www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2022.3.12

UDC 81’42:004.738.5
LBC 81.055.51.5

Submitted: 21.01.2022
Accepted: 21.03.2022

ON TEXT AUTHOR IDENTIFICATION IN NETWORK COMMUNICATION [1]

Tatyana V. Romanova

National Research University Higher School of Economics – Nizhny Novgorod, Nizhny Novgorod, Russia

Anna Yu. Khomenko

National Research University Higher School of Economics – Nizhny Novgorod, Nizhny Novgorod, Russia

Abstract. The article deals with general and particular patterns of solving the identification problem of attributional linguistics. Experimental research is carried out on the material of the texts of written network communication. The problem is solved for several discursive areas: online literature, corporate e-mail, short comments on posts on the entertainment portal. These discursive spheres correlate with several functional styles of speech: fiction, official and colloquial, which makes it possible to identify the trends in textual attribution within the material of Internet communication. A holistic approach with linguistic modelling as its means, is used for the analysis of speech material. The problem of determining text authorship is solved by comparing a disputed text, the author of which, according to the conditions of the experiment, is unknown, with texts, the authors of which are known. The application of specialized software “KhoRom”, has enabled the researchers to create mathematical models of the individual styles of the compared texts authors. The contrast of the resulting models is performed by considering Pearson linear correlation coefficient, coefficient of determination, and Students t-criterion. On the basis of these models and their study, a number of conclusions about statistical patterns of an authorship study in various discursive spheres of electronic communication are made. The revealed general and particular statistical regularities are represented in score tables.

Key words: attribution, language personality, linguistic model, mathematical model, network communication.

Citation. Romanova T.V., Khomenko A.Yu. On Text Author Identification in Network Communication. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2022, vol. 21, no. 3, pp. 143-157. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2022.3.12

УДК 81’42:004.738.5
ББК 81.055.51.5

Дата поступления статьи: 21.01.2022
Дата принятия статьи: 21.03.2022

ИДЕНТИФИКАЦИЯ АВТОРА ТЕКСТА В СЕТЕВОЙ КОММУНИКАЦИИ [1]

Татьяна Владимировна Романова

Национальный исследовательский университет «Высшая школа экономики» – Нижний Новгород, г. Нижний Новгород, Россия

Анна Юрьевна Хоменко

Национальный исследовательский университет «Высшая школа экономики» – Нижний Новгород, г. Нижний Новгород, Россия

Аннотация. Статья посвящена выявлению и описанию общих и частных закономерностей при решении идентификационной задачи атрибуционной лингвистики. Экспериментальное исследование выполнено на материале текстов письменной сетевой коммуникации. Задача идентификации автора решается для нескольких дискурсивных сфер: сетевой литературы, корпоративной электронной переписки и коротких комментариев к постам на развлекательном портале. Учет того, что данные дискурсивные сферы коррелируют с художественным, официально-деловым и разговорным функциональными стилями, по-

зволил обозначить векторы текстовой атрибуции в рамках интернет-коммуникации. Для анализа речевого материала используется холистический подход, инструментом которого является лингвистическое моделирование. Предложена оригинальная методика определения авторства при сравнении спорного текста, автор которого, по условиям эксперимента, неизвестен, с текстами-образцами, авторы которых известны. С применением программного обеспечения «ХоРом» созданы математические модели идиостилей авторов сопоставляемых текстов. Сравнение полученных моделей осуществлено посредством коэффициента линейной корреляции Пирсона, коэффициента детерминации и *t*-критерия Стьюдента. Сделаны выводы о статистических закономерностях при проведении авторoведческого исследования разных дискурсивных сфер электронной коммуникации. Установленные общие и частные статистические закономерности оформлены в виде рейтерских таблиц.

Ключевые слова: атрибуция, языковая личность, лингвистическая модель, математическая модель, сетевая коммуникация.

Цитирование. Романова Т. В., Хоменко А. Ю. Идентификация автора текста в сетевой коммуникации // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2022. – Т. 21, № 3. – С. 143–157. – DOI: https://doi.org/10.15688/jvolsu2.2022.3.12

Введение

Статья посвящена поиску решения традиционной для текстовой атрибуции идентификационной задачи на современном материале сетевой (электронной) коммуникации.

Сетевую (электронную, виртуальную, цифровую, интернет-, компьютерно-опосредованную) [Горошко, 2012] коммуникацию можно понимать как «использование людьми электронных сообщений (чаще мультимедийных) для формирования знаний и взаимопонимания в разнообразных средах, контекстах и культурах» [Розина, 2005, с. 32]. Следовательно, сетевая коммуникация – сложное, разножанровое явление, включающее большое количество дискурсивных сфер: коммуникацию в мессенджерах (групповую и индивидуальную) с помощью коротких текстовых сообщений, на интернет-платформах в рамках жанра комментария, эпистолярную (личную и официально-деловую) и пр.

Поскольку анонимность сетевого общения, возможность использования сетевых псевдонимов часто связаны с правонарушениями, совершаемыми в интернет-пространстве, актуальными являются вопросы атрибуции автора в такой коммуникации. Идентификация правонарушителей по речевым продуктам в сети – задача судебной авторoведческой (атрибуционной) экспертизы, методики которой находятся в стадии разработки и апробации. Целью работы стал поиск значимых закономерностей при определении авторства в электронной коммуникации.

Материал и методы

В рамках любого типа коммуникации в том или ином объеме свою репрезентацию получает языковая личность (далее – ЯЛ) пишущего, материально эксплицированная в идиостиле автора письменного текста. Представители таких направлений языкознания, как психолингвистика, когнитивная лингвистика, социолингвистика, судебная лингвистика и др., пользующиеся термином «языковая личность», обозначают им многоуровневое явление, стараясь анализировать и описывать ЯЛ с позиций холистического подхода [Виноградов, 1961; Белянин, 2000; McMenamin, 2002; Карасик, 2004; Галяшина, 2003; Coulthard, 2004; Shuy, 2005; Вул, 2007; Караулов, 2010; Романова, 2011; Litvinova, Sboev, Panicheva, 2018; Ионова, Огорелков, 2020]. Это, на наш взгляд, является естественным и оправданным, поскольку, как отмечал Б. Блох, материальный репрезентант языковой личности, индивидуальный авторский стиль, есть не то, что человек говорит в какой-то момент времени, а то, что он вообще *может* сказать на этом языке [Bloch, 1948, p. 7].

В рамках использования целостного подхода в настоящем исследовании предлагается применение методов модельной лингвистики, позволяющих создать целостную модель ЯЛ пишущего, отражающую все уровни языка и организации ЯЛ, а также компетенции, с ним связанные, и позволяющую решать идентификационную задачу закрытого класса (то есть с ограниченным количеством авторов текстов): получение информации соб-

ственно об авторе текста при попарном сравнении спорного текста, автор которого неизвестен, с текстами-образцами, автор / авторы которых заведомо известны.

