].

Научная новизна работы заключается в создании аутентичной атрибуционной методики, включающей в себя как методы традиционного анализа текста и исследования языковой личности с когнитивной точки зрения, так и машинный автоматический анализ речи пишущего. Важно, что настоящая методика не только разработана, но и формализована: на ее основе создан прототип атрибуционного программного обеспечения открытого доступа, основанный на принципах модельной лингвистики.

Теоретическая значимость исследования состоит в разработке архитектуры лингвистической модели, наиболее универсальной для атрибуции текстов любого объема и жанровой отнесенности и подходящей для формализации с помощью современных компьютерных инструментов.

Практическая ценность работы определяется возможностью применения полученной методики и программного обеспечения для целей аутентификации художественных произведений, идентификации автора письменного текста в судебном автороведении.

Основное содержание работы заключено в следующую структуру:

Введение.

Глава 1. Проблема атрибуции текста. История вопроса.

1.1. Атрибуция текста в конце XIX в. Зарождение стилостатистики.

1.2. Атрибуция текста в начале XX века. Отечественная и зарубежная школы.

1.3. Атрибуция текста в середине – второй половине XX века.

1.4. Атрибуция текста в конце XX – начале XXI в.

1.5. Современное состояние текстовой атрибуции.

Выводы к главе 1.

Глава 2. Лингвистические модели и их свойства.

2.1. Природа модели, вариации понятия, свойства моделей.

2.2. Классификация моделей по разным основаниям.

2.3. Критерии оценки качества модели с теоретической точки зрения.

Выводы к главе 2.

Глава 3. Оценка качества существующих квалификативных и квантитативных атрибуционных моделей.

3.1. Общетеоретическое рассмотрение проблемы.

3.2. Конкретные случаи применения формальных моделей.

3.3. Конкретные случаи применения квалификативных моделей.

3.3.1. Кейс 1.

3.3.2. Кейс 2.

3.3.3. Кейс 3.

3.4. Интегрирование квантитативных и квалификативных моделей.

3.4.1. Кейс 1.

3.4.2. Кейс 2.

3.4.3. Кейс 3.

3.4.4. Кейс 4.

Выводы к главе 3.

Глава 4. Создание прототипа интегративного полуавтоматизированного алгоритма идентификации автора письменного текста.

4.1. Описание алгоритма.

4.2. Теоретическая оценка качества разработанной модели.

4.3. Апробация алгоритма.

Приложения 1-7

Краткое содержание работы.

Примат именно интегративного подхода к решению задач текстовой атрибуции обусловлен возможностями, которые предоставляет междисциплинарность исследований. Методы интерпретативной лингвистики выявляют информацию об авторских компетенциях в традиционном понимании (тезаурус личности, ее прагматикон, уровни владения компетенциями письменной речи), а стилостатистика дает возможность сделать результаты интерпретативного анализа объективными. Более того, такой подход к анализу текста в теории должен быть универсальным и решать задачи атрибуции как в научных целях, так и в прагматических. Одновременно он должен решать проблему атрибуции текстов разного объема и жанровой отнесенности.

Настоящее исследование выдвигает концепцию прототипа программного обеспечения, основанную, с одной стороны, на анализе авторских компетенций с точки зрения структурированной языковой личности по Ю.Н. Караулову [Караулов 1987] и С.М. Вулу [Вул 1973, 2007], а с другой – на объективации качественных исследовательских данных количественными. При этом традиционный анализ языковой личности осуществляется не вручную, а с помощью автоматической обработки естественного языка. Такой подход дает возможность максимально автоматизировать процесс атрибуции и при этом получить достоверные результаты.

Алгоритм анализа начинается с того, что автор определяет на основании анализа теоретического материала ряд параметров языковой личности, которые заведомо в той или иной степени должны идентифицировать авторский идиостиль и одновременно могут быть извлечены из текста автоматически с минимальным предпроцессингом. Речь идет о том, что данные параметры

должны быть относительно универсальны для любого текста и их должно быть легко извлекать, используя некоторые предустановленные правила и минимальную текстовую обработку, осуществляемую не вручную экспертом (ручная разметка, выравнивание текстов и пр.), а автоматически (токенизация, присвоение pos-tags). Итак, приведенным выше условиям удовлетворили следующие параметры:

1) реализация прагматикона личности на синтаксическом уровне: вводные слова и конструкции, эксплицирующие субъективную модальность; конструкции со словами «большинство/меньшинство», целевые, выделительные и сравнительные обороты, репрезентирующие уровень освоения автором компетенций письменной речи и его отношение к действительности; синтаксические сращения, дающие представление в том числе о функциональной стилистической отнесенности текста; сравнительные придаточные, глагольные односоставные предложения, эксплицирующие репрезентацию действительности в текстовом материале; обращения;

2) описание тезауруса личности: в данный раздел были включены наиболее частотные сочетания слов, которые описывают грамматико-семантические особенности текста; ключевые лексемы текста; экспликаты аксиологических текстовых доминант дихотомии «свой/чужой»;

3) вербально-семантический уровень авторского лексикона: частеречная отнесенность слов текста (количество глаголов, прилагательных, существительных и пр. частей речи), сложные слова полуслитного написания; модальные частицы, междометия, наличие/отсутствие модального постфикса «-то», предпочтительные слова-интенсификаторы.

Обработка текстов осуществлялась при помощи языка программирования Python. На этапе предпроцессинга тексты разделяются на предложения с помощью стандартной библиотеки NLTK с уточнением использование русской модели для обработки текстов, тексты подвергаются токенизации, словам текста

присваиваются частеречные теги с грамматическими характеристиками с помощью Pymorthy2.

Для анализа синтаксических структур были прописаны правила, основанные на pos-tags, как то, например: экспликаты субъективной модальности (вводные слова): 1)__,Prnt,__ 2)<начало предложения> Prnt,__ со списком вводных слов; целевые обороты: с целью/из расчёта + INFN; глагольные односоставные предложения, например, определённо-личные: есть VERB в 1per или 2per в sing или plur в pres или futr в indc, нет подлежащего, то есть нет: NOUN или NPRO в nomn в sing или plur NUMR + NOUN7 в nomn в sing или plur много/мало/несколько + NOUN8 в gent/ gent2 в sing или plur у + NOUN NPRO в gent/ gent2 в sing или plur NOUN или NPRO в datv в sing или plur и пр.

Настоящие формулы были протестированы на обширном текстовом материале (учебные тексты для РКИ объемом 4000 предложений). Для поиска заданных грамматических моделей использовались регулярные выражения (модуль Re).

Этот же алгоритм поиска осуществляется при отборе параметров, имеющих морфологическую отнесенность, например, модального постфикса «-то»: POST-то, кроме NPRO, NPRO в nomn, gent, datv, accs. ablt, loct, voct, gen1, gen2, acc2, loc1, loc2 в sing или plur, APRO в nomn, gent, datv, accs. ablt, loct, voct, gen1, gen2, acc2, loc1, loc2 в sing или plur. После извлечения указанной морфолого-синтаксической информации из текстов реализуется подсчет абсолютной частоты встречаемости каждого признака, затем абсолютные частоты переводятся в относительные, что позволяет сравнивать тексты разных объемов. Подсчет ipm (instance per million) для лексического материала проводится стандартным способом: количество употреблений лексемы в тексте, поделенное на объем текста и умноженное на 1 миллион. Для синтаксических параметров количество каждого параметра делится на количество предложений в тексте.

