А.Ю. Хоменко

старший преподаватель,

Национальный исследовательский университет «Высшая школа экономики»,

Нижний Новгород, Россия

akhomenko@hse.ru

ЛИНГВИСТИЧЕСКОЕ МОДЕЛИРОВАНИЕ КАК ИНСТРУМЕНТ АТРИБУЦИИ ТЕКСТА[1]

Аннотация: Настоящее исследование посвящено разработке проблем текстовой атрибуции на основе постулатов модельной лингвистики. В работе речь пойдет о решении идентификационной задачи закрытого класса при попарном сравнении письменных текстов.

Ключевые слова: атрибуция письменного текста; автороведческая экспертиза; моделирование языковой личности; моделирование идиостиля автора.

Автор любого текста есть языковая личность. Феномен языковой личности изучается исследователями с точки зрения разных подходов: когнитивной лингвистики [20; 21; 35], лингводидактики [41], психолингвистики [5; 39], лингвокультурологии [10; 11], социолингвистики и антропологии [9; 62], судебной лингвистики [15; 19; 49; 55]. 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 [2, с.99–107]. Сама же атрибуция есть не что иное, как модель лингвистического исследования, «имитирующая те исследовательские процедуры, которые ведут лингвиста к обнаружению того

[1] Исследование поддержано Российским фондом фундаментальных исследований: грант РФФИ в рамках научного проекта «Аспиранты» (№ 19-312-90022).

или иного языкового явления» [2, с.99].

Атрибуционная лингвистика со времен Л. Кэмпбелла [48] и В. Лютославского [54] на западе и Н.А. Морозова [28] в России всегда шла двумя параллельными путями: путем стилеметрии и путем квалификативного анализа текста. Квантитативные методы сейчас получают наибольшее распространение [45; 50–53; 56–61] в то время, как квалификативные в основном используются в судебном авторoведении [1; 22; 36; 38] как дань традиции [12-14] и в связи с законодательством [31; 42]. Многие из квантитативных подходов продуктивны, тем не менее они рассматривают индивидуальный стиль как череду языковых вероятностей, а не как продукт формирования речевой способности индивида, что, на наш взгляд, является важной методологической проблемой. С помощью только квантитативных подходов, основанных на сборе данных о неких традиционных стилеметрических параметрах, пусть даже и в большом их количестве [46], 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 экспертным путем, основным недостатком которого, в свою очередь, является немалая доля субъективизма и прямая связь с уровнем компетенции исследователя. Так, для создания, с одной стороны, полной и адекватной, а с другой – объективной модели языковой личности на современном этапе развития атрибуционной лингвистики, на наш взгляд, необходима интеграция квалификативного и квантитативного анализов в одной модели исследования.

В работе предлагается именно интегративная модель, реализуемая по следующему алгоритму: 1) автоматическое извлечение из текста параметров, описывающих идиостиль с точки зрения прагматикона, тезауруса и лексикона автора (имитация экспертной работы в автоматическом режиме); 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. Так, в работе речь идет о концепции интегративной модели, которая соединяет два подхода, объективируя интерпретацию статистикой с последующим анализом статистических данных при автоматической имитации экспертной работы.

Гипотеза исследования заключается в том, что с помощью фиксированного свода формализованных правил можно создать интегративную атрибуционную модель, являющуюся объективной, всесторонне имитирующей оригинал и достаточно полной для успешного решения идентификационной задачи атрибуционной лингвистики на текстах разного объема и жанровой отнесенности в полуавтоматическом режиме.

По результатам разработки проблемы лингвистического моделирования, на основе теоретических выкладок [2–4; 7; 17; 33; 47; 26; 27; 29; 44] и пр. создана классификация модельных свойств для оценки эффективности модели с теоретической точки зрения (Таблица 1).

Таблица 1. 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 поставленной научной цели	низкий / средний / высокий уровень
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, точность, однозначность модели**	коррелирует с уровнем формализации математической модели	а) полный, 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 решения поставленной задачи	низкий / средний / высокий уровень
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

Эта таблица применяется при оценке работоспособности созданной атрибуционной модели с теоретической точки зрения.

В предлагаемой модели идентификационные параметры выделяются на всех трех уровнях языковой личности при её понимании в концепции Ю.Н. Караулова: вербально-семантическом; лингвокогнитивном; мотивационном [21, с.53]. Языковая личность понимается как результат ее формирования в определенной социальной среде: автобиографический, социолингвистический и юрислингвистический подходы [9; 12; 13; 49; 62].

