ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ «НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ «ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Хоменко Анна Юрьевна

ЛИНГВИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КАК ИНСТРУМЕНТ АТРИБУЦИИ ТЕКСТА

Резюме
диссертации на соискание ученой степени
кандидата филологических наук

Научный руководитель:
доктор филологических наук
профессор
Романова Татьяна Владимировна

Москва 2021

Общая характеристика исследования

Настоящее исследование посвящено разработке проблем текстовой атрибуции на основе постулатов модельной лингвистики. В диссертационном исследовании речь идёт о решении идентификационной задачи закрытого класса при попарном сравнении письменных текстов.

Автор любого текста есть языковая личность. Феномен языковой личности изучается исследователями с точки зрения разных подходов: когнитивной лингвистики [Карасик 2003; Караулов 2010; Романова 2011], лингводидактики [Тарнопольский, Кожушко 2019], психолингвистики [Седов 1999; Белянин 2000], лингвокультурологии [Воркачев 2001; Воробьев 2011], социолингвистики и антропологии [Виноградов 1961; Shuy 2005; Вул 2007], судебной лингвистики [McMenamin 2002; Галяшина 2003; Coulthard 2004; Ионова, Огорелков 2020]. Все эти подходы принимают постулат о том, что материальной репрезентацией языковой личности является идиостиль пишущего. Феномены языковой личности и идиостиля очень сложны и многоаспектны, поэтому эти сущности приходится моделировать. Обычно при их моделировании используются модели речевой деятельности, а именно модели анализа [Апресян 1966: 99 – 107]. Сама же атрибуция есть не что иное, как модель лингвистического исследования, «имитирующая те исследовательские процедуры, которые ведут лингвиста к обнаружению того или иного языкового явления» [Апресян 1966: 99].

Атрибуционная лингвистика со времен Л. Кэмпбелла [Campbell 1867] и В. Лютославского [Lutoslawski 1897] на западе и Н.А. Морозова [Морозов 1915] в России всегда шла двумя параллельными путями: путем стилеметрии и путем квалификативного анализа текста. Квантитативные методы сейчас получают наибольшее распространение [Korobov 2015; Murauer, Tschuggnall, Specht 2018; Muttenthaler, Lucas, Amann 2019; Litvinova, Sboev, Panicheva 2018; Custódio, Paraboni 2018; Gomzin, Laguta, Stroev 2018; Panicheva, Mirzagitova, Ledovaya 2018; Bacciu, Morgia, La 2019], в то время как квалификативные в основном используются в судебном автороведении [Рубцова, Ермолаева,

Безрукова и др. 2007; Абрамкина 2019; Ким 2019; Сааков 2018] как дань традиции [Вул 1973, 1982, 2007; Галяшина, Ермолова 2005] и в связи с законодательством [Федеральный закон от 31 мая 2001 г. N 73-ФЗ; Приказ от 27 декабря 2012 года N 237]. Многие из квантитативных подходов продуктивны, тем не менее они рассматривают индивидуальный стиль как череду языковых вероятностей, а не как продукт формирования речевой способности индивида. С помощью только квантитативных подходов, основанных на сборе данных о неких традиционных стилеметрических параметрах, пусть даже и в большом их количестве [Bhargava, Mehndiratta, Asawa 2013], невозможно создать полную, адекватно отражающую оригинал модель идиостиля автора текста, являющуюся экпликацией языковой личности. Полная, адекватно отражающая оригинал модель языковой личности может быть создана только с помощью обширного квалификативного анализа индивидуального стиля автора, основным недостатком которого, в свою очередь, является значительная доля субъективизма и зависимость от компетенции исследователя.

Разработанная в диссертационном исследовании модель реализуется по следующему алгоритму: 1) автоматическое извлечение из текста параметров, описывающих идиостиль с точки зрения прагматикона, тезауруса и лексикона автора; 2) поиск традиционных стиметрических данных; 3) присвоение веса каждому параметру; 4) построение математических моделей сравниваемых текстов; 5) сравнение математических моделей; 6) экспертный анализ статистических данных.

Актуальность исследования определяется тем, что в современном обществе особое значение имеет аутентификация письменного материала. Методики атрибуции становятся необходимыми в филологических экспертизах при определении авторства известных художественных произведений (например, текстов Ф.М.Достоевского, М.А.Шолохова), в судебных автороведческих экспертизах при решении диагностических и идентификационных задач, при решении научных задач, связанных с

идентификацией лиц по письменной речи и определением характеристик пола, возраста, социального статуса этих лиц. Каждая из перечисленных научных и практических сфер деятельности нуждается в использовании полных и всесторонних моделей атрибуции, дающих объективные результаты.

Объектом исследования являются собственно модели лингвистического исследования, а именно: модели текстовой атрибуции; методы и методики текстовой атрибуции, имеющие квантитативную, квалификативную и интегративную основы. **Предметом** исследования становится параметры текстовой атрибуции, позволяющие создать полную и адекватную модель идиостиля автора как экспликатора языковой личности пишущего. При исследовании идиостиля используются формализованные модели анализа, «конечное число правил, способных проанализировать бесконечное число предложений» [Апресян 1966: 107], в основе которых лежат принципы семантического синтаксиса [Падучева 1974], грамматики конструкций [Лингвистика конструкций 2010], построения структурных схем предложений русского языка [Русская грамматика, URL: http://rusgram.narod.ru/index.html]. В основе формализации уровневой структуры языковой личности в исследовании лежат постулаты теории Ю.Н. Караулова [Караулов 2010].

Создание интегративной модели текстовой атрибуции, основанной на сочетании квантитативного и квалификативного подходов к анализу текста, и разработка прототипа программного обеспечения на ее основе стали **целью** настоящего диссертационного исследования.

Гипотеза исследования заключается в том, что с помощью фиксированного свода формализованных правил можно создать интегративную атрибуционную модель, являющуюся достаточно полной, всесторонне имитирующей оригинал и одновременно объективной. Эта модель дает исследователю возможность успешно решать идентификационную задачу атрибуционной лингвистики на текстах разного объема и жанровой отнесенности.

Для достижения означенной цели были поставлены следующие **задачи**:

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

– сформулировать рабочее понятие модели, разработать ее структуру; применить их при создании автоматизированного программного комплекса по атрибуции текста;

– разработать проблему наиболее используемых квантитативных и квалификативных методов и методик анализа идиостиля как репрезентанта языковой личности автора; адаптировать их для применения в прототипе автоматизированного программного комплекса по атрибуции текста;

– создать свод формализованных правил для вычленения лингвистических характеристик, позволяющих идентифицировать автора письменного текста;

– собрать разножанровую коллекцию текстов в качестве материала для верификации результатов работы интегративной атрибуционной модели;

– создать эффективную модель для решения идентификационной задача текстовой атрибуции;

– апробировать модель на созданной коллекции разножанровых текстов; определить валидность выводов, получаемых при анализе текстов посредством разработанной интегративной модели;

– при необходимости разработать стратегию по повышению уровня эффективности модели.