Установление наиболее частотных сочетаний слов для текстов осуществляется после описанного выше предпроцессинга, при подсчете также учитывается отсутствие слова в списке стоп-слов из модуля NLTK, кириллическое написание и длина слова более 2 символов. В результате при сравнении двух текстов для каждого формируется список наиболее частотных сочетаний слов, числовой метрикой для которых также служит ipm.

Ключевые лексемы определяются с помощью алгоритма логарифмического правдоподобия при сравнении интересующего текста с референтным корпусом большого объема (использовался корпус «Opencorpora», URL: http://opencorpora.org, дата обращения: 08.02.2020, объемом на дату обращения 1540034 слова). В результате для каждого текста получаем список ключевых слов с числовой экспликацией значения меры логарифмического правдоподобия (loglikelihood score, или LL). В конечный список включаются лишь слова со значением LL более 50.

При анализе ключевых лексем и наиболее частотных сочетаний слов из полученных списков удаляются сочетания с личными именами и именами собственными, поскольку данные лексемы маркируют не собственно особенности авторских идиостилей, а тематическую отнесенность текстов.

Под экспликатами аксиологических текстовых доминант групп «свой/чужой» в настоящем исследовании понимается дисперсия местоимений «я/мы-группы», «ты/они-группы», то есть ведется подсчет местоимений всех разрядов в прямых и косвенных падежах по соответствующим группам.

Под словом-интенсификатором подразумевается лексема, используемая для определения степени семантической категории интенсивности. Чаще всего говорят о наречиях-интенсификаторах, круг их хоть и велик, но ограничен (*очень, сильно, адски* – из современного дискурса). Тем не менее категория интенсивности не исчерпывается исключительно наречным наполнением, например: *Какая красота!*, - в данном случае интенсификатором служит

местоимение *какая*. Так, в исследовании был создан свод правил для поиска структур с интенсификаторами; в список интенсификаторов входят как наречия с некоторыми грамматическими ограничениями (авторы не осуществляют поиск структур, где наречие не эксплицирует категорию интенсивности, например, является частью составного именного сказуемого: *Он чувствует себя хорошо*), так и некоторые прилагательные и местоимения в соответствующих грамматических конструкциях, как то: ADJ «настоящий» в nomn, accs в sing или plur + NOUN: *настоящий бардак*. Метрикой для каждого найденного слова в конечной модели служит ipm.

Для каждого текста также определяется ряд традиционных стилеметрических параметров: средняя длина слова, средняя длина предложения и количество предложений объемом более 8 слов, то есть длинных предложений, некоторые традиционные стилостатистические индексы (Флеша-Кинкейда, предметности, качественности (по Б.Н.Головину [Головин 1970])).

Далее все полученные данные сводятся в две математические модели, которые сравниваются между собой посредством коэффициента корреляции Пирсона, доказывающего или опровергающего гипотезу H_0 о том, что автором двух сравниваемых текстов является одно лицо. То есть, по сути, решается традиционная идентификационная задача, подразумевающая наличие спорного текста с неизвестным авторством и текста-образца, автор которого заведомо известен. Эти математические модели в некотором объеме описывают авторские индивидуальные стили, поэтому если стили разные, модели должны иметь статистически значимые различия, которые отражаются на отношениях между параметрами. Релевантность применения коэффициента корреляции Пирсона для сравнения математических моделей авторских идиостилей описана, например, у [Радбиль, Маркина 2019].

Что касается значения коэффициента корреляции для сравнения математических моделей идиостилей авторов и их языковых личностей, то

необходимо интерпретировать это значение не так, как оно интерпретируется в математической статистике безотносительно к текстовой атрибуции. Так, для традиционной математической статистики значимым считается коэффициент, равный 0,6, в то время как для текстовой атрибуции это значение не является значимым. По экспериментальным данным, чтобы признать тексты объемом от 20 000 знаков принадлежащими одному автору, коэффициент корреляции должен быть выше 87% [Радбиль, Маркина 2019: 164].

Описанная выше модель получила реализацию в виде прототипа программного продукта, размещенного в открытом доступе в сети Интернет, URL: http://khorom-attribution.ru/#/.

Архитектура программы выглядит следующим образом:

I. Модуль автоматического анализа текста:

1) традиционная стилостатистика:

А) индекс Флеша-Кинкейда[1];

Б) индекс туманности Ганнинга (или фог-индекс, Fog Index)[2];

В) индексы по Б.Н. Головину [Головин 1970]:

- коэффициент предметности (Рг) — отношение суммы существительных и местоимений к сумме прилагательных и глаголов.

- коэффициент качественности (Qu) — отношение суммы прилагательных и наречий к сумме глаголов и существительных.

- коэффициент активности (Ac) — отношение суммы глаголов и глагольных форм к количеству слов в тексте.

[1] Индекс удобочитаемости — «интуитивное понятие сложности / легкости текста для чтения и связанной с этим скорости чтения и понимания текста в лингвистике XX века было формализовано в виде индексов удобочитаемости (readability). В их основе лежит ряд презумпций — например, о том, что:
1. короткие предложения читать легче, чем длинные; 2. длинные слова затрудняют чтение;
3. читатель замедляется или «спотыкается», встречая низкочастотные и / или незнакомые ему слова и т. п.» [Ляшевкая, URL: https://ling.hse.ru/data/2016/12/15/1111563794/Readability%20talk.pdf].

[2] Индекс туманности Ганнинга – «показывает уровень удобочитаемости текста. Этот метод проверки комфорта восприятия текста назван по имени создателя Роберта Ганнинга. Вначале он предназначался для журналистов, чтобы избежать туманных формулировок в написанном. В настоящее время используется копирайтерами для определения степени простоты текста для читателей» [URL: https://ru.megaindex.com/support/faq/index-gunninga].

- коэффициент динамизма (Din) — отношение суммы глаголов и глагольных форм к сумме существительных, прилагательных и местоимений.

- коэффициент связности текста (Con) — отношение суммы предлогов и союзов к числу предложений [Радбиль, Маркина 2019: 159].

Г) идентификатор слов несловарного написания;

Д) определение длин:

- определение средней длины слова;

- определение средней длины предложения;

- определение количества предложений объемом более 8 слов, то есть длинных предложений;

2) автоматическая обработка текста, основанная на принципах когнитивной лингвистики:

А) реализация прагматикона личности на синтаксическом уровне (осуществляется по предустановленным правилам):

- наличие предложений с однородными рядами;

- предложения с обособленными приложениями;

- вставные конструкции;

- сопоставительные придаточные;

- сложные синтаксические конструкции;

- вводные слова и конструкции,

- целевые, выделительные и сравнительные обороты

- синтаксические сращения,

- сравнительные придаточные,

- глагольные односоставные предложения;

- обращения;

Б) описание тезауруса личности:

- наиболее частотные сочетания слов (словные n-граммы),

- ключевые лексемы текста;

- экспликаты аксиологических текстовых доминант дихотомии «свой/чужой»;

В) вербально-семантический уровень языковой личности:

- сложные слова полуслитного написания;
- модальные частицы;
- междометия;
- наличие/отсутствие модального постфикса «-то»;
- предпочтительные слова-интенсификаторы (рис. 1, 2).

