УДК 811.161.1'42
ББК Ш141.12-51 *ГСНТИ 16.21.21* *Код ВАК 10.02.19*
DOI 10.26170/pl20-03-22

А. Ю. Хоменко
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: 0000-0003-3564-6293 ☑

Е. Р. Бенькович
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

Д. И. Гайнутдинова
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

Л. Р. Гасанова
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

А. А. Костина
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

З. О. Мазунина
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

А. С. Николаева
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

Е. В. Пимонова
Национальный исследовательский университет «Высшая школа экономики», Нижний Новгород, Россия
ORCID ID: — ☑

☑ ***E-mail:*** *akhomenko@hse.ru; erbenkovich@edu.hse.ru; digaynutdinova@edu.hse.ru; lrgasanova@edu.hse.ru; aakostina_1@edu.hse.ru; zomazunina@edu.hse.ru; asnikolaeva_1@edu.hse.ru; evpimonova_1@edu.hse.ru.*

Автоматическая обработка текста и лингвистическое моделирование как способы решения проблем атрибуционной лингвистики

АННОТАЦИЯ. *В настоящей работе речь пойдет об апробации интегративной методики атрибуционного анализа текста на русском языке, основанной на соединении результатов интерпретативного исследования материала и объективации этих результатов посредством математической статистики. Исследование построено по следующему алгоритму: 1) автоматическое извлечение из текста параметров, описывающих идиостиль с точки зрения прагматикона, тезауруса и лексикона автора; 2) поиск традиционных стиметрических текстовых данных; 3) присвоение веса каждому параметру; 4) построение математических моделей сравниваемых текстов; 5) сравнение математических моделей с целью выявления уровня их корреляции между собой. Поиск параметров, описывающих модель авторского идиостиля, ведется на основании подхода к тексту как к продукту деятельности конкретной языковой личности. Языковая личность автора описывается с позиции подхода Ю. Н. Караулова. Автоматическое извлечение предустановленных параметров осуществляется с помощью алгоритмов, сконструированных на ЯП Python. Для апробации алгоритма использованы тексты нежанровой художественной прозы разной тематики с заведомо известным авторством: «Наши» С. Д. Довлатова и «Обертон» В. П. Астафьева. Исследованием доказана работоспособность разработанной методики.*

КЛЮЧЕВЫЕ СЛОВА: *текстовая атрибуция; языковая личность; автоматическая обработка текста; математические модели; русский язык.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Хоменко Анна Юрьевна, старший преподаватель, департамент прикладной лингвистики и иностранных языков, стажер-исследователь, лаборатория теории и практики систем поддержки принятия решений, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; эксперт-лингвист, эксперт-авторовед, эксперт-фоноскопист, Центр экспертизы и оценки «ЕСИН», г. Нижний Новгород, г. Москва; e-mail: akhomenko@hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Бенькович Елена Романовна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: erbenkovich@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Гайнутдинова Диана Ильдаровна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: digaynutdinova@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Гасанова Лейла Рафиг кызы, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: lrgasanova@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Костина Алина Анатольевна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: aakostina_1@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Мазунина Зоя Олеговна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: zomazunina@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Николаева Ангелина Сергеевна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: asnikolaeva_1@edu.hse.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Пимонова Елена Владимировна, студентка факультета гуманитарных наук, Национальный исследовательский университет «Высшая школа экономики»; 603155, Россия, г. Нижний Новгород, ул. Б. Печерская д.25/12, каб. 310; e-mail: evpimonova_1@edu.hse.ru.*

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

БЛАГОДАРНОСТИ. Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-31-27001 (19-312-90022).