Теоретическая значимость исследования состоит в разработке структуры лингвистической модели, наиболее универсальной для атрибуции текстов любого объема и жанровой отнесенности и подходящей для формализации с помощью современных компьютерных инструментов. Теоретический вклад автора в разрабатываемую проблему также связан с созданием моделей анализа (формализованных правил для поиска) лингвистических конструкций и иных идентификационных компонентов языковой личности в автоматическом режиме. Также в работе была создана

система оценки любых лингвистических моделей, в том числе атрибуционных.

Практическая значимость работы обусловлена тем, что функционал разработанных моделей анализа и модели исследования, а также созданного на их основе электронного ресурса много шире изначально заложенных возможностей. Наработки диссертационного исследования можно использовать не только для решения идентификационной задачи атрибуционной лингвистики, но и для исследования языковой личности писателей, журналистов, политиков и пр., при проведении диагностики языковой личности конкретного человека для решения задач психолингвистики, психологии, для обследования обобщенной языковой личности той или иной социальной группы, субкультуры и др. в целях решения задач социолингвистики, социологии.

Результаты исследования были использованы в преподавании таких научных дисциплин, как функциональные и когнитивные модели в лингвистике, лингвистическая экспертиза, компьютерные инструменты лингвистических исследований образовательной программы бакалавриата «Фундаментальная и прикладная лингвистика», магистерских образовательных программ «Политическая лингвистика» и «Applied Linguistics and Text Analytics» НИУ ВШЭ (кампус в Нижнем Новгороде).

Материалом для тестирования и апробации созданной атрибуционной модели исследования послужили разножанровые произведения, варьирующиеся по объему:

1) коллекция текстов художественной литературы, включающая тексты С.Д.Довлатова и В.П.Астафьева: 10 текстов, средний объем – 20,000 слов;

2) коллекция текстов современной сетевой беллетристики (авторские тексты, размещенные на ресурсе «Книга фанфиков»), включающая тексты 3 авторов-женщин, 4 авторов-мужчин: 187 текстов; средний объем – от 1,500 до 40,000 слов:

3) коллекция текстов сетевой публицистики (тексты электронной газеты «The Village»), включающая тексты 3 авторов-женщин, 3 авторов-мужчин: 600

текстов, средний объем – от 500 до 1,500 слов:

4) коллекция текстов электронных комментариев (тексты, размещенные на развлекательном портале «ЯПлакалЪ»), включающая тексты 3 авторов-женщин, 3 авторов-мужчин: 424 текста, средний объем – от 50 до 100 слов:

5) коллекция текстов корпоративной русскоязычной переписки, включающая тексты 2 авторов-женщин, 2 авторов-мужчин: всего 236 текстов (от 45 до 49 писем для одного автора); средний объем – от 50 до 500 слов.

Апробация исследования проходила на различных международных и всероссийских научных и научно-практических конференциях, как то:

2021 г.: 1) Международный научно-практический семинар «Лингвистическая диагностика: методы исследования личности», Москва, Государственный институт русского языка имени А. С. Пушкина; 2) 27-я Международная конференция по компьютерной лингвистике и интеллектуальным технологиям «Диалог-2021», Москва.

2020 г.: 1) X Международный конгресс по когнитивной лингвистике «Когнитивно-дискурсивная парадигма в лингвистике и смежных науках: современные проблемы и методология исследования», Екатеринбург, Уральский государственный педагогический университет; 2) Международная научная конференция «Русский язык в современном научном и образовательном пространстве», Москва, Российский университет дружбы народов; 3) VI Международная научно-практическая конференция «Язык. Право. Общество», Пенза, Пензенский государственный университет; 4) XLIX Международная научная филологическая конференция, посвященная памяти Людмилы Алексеевны Вербицкой (1936-2019), Санкт-Петербург, Санкт-Петербургский государственный университет.

2019 г.: 1) IX Международный конгресс по когнитивной лингвистике «Интегративные процессы в когнитивной лингвистике», Нижний Новгород, Высшая школа экономики; 2) Современная теоретическая лингвистика и проблемы судебной экспертизы, Москва, Государственный институт русского языка имени А. С. Пушкина; 3) Artificial Intelligence and Natural Language

(AINL), Тарту, University of Tartu; 4) Лингвополитическая персонология: дискурсивный поворот, Екатеринбург, Уральский государственный педагогический университет, Российская академия народного хозяйства и государственной службы при Президенте РФ, Уральский институт управления; 5) Языковая личность и эффективная коммуникация в современном поликультурном мире, Минск, Белорусский государственный университет.

2018 г.: 1) V Международная научно-практическая конференция «Язык. Право. Общество», Пенза, Пензенский государственный университет; 2) Массмедийная политическая коммуникация: методы и приемы лингвистического анализа и лингвистической экспертизы, Екатеринбург, Уральский государственный педагогический университет;

2014 г.: Проблемы языковой картины мира в синхронии и диахронии, Нижний Новгород, Мининский университет);

2013 г.: 1) Язык. Право. Общество, Пенза, Пензенский государственный университет; 2) Artificial Intelligence and Natural Language (AINL), Санкт-Петербург.

На защиту выносятся следующие положения:

1. Моделирование языковой личности автора и моделирование языковых структур (на основе принципов семантического синтаксиса, грамматики конструкций, модельных схем предложений), используемых в тексте, является эффективным инструментом атрибуции текста;

2. Сочетание квалификативных и квантитативных методов атрибуционного анализа позволяет создать достаточно полную для решения идентификационной задачи, всесторонне имитирующую оригинал и одновременно объективную атрибуционную модель.

3. Для теоретической оценки качества любой модели, в том числе и модели атрибуции или модели идиостиля как репрезентации языковой личности, актуальны следующие критерии:

- полнота модели;

- простота модели;
- точность модели;
- экономичность модели;
- адекватность модели;
- единство в своей раздельности;
- цельность модели;
- структурность модели;
- экспланаторность;
- эвристичность модели (как частный случай экпланаторности);
- коммуникативность модели (с точки зрения языка);
- дедуктивность модели;
- интерпретируемость модели;
- математичность, точность, однозначность модели;
- уровень формализации модели;
- уровень технически точного отражения объекта моделирования;
- уровень реально-жизненного отражения объекта моделирования;
- уровень субъективизма в модели;
- уровень существенности модельных признаков (уровень абстракции (идеализации) модели);
- уровень действенности;
- уровень функциональной и практической направленности модели;
- «гипотезная мощность»;
- эстетические свойства модели (опционально).

4. Указанные критерии позволяют провести исследование валидности модели как с точки зрения общих обязательных свойств, так и по принципу типа модели.