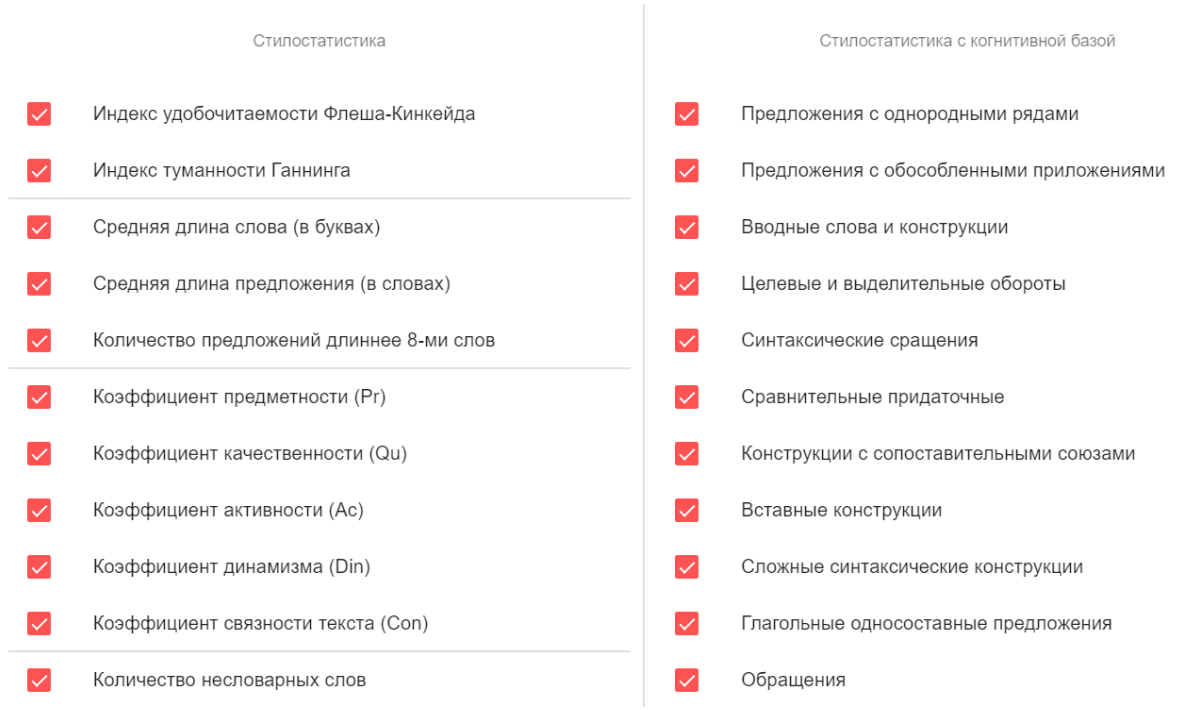

Рис. 1. Набор предустановленных текстовых параметров ресурса: http://khorom-attribution.ru/#/ (1)

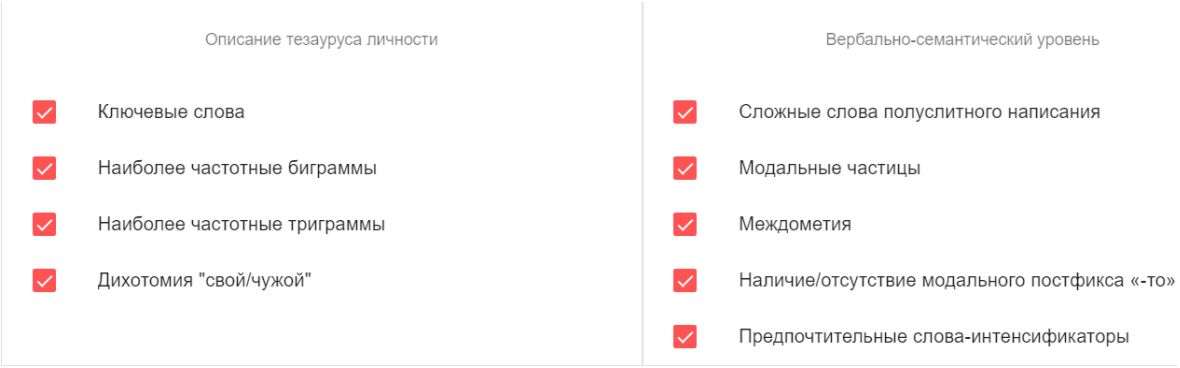

Рис. 2 Набор предустановленных текстовых параметров ресурса: http://khorom-attribution.ru/#/ (2)

II. Модули для пользователя:

1) Ввод данных (рис.3): на вход подаются два текста A и B; пользователь имеет возможность ввести текст с клавиатуры или загрузить его с компьютер (рис.4), а также выбрать жанр (рис. 5).

Рис. 3. Визуализация процедуры ввода данных ресурса: http://khorom-attribution.ru/#/

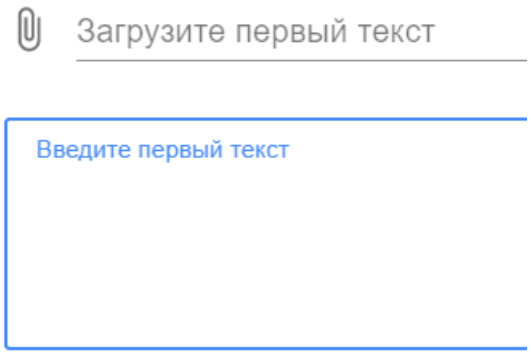

Рис. 4. Типы ввода данных ресурса: http://khorom-attribution.ru/#/

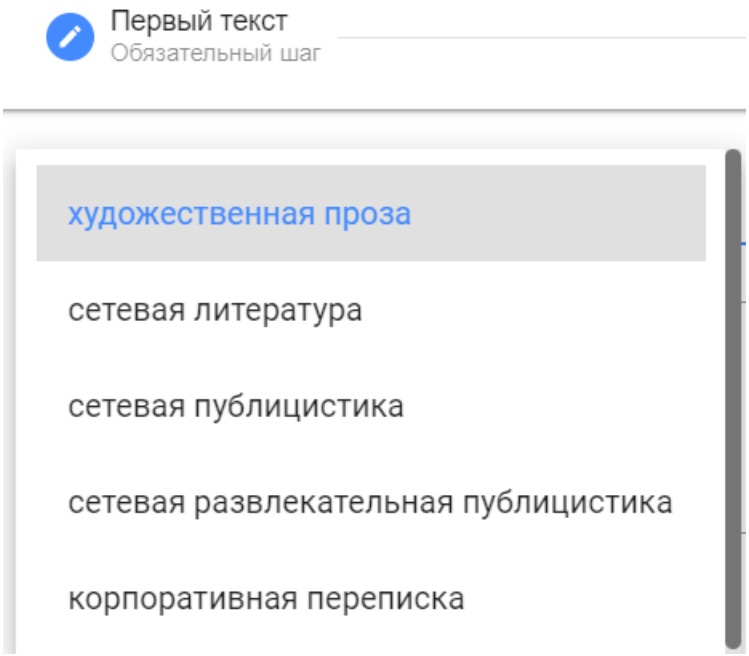

Рис. 5. Визуализация процедуры выбора жанра текста для ввода ресурса: http://khorom-attribution.ru/#/

2) Автоматическая обработка текстового материала:

А) для каждого текста считаются все параметры, затем всем параметрам присваиваются числовые значения (относительные частоты или значения мер loglikelihood и пр.), все они сводятся в две модели: модель для текста А и модель для текста В. Для этих двух моделей высчитывается коэффициент корреляции Пирсона;

Б) пользователь может строить модель не только на основе предустановленных параметров, но и имеет возможность выбирать те, которые считает наиболее релевантными для определенной пары текстов, те, которые, по его мнению, дают наибольший прирост информации (функция множественного выбора (рис. 6)). Этот функционал выгодно отличает разработанное программное обеспечение от подобных атрибуционных программ, основанных, например, на машинном обучении, где все параметры предустановлены не пользователем, а разработчиком. Это же делает настоящий ресурс не полностью автоматическим (что может быть крайне важным как для исследовательских

целей, так и для судебного автороведения, где полная автоматизация процесса идентификации недопустима).