Проблема атрибуции текста в современной лингвистике становится все более актуальной. Атрибуционная лингвистика со времен Л. Кэмпбелла [Campbell 1867] и В. Лютославского [Lutoslawski 1897] на Западе и Н. А. Морозова [Морозов 1916] в России всегда шла двумя параллельными путями: путем стилеметрии [Mendenhall 1887; Mosteller, Wallace 1964; Захаров 2000; Merriam 2003; Labbe, Labbe 2001; Juola, Sofko, Brennan 2006; Мартыненко 2015; Litvinova, Seredin, Litvinova, etc., 2017; Wright 2017; Karlgren, Esposito, Gratton, etc. 2018 и пр.] и путем качественного анализа текста [Вул 1973, 2007; Горошко 2003; Комиссаров 2001; McMenamin 2002; Галяшина 2003; Coulthard 2004 и пр.]. На современном этапе развития исследовательского поля две эти ветви интегрируются неохотно, а если и интегрируются, то их координация происходит посредством объяснения стилеметрических данных с точки зрения традиционной квалификативной лингвистики: объяснение длины предложения как отражения уровня компетенций автора в письменной речи [Степаненко 2017: 19—20], объяснение n-грамм как косвенной экспликации грамматических текстовых реалий [Захаров, Хохлова 2008: 41—42]. Безусловно, этот путь продуктивен, а все вероятности конкретного текста являются одновременно и вероятностями языка как системы идиолектов. Тем не менее данный путь всегда будет ограничен невозможностью создать достаточно полную модель языковой личности автора: анализ одного или нескольких аспектов идиостиля едва ли может репрезентировать языковую личность в полном объеме и всесторонне. Представляется логичным путь исследования глубинных синтаксических структур как базы для сравнения моделей индивидуальных авторских стилей. Разработкой данного направления занимается санкт-петербуржская школа прикладной лингвистики [Марусенко 1990; Родионова 2008 и пр.]. Данный подход, безусловно, работоспособен, но его реализация возможна, с одной стороны, только на объемных текстах, с другой — она очень трудоемка и сложна в техническом отношении. Более просто реализуемым выглядит подход, основанный на интеграции анализа традиционных стилостатистических параметров (длин слов и предложений, наиболее частотных n-грамм, служебных слов и POS-tags) и анализа авторских идиосинкразем, в основном ошибок разного рода, предложенный, например, М. Коппелом и Дж. Шлером [Koppel, Schler 2003].

На современном этапе существует острая необходимость интеграции качественного и количественного анализа в атрибуции. Становится все более очевидным, что построение моделей авторских идиостилей лишь на основании традиционных стилостатистических данных не может в полной мере удовлетворить атрибуционную лингвистику, в особенности лингвистику судебную. Для судебного авторoведения в соответствии с законодательством РФ использование только статистических методов анализа вообще является недопустимым в силу отсутствия у эксперта-автороведа глубоких познаний в области математической статистики, теории вероятности, big data [Приказ от 27 декабря 2012 года N 237; Федеральный закон от 31 мая 2001 г. N 73-ФЗ] и методических основ су-

дебного авторoведения [Рубцова, Ермолаева, Безрукова и др. 2007].

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

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

Алгоритм анализа начинается с того, что группа экспертов определяет на основании анализа теоретического материала ряд параметров языковой личности, которые заведомо в той или иной степени должны идентифицировать авторский идиостиль и одновременно могут быть извлечены из текста автоматически с минимальным предпроцессингом. Речь идет о том, что данные параметры должны быть относительно универсальны для любого текста и их должно быть легко извлекать, используя некоторые предустановленные правила и минимальную текстовую обработку, осуществляемую не вручную экспертом (ручная разметка, выравнивание текстов и пр.), а автоматически (токенизация, присвоение pos-tags). Итак, приведенным выше условиям удовлетворили следующие параметры:

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

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

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

Обработка текстов осуществлялась при помощи ЯП Python. На этапе предпроцессинга тексты разделяются на предложения с помощью стандартной библиотеки NLTK с уточнением использования русской модели для обработки текстов, тексты подвергаются токенизации, словам текста присваиваются частеречные теги с грамматическими характеристиками с помощью Pymorthy2.

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

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

Этот же алгоритм поиска осуществляется при отборе параметров, имеющих морфологическую отнесенность, например, модального постфикса «-то»: POST-то, кроме

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

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

Ключевые лексемы определяются с помощью алгоритма логарифмического правдоподобия при сравнении интересующего текста с референтным корпусом большого объема (использовался корпус «Opencorpora», URL: *http://opencorpora.org*, дата обращения: 08.02.2020, объемом на дату обращения 1 540 034 слова). В результате для каждого текста получаем список ключевых слов с числовой экспликацией значения меры логарифмического правдоподобия (log-likelihood score, или LL). В конечный список включаются лишь слова со значением LL более 50.