Моделирование языковой личности пишущего как совокупности нескольких уровней осуществляется нами с использованием интегративного подхода, сочетающего методы интерпретативной лингвистики для выявления характеристик авторского идиостиля в традиционном понимании (тезаурус личности, ее прагматикон и лексикон) и стилостатистики для объективации результатов интерпретативного анализа.

Методика реализуется по следующему алгоритму: 1) автоматическое извлечение из текста параметров, описывающих прагматикон, тезаурус и лексикон автора; 2) поиск традиционных стилеметрических данных; 3) присвоение веса каждому параметру; 4) построение математических моделей сравниваемых текстов; 5) сравнение математических моделей; 6) экспертный анализ статистических данных. Предлагаемый алгоритм анализа реализован с помощью интерфейсного программного обеспечения открытого доступа «ХоРом», созданного при участии авторов статьи.

В основе формализации уровневой структуры языковой личности в исследовании лежат постулаты теории Ю.Н. Караулова [Караулов, 2010]. Собственно процесс формализации строится на принципах семантического синтаксиса [Падучева, 1974] и в соответствии с правилами грамматики русского языка [Русская грамматика].

Структура языковой личности рассматривается как совокупность трех уровней: вербально-семантического, лингвокогнитивного, мотивационного [Караулов, 2010, с. 53].

Объективация данных, полученных с помощью интерпретативного анализа, осуществляется посредством создания математической модели каждого из текстов как продуктов репрезентации индивидуального стиля пишущего. Такая математическая модель представляет собой пул параметров всех уровней с относительной частотой реализации каждого параметра, в модель также включены традиционные стилостатистические параметры [Khomenko et al., 2021]. Две математические модели идиостилей авторов, репрезентированных в двух сравниваемых текстах, сопоставляются с помощью следующих метрик: коэффициента корреляции Пирсона (ККП), коэффициента детерминации линейной регрессии (КД) и *t*-критерия Стьюдента (*p*-value).

Для достижения поставленной цели авторы работы исследовали несколько дискурсивных сфер интернет-коммуникации:

– сетевую литературу, фанфикшен – современную беллетристику, локализующуюся как часть определенной субкультуры (фэндома). Использован материал портала «Книга фанфиков», включающий тексты 3 авторов-женщин, 4 авторов-мужчин, всего 187 текстов, средний объем текстов – от 1 500 до 40 000 слов;

– корпоративную электронную переписку. Использован материал анонимизированной корпоративной русскоязычной переписки, включающий тексты 2 авторов-женщин, 2 авторов-мужчин, всего 236 текстов (от 45 до 49 писем для одного автора), средний объем текстов – от 50 до 500 слов;

– короткие комментарии к постам на развлекательном портале. Использован материал ресурса «ЯПлакалЪ», включающий тексты 3 авторов-женщин, 3 авторов-мужчин, всего 424 текста, средний объем текстов – от 50 до 100 слов.

Со всеми текстовыми коллекциями можно ознакомиться в репозитории (Электронный репозиторий).

Выбранные дискурсивные сферы имеют корреляции с традиционными функциональными стилями: художественным, разговорным и официально-деловым, что позволяет найти общие и частные закономерности авторской атрибуции в электронной коммуникации.

Результаты и обсуждение

Сетевая литература

Сетевая литература (фанфикшен) отвечает в той или иной степени жанровым характеристикам художественной прозы, тем не менее идиостили авторов современной беллетристики, конечно, менее исследованы в сравнении с идиостилями классиков, что значительно осложняет процесс идентификации. Сетевую литературу можно воспринимать как коммуникацию в рамках схемы «от одного – ко многим» [Бондаренко, 2004, с. 157], а с уче-

том специфики фендомной организации (тексты продуцируются в основном в связи с одним художественным целым: фильмом, книгой, компьютерной игрой, комиксом) и специфики самой платформы размещения («Книга фанфиков» предполагает не только размещение текста произведения, но и возможность его комментирования внешними и внутренними пользователями, а также общепользовательскую корректуру материала) электронную беллетристику можно воспринимать и как коммуникацию по схемам «от многих – ко многим», «от многих – к одному», «от одного – к одному» [Бондаренко, 2004, с. 157].

В рамках дискурса современной сетевой художественной литературы в качестве примеров анализа приведем несколько обследованных текстовых пар (вместо фамилий и имен приводятся никнеймы, используемые авторами на ресурсе «Книга фанфиков»).

При анализе текстовой пары Аллесий «Третья игра: Сын Бессмертного» – Аллесий «Третья игра: Путь Шамана» (тексты одного жанра и одного автора с сопоставимым объемом: 22 412 слов vs 22 405 слов) получаем валидные результаты метрик с очень близкой стилостатистической составляющей модели – ККП: 1, КД: 1, *p*-value: 0,97 (табл. 1).

При этом параметры ЯЛ из структурированной интерпретативной модели более неоднородны, чем для прозы известных авторов (см. табл. 2).

Тем не менее эта неоднородность не позволяет сформировать ложноотрицательный вывод для пары анализируемых текстов. При сравнении моделей без «идеальной» стилостатистики только по интерпретативным параметрам также получаем валидный результат – ККП: 0.99, КД: 0.98, *p*-value: 0.95.

В данном случае автоматический алгоритм интегративной методики создает достаточно полную и адекватную модель для валидного вывода по гипотезе H_0 о том, что автором сравниваемых текстов является одно лицо при условии подбора релевантных для анализируемого дискурса параметров конечной модели, а также верной и, что важно, совокупной интерпретации статистических метрик.

При сравнении текстов одного жанра и одного автора с различным объемом (22 412 слов vs 812 слов) – Аллесий «Третья игра: Сын Бессмертного» и Аллесий «Каменный Дом» – получаем сходные с предыдущим экспериментом результаты. При полной модели коэффициенты доказывают гипотезу H_0 – ККП: 1, КД: 1, *p*-value: 0,95.

Из интерпретации модели становится очевидным, что стилостатистический компонент для индивидуального стиля Аллесия дает наибольший прирост значения меры корреляции моделей.