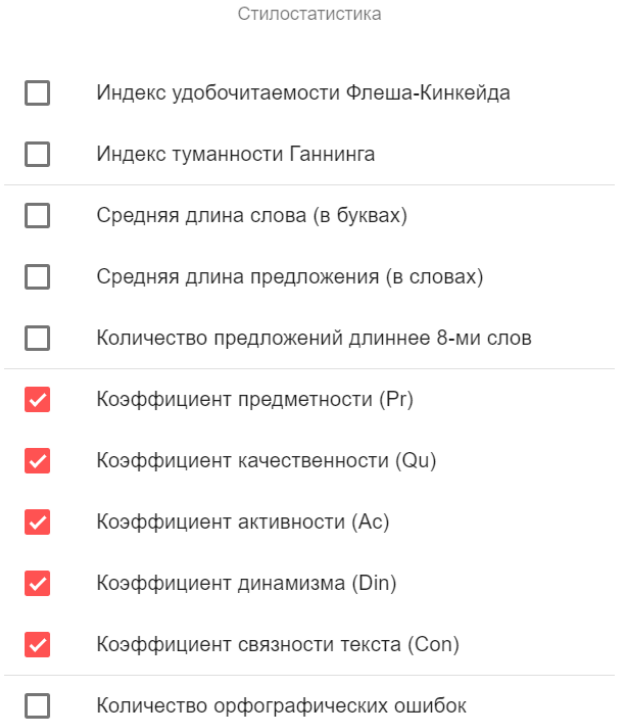

Рис. 6. Визуализация функции множественного выбора ресурса: http://khorom-attribution.ru/#/

3) Вывод данных.

А) в качестве результата выводится значения коэффициента корреляции Пирсона, значение линейной регрессии, критерия Стьюдента для моделей двух сравниваемых текстов, а также коэффициенты корреляции по отдельным параметрам значения каждого параметра для двух текстов совместно с числовым значением для каждого параметра (рис. 7):

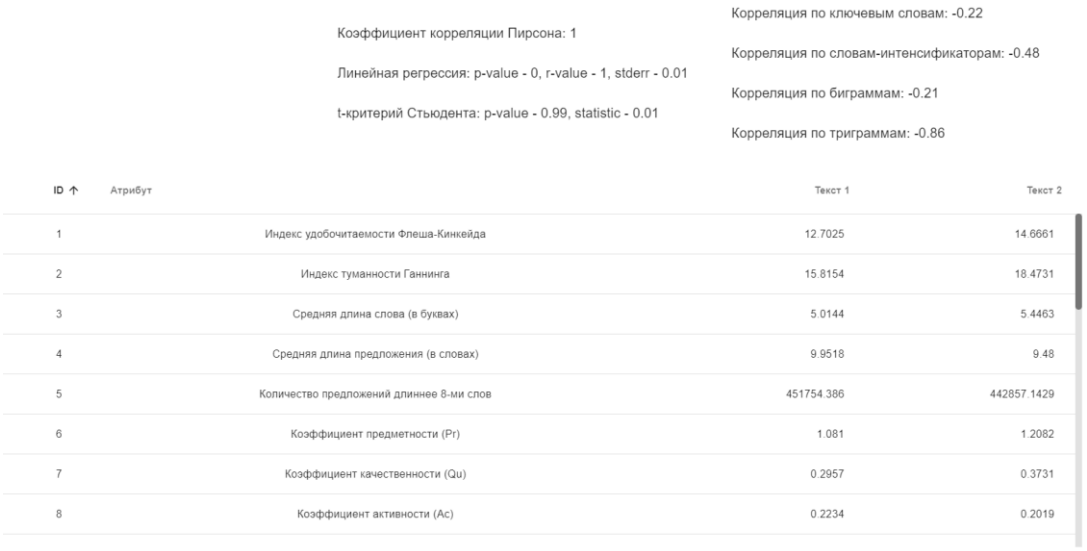

Рис. 7. Визуализация вывода результата ресурса: http://khorom-attribution.ru/#/

Б) для пользователя также запрограммирован модуль проверки: выбрав вкладку *Вспомогательные параметры* можно просмотреть все выделенные в тексте компоненты, для которых рассчитаны относительные частоты и иные метрики (рис. 8, 9).

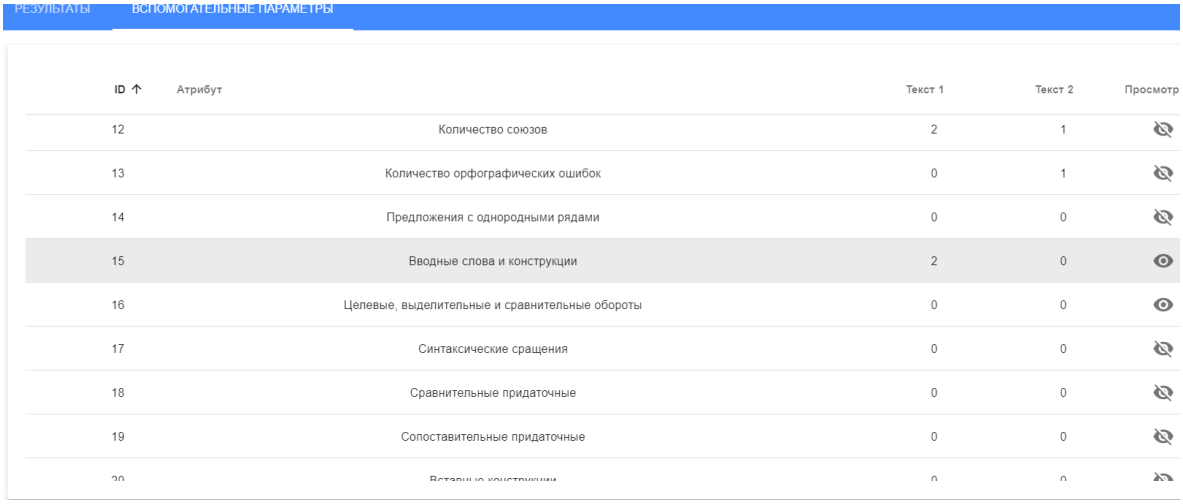

Рис.8. Визуализация функции просмотра подробного результата исследования на ресурсе: http://khorom-attribution.ru/#/ (1)