5. Оценка моделей с помощью предложенной классификации релевантна не только на этапе выбора модели для решения практической или теоретической задачи, но и для обследования уже созданной и апробированной (экспериментально проверенной) модели, с целью объяснения

продуктивности и непродуктивности компонентов этой модели.

6. Разработанная полуавтоматическая атрибуционная модель исследования языковой личности может быть успешно использована для решения идентификационной задачи автороведения для текстов разных жанров и объемов.

Структура работы. Работа состоит из Введения, четырёх глав, Заключения, Библиографического списка, насчитывающего 259 наименования, в том числе 74 англоязычное. Результаты исследования представлены в виде 20 таблиц, 17 рисунков и 14 приложений.

Основное содержание работы

Во **Введении** сформулированы цель и задачи исследования, его предмет и объект; обоснованы актуальность, новизна, теоретическая и практическая значимость, сформулированы положения, выносимые на защиту, приведены сведения об апробации научной работы.

В **Главе 1** разработана проблема лингвистического моделирования, на основе теоретических выкладок [Bloomfield 1926], [Ельмслев 2005], [Лосев 2004], [Апресян 1966], [Штофф 1966], [Ревзин 1977], [Баранов 2001], [Медведева 2010], [Павловская 2010], [Белоусов 2010], [Брюшинкин 2009] и др. описаны критерии для определения типа лингвистической модели (Таблица 1), а также создана классификация модельных свойств для оценки эффективности модели с теоретической точки зрения (Таблица 2).

В Таблице 1 представлен кумулятивный список разного рода моделей. Классификация моделей в нем имеет разные типологические основания, которые дают возможность исследования валидности модели с точки зрения типа модели и его релевантности для решаемых с её помощью задач.

Таблица 1. Критерии для определения типа модели

Критерии типизации моделей		
№	**Критерий (наименование)**	**Описание (объяснение)**
1.	**Модели речевой деятельности**	имитируют языковые процессы и явления

1.1.	***Несемантические***	имитируют грамматическую способность
1.2.	***Семантические***	имитируют способность понимать и строить осмысленные предложения
2.	**Аналитические**	содержат конечное число правил, способных проанализировать бесконечное число предложений
3.	**Синтезирующие**	содержат конечное число правил, способных построить бесконечное число предложений
3.1.	***Собственно синтезирующие***	от смысла к форме
3.2.	***Порождающие***	от базовой формы к разнообразию небазовых форм: включают аппарат для построения бесконечного числа правильных предложений данного языка и приписывания каждому из них некоторой структурной характеристики
4.	**Исследовательские модели**	имитируют деятельность исследователя по обнаружению языковых явлений
4.1.	***Модели дешифровки***	входной информацией является лишь текст в его естественной среде
4.2.	***Экспериментальные модели***	входная информация включает текст и некоторые правила, способные модифицировать текстовую информацию. Текст функционирует в искусственной среде и подвергается экспертом-исследователем разного рода экспериментам (преобразованиям)
4.2.1	*Собственно экспериментальные*	использующие на «входе» текст + множество правильных фраз данного языка
4.2.2.	*Дескриптивные, описательные*	использующие на «входе» текст + множество правильных фраз данного языка + множество семантических инвариантов
5.	**Метамодели**	имитируют теоретическую и экспериментальную оценку готовых моделей речевой деятельности или лингвистического исследования (описывают лингвистические теории)
6.	**Модели-исчисления**	система разрешений
7.	**Модели-алгоритмы**	система команд
8.	**Вероятностные**	используют принципы теории вероятности и математической статистики для принятия решений по модели (предсказания поведения объектов)
9.	**Структурные**	Детерминистские

10.	**Смешанные**	статистика + детерминизм (наиболее распространённые для решения практических задач)
11.	**Компонентные**	состоят из набора характеристик объекта, записанных в той или иной сигнатуре
12.	**Гомогенные**	состоят из описания/анализа одной сущностной стороны явления
13.	**Диахронические**	анализируют язык в его развитии
14.	**Синхронические**	анализируют язык определённого периода
15.	**Теоретические**	созданы для решения собственно теоретических проблем лингвистики (чаще всего это метамодели)
16.	**Прикладные**	созданы для решения практических задач заказчика
17.	**Графические**	представляют собой визуальный образ объекта моделирования; графическое представление использовано для конструирования структуры объекта
18.	**Символьные**	представляют собой набор символов, описывающих оригинальный объект
19.	**Компьютерные**	представляют собой набор символов определённой формализованной сигнатуры, алфавит, организованный в форме исчисления или алгоритма и реализованный на компьютере

Для теоретической оценки качества любой модели, в том числе и модели атрибуции или модели идиостиля как репрезентации языковой личности, актуальны критерии оценки моделей, представленные в Таблице 2.

Таблица 2. Критерии для оценки лингвистической модели

Критерии для оценки свойств моделей			
№	**Критерий (наименование)**	**Описание (объяснение)**	**Шкала оценки**
1.	**Полнота модели**	способность отражать всю необходимую информацию	низкий / средний / высокий уровень
2.	**Простота модели**	использования относительно небольшого количества средств (сигнатуры, правил) для достижения	низкий / средний / высокий уровень

		поставленной научной цели	
3.	**Точность модели**	возможность выполнения операций представляемым моделью формальным аппаратом	наличие / отсутствие
4.	**Экономичность модели**	экономичное использование энергетических и временных ресурсов при применении модели	низкий / средний / высокий уровень
5.	**Адекватность модели**	свойство максимальной схожести с объектом-оригиналом	низкий / средний / высокий уровень
6.	**Единство в своей раздельности**	модель всегда предполагает разбиение на части внутри целого (модельный кортеж всегда состоит из подмножеств)	возможность / отсутствие возможности разбиение на подмножества
7.	**Цельность модели**	модельный кортеж представляет собой неделимое множество	наличие / отсутствие целостности (обусловливается наличием / отсутствием связи между подмножествами элементов кортежа, создающей неделимую в своём единстве структуру)
8.	**Структурность модели**	перенесения структуры субстрата моделируемого объекта на другой субстрат	а) наличие / отсутствие; б) удачный / неудачный выбор «принимающего» субстрата;

			в) удачная / неудачная организация структуры модели
9.	**Экспланаторность**	«объяснительная мощность» модели; способность модели давать информацию о причинах наблюдаемых фактов и предсказывать новые	наличие / отсутствие
10.	**Эвристичность модели (как частный случай экпланаторности)**	способность модели к поиску новых знаний об объекте	низкий / средний / высокий уровень
11.	**Коммуникативность модели (с точки зрения языка)**	в основе любой лингвистической модели лежит не набор отвлечённых статистических закономерностей, формул, функций и цифр, а язык как инструмент общения	наличие / отсутствие
12.	**Дедуктивность модели**	наличие/отсутствие эмпирического изучения языковых фактов как основы моделирования; моделирование «снизу»: использования средств и методов классического языкознания для обследования языковых фактов	а) наличие / отсутствие; б) низкий / средний / высокий уровень оперирования собственно языковыми, лингвистическими методами анализа как основой для моделирования
13.	**Интерпретируемость модели**	интерпретация модели – возможность подстановки объектов некоторой предметной области вместо объектов (символов) модели	а) наличие / отсутствие; б) простота / сложность подстановки
14.	**Математичность,**	коррелирует с уровнем	а) полный,