Для извлечения лингвистической информации компьютерным способом (имитация работы эксперта в автоматическом режиме) все формальные правила, основанные на принципах семантического синтаксиса [30], грамматики конструкций [24], грамматики русского языка [37] при использовании структурных схем, были запрограммированы и интегрированы в электронный ресурс «ХоРом»: http://khorom-attribution.ru/#/[2].

Модуль пользователя ресурса имеет следующие функции: на вход

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

подаются два текста A и B; предварительно пользователь имеет возможность выбрать жанр текста (рис. 1). Наличие этой опции мотивировано варьированием правил поиска лингвистических структур в разных дискурсах.

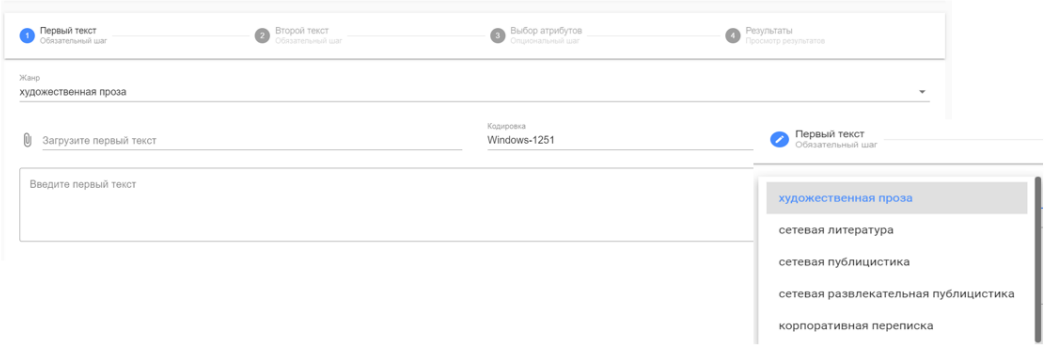

Рис. 1. Пользовательский функционал ресурса «ХоРом»: выбор жанра текста

Пользователь может строить модель не только на основе предустановленных параметров, но и имеет возможность выбирать те, которые считает наиболее релевантными для определенной пары текстов (рис. 2). Этот функционал отличает разработанное программное обеспечение от других, например, основанных на машинном обучении [59; 60], где все параметры предустановлены не пользователем, а разработчиком. Это же делает настоящий алгоритм не полностью автоматическим, а конечное решение модель оставляет за пользователем.

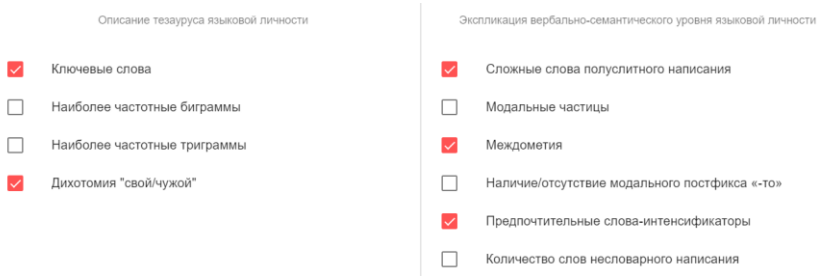

Рис. 2. Пользовательский функционал ресурса «ХоРом»: выбор параметров анализа

Пользователь не только может посмотреть, где в тексте расположены все включенные в подсчет частот реализации параметров, но также может исключить те, которые он считает шумом, а затем пересчитать данные для конечных моделей (рис. 3).