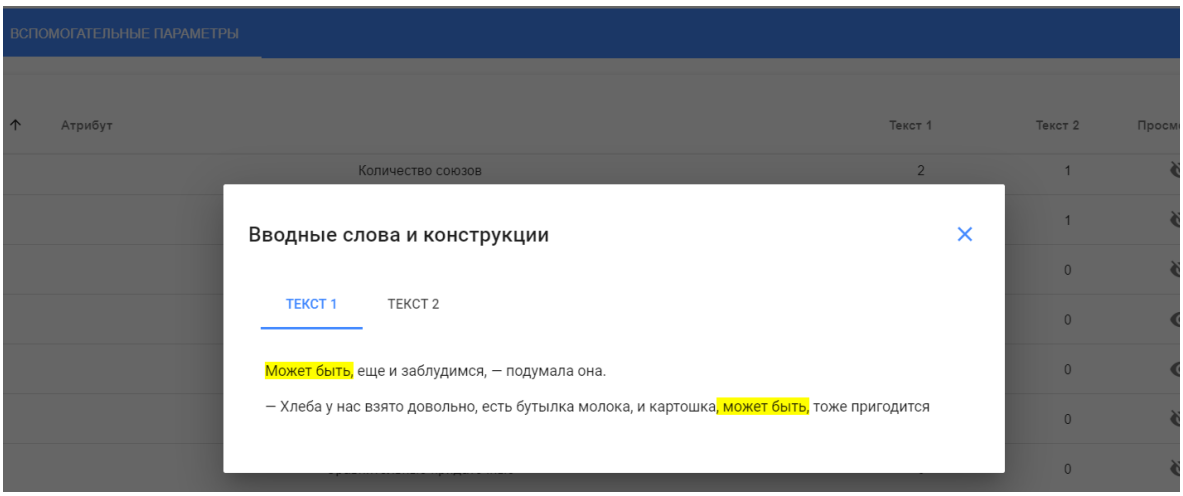

Рис.9. Визуализация функции просмотра подробного результата исследования на ресурсе: http://khorom-attribution.ru/#/ (2)

Интегративная методика, основанная на использовании подходов интерпретативной и когнитивной лингвистики в совокупности с методами традиционной стилеметрии, безусловно, даёт свои результаты. Важно, что интерпретативную часть анализа не обязательно должен делать специалист собственноручно, выделение идентификационных критериев, как показали экспериментальные исследования, можно автоматизировать, причем имеется возможность автоматизировать процесс без предварительной ручной обработки текстов и без применения синтаксических парсеров. Разработанное и описанное в настоящем диссертационном исследовании программное обеспечение построено на принципах когнитивного анализа языковой личности автора текста, что позволяет просто понять его представителю любой традиционной русскоязычной научной школы. Объективация данных интерпретативной части при работе ресурса происходит посредством простых метрик – коэффициента корреляции Пирсон, линейной регрессии и коэффициента Стьюдента. Традиционные стилостатистические параметры (индексы и коэффициенты) также интуитивно понятны пользователю без инженерного образования. Все это упрощает процесс пользования ресурсом и понимание атрибуционной модели, лежащей в его основе.

С теоретической, фундаментальной точки зрения получаемые с помощью ресурса модели авторских идиостилей и языковых личностей весьма эффективны для решения атибуционных задач. Так, обратимся к разработанной автором диссертационного исследования системе оценки качества моделей. Поведем анализ описанных моделей в рамках указанной классификации (Таблицы 1, 2):

Таблица 1. Оценка качества модели идиостиля автора текста (для каждого нового текста вне сравнении двух текстовых моделей), создаваемой ресурсом: http://khorom-attribution.ru/#/

№	Критерий (наименование)	Шкала оценки
1.	**Полнота модели**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных параметров когнитивного пула. Тем не менее уровень полноты настоящей модели высок по сравнению с уровнем полноты любой модели, созданной только на основе традиционных стилостатистических параметров, или модели, созданной на основе интерпретативного подхода.
2.	**Простота модели**	высокий уровень: модель состоит из традиционных стилостатистических параметров и компонентов, широко распространённых в структурной и когнитивной лингвистике (предложения разных типов, дихотомия «свой/чужой», ключевые слова и пр.)
3.	**Точность модели**	наличие: возможность выполнения операций представляемым моделью формальным аппаратом имеет место, функционал модели (описание идиостиля автора как фрагмента его языковой личности) однозначен и выполняется моделью
4.	**Экономичность модели**	высокий уровень: модель компактна и экономична
5.	**Адекватность модели**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных

		параметров когнитивного пула. Тем не менее уровень адекватности настоящей модели высок по сравнению с уровнем адекватности любой модели, созданной только на основе традиционных стилостатистических параметров, или модели, созданной на основе интерпретативного подхода
6.	**Единство в своей раздельности**	наличие возможности разбиения на подмножества (наглядно показано в пользовательском модуле ресурса: совокупность разноуровневых параметров – стилостатистика, стилостатистика с когнитивной базой с параметрами разных уровней языковой личности внутри)
7.	**Цельность модели**	наличие: наличие связи между подмножествами элементов кортежа, создающей неделимую в своём единстве структуру. Цельность наглядно продемонстрирована при делении модели на блоки: стилостатистика, стилостатистика с когнитивной базой.
8.	**Структурность модели**	а) наличие (компонентная структура модели ярко выражена: внутри каждого блока наличествует иерархичность: уровень → компонент уровня → компоненты подуровня – стилостатистика с когнитивной базой → вербально-семантический уровень → слова-интенсификаторы); б) удачный выбор «принимающего» субстрата: состоятельность облечения текстовых данные в числовую форму неоспорима еще со времен зарождения стилостатистики [Lutoslawski W. 1897]; в) удачная организация структуры модели доказывается компонентностью структуры, иерархичной организацией компонентов, а также цельностью модели
9.	**Экспланаторность**	наличие: модель идиостиля, репрезентированного в тексте, посредством присвоения каждому параметру своего веса даёт возможность лучше понять объект

		анализа, объясняет особенности идиостиля конкретного автора (насколько динамичным или связным является его текст), позволяет понять структуру языковой личности (преобладание конкретных компонентов на вербально-семантическом, тезаурусном или прагматическом уровнях)
10.	**Эвристичность модели (как частный случай экпланаторности)**	высокий уровень: модель позволяет получать новые знания об объекте, то есть о языковой личности автора текста посредством анализа его идиостиля, позволяет предсказывать поведение объекта (например, при очень низком индексе удобочитаемости ряда текстов одного автора следует предположить, что любой последующий текст из генеральной совокупности текстов одного автора также будет обладать низким уровнем удобочитаемости) и моделировать сходные объекты на основе заданных параметров (с помощью нейронных сетей или методов машинного обучения)
11.	**Коммуникативность модели (с точки зрения языка)**	наличие: модель языковая, лингвистическая
12.	**Дедуктивность модели**	а) наличие: для формирования компонентов модели были использованы имеющиеся научные выкладки (от В. Лютославского и Н.А. Морозова до Д. Райта и Т. Литвиновой); б) высокий уровень оперирования собственно языковыми, лингвистическими методами анализа как основой для моделирования обусловлен освоением большого объема междисциплинарного материала в процессе создания модели: методы интерпретативной, структурной, когнитивной лингвистики, математической статистики, стилостатистики
13.	**Интерпретируемость модели**	а) наличие; б) простота подстановки: можно переименовать компоненты модели (поменять сигнатуру) без вредя для ее структуры