Тем не менее экспериментальным путем в исследовании удалось выяснить, что стилостатистический компонент является фактором шума для текстов современной сетевой литературы. Данный вывод позволил сделать ряд экспериментов над текстами одного и различных авторов: стилостатистика в них весьма сходна. Медианное значение каждого из параметров стилостатистического пула для сетевой литературы представлено ниже:

Таблица 1. **Стилостатистический пул параметров анализа ЯЛ. Автор – Алессий**

Table 1. **Stylostatistic parameters set for the analysis of Alessiy language personality**

№	Параметр	Аллесий «Третья игра: Сын Бессмертного»	Аллесий «Третья игра: Путь Шамана»
1	Индекс удобочитаемости Флеша-Кинкейда	13.2395	13.6227
2	Индекс туманности Ганнинга	15.975	16.4649
3	Средняя длина слова (в буквах)	5.1227	5.1099
4	Средняя длина предложения (в словах)	9.2341	9.5605
5	Количество предложений длиннее 8 слов	428089.4137	442993.9077
6	Коэффициент предметности (Pr)	1.2382	1.2306
7	Коэффициент качественности (Qu)	0.5196	0.5268
8	Коэффициент активности (Ac)	0.1698	0.181
9	Коэффициент динамизма (Din)	0.3845	0.4204
10	Коэффициент связности текста (Con)	1.8794	1.9909

Таблица 2. **Интерпретативный пул параметров анализа ЯЛ. Автор – Алессий**

Table 2. **Qualifying parameters set for the analysis of Alessiy language personality**

№	Параметр	Аллесий «Третья игра: Сын Бессмертного»	Аллесий «Третья игра: Путь Шамана»
1	Количество слов несловарного написания	4384.7629	5370.9604
2	Предложения с однородными рядами	2466.4292	1957.2144
3	Предложения с обособленными приложениями	319.7223	273.0997
4	Вводные слова и конструкции	7810.359	6417.8425
5	Целевые и выделительные обороты	319.7223	546.1994
6	Конструкции с семантикой сравнения	4110.7153	4415.1115
7	Синтаксические сращения	0	0
8	Сравнительные придаточные	6074.7237	7146.1083
9	Конструкции с сопоставительными союзами	1872.6592	1775.1479
10	Вставные конструкции	5206.906	2002.731
11	Сложные синтаксические конструкции	41929.2957	41829.7679
12	Глагольные односоставные предложения	10139.7643	10559.8543
13	Обращения	1141.8654	409.6495
14	Местоимения «я, мы»-группы	44898.1456	53072.3714
15	Местоимения «ты, вы»-группы	33296.7936	37778.7893
16	Сложные слова полуслитного написания	913.4923	1183.432
17	Модальные частицы	28546.6338	26172.0528
18	Междометия	1735.6353	819.299
19	Наличие / отсутствие модального постфикса «-то»	1370.2384	1228.9486

1. Индекс удобочитаемости Флеша-Кинкейда: 14.7924.

2. Индекс туманности Ганнинга: 18.02705.

3. Средняя длина слова (в буквах): 5.25835.

4. Средняя длина предложения (в словах): 10.7283.

5. Количество предложений длиннее 8 слов: 510434.1406.

6. Коэффициент предметности (Pr): 1.2386.

7. Коэффициент качественности (Qu): 0.49.

8. Коэффициент активности (Ac): 0.1824.

9. Коэффициент динамизма (Din): 0.39525.

10. Коэффициент связности текста (Con): 2.0962.

При сравнении текстовой пары одного жанра (тематика не учитывается) и разных авторов с сопоставимым объемом: Аллесий «Третья игра: Сын Бессмертного» – Tigrewurmut «Звездные Войны: Сила или Жизнь» (22 412 слов vs 22 458 слов) для построения моделей языковых личностей авторов использовались только интерпретативные параметры. Результат оправдывает ожидания и позволяет сделать валидные выводы – ККП: 0,98, КД: 0,96, *p*-value: 0,91.

Со статистической точки зрения значения метрик кажутся высокими. Тем не менее пристальное изучение коэффициентной наполненности самих моделей ЯЛ в совокупности с данными ряда экспериментов, подобных описанному выше, подтверждает, что для сетевой литературы полученные значения метрик можно считать достаточно низкими, чтобы отвергнуть гипотезу H_0. В результате экспериментов удалось выявить следующие статистические закономерности для установления авторства по текстам сетевой литературы (см. табл. 3).

Приведем еще несколько примеров сравнения текстовых пар (тексты одного жанра и разных авторов с разным объемом):

– Аллесий «Третья игра: Сын Бессмертного» и Миха Француз «Следующая ступень» (22 412 слов vs 12 340 слов) – ККП: 0,86, КД: 0,74, *p*-value: 0,51;

– Аллесий «Третья игра: Сын Бессмертного» и Кицунэ Миято «Желчь броненосца» (22 412 слов vs 2 505 слов) – ККП: 0,69, КД: 0,45, *p*-value: 0,98;

– Аллесий «Третья игра: Сын Бессмертного» и Ктая «Взгляд сквозь щели канона» (22 412 слов vs 1 040 слов) – ККП: 0,37, КД: 0,14, p-value: 0,82.

Таблица 3. **Рейтерская таблица наиболее частых случаев распределения значений метрик в сетевой литературе ***

Table 3. **Score table for the most common metrics values in network literature ***

ККП	КД линейной регрессии	*t*-критерий Стьюдента (*p*-value)	Автором сравниваемых текстов, вероятно **, является одно лицо	Авторами сравниваемых текстов, вероятно, не является одно лицо	Комментарий
Не ниже 0,99; обычно достигает 1,00	Не ниже 0,98; обычно достигает 1,00	Не ниже 0,95; обычно около 0,97	+	–	Значения всех метрик должны быть весьма высоки
Обычно около 0,86; изредка достигает высоких значений (до 0,98)	Обычно около 0,74; изредка достигает высоких значений (до 0,96)	Обычно около 0,50; редко может достигать 0,91, но не превышает это значение	–	+	Значения метрик могут быть высокими, но важно обращать внимание именно на их сочетания; основной метрикой для беллетристики, как и для художественной литературы, остается *p*-value *t*-критерия Стьюдента: если эта метрика имеет низкое значение, то авторами сравниваемых текстов, скорее всего, являются разные люди. Средние значения других метрик ниже, чем при едином авторстве текстов, но могут быть и достаточно высокими. Важно обращать внимание на их совокупность. Иногда даже *p*-value *t*-критерия Стьюдента достигает значения около 0,91. Если это происходит, при принятии решения необходимо применять принципы холистического подхода
Обычно низкие значения около 0,37; иногда достигается значение 0,69	Обычно низкие значения около 0,14; иногда достигается значение 0,45	Может быть около 0,82; изредка достигает 0,98	–	+	В ходе исследования были обнаружены случаи появления весьма высокого значения *p*-value *t*-критерия Стьюдента при сравнении текстов разных авторов, тем не менее в этих случаях неизменно низкими были значения других метрик. В описанной ситуации следует обращать внимание на совокупные значения метрик при принятии атрибуционного решения

Примечания. * – если в приведенных рейтерских таблицах не указаны некоторые значения и «вилки» значений рассматриваемых метрик, это говорит о том, что в случае появления таких значений при формировании атрибуционных выводов исследователь должен обращать внимание на значения наиболее релевантных параметров для анализируемой модели, основываться на своем исследовательском опыте; ** – вероятностный характер вывода связан с тем, что в каждом конкретном случае в соответствии с разработанной методикой решение о конечном авторстве принимает исследователь.

Notes. * – if some determined metrics values and variables of values are not indicated in the given evaluation tables, the researcher should pay attention to the values of the most relevant parameters for the analyzed model, using his/her research experience. ** – probabilistic conclusion is due to the fact that in each specific case, in accordance with the developed method, the decision on the final authorship is made by the researcher.

Исходя из экспериментальных данных, приведенных в таблице 2, в рассматриваемых текстовых парах гипотеза H_0 может быть отвергнута при условии верной интерпретации статистических данных, что является валидным выводом.