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

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

Под словом-интенсификатором подразумевается лексема, используемая для определения степени семантической категории интенсивности. Чаще всего говорят о наречиях-интенсификаторах, круг их хоть и велик, но ограничен (*очень*, *сильно*, *адски* — из современного дискурса). Тем не менее категория интенсивности не исчерпывается исключительно наречным наполнением, например: *Какая красота!* — в данном случае интенсификатором служит местоимение *какая*. Так, в исследовании был создан свод правил для поиска структур с интенсификаторами; в список интенсификаторов входят как наречия с некоторыми грамматическими ограничениями (авторы не осуществляют поиск структур, где наречие не эксплицирует категорию интенсивности, например, является частью составного именного сказуемого: *Он чувствует себя хорошо*), так и некоторые прилагательные и местоимения в соответствующих грамматических конструкциях, как то: ADJ «настоящий» в nomn, accs в sing или plur + NOUN: *настоящий бардак.* Метрикой для каждого найденного слова в конечной модели служит ipm. Для поиска интенсификаторов в данном случае предпроцессинг осуществлялся модулем для токенизации Razdel (URL: *https://github.com/natasha/razdel*, дата обращения:10.02.2010), использующим правила. При тестировании он показал лучшие результаты, чем инструменты токенизации NLTK.

Для каждого текста исследователи также определили ряд традиционных стилеметрических параметров: среднюю длину слова, среднюю длину предложения и количество предложений объемом более 8 слов, т. е. длинных предложений.

Далее все полученные данные сводятся в две математические модели, которые сравниваются между собой посредством коэффициента корреляции Пирсона, доказывающего или опровергающего гипотезу H_0 о том, что автором двух сравниваемых текстов является одно лицо. Эти математические модели в некотором объеме описывают авторские индивидуальные стили, посему если стили разные, модели должны иметь статистически значимые различия, которые отражаются на отношениях между параметрами. Релевантность применения коэффициента корреляции Пирсона для сравнения математических моделей авторских идиостилей описана, например, в исследовании [Радбиль, Маркина 2019]. Исходя из экспериментальных данных и теоретического осмысления настоящей метрики, представляется, правда, что коэффициент будет однозначно эффективен для текстов большого объема, не менее 20 000 слов, но неясно его «поведение» на текстах малого объема.

Итак, для проверки работоспособности предложенного выше алгоритма и доказа-

тельства постулата об эффективности интегративной методики текстовой атрибуции авторы проанализировали два текста нежанровой художественной прозы объемом более 20 000 слов с заведомо известным авторством и разной тематической отнесенностью: 1) С. Д. Довлатов — «Наши» (1983 г.), объем — 21 230 слов; 2) В. П. Астафьев — «Обертон» (1996 г.), объем — 26 070 слов. При этом гипотеза H_0 заключается в том, что автором двух текстов является одно лицо; H_1 — авторы двух текстов — разные лица (соответствует действительности). Математические модели идиостилей, реализуемых в текстах, представлены в таблице 1.

Таблица 1*

Математические модели идиостилей С. Д. Довлатова и В. П. Астафьева

Идентификационные параметры	С. Д. Довлатов, «Наши», относительная частота, в единицах измерения	В. П. Астафьев, «Обертон», относительная частота, в единицах измерения
I. Прагматикон личности		
Вводные слова	2186,421	3626,943
Конструкции с «большинство/меньшинство»	76,71653	47,10316
Целевые, выделительные и сравнительные обороты	6379,585	14492,75
Синтаксические сращения	230,1496	0
Сравнительные придаточные	14354,07	13285,02
Глагольные односоставные предложения	122807	824879,2
II. Тезаурус личности		
1) частотные сочетания слов		
весь ещё	306,8661	0
здоровый тело	0	188,4126
тело соответствующий	0	188,4126
соответствующий дух	0	188,4126
именно это	0	188,4126
это учить	0	188,4126
2) ключевые лексемы текста		
я	842,01	1258,28
ты	277,14	334,18
сортировка	149,51	0
девка	130,22	0
хата	110,93	0
начальник	82,24	0
домой	60,33	58,03
мать	58,49	182,96
командир	50,39	0
дядя	0	486,66
дед	0	352
брат	0	330,23
тётка	0	287,39
отец	0	206,95
сказать	0	186,17
мы	0	70,85
пить	0	65,66
думать	0	57,65
водка	0	53,64
любить	0	50,95
3) экспликаты аксиологических текстовых доминант групп «свой/чужой»		
Свой	0,0297	0,0406
Чужой	0,0184	0,0322
Я	0,0211	0,0286
Мы	0,0038	0,0063
Ты	0,0047	0,0057
Он	0,0064	0,0208
Она	0,0078	0,0084
Они	0,0042	0,0029