14.	**Математичность, точность, однозначность модели**	а) полный, целостный аппарат формализации модели; б) удачная работа этого аппарата как основа для машинной реализации модели
15.	**Уровень формализации модели**	математический уровень
16.	**Уровень технически-точного отражения объекта моделирования**	высокий уровень: удачный способ формализации модели, выбор сигнатуры (метаязык прост и понятен пользователю); удачная машинная реализация (простой, понятный интерфейс)
17.	**Уровень реально-жизненного отражения объекта моделирования**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных параметров когнитивного пула. Тем не менее уровень отображения объекта настоящей модели значительно выше, чем уровень отображения объекта любой модели, созданной только на основе традиционных стилостатистических параметров, или модели, созданной на основе интерпретативного подхода
18.	**Уровень субъективизма в модели**	низкий уровень: в модели применяется объективация методами математической статистики данных, полученных с помощью интерпретативного лингвистического подхода. Всем параметрам присваиваются веса в виде относительных частот или специализированных метрик, исключающих субъективизм в подсчетах
19.	**Уровень существенности модельных признаков (уровень абстракции (идеализации) модели)**	высокий уровень: за счет сочетания параметров традиционного стилостатистического и когнитивного пулов модель реализует принцип «золотого сечения» для текстовой атрибуции. Так, модель не является очень абстрактной, состоящей только из цифр текстовой статистики (длины, индексы и пр.), но и не является компонентно слишком разбитой (в ней не описывается каждая особенность авторского стиля конкретного текста, выделены лишь наиболее яркие, имеющие

		относительно универсальную, наиболее высокую идентификационную мощность компоненты)
20.	**Уровень действенности**	высокий уровень: модель описывает достаточно компонентов идиостиля автора, являющегося составной частью его (автора) языковой личности, чтобы этого автора идентифицировать
21.	**Уровень функциональной и практической направленности модели**	а) полностью соответствует целевому использованию: б) обладает высоким уровнем соответствия
22.	**«Гипотезная мощность»**	наличие: модель каждого текста, вступая в отношения дихотомии с моделью текста, с которой она сравнивается (текст 1 и текст 2, текст А и текст Б), оформляет гипотезу H_0 или H_1 о том, что автором двух текстов является одно лицо или разные лица соответственно
23.	**Эстетические свойства модели (опционально)**	наличие: модель выглядит гармоничной и понятной

Таблица 2. Оценка качества модели идентификации автора текста (собственно процесса определения верности/неверности гипотезы H_0), ресурс: http://khorom-attribution.ru/#/

№	Критерий (наименование)	Шкала оценки
1.	**Полнота модели**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных параметров когнитивного пула. Тем не менее уровень полноты настоящей модели высок по сравнению с уровнем полноты любой модели, созданной только на основе традиционных стилостатистических

		параметров, или модели, созданной на основе интерпретативного подхода
2.	**Простота модели**	высокий уровень: модель интуитивно понятна пользователю, поскольку построена по принципу сравнения спорного текста (текста, автор которого неизвестен) и текста-образца (текста, автор которого заведомо известен)
3.	**Точность модели**	наличие: возможность выполнения операций представляемым моделью формальным аппаратом имеет место, функционал модели (сравнение идиостилей двух авторов как фрагментов их языковых личностей на предмет схожести этих моделей) однозначен и выполняется моделью
4.	**Экономичность модели**	высокий уровень: модель компактна и экономична
5.	**Адекватность модели**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных параметров когнитивного пула. Тем не менее уровень адекватности настоящей модели высок по сравнению с уровнем адекватности любой модели, созданной только на основе традиционных стилостатистических параметров, или модели, созданной на основе интерпретативного подхода
6.	**Единство в своей раздельности**	наличие возможности разбиения на подмножества (наглядно показано в пользовательском модуле ресурса: два текста для сравнения – текст с неизвестным авторством и текст-образец с известным авторством; внутри каждой модели совокупность разноуровневых параметров)

7.	**Цельность модели**	наличие: наличие связи между подмножествами элементов кортежа, создающей неделимую в своём единстве структуру. Цельность наглядно продемонстрирована при делении модели на блоки: модель текста 1 (А), модель текста 2 (Б), стилостатистика, стилостатистика с когнитивной базой – для каждой модели.
8.	**Структурность модели**	а) наличие (компонентная структура модели ярко выражена: модель текста 1 (А), модель текста 2 (Б); внутри каждого блока наличествует иерархичность: уровень → компонент уровня → компоненты подуровня – стилостатистика с когнитивной базой → вербально-семантический уровень → слова-интенсификаторы); б) удачный выбор «принимающего» субстрата: состоятельность облечения текстовых данные в числовую форму неоспорима еще со времен зарождения стилостатистики [Lutoslawski W. 1897]; в) удачная организация структуры модели доказывается компонентностью структуры, иерархичной организацией компонентов, а также цельностью модели
9.	**Экспланаторность**	высокий уровень: модель не только позволяет выяснить, является ли то или иное лицо автором конкретного текста, но и даёт информацию о причинах наблюдаемого факта, посредством своих метрик она разъясняет, почему подтверждает гипотеза H_0 или H_1. Это

		происходит потому, что модель позволяет увидеть числовые данные для каждого объекта. Например, модель подтверждает гипотезу H_0, а ее расшифровка показывает, что для двух текстов практически идентичны значения всех индексов и коэффициентов, а также близки значения параметров когнитивного пула: ключевые слова, слова-интенсификаторы, типы предложений
10.	**Эвристичность модели (как частный случай экпланаторности)**	наличие: модель позволяет получать новые знания об объектах. Так, до начала применения модели исследователь не знает, одно или разные лица являются автором двух сравниваемых текстов, после применения модели у исследователя появляется данная информация
11.	**Коммуникативность модели (с точки зрения языка)**	наличие: модель языковая, лингвистическая
12.	**Дедуктивность модели**	а) наличие: для формирования компонентов модели были использованы имеющиеся научные выкладки (от В. Лютославского и Н.А. Морозова до Д. Райта и Т. Литвиновой); б) высокий уровень оперирования собственно языковыми, лингвистическими методами анализа как основой для моделирования обусловлен освоением большого объема междисциплинарного материала в процессе создания модели: методы интерпретативной, структурной, когнитивной лингвистики, математической статистики, стилостатистики

13.	**Интерпретируемость модели**	а) наличие; б) простота подстановки: можно переименовать компоненты модели (поменять сигнатуру) без вредя для ее структуры
14.	**Математичность, точность, однозначность модели**	а) полный, целостный аппарат формализации модели; б) удачная работа этого аппарата как основа для машинной реализации модели
15.	**Уровень формализации модели**	математический уровень
16.	**Уровень технически-точного отражения объекта моделирования**	высокий уровень: удачный способ формализации модели, выбор сигнатуры (метаязык прост и понятен пользователю); удачная машинная реализация (простой, понятный интерфейс)
17.	**Уровень реально-жизненного отражения объекта моделирования**	средний уровень: повышение уровня возможно при увеличении числа запрограммированных параметров когнитивного пула. Тем не менее уровень отображения объекта настоящей модели высок по сравнению с уровнем отображения объекта любой модели, созданной только на основе традиционных стилостатистических параметров, или модели, созданной на основе интерпретативного подхода
18.	**Уровень субъективизма в модели**	низкий уровень: в модели применяется объективация методами математической статистики данных, полученных с помощью интерпретативного лингвистического подхода. Всем параметрам присваиваются веса в виде относительных частот или специализированных метрик, модели сравниваются классическим статистическим способом, с

		помощью коэффициента корреляции Пирсона
19.	**Уровень существенности модельных признаков (уровень абстракции (идеализации) модели)**	высокий уровень: за счет сочетания параметров традиционного стилостатистического и когнитивного пулов модель реализует принцип «золотого сечения» для текстовой атрибуции. Так, модель не является очень абстрактной, состоящей только из цифр текстовой статистики (длины, индексы и пр.), но и не является компонентно слишком разбитой (в ней не описывается каждая особенность авторского стиля конкретного текста, выделены лишь наиболее яркие, имеющие относительно универсальную, наиболее высокую идентификационную мощность компоненты)
20.	**Уровень действенности**	высокий уровень: модель описывает достаточно компонентов идиостиля автора, являющегося составной частью его (автора) языковой личности, чтобы этого автора идентифицировать; более того, модель использует удачную для сравнения моделей идиостилей метрику
21.	**Уровень функциональной и практической направленности модели**	а) полностью соответствует целевому использованию: б) обладает высоким уровнем соответствия
22.	**«Гипотезная мощность»**	наличие: модель идентификации автора при наличии двух текстов (спорного текста и текста-образца) априорно включает в себя гипотезу H_0 и H_1 о том, что автором двух текстов является одно лицо или разные лица соответственно