	точность, однозначность модели	формализации математической модели	целостный / неполный аппарат формализации модели; б) удачная / неудачная работа этого аппарата как основа для машинной реализации модели
15.	**Уровень формализации модели**	представляет собой структуру, описанную условной сигнатурой языка, либо структуру, описанную с помощью математического, числового, формульного аппарата	дескриптивный уровень / математический уровень
16.	**Уровень технически-точного отражения объекта моделирования**	удачный / неудачный способ формализации модели, выбор сигнатуры; удачная / неудачная машинная реализация (при наличии)	низкий / средний / высокий уровень
17.	**Уровень реально-жизненного отражения объекта моделирования**	насколько объёмно и полно структура модели передаёт объект-оригинал	низкий / средний / высокий уровень
18.	**Уровень субъективизма в модели**	наличие / отсутствие личностных оценок и суждений исследователя в структуре модели	низкий / средний / высокий уровень
19.	**Уровень существенности модельных признаков (уровень абстракции (идеализации) модели)**	удачное / неудачное нивелирование, элиминация в модели языковых фактов, не имеющих значения для конкретной задачи	низкий / средний / высокий уровень
20.	**Уровень действенности**	оценка работоспособности модели в условиях	низкий / средний / высокий уровень

		решения поставленной задачи	
21.	**Уровень функциональной и практической направленности модели**	соответствие модели её целевому использованию. Цель создания модели может быть: собственно лингвистическая, практическая, математическая и пр.	а) соответствует / не соответствует целевому использованию; б) низкий / средний / высокий уровень соответствия
22.	**«Гипотезная мощность»**	имеется ли в основе модели какая-либо гипотеза	наличие / отсутствие
23.	**Эстетические свойства модели (опционально)**	гармоничность организации структуры модели	наличие / отсутствие

В **Главе 2** проводится анализ отечественных и зарубежных методов и методик, алгоритмов и программных комплексов, предназначенных для решения атрибуционных задач, а также анализ атрибуционных параметров различных уровней, используемых в разных подходах. Оценивается работоспособность перечисленных компонентов, описываются положительные и отрицательные стороны их применения.

В **Главе 3** вычленены наиболее эффективные компоненты и атрибуционные параметры рассмотренных в Главе 2 методик (М. Коппела, Дж. Шлера, А.Н.Баранова, Дж. Ангори и пр.) и программных комплексов (RusProfilingLab, КАТ, «Авторовед» и пр.), описаны результаты анализа разнородного текстового материала 10 блоков общим объемом около 116 тыс. слов (тексты официально-делового стиля, корпоративная переписка, короткие текстовые сообщения, публицистика и пр.) с целью определения ряда параметров языковой личности, которые заведомо являются важными компонентами авторского идиостиля и одновременно могут быть извлечены из текста автоматически с минимальным препроцессингом. Эти параметры легли в основу созданной атрибуционной модели.

Идентификационные параметры для модели выделяются на всех трех

уровнях языковой личности при её понимании в концепции Ю.Н. Караулова [Караулов 2010: 53]. Языковая личность понимается как результат ее формирования в определенной социальной среде: автобиографический, социолингвистический и юрислингвистический подходы [Виноградов 1961; Coulthard 2004; Shuy 2005; Вул 2007].

В **Главе 4** описывается созданный на основе сочетания идентификационных параметров и лингвистического моделирования прототип программного продукта как части полуавтоматической интегративной модели текстовой атрибуции. Приводятся результаты апробации алгоритма на разножанровом материале. Для извлечения компьютерным способом все формальные правила были запрограммированы и интегрированы в лингвистический ресурс «ХоРом»: http://khorom-attribution.ru/#/[1]. Подробно некоторые правила поиска лингвистических структур описаны в статье [Khomenko, Baranova, Romanov, Zadvornov 2021].

В результате эмпирического исследования на вербально-семантическом уровне были запрограммированы для поиска такие стилостатистические параметры, как частеречная отнесенность слов (количество знаменательных частей речи, соотношение разных частей речи – индекс удобочитаемости, коэффициент предметности и пр. [Головин 1970], индекс туманности Ганнинга, индекс Флеша-Кинкейда с коэффициентом для русского языка [Solnyshkins, Guryanov, Gafiyatova, Varlamova 2018: 679]), средние длины слов; а также такие параметры идиолекта [Литвинова 2019], как наличие/отсутствие сложных слов полуслитного написания, модальные частицы, междометия, наличие/отсутствие модального постфикса «-то», предпочтительные слова-интенсификаторы. Всего было использовано 10 стандартных алгоритмов и 32 уникальных, созданных непосредственно для целей исследования формализованных правила для извлечения лингвистических структур. Формализованный поиск единиц этого уровня осуществляется в соответствии

[1] Ресурс «ХоРом» разработан командой исследователей под руководством Хоменко А.Ю. Главный инженер проекта: Баранова Ю.Н.

с морфологической природой этих единиц, материально эксплицированной посредством присвоения им морфологических тегов.

Вербально-семантический уровень (уровень идиолекта в соответствии с концепцией [Кристалл, Дейви 1980], [Федотова 2013], [Буров 2017]) наиболее легко формализуем, поскольку сам по себе имеет «более формальные» языковые характеристики, которые априори считаются стабильными [Литвинова 2019].

Для репрезентации фрагмента тезаурусного уровня выбраны такие параметры, как ключевые лексемы, наиболее частотные словные триграммы и биграммы, экспликаторы аксиологических текстовых доминант дихотомии «свой/чужой».

Тезаурусный уровень наиболее труден для формализации. Можно автоматически создать материальную экспликацию авторского тезауруса [Бессмертный, Нугуманова 2012], тем не менее определить, как лексемы в тезаурусе «выстраиваются в упорядоченную, достаточно строгую иерархическую систему, в какой-то степени (непрямой) отражающую структуру мира» [Караулов 2010: 52], крайне сложно. Этот уровень репрезентирован наименьшим количеством параметров (всего 3 стандартных алгоритма и одно аутентичное правило для извлечения лингвистической информации) именно в силу стремления не просто формализовать некоторые компоненты языковой личности с целью ее компьютерной репрезентации, но и сделать конечную модель интерпретируемой.