Итак, установлено, что для дискурсивной сферы сетевой литературы неинформативным является стилостатистический пул, поскольку, по экспериментальным данным, значения стилостатистических параметров близки для всех обследованных текстов. Тем не менее даже без стилеметрического блока интегративная методика дает возможность создать достаточно полные модели языковых личностей авторов сравниваемых текстов, объективно и адекватно имитирующую оригиналы, что позволяет решить идентификационную задачу атрибуционной лингвистики при верном анализе статистики, получаемой в результате математического моделирования идиостилей авторов сравниваемых текстов.

Корпоративная переписка

Корпоративная переписка исследуется в традиционном ключе атрибуционной лингвистики вслед за работами, основанными на анализе, например, корпуса писем «Энрон» (англ.: The Enron Email Corpus) [Friginal, Hardy, 2014; Wright, 2017] и пр. и/или имеющими место при решении практически ориентированных задач [Резанова, Романов, Мещеряков, 2013].

Тексты корпоративной переписки воспринимаются и обследуются не по отдельности, а в рамках массива, который при сопоставлении текстов одного автора делится примерно на равные сравниваемые части. Каждому автору в корпусе было присвоено кодовое имя.

Результаты экспериментов позволили выяснить следующее: при сравнении идиостиля одного автора (как мужчины, так и женщины), репрезентированного в разных текстах, корреляционные характеристики ЯЛ демонстрируют весьма высокие показатели:

– Сергей – Сергей (совокупность текстового материала составляет 1 503 слова, между собой сравниваются примерно равные части массива) – ККП: 1, КД: 1, *p*-value: 0,91;

– Мария – Мария (совокупность текстового материала составляет 991 слово) – ККП: 1, КД: 1, *p*-value: 0,89.

Высокой релевантностью для моделей ЯЛ авторов корпоративных писем обладает стилостатистический пул. Тем не менее объема текстового материала в корпоративных письмах достаточно и для создания объективно имитирующей объект-оригинал модели с помощью данных интерпретативного пула.

При сравнении ЯЛ Марии и Сергея (авторы разной гендерной принадлежности) результаты подсчетов по всем статистическим метрикам существенно разнятся как в блоке стилеметрии, так и в блоке когнитивно и психолингвистически маркированных параметров – ККП: 0,99, КД: 0,98, *p*-value: 0,48. Несмотря на относительно высокие значения ККП и КД, наибольший прирост информации показывает *t*-статистика, значения которой являются слишком низким, чтобы признать гипотезу H_0 подтвержденной. Значения коэффициентов хотя и высоки, но не достигают 100 %, что, исходя из экспериментов для данной дискурсивной сферы, необходимо для признания гипотезы H_0 верной. Важно, что корреляция по ключевым словам в этом случае практически равна нулю, а по словам-интенсификаторам – и вовсе отсутствует.

При сравнении корпоративной переписки двух разных авторов-мужчин (текстовая пара: Иван – Петр: совокупность текстового материала для Ивана составляет 1 145 слов; совокупность текстового материала для Петра составляет 2 264 слова) видим наглядные различия в структурах ЯЛ, выраженные в числовом отношении – ККП: 0,95 (по экспериментальным данным для этой дискурсивной сферы настоящее значение коэффициента является достаточно низким), КД: 0,90, *p*-value: 0,51 (наиболее показательный коэффициент для данного материала равно, как и для дискурса корпоративной переписки в целом).

При итоговом сравнении видим большое количество «выбросов», сильно разнящихся значений в парах, что говорит о разной стилистике текстов (см. табл. 4).

При сравнении «женской» пары текстов (Мария – Василиса: совокупность текстового материала для Василисы составляет 5 777 слов) конечные коэффициенты получаются значительно более спорными, чем для текстовой

Таблица 4. **Математические модели ЯЛ. Иван и Петр**

Table 4. **Mathematical models of Ivan and Peter language personalities**

№	Параметр	Иван	Петр
1	Индекс удобочитаемости Флеша-Кинкейда	17.9679	21.6623
2	Индекс туманности Ганнинга	21.8419	24.3697
3	Средняя длина слова (в буквах)	6.2608	6.3731
4	Средняя длина предложения (в словах)	**6.1605**	**13.0714**
5	Количество предложений длиннее 8 слов	**166666.6667**	**607142.8571**
6	Коэффициент предметности (Pr)	2.5455	1.7319
7	Коэффициент качественности (Qu)	0.2371	0.3298
8	Коэффициент активности (Ac)	0.1102	0.1466
9	Коэффициент динамизма (Din)	0.1695	0.2572
10	Коэффициент связности текста (Con)	1.0741	2.4405
11	Количество слов несловарного написания	**8016.0321**	**19125.6831**
12	Предложения с однородными рядами	0	1821.4936
13	Предложения с обособленными приложениями	0	0
14	Вводные слова и конструкции	**6012.024**	**0**
15	Целевые и выделительные обороты	0	0
16	Конструкции с семантикой сравнения	0	0
17	Синтаксические сращения	0	0
18	Сравнительные придаточные	**0**	**910.7468**
19	Конструкции с сопоставительными союзами	0	0
20	Вставные конструкции	15030.0601	13661.2022
21	Сложные синтаксические конструкции	**13026.0521**	**17304.1894**
22	Глагольные односоставные предложения	**17034.0681**	**10018.2149**
23	Обращения	**48096.1924**	**910.7468**
24	Местоимения «я, мы»-группы	**5010.02**	**12750.4554**
25	Местоимения «ты, вы»-группы	**23046.0922**	**32786.8852**
26	Сложные слова полуслитного написания	0	0
27	Модальные частицы	**5010.02**	**3642.9872**
28	Междометия	1002.004	0
29	Наличие / отсутствие модального постфикса «-то»	0	0

пары авторов-мужчин – ККП: 1, КД: 1, *p*-value: 0,63 (в данном случае вновь только *t*-статистика дает валидные результаты).

Несмотря на то что модели существенно различаются в собственно интерпретативном аспекте (имеет место немалое количество «волн» и «выбросов»), они достаточно близки в стилеметрии, что и дает высокие показатели корреляции Пирсона и коэффициента детерминации.

Сходство стилостатистической компоненты ЯЛ авторов-женщин может быть объяснимо с точки зрения гендерной лингвистики [Калугина, 2013].

При анализе текстов смешанной пары Мария – Петр коэффициенты в совокупном представлении и *t*-статистика как центральная метрика для дискурса корпоративной переписки помогают опровергнуть гипотезу H_0 – ККП: 0,99, КД: 0,98, *p*-value: 0,63.

Обследование моделей индивидуальных стилей авторов показывает существенные различия числовой структуры их состава, которые свидетельствуют о разных языковых личностях пишущих (см. табл. 5).

Итак, по результатам экспериментов удалось установить следующие закономерности, релевантные для авторизации текстов корпоративной переписки (см. табл. 6).