23.	**Эстетические свойства модели (опционально)**	наличие: модель выглядит гармоничной, сбалансированной и понятной

Из приведённого выше видим, что была разработана алгоритмическая (последовательность приказов (команд)), по Ю.Д. Апресяну, модель. Отладка (тестирование) данной модели проводится на следующей коллекции текстов разных жанров и регистров:

1) подкорпус текстов сетевой газеты «The Village» (https://www.the-village.ru/), разбитый по авторам[3], включает в себя тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов

А) Юлия Рузманова: https://www.the-village.ru/users/978109/posts. 100 случайных текстов.

Б) Алена Дергачева: https://www.the-village.ru/users/1444805/posts. 100 случайных текстов.

В) Ольга Карасева: https://www.the-village.ru/users/1376462/posts. 100 случайных текстов.

Г) Андрей Яковлев: https://www.the-village.ru/users/1268495/posts. 100 случайных текстов.

Д) Лёва Левченко: https://www.the-village.ru/users/1345043/posts. 100 случайных текстов.

Е) Кирилл Руков: https://www.the-village.ru/users/1372356/posts. 100 случайных текстов;

2) подкорпус текстов развлекательного портала «ЯПлакалъ» (https://www.yaplakal.com/), разбитый по авторам[4], включает в себя тексты 3 авторов-женщин, 3 авторов-мужчин; всего 600 текстов:

[3] Выборка авторов является целевой и регулируется гендером автора (3 автора-женщины, 3 автора-мужчины) и количеством публикаций на сайте (не менее 100 публикаций).

[4] Выборка авторов является целевой и регулируется гендером автора (3 автора-женщины, 3 автора-мужчины) и количеством публикаций на сайте (не менее 100 публикаций).

А) автор под псевдонимом KalinAKalina (https://www.yaplakal.com/members/member258716.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=e00f165756c5fb34e7070249db899089&search_in=titles. 100 случайных текстов.

Б) автор под псевдонимом Изюбрина (https://www.yaplakal.com/members/member251139.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=39e2a40544acd7aa0e33eda51311d25c&search_in=titles. 100 случайных текстов.

В) автор под псевдонимом motya (https://www.yaplakal.com/members/member15978.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=e8d6a1bb59846b351859b207270278b6&search_in=titles. 100 случайных текстов.

Г) автор под псевдонимом SESHOK (https://www.yaplakal.com/members/member328581.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=90b5c6fa48664bf9f471a68e5e415b52&search_in=titles. 100 случайных текстов.

Д) автор под псевдонимом OBrian (https://www.yaplakal.com/members/member232578.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=484c23c05e727845097ba93c98fec9ac&search_in=titles&result_type=&hl=&st=25. 100 случайных текстов.

Е) автор под псевдонимом InGrib (https://www.yaplakal.com/members/member401507.html): https://www.yaplakal.com/?act=Search&nav=tu&CODE=show&searchid=608ec6cbc40e320ffeee10d69f4292ec&search_in=titles. 100 случайных текстов.

3) Подкорпус текстов ресурса сетевой литературы «Книга фанфиков» (https://ficbook.net/), разбитый по авторам[5], включает в себя тексты 3 авторов-женщин, 4 авторов-мужчин; всего 190 текстов;

А) автор под псевдонимом Хана_Вишнёвая (https://ficbook.net/authors/3191?show=about#profile-tabs): https://ficbook.net/authors/3191/profile/works#profile-tabs. 50 случайных текстов.

Б) автор под псевдонимом Кицунэ Миято (https://ficbook.net/authors/486110?show=about#profile-tabs): https://ficbook.net/authors/486110/profile/works#profile-tabs. 50 случайных текстов.

В) автор под псевдонимом Ктая (https://ficbook.net/authors/16464?show=about#profile-tabs): https://ficbook.net/authors/16464/profile/works#profile-tabs. 50 случайных текстов.

Г) автор под псевдонимом миха француз (https://ficbook.net/authors/1179435): https://ficbook.net/authors/1179435/profile/works#profile-tabs. 10 случайных текстов.

Д) автор под псевдонимом Аллесий (https://ficbook.net/authors/1644702?show=about#profile-tabs): https://ficbook.net/authors/1644702/profile/works#profile-tabs. 10 случайных текстов.

Е) автор под псевдонимом Tigrewurmut (https://ficbook.net/authors/1469649?show=about#profile-tabs): https://ficbook.net/authors/1469649/profile/works#profile-tabs. 10 случайных текстов.

[5] Выборка авторов является целевой и регулируется гендером автора (3 автора-женщины, 4 автора-мужчины) и его популярностью на ресурсе (все авторы взяты из 100 наиболее популярных авторов на дату 15.07.2020: https://ficbook.net/authors?tab=popular#popular).

Ж) автор под псевдонимом Rakot (https://ficbook.net/authors/407214/profile/works#profile-tabs): https://ficbook.net/authors/407214/profile/works#profile-tabs: все тексты.

4) Подкорпус текстов художественной литературы (нежанровая проза), включающий тексты С.Д.Довлатова и В.П.Астафьева (все доступные оцифрованные тексты авторов художественной направленности);

5) Подкорпус текстов корпоративной русскоязычной переписки, разбитый по авторам, включает в себя тексты 2 авторов-женщин, 2 авторов-мужчин за один год (2018 г.).

В силу того, что любая модель строится «на основе гипотезы о возможном устройстве оригинала и представляет собой функциональный аналог оригинала, что позволяет переносить знания с модели на оригинал» [Медведева 2010: 4], справедливым будет заметить, что «критерием адекватности модели служит практический эксперимент» [там же]. Именно экспериментальным способом была проверена работоспособность (точность и эффективность) созданного ресурса, в основе которого лежит смешанная (вероятностная + детерминистская), специфическая, прикладная, компьютерно-символьная атрибуционная модель.

Апробация исследования.