Прагматикон языковой личности формализован посредством следующего набора параметров: вводные слова и конструкции, эксплицирующие субъективную модальность; целевые, выделительные обороты, конструкции с сопоставительными союзами, репрезентирующие уровень освоения автором компетенций письменной речи и его коммуникативные стратегии и тактики; синтаксические сращения, дающие представление в том числе об авторских предпочтениях в функционально-стилистической отнесенности текста; предложения с обособленными

приложениями; сложные синтаксических конструкции; сравнительные придаточные, глагольные односоставные предложения, эксплицирующие функциональный тип повествования; наличие/отсутствие и виды обращений как контактоустанавливающего элемента. Всего выделяется 11 конструкций и создано 107 аутентичных правил для извлечения информации из текста.

Для анализа синтаксических структур были прописаны правила, основанные на POS-tags, а также на том, какие синтаксические отношения имеют место в предложении [Падучева 1974] и какую грамматическую конструкцию реализуют те или иные его компоненты [Лингвистика конструкций 2010].

Уровень прагматикона автора не слишком сложен для формализации, но результаты работы алгоритма на этом уровне требуют тщательной проверки и интерпретации исследователем.

После извлечения всех параметров, связанных со словесными структурами, реализуется подсчет ipm (instance per million). Для синтаксических параметров количество каждого параметра делится на количество предложений в тексте. Это значение присваивается каждому параметру как его относительная частота встречаемости.

В качестве результата работы алгоритма выводится значения коэффициента корреляции Пирсона, значение линейной регрессии (оценивать следует коэффициент детерминации), критерия Стьюдента для моделей двух сравниваемых текстов, а также значения метрик каждого параметра для двух текстов, метрик, доказывающих или опровергающего гипотезу H_0 о том, что автором двух сравниваемых текстов, вероятно[2], является одно лицо.

Важно, что данный блок не является конечным шагом в разработанной модели. Текстовую статистику необходимо интерпретировать, в разработанной модели именно эксперт, основываясь на статистических данных и на их интерпретации, принимает конечное решение об атрибуции текста, используя

[2] Вероятностный характер вывода связан с тем, что в каждом конкретном случае в соответствии с разработанной моделью решение о конечном авторстве принимает исследователь.

рейтерские таблицы, подготовленные по результатам исследования (Таблица 3), и свой экспертный опыт.

Таблица 3. Пример рейтерской таблицы для оценки результатов работы атрибуционной модели

Тип дискурса	**коэффициент корреляции Пирсона**	**коэффициент детерминации линейной регрессии**	**t-критерий Стьюдента (p-value)**	**Автором сравниваемых текстов, вероятно[3], является одно лицо**	**Авторами сравниваемых текстов, вероятно, не является одно лицо**	**комментарий**
Художественная проза	не ниже 0,97; обычно: 1,00	не ниже 0,94; обычно: 1,00	не ниже: 0,91; редко выше: 0,93 Итог: около 0,90	+	-	значения всех метрик примерно от 0,90 до 1,00
Художественная проза	1,00	1,00	не ниже: 0,84	+	-	при анализе текстов художественной прозы встречаются ситуации, когда значение наиболее важной для объемных произведений метрики p-value t-критерия Стьюдента достаточно низко. В таких случаях, чтобы была возможность признать автором сравниваемых текстов одно лицо, значения других метрик должно достигать единицы
Художественная проза	может достигать	может достигать	около 0,50	-	+	если для текстов

[3] Вероятностный характер вывода связан с тем, что в каждом конкретном случае в соответствии с разработанной методикой решение о конечном авторстве принимает исследователь.

	0,97	0,94; обычно около 0,84				художественной прозы значение p-value t-критерия Стьюдента низко (около 0,50), то высокие значения иных метрик при их наличии можно не брать в расчет

Оценка результатов работы методики проводилась с двух точек зрения:

1) полученные модели языковой личности рассматривались с точки зрения шкалы теоретической оценки моделей, представленной в Таблице 2. Посредством этих оценок удалось доказать гипотезу о том, что с помощью фиксированного свода формализованных правил можно создать интегративную атрибуционную модель, являющуюся достаточно полной, всесторонне имитирующей оригинал и одновременно объективной для решения идентификационной проблемы закрытого класса при попарном сравнении текстов;

2) полученные модели также оценивались с точки зрения решения с их помощью идентификационной задачи атрибуционной лингвистики. Тестирование и апробация созданного алгоритма проходили на текстовых коллекциях, означенных в **материале исследования.** Удалось получить следующие **результаты**:

1) коллекция текстов художественной литературы: доля правильных ответов алгоритма (accuracy), точность (precision) и полнота (recall) равны 100%, F-мера – 1[4];

2) коллекция текстов сетевой беллетристики: доля правильных ответов алгоритма – 83%, точность – 67%, полнота – 100%, F-мера – 0,8;

3) коллекция текстов сетевой публицистики: доля правильных ответов алгоритма, точность и полнота равны 100%, F-мера – 1;

[4] Здесь и далее значения метрик указаны в связи с интерпретацией статистических данных с помощью методических рекомендаций и рейтерских таблиц, разработанных для целей анализа.

4) коллекция текстов развлекательной публицистики: доля правильных ответов алгоритма – 40%, точность – 0, полнота– 0, F-мера – 0;

5) коллекция текстов корпоративной русскоязычной переписки: доля правильных ответов алгоритма – 83%, точность – 67%, полнота – 100%, F-мера – 0,8.

В результате исследования удалось также сделать следующие **выводы**:

1) для дискурса художественной прозы (как прозы признанных авторов, так и беллетристики) наиболее информативной метрикой является t-статистика Стьюдента;

2) для жанра современной беллетристики неинформативным является стилостатистический пул, поскольку, по экспериментальным данным, значения стилостатистических параметров близки для всех обследованных текстов;

3) для признания гипотезы H_0 верной при анализе публицистических текстов значения коэффициентов корреляции и детерминации должны достигать единицы. Необходимость такого высокого уровня значений связана с объемом текстового материала и его спецификой. Важно, что для публицистического дискурса следует признать значительно менее релевантным t-статистику, которая для художественного дискурса является наиболее информативным показателем. Что касается гендерной дифференциации материала, стоит заметить, что «женские» публицистические тексты более коррелируют с «женскими», равно как и «мужские» с «мужскими»;

4) для коротких текстовых сообщений: корпоративная переписка, комментарии в сети интернет, – необходимо создание репрезентативной выборки из совокупности текстов объемом не менее 500 слов. Ограничение в 100 слов, выведенное еще С.М. Вулом и имеющее место до сих пор в судебном автороведении [Рубцова, Ермолаева, Безрукова и др. 2007] как объем, необходимый для определения авторства текстов при встраивании в анализ статистической информации, должно быть увеличен. Для улучшения работы

алгоритма на данном материале в настоящий момент разрабатываются дополнительные параметры для построения моделей идиостиля. Они связаны с так называемым дигитальным почерком. Это графический литуратив, графическая гибридизация, использование элементов текста, написанных заглавными буквами, эмотиконы и прочие графические символы, выражающие эмоциональность речи;

5) разножанровые произведения тоже можно валидно обследовать с помощью разработанной интегративной модели (можно, например, сравнить текст электронного сообщения с публицистической статьей): доля правильных ответов алгоритма – 83%, точность – 67%, полнота – 100%, F-мера – 0,8.