Важным наблюдением при анализе этого блока стал вывод о том, что наиболее значимой метрикой для определения авторства текста корпоративной переписки по статистическим данным холистической модели ЯЛ является *t*-критерий Стьюдента и его значение. Анализировать корпоративные письма следует на данном этапе развития методики массивами не менее 500 слов. Ограничение в 100 слов, выведенное еще С.М. Вулом и принятое до сих пор в судебном автороведении [Рубцова и др., 2007] как объем, необходимый для определения авторства текстов, при встраивании в анализ статистической информации должно быть увеличено.

Таблица 5. **Математические модели ЯЛ. Мария и Петр**

Table 5. **Mathematical models of Maria and Peter language personalities**

№	Параметр	Мария	Петр
1	Индекс удобочитаемости Флеша-Кинкейда	**17.2301**	**23.8627**
2	Индекс туманности Ганнинга	21.7719	27.4309
3	Средняя длина слова (в буквах)	6.0395	7.0715
4	Средняя длина предложения (в словах)	**7.1235**	**12.75**
5	Количество предложений длиннее 8 слов	**271604.9383**	**634146.3415**
6	Коэффициент предметности (Pr)	2.2214	2.234
7	Коэффициент качественности (Qu)	0.2905	0.2667
8	Коэффициент активности (Ac)	0.1144	0.1196
9	Коэффициент динамизма (Din)	0.1823	0.1888
10	Коэффициент связности текста (Con)	**1.358**	**2.5793**
11	Количество слов несловарного написания	**12131.7158**	**6695.3611**
12	Предложения с однородными рядами	**0**	**1434.7202**
13	Предложения с обособленными приложениями	**0**	**956.4802**
14	Вводные слова и конструкции	**1733.1023**	**956.4802**
15	Целевые и выделительные обороты	0	0
16	Конструкции с семантикой сравнения	0	0
17	Синтаксические сращения	0	0
18	Сравнительные придаточные	**0**	**956.4802**
19	Конструкции с сопоставительными союзами	**3466.2045**	**956.4802**
20	Вставные конструкции	**12131.7158**	**20564.3233**
21	Сложные синтаксические конструкции	**24263.4315**	**11956.0019**
22	Глагольные односоставные предложения	**13864.818**	**8608.3214**
23	Обращения	**19064.1248**	**6695.3611**
24	Местоимения «я, мы»-группы	**10398.6135**	**9086.5615**
25	Местоимения «ты, вы»-группы	**24263.4315**	**17216.6428**
26	Сложные слова полуслитного написания	**0**	**1434.7202**
27	Модальные частицы	**3466.2045**	**1912.9603**
28	Междометия	**0**	**1912.9603**
29	Наличие / отсутствие модального постфикса «-то»	0	0

Таблица 6. **Рейтерская таблица наиболее частых случаев распределения значений метрик в корпоративной переписке**

Table 6. **Score table for the most common metrics values in business e-correspondence**

ККП	КД линейной регрессии	*t*-критерий Стьюдента (*p*-value)	Автором сравниваемых текстов, вероятно, является одно лицо	Авторами сравниваемых текстов, вероятно, не является одно лицо	Комментарий
Достигает 1,00	Достигает 1,00	Не ниже 0,89	+	–	Для коротких текстов важно высокое значение (обычно достигающее именно 1,00) ККП и КД. При этом *p*-value *t*-критерия Стьюдента тоже не должно быть низким
Может быть около 0,95; иногда даже достигать 1,00	Может быть около 0,90; иногда даже достигать 1,00	Низкие значения, около 0,50–0,51 (редко, но может достигать 0,63)	–	+	Иногда значения ККП и КД могут быть высокими (даже достигать 1,00), но в сочетании с низким значением *p*-value *t*-критерия Стьюдента эти значения не позволяют сделать вывод о едином авторстве сравниваемых документов

Комментарии на развлекательном ресурсе

В рамках исследования был проведен анализ коротких (до 200 слов) «постов» и комментариев, размещенных на развлекательном портале «ЯПлакалЪ».

Результаты анализа показали, что для коротких текстов (до 500 слов), стилостатистический и интерпретативный блоки алгоритма не позволяют создать модели ЯЛ, которые необходимы для получения валидных атрибуционных результатов. Ограничение в 100 слов для коротких комментариев также должно быть значительно увеличено.

Наличие данного ограничения по объему не значит, что тексты сетевых комментариев не могут быть обследованы с помощью предлагаемого алгоритма. Так, задача идентификации автора по коротким сообщениям (постам) в сети Интернет может быть решена посредством создания репрезентативной выборки объемом не менее 500 слов, то есть удобной, которая создана как по стохастическому алгоритму, так и с помощью некоторых правил.

В экспертной и исследовательской практике решение задачи авторизации коротких сообщений стоит очень остро, путь использования совокупности текстов короткого объема как единого массива применяется как в судебной авторoведческой практике [Хоменко, 2019], так и в практике фоноскопических экспертиз [Хоменко и др., 2014]). Это послужило основанием для того, чтобы делать репрезентативные выборки из полученных совокупностей текстов.

Так, в рамках настоящего эксперимента были созданы текстовые выборки для ряда авторов (вместо имен и фамилий приводятся никнеймы, используемые авторами на ресурсе «ЯПлакалЪ») объемом от 500 до 550 слов. Сравнение языковых личностей авторов и их индивидуальных стилей проводилось на основе этих выборок.

При сопоставлении текстов одного автора-мужчины (SESHOK (519 слов) – SESHOK (516 слов)) корреляционный и детерминационный коэффициенты равны единице при *t*-статистике 0,98. Тем не менее при исследовании моделей и их числовых характеристик установлено, что репрезентация идиостиля автора неодинакова, числовые значения существенно отличаются, а при сравнении моделей двух ЯЛ наблюдается большое количество «выбросов» (см. табл. 7).

При сопоставлении текстов одного автора-женщины (KalinAKalina (542 слова) – KalinAKalina (543 слова)) наблюдается весьма сходная картина: корреляционный и детерминационный коэффициенты равны единице при *t*-статистике 0,90 и наличие сильных «выбросов» при сравнении моделей.

Сопоставление текстовых пар SESHOK – SESHOK, KalinAKalina – KalinAKalina только по интерпретативным параметрам (без стилеметрических) значения метрик резко снижаются. Такой уровень заставляет отвергать гипотезу H_0, что, естественно, невалидно.

При сопоставлении текстовой пары SESHOK (519 слов) – OBrian (546 слов) (тексты разных авторов-мужчин) с использованием параметров стилостатистического и интерпретативного блоков результаты корреляции и детерминации получаются не ниже, чем при сравнении текстовых пар одного автора – ККП: 1, КД: 1, *p*-value: 0,96.

Если исключить стилостатистику из конечных моделей, то значения метрик резко снижаются – ККП: 0,78, КД: 0,61, *p*-value: 0,73. Тем не менее проведенные ранее эксперименты показали, что то же самое происходит и при сравнении текстов одного автора, значит, данные результаты непоказательны.