Окончание таблицы 1

Идентификационные параметры	С. Д. Довлатов, «Наши», относительная частота, в единицах измерения	В. П. Астафьев, «Обертон», относительная частота, в единицах измерения
III. Вербально-семантический уровень авторского лексикона:		
1) частеречная отнесенность слов текста		
Предлоги	119025,7	22609,51
Прилагательные	103644	28026,38
Существительные	301841,2	79886,95
Глаголы	134292,3	43805,93
Местоимения-существительные	64096,66	20301,46
Союзы	97314,92	16721,62
Инфинитив	22132,72	4145,078
Наречие	58688,15	16344,8
Числительные	2685,079	2637,777
2) слова-интенсификаторы		
Больно	115,0748	0
вовсе+не	115,0748	0
Довольно	38,35827	282,6189
Едва	153,4331	94,20631
Изрядно	38,35827	0
Какой	1035,673	188,4126
Настоящий	76,71653	94,20631
Невероятный	38,35827	47,10316
Страшно	76,71653	0
Сущий	38,35827	0
Так	997,3149	94,20631
Абсолютно	0	47,10316
3) иные морфолого-лексические характеристики		
Модальные частицы	8937,476	9185,115
Модальный постфикс «-то»	3759,11	1695,714
Междометия	1074,031	423,9284
Модальные частицы	8937,476	9185,115
Междометия	1074,031	423,9284
Сложные слова полуслитного написания	1572,689	471,0316
IV. Традиционные стилеметрические параметры		
Предложения, превышающие 8 слов	77086,66	1207,729
Средняя длина предложения	13,96	6,49
Средняя длина слова	5,36	5,54

* В статье модель приведена не полностью для сохранения эргономики публикации. В модели элиминирован полный список ключевых слов и интенсификаторов.

Конечные математические модели имеют 137 параметров сравнения. Коэффициент корреляции Пирсона для сравниваемых моделей равен 0,395, что, безусловно, говорит о различном авторстве двух текстов, поскольку, чем ближе коэффициент корреляции к 1, тем более сходны модели, соответственно, авторские стили. По экспериментальным данным, чтобы признать тексты принадлежащими одному автору, коэффициент корреляции должен быть выше 87 % [Радбиль, Маркина 2019: 164].

Итак, следует отметить, что интегративная методика, основанная на использовании подходов интерпретативной и когнитивной лингвистики в совокупности с методами традиционной стилеметрии, безусловно, дает свои результаты. В настоящей статье было продемонстрировано, что интерпретативную часть анализа не обязательно должен делать специалист собственноручно, выделение идентификационных критериев можно автоматизировать, причем важно, что имеется возможность автоматизировать процесс без предварительной ручной обработки текстов и без применения синтаксических парсеров. Эта особенность важна для создания прототипа программного обеспечения, которое можно было бы применять в том числе для решения задач судебной лингвистики, поскольку эксперт-авторовед не всегда обладает необходимыми знаниями в области корпусной лингвистики, статистики и пр. Интеграция всех модулей анализа, описанных выше, в одну программную оболочку сделает возможным автоматизацию некоторых частей, а возможно, и полностью атрибуционного анализа. Важно, что

данный алгоритм показал свою эффективность также и на текстах меньшего объема и иной функционально-стилистической отнесенности: сходный анализ был проведен для текстов корпоративной русскоязычной переписки.

ЛИТЕРАТУРА

1. Вул, С. М. Криминалистическое исследование признаков письменной речи / С. М. Вул. — Киев, 1973. — 44 с. — Текст : непосредственный.