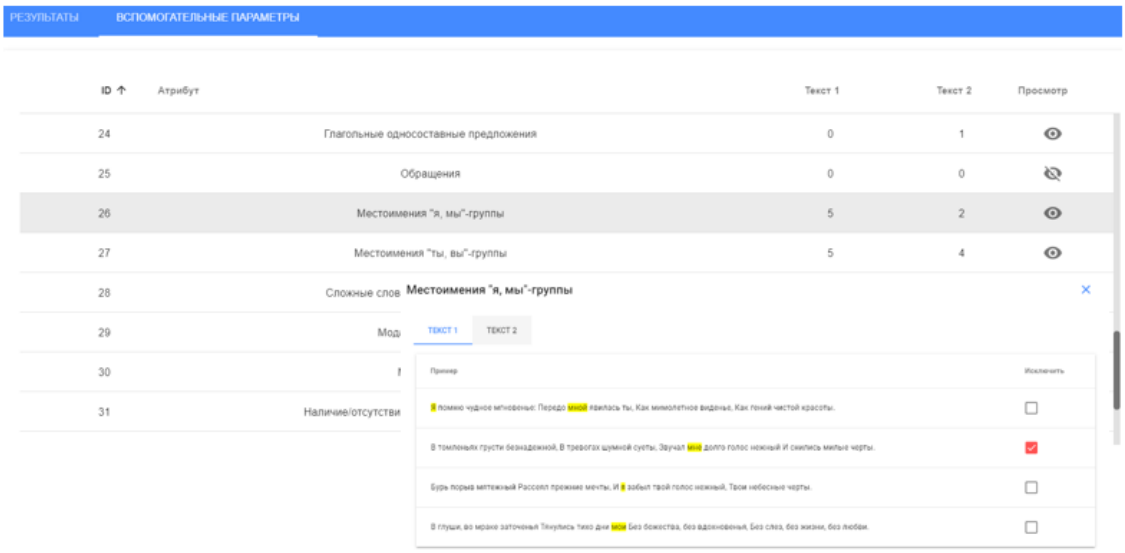

Рис. 3. Пользовательский функционал ресурса «ХоРом»: ручная отладка работы модели

В результате эмпирического исследования валидности идентификационных параметров 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. [16], индекс туманности Ганнинга, индекс Флеша-Кинкейда с коэффициентом для русского языка [63,с.679]), средние длины слов; а также такие параметры идиолекта [25], 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 с морфологической природой этих единиц, материально эксплицированной посредством присвоения им морфологических тегов. Например, поиск элементов с модальным постфиксом «-то» осуществляется в соответствии со следующим алгоритмом:

1) + Prnt-то

2) – SPRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl

3) – APRO, nom / gen / dat / acc/ ins / loc / voc / gen2 / acc2 / loc2, sin / pl

где *Prnt* – любая часть речи; «+» в начале схемы – наличие элемента в предложении; «–» в начале схемы – отсутствие элемента в предложении; «/» - обозначение «или»; SPRO – местоимение-существительное; APRO – местоимение-существительное; nom – именительный падеж; gen/gen2 – родительный и родительный второй соответственно; dat - дательный падеж; acc / acc2 – винительный и винительный второй соответственно; ins – творительный падеж; loc / loc2 – предложный и предложный второй соответственно; voc – звательная форма; sg / pl – единственное и множественное число соответственно[3].

Схему можно прочитать следующим образом: осуществляется поиск любой части речи, имеющей модальный постфикс «-то», кроме местоимений, обладающих семантикой и синтаксической функцией существительных, и местоимений, обладающих семантикой и синтаксической функцией прилагательных, в любом падеже множественного или единственного числа.

При поиске слов-интенсификаторов использовалось понимание этого термина в парадигме определения степени семантической категории интенсивности [34]. Чаще всего говорят о наречиях-интенсификаторах, но несмотря на высокую частоту употребления этой части речи для реализации категории интенсивности, она не исчерпывается исключительно наречиями. Всего для поиска экспликаторов этой категории написано 16 специфических правил, а также создан список из 93 интенсификаторов.

Вербально-семантический уровень (уровень идиолекта в соответствии с концепцией [8; 23; 43]) легко формализуем, поскольку сам по себе имеет «более формальные» языковые характеристики, которые априори считаются стабильными [25].

Для репрезентации фрагмента тезаурусного уровня выбраны такие

[3] Здесь и далее используется номенклатура, соответствующая частеречному и морфологическому тегированию в Национальном корпусе русского языка: https://ruscorpora.ru/new/corpora-morph.html.

параметры, как ключевые лексемы, наиболее частотные словные триграммы и биграммы, экспликаторы аксиологических текстовых доминант дихотомии «свой/чужой».

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 (LL). В конечный список включаются лишь слова со значением LL более 50 (экспериментально полученный коэффициент). Поиск словных биграмм и триграмм основан на абсолютной частотности встречаемости слов рядом друг с другом и осуществляется с помощью встроенной функций Python 3.6. При подсчете реализаций этого параметра учитывается отсутствие слова в списке стоп-слов, кириллическое написание и длина слова более 2 символов. При анализе ключевых лексем и наиболее частотных сочетаний слов из полученных списков удаляются сочетания с именами собственными, поскольку данные лексемы маркируют не собственно особенности авторских идиостилей, а тематическую отнесенность текстов. Под экспликаторами 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 [40].