При сопоставлении текстов разных авторов-женщин и текстов авторов разной гендерной принадлежности:

– KalinAKalina – motya (автор-женщина) (со стилеметрией – 1:1:0,97; только интерпретативная база – 0,81:0,66:0,98);

– SESHOK – KalinAKalina (со стилеметрией – 1:1:0,98; только интерпретативная база – 0,9:0,81:0,77) – ситуация не меняется.

Объяснение этого можно предложить в результате анализа модели без стилостатистических данных (см. табл. 8).

Таблица 7. **Математические модели ЯЛ SESHOK**

Table 7. **Mathematical models of SESHOK language personality**

№	Параметр	SESHOK	SESHOK
1	Индекс удобочитаемости Флеша-Кинкейда	20.2671	27.7102
2	Индекс туманности Ганнинга	21.9434	29.8649
3	Средняя длина слова (в буквах)	5.88	6.5984
4	Средняя длина предложения (в словах)	20.12	26.8421
5	Количество предложений длиннее 8 слов	880000	947368.4211
6	Коэффициент предметности (Pr)	1.4809	1.5481
7	Коэффициент качественности (Qu)	0.3598	0.4333
8	Коэффициент активности (Ac)	0.1451	0.1373
9	Коэффициент динамизма (Din)	0.2786	0.2405
10	Коэффициент связности текста (Con)	4	5.1053
11	Количество слов несловарного написания	37773.3598	1960.7843
12	Предложения с однородными рядами	0	0
13	Предложения с обособленными приложениями	0	0
14	Вводные слова и конструкции	5964.2147	0
15	Целевые и выделительные обороты	1988.0716	0
16	Конструкции с семантикой сравнения	0	7843.1373
17	Синтаксические сращения	0	0
18	Сравнительные придаточные	3976.1431	11764.7059
19	Конструкции с сопоставительными союзами	1988.0716	1960.7843
20	Вставные конструкции	0	9803.9216
21	Сложные синтаксические конструкции	29821.0736	31372.549
22	Глагольные односоставные предложения	0	1960.7843
23	Обращения	0	1960.7843
24	Местоимения «я, мы»-группы	13916.501	7843.1373
25	Местоимения «ты, вы»-группы	19880.7157	23529.4118
26	Сложные слова полуслитного написания	3976.1431	0
27	Модальные частицы	25844.9304	17647.0588
28	Междометия	0	0
29	Наличие / отсутствие модального постфикса «-то»	0	0

Таблица 8. **Интерпретативный пул параметров анализа ЯЛ. KalinAKalina и motya**

Table 8. **Qualifying parameters set for the analysis of KalinAKalina and motya language personalities**

№	Параметр	KalinAKalina	motya
1	Количество слов несловарного написания	25291.8288	13487.4759
2	Предложения с однородными рядами	0	0
3	Предложения с обособленными приложениями	0	0
4	Вводные слова и конструкции	0	0
5	Целевые и выделительные обороты	0	1926.7823
6	Конструкции с семантикой сравнения	0	1926.7823
7	Синтаксические сращения	0	0
8	Сравнительные придаточные	11673.1518	7707.1291
9	Конструкции с сопоставительными союзами	0	1926.7823
10	Вставные конструкции	5836.5759	3853.5645
11	Сложные синтаксические конструкции	35019.4553	40462.4277
12	Глагольные односоставные предложения	0	0
13	Обращения	0	0
14	Местоимения «я, мы»-группы	17509.7276	0
15	Местоимения «ты, вы»-группы	9727.6265	23121.3873
16	Сложные слова полуслитного написания	0	1926.7823
17	Модальные частицы	19455.2529	5780.3468
18	Междометия	0	0
19	Наличие / отсутствие модального постфикса «-то»	0	0

Таблица демонстрирует, что проблема заключается не только в «выбросах», но и в том, что резко снижается количество параметров для анализа: многие из них не представлены в обоих текстах. Безусловно, это отражение стилистических особенностей дискурса комментариев из сети Интернет: в таких текстах нет развернутого синтаксиса с осложненными предложениями, синтаксических сращений и пр. В данном виде коммуникации используются другие средства текстовой выразительности, которые пока не попадают в нашу параметризованную модель. Таким образом, в моделях нет объектов для сравнения по предлагаемому алгоритму. Однако это не означает, что интегративная методика, основанная на сочетании качественных, преимущественно интерпретативных, и количественных методов анализа нерелевантна для коротких электронных текстов. Данный подход может быть использован для определения авторства текста электронной коммуникации малого объема. Тем не менее для правильной работы алгоритма он требует дополнений в аспекте особой параметризации, добавления к параметрам, например, литуратива, графической гибридизации, архаичных аффиксов, использования элементов текста, написанных заглавными буквами, эмотиконов и прочих графических символов, обозначающих эмоциональность речи.

Приведенные выше параметры являются особенностями так называемого дигитального почерка. Именно эти особенности в современной электронной переписке дают большой прирост информации для определения авторства. Цифровая языковая личность, естественно, имеет интегративную основу: когнитивные схемы, формирующие речевые навыки, соседствуют в ней с вероятностями самого языка и последовательностью их реализации в речи специфического дискурса. Именно поэтому схема анализа, основанная на интеграции исследования конкретных речевых проявлений в рамках языковых вероятностей, может быть применена для определения авторства коротких электронных текстов при должной параметризации структуры языковой личности. Тем не менее для коротких текстов продуктивен и идиосинкратический подход, основанный не на холистическом методе, а на методе поиска частных признаков идиостиля автора.

Заключение

Атрибуция автора на материале текстов сетевой коммуникации – задача, требующая вдумчивого использования автороведческих методик для разных дискурсивных сфер. Применение интегративной методики анализа идиостиля автора письменного текста как экспликатора ЯЛ релевантно для текстов среднего и большого объема. По отношению к таким текстам можно использовать холистический подход, связанный с анализом ЯЛ в максимальной совокупности репрезентации ее компетенций, поскольку последние находят свои реализации на значительных текстовых объемах. На малых текстовых объемах пока более релевантен идиосинкратический подход к исследованию ЯЛ и ее идентификации.

Важные сведения дает изучение текстовой статистики. Так, в работе приведены рейтерские таблицы, которые могут быть использованы как основа для идентификационного исследования ЯЛ с применением ее математического моделирования. Для разных дискурсивных сфер более или менее релевантными являются следующие метрики анализа:

– для жанра современной беллетристики неинформативным является стилостатистический пул, поскольку значения стилостатистических параметров в этой коллекции близки для всех обследованных текстов;

– для корпоративной переписки и комментариев в сети Интернет необходима репрезентативная выборка из совокупности текстов объемом не менее 500 слов. Для улучшения работы алгоритма на данном материале в настоящий момент разрабатываются дополнительные параметры для построения моделей идиостиля;

– для корпоративной переписки наиболее значимым в сравнении с прочими является анализ значений *t*-статистики Стьюдента, именно она дает наиболее валидные результаты.