Так, следует говорить о том, что созданная модель дает исследователю возможность успешно решать идентификационную задачу атрибуционной лингвистики на текстах разного объема и жанровой отнесенности.

При использовании алгоритма самым ценным становятся не выводные данные, а собственно модели идиостилей как репрезентации языковых личностей пишущих, созданные с его помощью. Эти модели являются понятными и простыми, легко интерпретируемыми экспертным путем, с одной стороны, и полными и адекватно имитирующими объект-оригинал – с другой. Пример такой модели представлен в Таблице 4.

Таблица 4.

Модели: В. Астафьев, «Последний поклон» – В. Астафьев, «Звездопад»

коэффициент корреляции Пирсона = 100%;

коэффициент детерминации = 1;

p-value для t-критерий Стьюдента – 0,91

Параметры с относительной частотой появления		
Номер и название параметра	**Текст 1**	**Текст 2**
1. Индекс удобочитаемости Флеша-Кинкейда	15.5663	13.1349
2. Индекс туманности Ганнинга	18.9063	16.3078
3. Средняя длина слова (в буквах)	5.2664	4.9911
4. Средняя длина предложения (в словах)	13.3458	9.4829
5. Количество предложений длиннее 8-ми слов	533864.7494	404742.0965

6. Коэффициент предметности (Pr)	1.282	1.1534
7. Коэффициент качественности (Qu)	0.3839	0.48
8. Коэффициент активности (Ac)	0.198	0.2041
9. Коэффициент динамизма (Din)	0.4192	0.4701
10. Коэффициент связности текста (Con)	2.7853	2.0499
11. Количество слов несловарного написания	1405.3771	1710.7514
12. Предложения с однородными рядами	5578.5664	2324.8673
13. Предложения с обособленными приложениями	382.5749	833.443
14. Вводные слова и конструкции	3923.3445	6272.7552
15. Целевые и выделительные обороты	210.8066	87.7308
16. Конструкции с семантикой сравнения	4934.4353	5088.3888
17. Синтаксические сращения	113.2109	175.4617
18. Сравнительные придаточные	9365.277	9299.4692
19. Конструкции с сопоставительными союзами	148.3454	438.6542
20. Вставные конструкции	66.365	2544.1944
21. Сложные синтаксические конструкции	34849.449	42198.5349
22. Глагольные односоставные предложения	7046.4048	10747.0281
23. Обращения	1077.4558	1886.2131
24. Местоимения "я, мы"-группы	32444.6926	53691.2752
25. Местоимения "ты, вы"-группы	28431.5601	38031.3199
26. Сложные слова полуслитного написания	1495.1651	1096.6355
27. Модальные частицы	13561.8893	18598.9385
28. Междометия	620.7082	1272.0972
29. Наличие/отсутствие модального постфикса «-то»	2908.3499	2588.0598

Разработанная атрибуционная модель строится на основании перечисленных выше параметров, эти параметры выбраны в связи с относительной простой их автоматического извлечения и эмпирически доказанной значимостью для языковой личности пишущего [Хоменко 2014а; Степаненко 2017; Хоменко 2018а, б; Радбиль, Маркина 2019; Хоменко 2019а,в,г; Хоменко 2020а,б,в; Khomenko, Baranova, Romanov, Zadvornov 2021] и др. Представленный набор параметров не является конечным, это открытый список, которые может быть дополнен любыми релевантными характеристиками или метриками. Тем не менее важно, что каждый из параметров должен быть вписан в общую структуру языковой личности, он должен быть интерпретируемым, объяснимым с точки зрения лингвистического знания и/или текстовой статистики. Параметры не должны

использоваться изолированно, они должны исследоваться в совокупности в рамках предложенной структуры языковой личности. Для того чтобы удовлетворять условиям полноты, всесторонности и объективности, атрибуционная модель должна включать параметры всех уровней языковой личности и компоненты всех уровней языка.

Результаты работы атрибуционной модели можно сравнить с результатами работы других моделей, основанных на машинном обучении и нейронных сетях. Результаты такого сравнения приведены, например, в [Khomenko, Baranova, Romanov, Zadvornov 2021]. Тем не менее следует отметить, что сравнение разрабатываемой в диссертации модели с алгоритмами, основанными на машинном обучении и нейронных сетях, можно считать практически нерелевантным, поскольку представленная модель атрибуции имеет базис, отличающий ее от полностью автоматических система: эта модель всегда требует интерпретации исследователем.

В **Заключении** подводятся итоги исследования, сформулированы рекомендации по улучшению структуры моделей, лежащих в основе атрибуционного анализа. В **Приложениях** содержится эмпирический материал.

Основное содержание работы отражено в следующих публикациях

изданиях, индексируемых в Scopus:

1) 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;

в изданиях, включенных в список журналов высокого уровня, подготовленный в НИУ ВШЭ:

2) Хоменко А. Ю., Корсаков К. Судебный PR в России: понятие и его суть, явление и его социально-государственная значимость // Вестник Воронежского государственного университета. Серия: Право. 2019. № 4 . С. 284-300.

3) 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;

в прочих изданиях:

4) Khomenko A., Baranova Yu., Romanov A., Zadvornov K. Linguistic Modeling as a Basis for Creating Authorship Attribution Software [Лингвистическое моделирование как основа для создания электронного атрибуционного ресурса]. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2021” Moscow, June 16–19, 2021, URL: http://www.dialog-21.ru/media/5315/khomenkoaplusetal048.pdf.

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

5) Романова Т. В., Хоменко А. Ю. Лингвистическое моделирование как инструмент атрибуции текста // В кн.: Русский язык в современном научном и образовательном пространстве. Издательство РУДН, 2020. С. 250-252.

6) Хоменко А. Ю. Возможность идентификации лица по разнородному материалу - устной и письменной речи – в условиях одной экспертной задачи // В кн.: Лингвополитическая персонология: дискурсивный поворот : материалы Междунар. науч. конф. (Екатеринбург, 29—30 нояб. 2019 г.). [б.и.], 2019. С. 214-218.

7) Хоменко А. Ю., Зинина А. М. Глобальный английский и интернет-сленг. Влияние на языковую личность современной молодежи (на материале русского и китайского языков) // В кн.: 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.

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

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

10) 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.

11) 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.

12) 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. 192-198.

13) 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.

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

15) 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. 293-297.

16) 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.

17) 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.