Тезаурусный уровень наиболее труден для формализации. Можно автоматически создать материальную экспликацию авторского тезауруса [6], тем не менее определить, как лексемы в тезаурусе «выстраиваются в упорядоченную, достаточно строгую иерархическую систему, в какой-то степени (непрямой) отражающую структуру мира» [21,с. 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 [30] и какую грамматическую конструкцию реализуют те или иные его компоненты [24]. Например, для вычленения из текста вводных слов формализованное правило (алгоритм поиска) выглядит следующим образом:

– для машинного представления создается словарь из всех возможных вводных слов русского языка;

– прописывается грамматико-пунктуационное правило, которое позволяет выделить из текста именно вводную конструкцию, а не омонимичную ей:

1) __, Prnt,__

2) <начало предложения> Prnt,

где *Prnt* – любая часть речи; __ – некоторая часть предложения, <*начало предложения*> – обозначение начала предложения.

Правило для поиска целевых оборотов основано на понятии семантической валентности [30,с.44] и грамматики предложных конструкций [24]. Так, составные предлоги «с целью/из расчета» требуют инфинитива (условие валентности) при реализации целевого оборота, значит, формализованное правило для поиска таких структур будет выглядеть следующим образом: «с целью/из расчета» + INF, где обозначение *INF* использовано для инфинитива.

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 конечные модели также включены значения стилеметрических метрик (Таблица 2).

Таблица 2.

Пример модели идиостиля, репрезентированного в текстах «Последний поклон» и «Звездопад» В. Астафьева

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

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[4], является одно лицо (рис. 4).

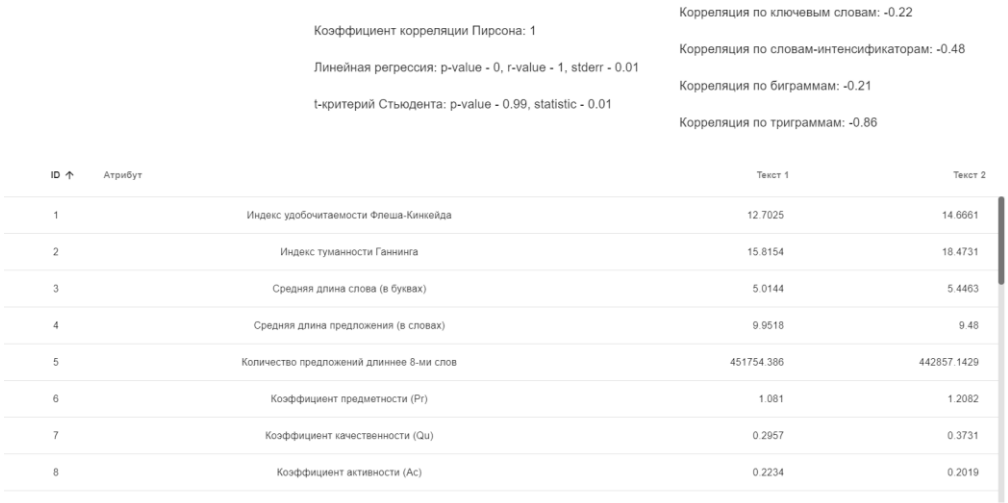

Рис. 4. Пользовательский функционал ресурса «ХоРом»: результаты сравнения моделей

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

Важно, что данный блок не является конечным шагом в разработанной модели. Текстовую статистику необходимо интерпретировать. Так, для традиционной математической статистики значимым считается коэффициент корреляции более 65% [18], в случае обработки текстовой информации о сходстве моделей следует говорить при коэффициенте корреляции 86% и выше [32]. Программное обеспечение намеренно не выдает результат в виде выводного знания формата «*Автором двух сравниваемых текстов является одно лицо/ Авторами двух сравниваемых текстов являются разные лица*», поскольку в разработанной модели именно эксперт, основываясь на статистических данных и на их интерпретации, принимает конечное решение об атрибуции текста, используя 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[5], 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

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

						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 0,97	может достигать 0,94; обычно около 0,84	около 0,50	-	+	если для текстов 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 модели проводилась с двух точек зрения:

1) полученные модели языковой личности рассматривались с точки зрения шкалы теоретической оценки моделей, представленной в Таблице 1. Посредством этих оценок удалось доказать гипотезу о том, что с помощью фиксированного свода формализованных правил можно создать интегративную атрибуционную модель, являющуюся полной, всесторонне имитирующей оригинал и одновременно объективной;