Представляется важным отметить, что анализ статистики призван лишь помочь эксперту сделать некоторые первичные выводы

о природе идиостилей сравниваемых текстов. Полный анализ ЯЛ с целью ее идентификации следует проводить в рамках целостного подхода с помощью автобиографического, социолингвистического и юрислингвистического методов.

ПРИМЕЧАНИЕ

[1] Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-312-90022 «Лингвистическое моделирование как инструмент атрибуции текста».

The reported study was funded by RFBR, project number 19-312-90022 "Linguistic Modeling as a Technique in Authorship Attribution".

СПИСОК ЛИТЕРАТУРЫ

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

Бондаренко С. В., 2004. Социальная структура виртуальных сетевых сообществ : дис. ... д-ра социол. наук. Ростов н/Д. 399 с.

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

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

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

Горошко Е. С., 2012. Современная интернет-коммуникация: структура и основные параметры // Интернет-коммуникация как новая речевая формация. М. : Наука : Флинта. С. 9–52.

Ионова С. В., Огорелков И. В., 2020. Речевая диагностика личности по гендерному признаку в автороведении: квантитативный подход // Вестник Волгоградского государственного университета. Серия 2, Языкознание. Т. 19, № 1. С. 115–127. DOI: https://doi.org/10.15688/jvolsu2.2020.1.10

Калугина Е. Н., 2013. Гендер в антропоориентированных науках // Актуальные проблемы гуманитарных и естественных наук. № 5. С. 251–254.

Карасик В. И., 2004. Языковой круг: личность, концепты, дискурс. М. : ГНОЗИС. 389 с.

Караулов Ю. Н., 2010. Русский язык и языковая личность. М. : ЛКИ. 264 с.

Падучева Е. В., 1974. О семантике синтаксиса. М. : Наука. 291 с.

Резанова З. И., Романов А. С., Мещеряков Р. В., 2013. Задачи авторской атрибуции текста в аспекте гендерной принадлежности (к проблеме междисциплинарного взаимодействия лингвистики и информатики) // Вестник Томского государственного университета. № 370. С. 24–28.

Розина И. Н., 2005. Педагогическая компьютерно-опосредованная коммуникация: теория и практика. М. : Логос. 437 с.

Романова Т. В., 2011. Человек и время: Язык. Дискурс. Языковая личность. Н. Новгород : Нижегор. гос. лингвист. ун-т им. Н.А. Добролюбова. 310 с.

Рубцова И. И., Ермолова Е. И., Безрукова А. И., Огорелков И. В., Захаров М. П., 2007. Комплексная методика производства автороведческих экспертиз : метод. рекомендации. М. : ЭКУ МВД России. 192 с.

Русская грамматика : науч. тр. : в 2 т. URL: http://rusgram.narod.ru/index.html

Хоменко А. Ю., 2019. Лингвистическое атрибуционное исследование коротких письменных текстов: качественные и количественные методы // Политическая лингвистика. № 2 (74). С. 177–187. DOI: 10.26170/pl19-02-20

Хоменко А. Ю., Зиновьев Д. Е., Цыганов А. А., Калинина В. В., 2014. Многообъектные исследования в судебной фоноскопической экспертизе // Язык. Право. Общество : сб. ст. II Междунар. науч.-практ. конф. Пенза : Изд-во ПГУ. С. 404–417.

Bloch B., 1948. A Set of Postulates for Phonemic Analysis // Language. № 24 (1). P. 3–46.

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

Friginal E., Hardy J., 2014. Corpus-Based Sociolinguistics: A Guide for Students. L. : Taylor & Francis. 167 p.

Khomenko A., Baranova Y., Romanov A., Zadvornov K., 2021. Linguistic Modeling as a Basis for Creating Authorship Attribution Software // Computational Linguistics and Intellectual Technologies. Iss. 20 (27). P. 1063–1074.

Litvinova T., Sboev A., Panicheva P., 2018. Profiling the Age of Russian Bloggers // Artificial Intelligence and Natural Language : Proceedings of the 7th International Conference (Saint Petersburg, October 17–19, 2018). Saint Petersburg : Springer. P. 167–177.

McMenamin G.R., 2002. Forensic Linguistics: Advances in Forensic Stylistics. Boca Raton : CRC Press LLC. 331 p.

Shuy R. W., 2005. Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language. N. Y. : Oxford University Press. 194 p.

Wright D., 2017. Using Word N-Grams to Identify Authors and Idiolects: A Corpus Approach to a Forensic Linguistic Problem // International Journal of Corpus Linguistics. № 22 (2). P. 212–241. URL: https://benjamins.com/#catalog/journals/ijcl.22.2.03wri/details

ИСТОЧНИКИ

Атрибуционное программное обеспечение «ХоРом». 2022. URL: http://khorom-attribution.ru/#/

Книга фанфиков. 2022. URL: https://ficbook.net/

Развлекательный портал «ЯПлакалЪ». 2022. URL: https://www.yaplakal.com/

Электронный репозиторий. 2022. URL: https://github.com/KhomenkoAnna/attr_lingvo

REFERENCES

Belyanin V.P., 2000. *Osnovy psiholingvisticheskoy diagnostiki: modeli mira v literature* [Fundamentals of Psycholinguistic Diagnostics: Models of the World in Literature]. Moscow, Trivola Publ. 247 p.

Bondarenko S.V., 2004. *Socialnaya struktura virtualnykh setevykh soobshchestv: dis. ... d-ra sotsiol. nauk* [Social Structure of Virtual Network Communities. Dr. soc. sci. diss.]. Rostov-on-Don. 399 p.

Vinogradov V.V., 1961. *Problema avtorstva i teoriya stiley* [The Problem of Authorship and the Theory of Styles]. Moscow, Goslitizdat Publ. 614 p.

Vul S.M. 2007. *Sudebno-avtorovedcheskaya identifikatsionnaya ekspertiza: metodicheskie osnovy* [Forensic Identification Expertise: Methodological Foundations]. Kharkiv, KhNIISE. 64 p.

Galyashina E.I. 2003. *Osnovy sudebnogo rechevedeniya* [Fundamentals of Judicial Speech Science]. Moscow, STENSI Publ. 236 p.

Goroshko E.S., 2012. Sovremennaya internet-kommunikatsiya: struktura i osnovnye parametry [Modern Internet Communication: Structure and Main Parameters]. *Internet-kommunikaciya kak novaya rechevaya formatsiya* [Internet Communication As a New Speech Formation]. Moscow, Nauka Publ., Flinta Publ., pp. 9-52.

Ionova S.V., Ogorelkov I.V., 2020. Rechevaya diagnostika lichnosti po gendernomu priznaku v avtorovedenii: kvantitativnyy podkhod [Personality Speech Diagnostics in Author Identification Based on Gender Parameter: Quantitative Approach]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], vol. 19, no. 1, pp. 115-127. DOI: https://doi.org/10.15688/jvolsu2.2020.1.10

Kalugina E.N., 2013. Gender v antropoorientirovannykh naukakh [Gender in Anthropocentric Sciences]. *Aktualnye problemy gumanitarnykh i estestvennykh nauk* [Actual Problems of the Humanities and Natural Sciences], no. 5, pp. 251-254.