Диссертационное исследование поддержано Российским фондом фундаментальных исследований: грант РФФИ в рамках научного проекта № 19-31-27001 «Аспиранты» (№ 19-312-90022).

Список литературы

Абрамкина Е. Е. Протокол допроса как объект автороведческой идентификационной экспертизы: задачи, проблемы, методика анализа //

Сборник материалов Международной научная конференция «Современная теоретическая лингвистика и проблемы судебной экспертизы», г. Москва, 1-2 октября 2019 г. – М.: Государственный институт русского языка им. А.С. Пушкина, 2019. С. 492-506.

Апресян Ю.Д. Идеи и методы современной структурной лингвистики. – М., 1966. 305 с.

Баранов А.Н. Введение в прикладную лингвистику: Учебное пособие. – М.: Эдиториал УРСС, 2001. — 360 с.

Белоусов К.И. Модельная лингвистика и проблемы моделирования языковой реальности. Модельная лингвистика и проблемы моделирования языковой реальности // Вестник Оренбургского государственного университета, 2010, № 11 (117). С.94-97.

Белянин В.П. Основы психолингвистической диагностики: модели мира в литературе / В.П. Белянин; Российская акад. наук, Ин-т языкознания, Фонд Чтения им. Н. А. Рубакина. – М.: Тривола, 2000. - 247 с.

Бессмертный И.А., Нугуманова А.Б. Метод автоматического построения тезаурусов на основе статистической обработки текстов на естественном языке // Известия томского политехнического университета. 2012. № 5. С. 125-130.

Брюшинкин В.Н. Когнитивный подход к аргументации // РАЦИО.ru. 2009. № 2. С. 2-22.

Буров А.А. К вопросу об идиостиле современного оратора как языковой личности // Записки Горного института. Т.160. Часть 1. 2017. С.8-9.

Виноградов В. В. Проблема авторства и теория стилей / В. В. Виноградов. – М.: Гослитиздат, 1961. – 614 с.

Воркачев С.Г. Лингвокультурология, языковая личность, концепт: становление антропоцентрической парадигмы в языкознании // Филологические науки. 2001. № 1. С. 64-72.

Воробьев В.В. Языковая личность в лингвокультурологии // Тез. докл. Языковая личность: Лингвистика. Лингвокультурология.

Лингводидактика. БашГУ. Ноябрь 2011 г. Уфа: РИЦ БашГУ, 2011. С. 234-237.

Вул С.М. Криминалистическое исследование признаков письменной речи / М-во юстиции УССР. Харьк. науч.-исслед. ин-т судебных экспертиз им. заслуж. проф. Н. С. Бокариуса. – Киев: [М-во юстиции УССР], 1973. - 44 с.

Вул С.М. Особенности оценки следователем и судом заключения идентификационной судебно-автороведческой экспертизы // Криминалистика и судебная экспертизы, 1982. №24. С.81-84.

Вул. С.М. Судебно-автороведческая идентификационная экспертиза: методические основы: Методическое пособие. — Харьков: ХНИИСЭ, 2007. — 64 с.

Галяшина Е.И. Основы судебного речеведения: Монография / Под ред. проф. М. В. Горбаневского. – М.: СТЭНСИ, 2003. – 236 с.

Галяшина Е.И., Ермолова Е.И. Перспективы развития автороведческой экспертизы в России // Судебная экспертизы, 2005. № 3. С. 5-10.

Головин Б. Н. Язык и статистика. – М.: Просвещение, 1970. – 190 с.

Ельмслев Л. Пролегомены к теории языка / Л. Ельмслев ; пер. с англ. [В. А. Звегинцев и др.]. - Москва : URSS, 2005. - 243, Пер.: Hjelmslev, Louis Prolegomena to a theory of language.

Ионова С. В., Огорелков И. В. Речевая диагностика личности по гендерному признаку в автороведении: квантитативный подход // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2020. – Т. 19, № 1. С. 115–127.

Карасик В.И. Языковой круг: личность, концепты, дискурс / В. И. Карасик; Науч.-исслед. лаб. «Аксиол. Лингвистика». - М.: ГНОЗИС, 2004 (ГУП Смол. обл. тип. им. В.И. Смирнова). – 389 с.

Караулов Ю.Н. Русский язык и языковая личность. Изд. 7-е. — М.: Издательство ЛКИ, 2010. — 264 с.

Ким Л. Г. Холистические принципы проведения идентификационной

авторoведческой экспертизы разножанровых текстов // Международная научная конференция «Современная теоретическая лингвистика и проблемы судебной экспертизы». М.: Государственный институт русского языка им. А.С. Пушкина, 2019. С. 507-516.

Кристалл Д., Дейви Д. Стилистический анализ // Новое в зарубежной лингвистике,1980. №9. С.148-165.

Лингвистика конструкций / Отв. ред. Е. В. Рахилина. — М.: «Издательский центр «Азбуковник», 2010. — 584 с.

Литвинова Т. А. Идиолект как объект корпусной идиолектологии: к становлению нового лингвистического направления. // Ученые записки Новгородского государственного университета имени Ярослава Мудрого. No 7 (25). С. 1–5.

Лосев А.Ф. Введение в общую теорию языковых моделей. / Под ред. И.А. Василенко. Изд. 2-е, стереотипное. – М.: Едиториал УРСС, 2004. – 296 с.

Медведева Т.Н. Формальные модели в лингвистике : Учебное пособие / Т.Н. Медведева. – Саратов: Научная книга, 2010. – 56 с.

Морозов Н. А. Лингвистические спектры: Средство для отличения плагиатов от истинных произведений того или другого известного автора : Стилеметрический этюд / Н. А. Морозов // Известия Отдела русского языка и словесности Императорской Академии наук. 1915. Т. 20, кн. 4. С. 93–127.

Павловская Н.Ю. Моделирование языковых категорий в свете когнитивно ориентированной парадигмы научного знания. // Модели в современной науке: единство и многообразие: сб. науч.тр. / под ред. С.С. Ваулиной, В.И. Грешных. – Калининград: Изд-во РГУ им. И.Канта, 2010 – 472 с. С. 73 – 80.

Падучева Е.В. О семантике синтаксиса. М.: Наука, 1974.

Приказ от 27 декабря 2012 года N 237 «Об утверждении Перечня родов (видов) судебных экспертиз, выполняемых в федеральных бюджетных судебно-экспертных учреждениях Минюста России, и Перечня экспертных

специальностей, по которым представляется право самостоятельного производства судебных экспертиз в федеральных бюджетных судебно-экспертных учреждениях Минюста России» (с изменениями на 13 сентября 2018 года). Официальный интернет-портал правовой информации. – URL: www.pravo.gov.ru (дата обращения: 03.05.2020).

Радбиль Т. Б., Маркина М. В. Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов/ Т. Б. Радбиль, М. В. Маркина // Политическая лингвистика. 2019. № 2 (74). С. 156-166.