2. Вул, С. М. Судебно-автороведческая идентификационная экспертиза: методические основы : методическое пособие / С. М. Вул. — Харьков : ХНИИСЭ, 2007. — 64 с. — Текст : непосредственный.

3. Галяшина, Е. И. Основы судебного речеведения / Е. И. Галяшина. — Москва, 2003. — 236 с. — Текст : непосредственный.

4. Горошко, Е. И. Судебно-автороведческая классификационная экспертиза: проблемы установления пола автора документа / Е. И. Горошко. — Текст : непосредственный // Теория и практика судебной экспертизы и криминалистики. — Харьков : Право, 2003. — Вып. 3. — С. 221—226.

5. Захаров, В. Н. Программа систем поддержки атрибуции текстов статей Ф. М. Достоевского / В. Н. Захаров ; соавт.: А. А. Леонтьев, А. А. Рогов, Ю. В. Сидоров. — Текст : непосредственный // Труд / ПетрГУ. — Петрозаводск, 2000. — Вып. 9. — С. 113—122. — (Сер. «Прикладная математик и информатика»).

6. Захаров, В. П. Статистический метод выявления коллокаций / В. П. Захаров, М. В. Хохлова. — Текст : непосредственный // Языковая инженерия: в поиске смыслов : доклады семинара «Лингвистические информационные технологии в Интернете» : XI Всероссийская объединенная конференция «Интернет и современное общество». — Санкт-Петербург : Изд-во Санкт-Петербургского университета, 2008. — С. 40—54.

7. Караулов, Ю. Н. Русский язык и языковая личность / Ю. Н. Караулов. — Москва : Наука, 1987. — 264 с. — Текст : непосредственный.

8. Комиссаров, А. Ю. Криминалистическое исследование письменной речи : учеб. пособие / А. Ю. Комиссаров. — Москва : ЭКЦ МВД России, 2000. — 126 с. — Текст : непосредственный.

9. Мартыненко, Г. Я. Стилеметрия: возникновение и становление в контексте междисциплинарного взаимодействия / Г. Я. Мартыненко. — Текст : непосредственный // Структурная и прикладная лингвистика : межвуз. сб. / под ред. А. С. Герда и И. С. Николаева. — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2015. — Вып. 11. — С. 9—28.

10. Марусенко, М. А. Атрибуция анонимных и псевдонимных литературных произведений методами распознавания образов / М. А. Марусенко. — Ленинград : Изд-во Ленингр. ун-та, 1990. — 164 с. — Текст : непосредственный.

11. Морозов, Н. А. Лингвистические спектры: средство для отличения плагиатов от истин. произведений того или др. известного авт. / Н. А. Морозов. — Петроград : Тип. Имп. Акад. наук, 1916. — 42 с. — URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3 (дата обращения: 05.07.2019). — Текст : электронный.

12. 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: 03.07.2019). — Текст : электронный.

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

14. Родионова, Е. С. Методы атрибуции художественных текстов / Е. С. Родионова. — Текст : непосредственный // Структурная и прикладная лингвистика : межвуз. сб. — Санкт-Петербург : Изд-во С.-Петерб. ун-та, 2008. — Вып. 7 / под ред. А. С. Герда. — С. 118—127.

15. Рубцова, И. И. Комплексная методика производства автороведческих экспертиз : методические рекомендации / И. И. Рубцова, Е. И. Ермолаева, А. И. Безрукова и др. — Москва : ЭКУ МВД России, 2007. — 192 с. — Текст : непосредственный.

16. Степаненко, А. А. Гендерная атрибуция текстов компьютерной коммуникации: статистический анализ использования местоимений / А. А. Степаненко. — DOI 10.17223/15617793/415/3. — Текст : непосредственный // Вестник Томского государственного университета. — 2017. — № 415. — С. 17—25.

17. Федеральный закон от 31 мая 2001 г. N 73-ФЗ «О государственной судебно-экспертной деятельности в Российской Федерации» // Российская газета. — 2001. — N 256 от 31 дек. — URL: https://base.garant.ru/12123142/ (дата обращения: 03.07.2019). — Текст : электронный.

18. Campbell, L. The Sophisties and Polilicus of Plato / L. Campbell. — Oxford : Clarendon, 1867. — 170 p. — Text : unmediated.