2) полученные модели также оценивались с точки зрения решения с их помощью идентификационной задачи атрибуционной лингвистики. Тестирование и апробация созданного алгоритма проходили на разножанровых текстовых коллекциях. **Результаты** апробации представлены ниже:

- коллекция текстов художественной литературы, включающая тексты

С.Д.Довлатова и В.П.Астафьева: 10 текстов, средний объем – 20,000 слов). Доля правильных ответов алгоритма (accuracy), точность (precision) и полнота (recall) равны 100%, F-мера – 1[6];

- 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. Accuracy – 83%, precision – 67%, recall – 100%, F-мера – 0,8;

- 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. Accuracy, precision и recall равны 100%, F-мера – 1;

- 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. Accuracy – 40%, precision – 0, recall – 0, F-мера – 0;

- 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. Accuracy – 83%, precision – 67%, recall – 100%, F-мера – 0,8.

Посредством 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;

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

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 авторoведении [36] как объем, необходимый для определения авторства текстов, при встраивании в анализ статистической информации должен 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): accuracy – 83%, precision – 67%, recall – 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 (пример – Таблица 2). Эти модели являются понятными и простыми, легко интерпретируемыми экспертным путем, с одной стороны, и полными и адекватно имитирующими объект-оригинал – с другой.

Результаты работы атрибуционной модели можно сравнить с результатами работы других моделей, основанных на машинном обучении и нейронных сетях. Так, экспериментальный результат работы программного обеспечения «Автороведъ» [61], разработанного для решения задач в том числе судебной лингвистики, на материале художественной прозы известных авторов и публицистики этих же авторов – это точность классификации 96,6%. «Автороведъ» использует метод опорных векторов и логистическую регрессию для решения задачи атрибуции. Этот алгоритм оказывается чувствительным к объему анализируемого текстового материала. Так, он не всегда успешно идентифицирует автора при исследовании текстов большого объема (В.Астафьева: «Затеси», «Последний поклон», «Ловля пескарей в Грузии»; С.Довлатова: «Ариэль», «Записки надзирателя», «Соло на ундервуде», «Компромисс»), отличного от того, на котором проходило обучение программы. Для улучшения точности работы алгоритма тексты подвергаются семплированию[7]. Предлагаемая в работе атрибуционная модель менее чувствительна к текстовому объему, поскольку разница в нем нивелируется относительным характером используемых частот и правильной параметризацией модели анализа для каждой конкретной текстовой пары.

Следует отметить, что сравнение разработанной модели с другими 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.

[7] Данные о работе программы «Автороведъ» предоставлены А.С. Романовым. Автор исследования выражает благодарность за предоставленные данные.

Следует отметить, что функционал разработанных моделей анализа и модели исследования, а также созданного на их основе электронного ресурса много шире изначально заложенных возможностей. Наработки 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. Важно, что при использовании разработанной методики в любом из представленных выше случаев модель языковой личности будет отвечать принципам полноты, простоты, адекватности, технически точного и объективного описания оригинала, она будет экспланаторной, коммуникативной и интерпретируемой.

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

Литература

1. 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.

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

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

4. 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.

5. 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.

6. 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.

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

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

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

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

11. 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.

12. 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.

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

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

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

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

17. 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.

18. Ивченко Г.И., Медведев Ю.И. Математическая статистика: Учебное пособие. – М.: Высшая школа, 1984. – 600 с.

19. 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.

20. 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.

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

22. Ким Л. Г. Холистические принципы проведения идентификационной автороведческой 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.

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

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

25. 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.

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

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

28. 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.

29. 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.

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

31. 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).

32. 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.

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

34. 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

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

36. 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.

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

38. 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.

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

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

41. 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.

42. 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).

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

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

45. 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.).

46. 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).

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

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

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

50. 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.).

51. 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.).

52. 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.

53. 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.

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

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

56. 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.).

57. 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.).

58. 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.

59. 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.

60. 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.

61. 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.

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

63. 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.

A. Yu. Khomenko

LINGUISTIC MODELING AS A TECHNIQUE IN AUTHORSHIP ATTRIBUTION

Abstract: This paper discusses the testing of an integrative attribution analysis method for texts in the Russian language. It is based on interpretative language study with its objectification through the usage of mathematical statistics methods. The algorithm solves the identification problem of authorship attribution.

Key words: written text attribution; forensic text attribution; language personality modeling; author's individual style modeling.