2020: X Международный конгресс по когнитивной лингвистике «Когнитивно-дискурсивная парадигма в лингвистике и смежных науках: современные проблемы и методология исследования» (Екатеринбург). Доклад: Компьютерные методы анализа для определения гендерной принадлежности текста. Опыт практического исследования

2019 г.: 1) IX Международный конгресс по когнитивной лингвистике «Интегративные процессы в когнитивной лингвистике» (Нижний Новгород). Доклад: Моделирование когнитивных структур на основе методик автороведческого анализа; 2) Современная теоретическая лингвистика и

проблемы судебной экспертизы (Москва). Доклад: Лингвистическая атрибуционная экспертиза в отечественной и зарубежной школах. перспективы развития автороведческих методик в России; 3) Artificial Intelligence and Natural Language (AINL) (Тарту). Доклад: Linguistic Modeling as a Technique in Forensic Authorship Attribution; 4) Лингвополитическая персонология: дискурсивный поворот (Екатеринбург). Доклад: Возможность идентификации лица по разнородному материалу — устной и письменной речи — в условиях одной экспертной задачи; 5) Языковая личность и эффективная коммуникация в современном поликультурном мире (Минск). Доклад: глобальный английский и интернет-сленг. Влияние на языковую личность современной молодежи (на материале русского и китайского языков)

2018: 1) V Международная научно-практическая конференция «Язык. Право. Общество» (Пенза). Доклад: атрибуция текстов малого объёма. статистические закономерности; 2) Массмедийная политическая коммуникация: методы и приемы лингвистического анализа и лингвистической экспертизы (Екатеринбург). Доклад: Лингвистическая атрибуционная экспертиза нехудожественного письменного текста;

2017: Судебно-психологическая экспертиза и комплексные судебные исследования видеозаписей. Доклад: Определение лингвистических признаков подготовленной и неподготовленной речи на материале установленного дословного содержания видеозаписи при производстве комплексной судебно-лингво-фоноскопической экспертизы

2014; 1) Русский язык – язык науки, культуры, коммуникации (Москва). Доклад: Использование корпусного подхода и машинных методов обработки текста для обучения русскому языку как иностранному; 2) Проблемы языковой картины мира в синхронии и диахронии (Нижний Новгород). Доклад: Анализ языковой личности автора текста с применением методов математической статистики как способ установления авторства текста;

2013: 1) Язык. Право. Общество (Пенза). Доклад: Апробация методов математической статистики при атрибуции текста в рамках судебного автороведения; 2) Artificial Intelligence and Natural Language (AINL) 2013 (Санкт-Петербург). Доклад: Алгоритм автоматизации идентификации автора письменного речевого произведения в рамках судебного автороведения.

Основные положения диссертации отражены в следующих публикациях:

1) Хоменко А.Ю. Компьютерные методы анализа для определения гендерной принадлежности текста. Опыт практического исследования. Когнитивные исследования языка. Вып. XXXVIII: Когнитивно-дискурсивная парадигма в лингвистике и смежных науках: современные проблемы и методология исследования: материала X Конгресса по когнитивной лингвистике 17-20 сентября 2020 г./ Отв. ред. вып.: А.П.Чудинов: Екатеринбург: Уральский государственный педагогический университет, 2019. С. 892-896. URL: https://uspu.ru/upload/medialibrary/42a/42ae59185a3e07d02e4b39be5ac98f5b.pdf.

2) Хоменко, А. Ю. Автоматическая обработка текста и лингвистическое моделирование как способы решения проблем атрибуционной лингвистики / А. Ю. Хоменко, Е. Р. Бенькович, Д. И. Гайнутдинова, Л. Р. Гасанова, А. А. Костина, З. О. Мазунина, А. С. Николаева, Е. В. Пимонова // Политическая лингвистика. — 2020. — № 3 (81). — С. 215-224. — DOI 10.26170/pl20-03-22

3) Романова Т. В., Хоменко А. Ю. Функционирование элементов семантического поля социальная значимость в русском и английском языках по данным словарных и корпусных источников // Вестник Санкт-Петербургского университета. Язык и литература. 2020. Т. 17. № 1. С. 49-73. doi

4) Хоменко А. Ю. Возможность идентификации лица по разнородному материалу - устной и письменной речи – в условиях одной экспертной задачи // Лингвополитическая персонология: дискурсивный поворот : материалы Междунар. науч. конф. (Екатеринбург, 29—30 нояб. 2019 г.). [б.и.], 2019. С. 214-

5) Хоменко А. Ю., Зинина А. М. Глобальный английский и интернет-сленг. Влияние на языковую личность современной молодежи (на материале русского и китайского языков) // Языковая личность и эффективная коммуникация в современном поликультурном мире : материалы V Междунар. науч.-практ. конф., посвящ. 20-летию основания каф. теории и практики перевода фак. социокультур. коммуникаций Белорус. гос. ун-та, Минск, 24–25 окт. 2019 г. Мн. : Белорусский государственный университет, 2019. С. 152-158.

6) Хоменко А. Ю. Лингвистическая атрибуционная экспертиза в отечественной и зарубежной школах. перспективы развития автороведческих методик в России // Международная научная конференция «Современная теоретическая лингвистика и проблемы судебной экспертизы». М. : Государственный институт русского языка им. А.С. Пушкина, 2019. С. 536-550.

7) Хоменко А. Ю. Лингвистическое атрибуционное исследование коротких письменных текстов: качественные и количественные методы // Политическая лингвистика. 2019. № 2 (74). С. 177-187. doi

8) Хоменко А. Ю. Моделирование когнитивных структур на основе методик автороведческого анализа // В кн.: Когнитивные исследования языка. Вып. XXXVII: Интегративные процессы в когнитивной лингвистике: материалы международного конгресса по когнитивной лингвистике / Отв. ред.: Т. В. Романова. Т. XXXVII: Интегративные процессы в когнитивной лингвистике: материалы международного конгресса по когнитивной лингвистике. Деком, 2019. С. 1069-1074.

9) Хоменко А. Ю. Атрибуция текстов малого объёма. Статистические закономерности // В кн.: Язык. Право. Общество: сб. ст. V Междунар. науч.-практ. конф. Пенза : ПГУ, 2018. С. 123-127.

10) Хоменко А. Ю. Лингвистическое моделирование как инструмент выявления искажений речевых навыков автора письменного речевого

произведения. Опыт практического исследования // Вопросы психолингвистики. 2018. № 2 (36). С. 209-226. doi

11) Хоменко А. Ю. Определение лингвистических признаков подготовленной и неподготовленной речи на материале установленного дословного содержания видеозаписи при производстве комплексной судебно-лингво-фоноскопической экспертизы // В кн.: Судебно-психологическая экспертиза и комплексные судебные исследования видеозаписей: сб. ст. Междунар. науч.-практ. конф. (г. Москва, 16 марта 2017 г.). РГУП, 2017. С. 192-

12) Хоменко А. Ю., Романова Т. В. Алгоритм автоматической атрибуции письменного текста в лингвокриминалистике // В кн.: Современные проблемы в области экономики, менеджмента, бизнес-информатики, юриспруденции и социально-гуманитарных наук: материалы XII-ой научно-практической конференции студентов и преподавателей НФ НИУ ВШЭ / Сост.: Е. А. Асланян. Н. Новгород : Национальный исследовательский университет Высшая школа экономики в Нижнем Новгороде, 2014. С. 220-223.

13) Хоменко А. Ю. Алгоритм для автоматической идентификации автора письменного речевого произведения в судебном автороведении // Юрислингвистика. 2014. № 3. С. 83-93.

14) Хоменко А. Ю. Анализ языковой личности автора текста с применением методов математической статистики как способ установления авторства текста. Проблемы языковой картины мира в синхронии и диахронии // В кн.: Проблемы языковой картины мира в синхронии и диахронии. Сборник статей по материал Всероссийской научной конференции молодых учёных Вып. 12. Н. Новгород : Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования "Нижегородский государственный педагогический университет имени Козьмы Минина", 2014. С.

15) Хоменко А. Ю. К вопросу об исследовании письменного речевого произведения в рамках автороведческой экспертизы на предмет его оригинальности // Политическая лингвистика. 2014. № 4. С. 306-312.

16) Хоменко А. Ю. Апробация методов математической статистики при атрибуции текста в рамках судебного автороведения // В кн.: Язык. Право. Общество: сб.ст.Всерос.науч.-практ.конф. Пенза : Издательство ПГУ, 2013. Гл. 52. С. 356-369.