19. Coulthard, M. Author identification, idiolect, and linguistic uniqueness / M. Coulthard. — Text : unmediated // Applied Linguistics. — 2004. — No 24 (4). — P. 431—447.

20. Juola, P. A Prototype for Authorship Attribution Studies / P. Juola, J. Sofko, P. Brennan. — Text : electronic // Literary and Linguistic Computing. — 2006. — Vol. 21. — Iss. 2. — 1 June. — P. 169—178. — URL: https://doi.org/10.1093/llc/fql0 (date of access: 05.07.2019).

21. Karlgren, J. Authorship Profiling Without Using Topical Information—Notebook for PAN at CLEF, 2018 / J. Karlgren, L. Esposito, Ch. Gratton, P. Kanerva. — URL: https://pdfs.semanticscholar.org/ee57/5920182cdc6de1337f71b07a25e830022459.pdf?_ga=2.139547835.909834531.1562339431-1809262388.1562339431 (date of access: 05.07.2019). — Text : electronic.

22. Koppel, M. Exploiting Stylistic Idiosyncrasies for Authorship Attribution / M. Koppel, J. Schler. — Text : unmediated // Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. — 2003. — No 69. — P. 72—80.

23. Labbe, C. Inter-Textual Distance and Authorship Attribution / C. Labbe, D. Labbe. — Text : unmediated // Corneille and Molière. Journal of Quantitative Linguistics. — Taylor & Francis (Routledge), 2001. — No 8 (3). — P. 213—231.

24. Litvinova, T. Gender identification in Russian written texts / T. Litvinova, P. Seredin, O. Litvinova, O. Zagorovskaya. — Text : electronic // XLinguae. — 2017. — Vol. 10. — Iss. 3. — P. 176—183. — URL: http://xlinguae.eu/files/XLinguae3_2017_14.pdf (date of access: 05.07.2019).

25. Lutoslawski, W. The origin and growth of Plato's logic / W. Lutoslawski. — London, 1997. — 613 p. — Text : unmediated.

26. McMenamin, G. R. Forensic Linguistics: advances in forensic stylistics / G. R. McMenamin. — 2002. — 331 p. — Text : unmediated.

27. Mendenhall, T. The characteristic curves of composition / T. Mendenhall. — Text : unmediated // Science. — 1987. — No 9. — P. 237—249.

28. Merriam, T. An Application of Authorship Attribution by Intertextual Distance in English / T. Merriam. — Text : unmediated // Corpus. — 2003. — N 2. — P. 142—168.

29. Mosteller, F. Applied Bayesian and Classical Inference: The Case of the Federalist Papers / F. Mosteller, D. L. Wallace. — Addison-Wesley, Reading, MA, 1984. — Text : unmediated.

30. Wright, D. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem / D. Wright. — Text : electronic // International Journal of Corpus Linguistics. — 2017. — 22 (2). — P. 212—241. — URL: https://core.ac.uk/download/pdf/84587040.pdf (date of access: 05.07.2019).

Хоменко А. Ю., Бенькович Е. Р., Гайнутдинова Д. И., Гасанова Л. Р., Костина А. А., Мазунина З. О., Николаева А. С., Пимонова Е. В. Политическая лингвистика. 2020. № 3 (81). С. 215—224.

A. Yu. Khomenko
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: 0000-0003-3564-6293 ☑

E. R. Ben'kovich
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

D. I. Gainutdinova
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

L. R. kyzy Gasanova
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

A. A. Kostina
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

Z. O. Mazunina
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

A. S. Nikolaeva
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

E. V. Pimonova
National Research University "Higher School of Economics", Nizhniy Novgorod, Russia
ORCID ID: — ☑

☑ ***E-mail:*** *akhomenko@hse.ru; erbenkovich@edu.hse.ru; digaynutdinova@edu.hse.ru; lrgasanova@edu.hse.ru; aakostina_1@edu.hse.ru; zomazunina@edu.hse.ru; asnikolaeva_1@edu.hse.ru; evpimonova_1@edu.hse.ru.*

Automatic Text Processing and Linguistic Modeling as Instruments for Solving Problems of Text Attribution