Ревзин И. И. Современная структурная лингвистика: Проблемы и методы / И.И. Ревзин ; АН СССР, Ин-т славяноведения и балканистики. - Москва : Наука, 1977. - 263 с.

Родионова С. Е. Семантика интенсивности и ее выражение в современном русском языке / С. Е. Родионова // Проблемы функциональной грамматики. Полевые структуры. СПб.: Издательство "Наука", 2005. С. 150-169

Романова Т.В. Человек и время: Язык. Дискурс. Языковая личность. Н. Новгород: Нижегородский государственный лингвистический ун-т им. Н.А. Добролюбова, 2011.

Рубцова И.И., Ермолаева Е.И., Безрукова А.И. и др. Комплексная методика производства автороведческих экспертиз: Методические рекомендации. – М:ЭКУ МВД России, 2007. — 192 с.

Русская грамматика: научные труды: в 2 т. URL: http://rusgram.narod.ru/index.html (дата обращения: 12.01.2021 г.).

Сааков Т.А. К вопросу о понятии демографических характеристик автора в судебной автороведческой экспертизе. // Язык. Право. Общество: сб. ст. V Междунар. науч.-практ. конф. Пенза : ПГУ, 2018. С. 99-102.

Седов К.Ф. Становление структуры устного дискурса как выражение эволюции языковой личности: дис. ... д-ра филол. наук. Саратов, 1999.

Степаненко А.А. Гендерная атрибуция текстов компьютерной коммуникации:

статистический анализ использования местоимений. // Вестник Томского государственного университета. 2017. № 415. С. 17–25.

Тарнопольский О.Б., Кожушко С.П. Культурно-обусловленные составляющие и методика формирования вторичной языковой личности у изучающих иностранный язык в его вузовском курсе // 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. Минск: БГУ, 2019. С. 300-304.

Федеральный закон от 31 мая 2001 г. N 73-ФЗ «О государственной судебно-экспертной деятельности в Российской Федерации», Российская газета, N 256 от 31.12.2001. – URL: https://base.garant.ru/12123142/ (дата обращения: 03.05.2020).

Федотова М. А. К вопросу о разграничении понятий идиостиль и идиолект языковой личности // Записки романо-германской филологии,2013. №1. С.220-226.

Штофф В.А. Моделирование и философия / В.А. Штофф. – М.; Л.: Наука, 1966. – 304 с.

Bacciu A., Morgia M. La, Mei A., Nemmi E. N., Neri V., Stefa J. CrossDomain Authorship Attribution Combining Instance-Based and Profile-Based Features // Notebook for PAN at CLEF 2019, 2019. – URL: http://ceur-ws.org/Vol2380/paper_220.pdf (дата обращения: 05.07.2020 г.).

Bhargava M., Mehndiratta P., Asawa K. Stylometric Analysis for Authorship Attribution on Twitter Author's // International Conference on Big Data Analytics, 2013. – URL: https://www.researchgate.net/publication/299669552 (дата обращения: 03.05.2021).

Bloomfield L. A set of postulates for the science of language // Language, 1926. №2 (3). Pp. 153–164.

Campbell L. The Sophisties and Polilicus of Plato / L. Campbell. – Oxford : Clarendon, 1867. – 170 p.

Coulthard M. Author Identification, Idiolect, and Linguistic Uniqueness // Applied Linguistics. 2004. № 24 (4). Pp. 431-447.

Custódio J. E., Paraboni I. EACH-USP Ensemble Cross-domain Authorship Attribution // Notebook for PAN at CLEF 2018, 2018. – URL: http://ceur-ws.org/Vol-2125/paper_76.pdf (дата обращения: 05.07.2020 г.).

Gomzin, A., Laguta, A., Stroev, V., Turdakov, D. Detection of author's educational level and age based on comments analysis // Paper presented at Dialogue 2018, Moscow, 30 May–2 June 2018. – URL: http://www.dialog-21.ru/media/4279/gomzin_turdakov.pdf (2018) (дата обращения: 05.07.2020 г.).

Korobov M.: Morphological analyzer and generator for Russian and Ukrainian languages / Khachay, M.Y., Konstantinova, N., Panchenko, A., Ignatov, D.I., Labunets, V.G. (eds.) // AIST 2015. CCIS, vol. 542, 2015. Pp. 320–332.

Litvinova T. A., Sboev A. G., Panicheva P. V. (2018), Profiling the Age of Russian Bloggers // Proceedings of the 7th International Conference, AINL 2018, St. Petersburg, 2018. Pp. 167–177.

Lutoslawski W. The origin and growth of Plato's logic. London. Longmans, Green and Co.1897.

McMenamin G.R. Forensic Linguistics: advances in forensic stylistics. / G.R. McMenamin. 2002. – 331 p.

Murauer B., Tschuggnall M., Specht G. Dynamic Parameter Search for Cross-Domain Authorship Attribution // Notebook for PAN at CLEF 2018, 2018. – URL: http://ceur-ws.org/Vol-2125/paper_84.pdf (дата обращения: 05.07.2020 г.).

Muttenthaler L., Lucas G., Amann J. Authorship Attribution in Fan-Fictional Texts given variable length Character and Word N-Grams // Notebook for PAN at CLEF 2019, 2019. – URL: http://ceur-ws.org/Vol-2380/paper_49.pdf (дата обращения: 05.07.2020 г.).

Panicheva, P., Mirzagitova, A., Ledovaya, Y.: Semantic feature aggregation for gender identification in Russian Facebook. // Filchenkov, A., Pivovarova, L., Žižka, J. (eds.) AINL 2017. CCIS, 2018. Vol. 789. Pp. 3–15.

Pimonova E., Durandin O., Malafeev A. Doc2vec or better interpretability? A method study for authorship attribution // Paper presented at Dialogue 2020, Moscow, June 15–20, 2020, 2020.

Romanov A., Kurtukova A., Fedotova A., Shelupanov A., Goncharov V. Authorship Identification of a Russian-Language Text Using Support Vector Machine and Deep Neural Networks // Future Internet. – 2021. – Volume 13. – Issue 1. – 3. – 16 p.

Romanov A.S., Sobolev A.A., Kurtukova A.V, Fedotova A.M., Shelupanov A.A. Determining the Age of the Author of the Text Based on Deep Neural Network Models // Information. – 2020. –Volume 11. – Issue 12. – 589. – 12 p.

Shuy R. W. Creating language crimes: How law enforcement uses (and misuses) language. New York: Oxford University Press, 2005. – 194 p.

Solnyshkins M., Guryanov I., Gafiyatova E., Varlamova E. (2018). Readability metrics: the case of Russian educational texts // Proceedings of ADVED 2018-4th International Conference on Advances in Education and Social Sciences. – Istanbul, Turkey, 2018.