Karasik V.I., 2004. *Yazykovoy krug: lichnost, kontsepty, diskurs* [Language Circle: Personality, Concepts, Discourse]. Moscow, GNOSIS Publ. 389 p.

Karaulov Yu.N. 2010. *Russkiy yazyk i yazykovaya lichnost* [Russian Language and Linguistic Personality]. Moscow, LKI. 264 p.

Paducheva E.V., 1974. *O semantike sintaksisa* [On the Semantics of Syntax]. Moscow, Nauka Publ. 291 p.

Rezanova Z.I., Romanov A.S., Meshcheryakov R.V., 2013. Zadachi avtorskoy atributsii teksta v aspekte gendernoy prinadlezhnosti (k probleme mezhdisciplinarnogo vzaimodeystviya lingvistiki i informatiki) [Tasks of the Author's Attribution of the Text in Terms of Gender (To the Problem of Interdisciplinary Interaction of Linguistics and Informatics)]. *Vestnik Tomskogo gosudarstvennogo universiteta* [Tomsk State University Journal], no. 370, pp. 24-28.

Rozina I.N., 2005. *Pedagogicheskaya kompyuterno-oposredovannaya kommunikatsiya: teoriya i praktika* [Pedagogical Computer-Mediated Communication: Theory and Practice]. Moscow, Logos Publ. 437 p.

Romanova T.V., 2011. *Chelovek i vremya: Yazyk. Diskurs. Yazykovaya lichnost* [Man and Time: Language. Discourse. Linguistic personality]. Nizhny Novgorod, Nizhegor. gos. lingvist. un-t im. N.A. Dobrolyubova. 310 p.

Rubcova I.I., Ermolaeva E.I., Bezrukova A.I., Ogorelkov I.V., Zakharov M.P., 2007. *Kompleksnaya metodika proizvodstva avtorovedcheskikh ekspertiz: metod. rekomendatsii* [Comprehensive Methodology for the Production of Author's Examinations. Methodological Recommendations]. Moscow, EKU MVD Rossii. 192 p.

Russkaya grammatika. V 2 t.: nauch. tr. [Russian Grammar. In 2 Vols. Scientific Papers]. URL: http://rusgram.narod.ru/index.html

Khomenko A.Yu., 2019. Lingvisticheskoe atributsionnoe issledovanie korotkikh pismennykh tekstov: kachestvennye i kolichestvennye metody [Linguistic Attributional Examination of Short Written Texts: Qualitative and Quantitative Methods]. *Politicheskaya lingvistika* [Political

Linguistics], no. 2 (74), pp. 177-187. DOI: 10.26170/pl19-02-20

Khomenko A.Yu., Zinovev D.E., Tsyganov A.A., Kalinina V.V., 2014. Mnogoobyektnye issledovaniya v sudebnoy fonoskopicheskoy ekspertize [Multi-Object Research in Forensic Phonoscopic Examination]. *Yazyk. Pravo. Obshchestvo: sb. st. II Mezhdunar. nauch.-prakt. konf.* [Language. Law. Society. Collection of Articles of the 2nd International Scientific and Practical Conference]. Penza, Izd-vo PGU, pp. 404-417.

Bloch B., 1948. A Set of Postulates for Phonemic Analysis. *Language*, no. 24 (1), pp. 3-46.

Coulthard M., 2004. Author Identification, Idiolect, and Linguistic Uniqueness. *Applied Linguistics*, no. 24 (4), pp. 431-447.

Friginal E., Hardy J., 2014. *Corpus-Based Sociolinguistics: A Guide for Students*. London, Taylor & Francis. 167 p.

Khomenko A., Baranova Y., Romanov A., Zadvornov K., 2021. Linguistic Modeling as a Basis for Creating Authorship Attribution Software. *Computational Linguistics and Intellectual Technologies*, iss. 20 (27), pp. 1063-1074.

Litvinova T., Sboev A., Panicheva P., 2018. Profiling the Age of Russian Bloggers. *Artificial Intelligence and Natural Language: Proceedings of the 7th International Conference (Saint Petersburg, October 17–19, 2018)*. Saint Petersburg, Springer, pp. 167-177.

McMenamin G.R., 2002. *Forensic Linguistics: Advances in Forensic Stylistics*. Boca Raton, CRC Press LLC. 331 p.

Shuy R.W., 2005. *Creating Language Crimes: How Law Enforcement Uses (and Misuses) Language*. New York, Oxford University Press. 194 p.

Wright D., 2017. Using Word N-Grams to Identify Authors and Idiolects: A Corpus Approach to a Forensic Linguistic Problem. *International Journal of Corpus Linguistics*, no. 22 (2), pp. 212-241. URL: https://benjamins.com/#catalog/journals/ijcl.22.2.03wri/details

SOURCES

Atributsionnoye programmnoye obespecheniye «KhoRom»[Attribution Software "KhoRom"]. 2022. URL: http://khorom-attribution.ru/#/

Kniga fanfikov [Ficbook]. 2022. URL: https://ficbook.net/

Razvlekatel'nyy portal «YAplakal» [Entertainment Portal "Yaplakal"]. 2022. URL: https://www.yaplakal.com/

Elektronnyy repozitoriy [Electronic Repository]. 2022. URL: https://github.com/KhomenkoAnna/attr_lingvo

Information About the Authors

Tatyana V. Romanova, Doctor of Sciences (Philology), Professor, Department of Applied Linguistics and Foreign Languages, National Research University Higher School of Economics – Nizhny Novgorod, Bolshaya Pecherskaya St, 25/12, 603155 Nizhny Novgorod, Russia, tvromanova@hse.ru, https://orcid.org/0000-0002-1833-2711

Anna Yu. Khomenko, Candidate of Sciences (Philology), Senior Lecturer, Department of Applied Linguistics and Foreign Languages, National Research University Higher School of Economics – Nizhny Novgorod, Bolshaya Pecherskaya St, 25/12, 603155 Nizhny Novgorod, Russia, akhomenko@hse.ru, https://orcid.org/0000-0003-3564-6293

Информация об авторах

Татьяна Владимировна Романова, доктор филологических наук, профессор департамента прикладной лингвистики и иностранных языков, Национальный исследовательский университет «Высшая школа экономики» – Нижний Новгород, ул. Большая Печерская, 25/12, 603155 г. Нижний Новгород, Россия, tvromanova@hse.ru, https://orcid.org/0000-0002-1833-2711

Анна Юрьевна Хоменко, кандидат филологических наук, старший преподаватель департамента прикладной лингвистики и иностранных языков, Национальный исследовательский университет «Высшая школа экономики» – Нижний Новгород, ул. Большая Печерская, 25/12, 603155 г. Нижний Новгород, Россия, akhomenko@hse.ru, https://orcid.org/0000-0003-3564-6293