ABSTRACT. *This paper focuses on the approbation of an integrative method of attribution text analysis in Russian, based on a combination of the results of an interpretive study of the material and objectification of these results through mathematical statistics. The study has been conducted according to the following algorithm: 1) automatic extraction of text parameters describing the idiostyle from the point of view of the author's pragmaticon, thesaurus, and lexicon; 2) automatic search for traditional stylometric text data (length of sentences, words, etc.); 3) weight assignment to each parameter; 4) creation of mathematical models of compared texts; 5) comparison of mathematical models in order to identify the level of their correlation with each other. The search for parameters describing the authors' individual style is carried out on the basis of the approach to the text as a product of a specific language personality. The author's language personality is described according Yu. N. Karaulov's approach. Automatic extraction of predefined parameters is performed using the algorithms designed in Python. To test the algorithm, texts of non-genre fiction of different themes and obviously known authorship were used: «Nashi» by S.D. Dovlatov and «The Overtone» by V.P. Astaf'iev. The study proves the efficiency of the methodology developed.*

KEYWORDS: *text attribution; linguistic personality; automatic text processing; mathematical models; Russian.*

AUTHOR'S INFORMATION: *Khomenko Anna Yur'evna, Senior Lecturer, Department of Applied Linguistics and Foreign Languages, Assistant Researcher of the Laboratory of the Theory and Practice of the Systems of Support in Making Decisions, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia; Expert-Linguist, Copyright Expert, Expert in Phonoscopy, Center for Expertise and Evaluation "ESIN", Nizhniy Novgorod, Moscow.*

AUTHOR'S INFORMATION: *Ben'kovich Elena Romanovna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Gainutdinova Diana Il'darovna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Gasanova Leyla Rafig kyzy, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Kostina Alina Anatol'evna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Mazunina Zoya Olegovna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Nikolaeva Angelina Sergeevna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

AUTHOR'S INFORMATION: *Pimonova Elena Vladimirovna, Student of the Humanities Faculty, National Research University "Higher School of Economics", Nizhniy Novgorod, Russia.*

FOR CITATION: *Khomenko, A. Yu.* Automatic Text Processing and Linguistic Modeling as Instruments for Solving Problems of Text Attribution / A. Yu. Khomenko, E. R. Ben'kovich, D. I. Gainutdinova, L. R. Gasanova, A. A. Kostina, Z. O. Mazunina, A. S. Nikolaeva, E. V. Pimonova // Political Linguistics. — 2020. — No 3 (81). — P. 215-224. — DOI 10.26170/pl20-03-22.

ACKNOWLEDGMENTS. Research has been accomplished with financial support of the Russian Foundation for Basic Research (RFBR) within scientific project No. 19-31-27001 (19-312-90022).

REFERENCES

1. Vul, S. M. Forensic Investigation of Signs of Writing. — Kiev, 1973. [Kriminalisticheskoe issledovanie priznakov pis'men¬noy rechi. — Kiev, 1973]. — (In Rus.)

2. Vul, S.M. Forensic Attribution Identification Examination: Methodological Basics: Methodological Manual. — Kharkov, Kharkov Scientific Research Institute of Forensic Expertise, 2007. 64 p. [Sudebno-avtorovedcheskaya identifikacionnaya ekspertiza: metodicheskie osnovy: Metodicheskoe posobie — Kharkov, Khar'kovskij nauchno-issledovatel'skij institut sudebnyh ekspertiz, 2007. 64 s.] — (In Rus.)

3. Galyashina, E. I. Basics of judicial speech — Moscow, 2003. 236 p. [Osnovy sudebnogo rechevedeniya, Moskva, 2003. 236 s.] — (In Rus.)

4. Goroshko, E. I. Forensic classification examination: problems of establishing the gender of the document author. Theory and practice of forensics examination — Har'kov, Pravo, Vol. 3, 2003. Pp. 221-226. [Sudebno-avtorovedcheskaya klassifikacionnaya ekspertiza: problemy ustanovleniya pola avtora dokumenta], [Teoriya i praktika sudebnoj ekspertizy i kriminalistiki — Har'kov, Pravo, Vyp. 3, 2003. S. 221-226.] — (In Rus.)

5. Zaharov V. P., Hohlova M. V. Statistical method for collocations detection, XI All-Russian Joint Conference «Internet and Modern Society» — St. Petersburg, 2008. pp. 40-54. [tatisticheskij metod vyyavleniya kollokacij. // Yazykovaya inzheneriya: v poiske smyslov: Doklady seminara «Lingvisticheskie informacionnye tekhnologii v Internete»: XI Vserossijskaya ob"edinennaya konferenciya «Internet i sovremennoe obshchestvo»: Izdatel'stvo Sankt-Peterburgskogo universiteta, 2008. — S. 40-54.] — (In Rus.)

6. Zaharov, V. N. The program of supporting systems for the attribution of articles by F. M. Dostoevsky // Trud, Petrozavodsk, Vol. 9, ser. Applied Mathematics and Computer Science, 2000, pp.113-122. [Programma sistem podderzhki atribucii tekstov statej F. M. Dostoevskogo // Trud / PetrGU. — Petrozavodsk, 2000. — Vyp. 9. Ser. «Prikladnaya matematik i informatika». — S.113-122. — Soavt.: Leont'ev A. A., Rogov A. A., Sidorov YU. V.] — (In Rus.)

7. Karaulov, Yu. N. The Russian Language and the Language Personality — Moscow, Nauka, 1987. 264 p. [Russkij yazyk i yazykovaya lichnost' — Moskva, Nauka, 1987. 264 s.] — (In Rus.)

8. Komissarov, A. Yu. Forensic study of written language — Moscow, Ministry of Internal Affairs of Russia, 2000. 126 p. [Kriminalisticheskoe issledovanie pis'mennoj rechi: ucheb. Posobie— Moskva, Ministerstvo vnutrennih del, 2000. 126 s.] — (In Rus.)

9. Martynenko, G. Ya. Stylometry: emergence and formation in the context of interdisciplinary interaction // Structural and applied linguistics. Vol. 11: Intercollegiate Compendium. Sat, St. Petersburg, St. Petersburg University, 2015. Pp. 9 — 28. [Stilemetriya: vozniknovenie i stanovlenie v kontekste mezhdisciplinarnogo vzaimodejstviya. Strukturnaya i prikladnaya lingvistikaVyp. 11: mezhvuz. sb. / pod red. A. S. Gerda i I. S. Nikolaeva. — SPb.: Izd-vo S.-Peterb. un-ta, 2015. — 304 s. S. 9 — 28.] — (In Rus.)

10. Marusenko, M. A. Attribution of Anonymous and Pseudonymous Texts as a Typical Pattern Recognition Problem // Historiography and Source Study of National History. — St. Petersburg, 2003. Vol. 3. [Atributsiya anonimnykh i psevdonimnykh tekstov kak tipichnaya zadacha raspoznavaniya obrazov // Istoriografiya i isto¬chnikovedenie otechestvennoy istorii. — SPb, 2003. Vyp. 3]. — (In Rus.)

11. Morozov, N. A. Linguistic Specters: a means for distinguishing of plagiarism and original works for famous authors — Petrograd, Type of Imp. Acad. Sciences, 1916. 42 p. [Lingvisticheskie spektry: sredstvo dlya otlicheniya plagiatov ot istin. proizvedeniy togo ili dr. izvestnogo avt. — Petrograd : tip. Imp. Akad. nauk, 1916. 42 s.] URL:http://www.textology.ru/library/book.aspx?bookId=1&textId=3 (accessed: 05.07.2019).

12. Order of December 27, 2012 N 237 «On approval of the List of types of forensic examinations performed in federal budgetary forensic institutions of the Ministry of Justice of Russia, and the List of expert specialties for which the right to independently conduct forensic examinations in federal budgetary judicial expert institutions of the Ministry of Justice of Russia» [Electronic resource] [Prikaz ot 27 dekabrya 2012 goda N 237 «Ob utverzhdenii Perechnya rodov (vidov) sudebnyh ekspertiz, vypolnyaemyh v federal'nyh byudzhetnyh sudebno-ekspertnyh uchrezhdeniyah Minyusta Rossii, i Perechnya ekspertnyh special'nostej, po kotorym predstavlyaetsya pravo samostoyatel'nogo proizvodstva sudebnyh ekspertiz v federal'nyh byudzhetnyh sudebno-ekspertnyh uchrezhdeniyah Minyusta Rossii»] URL: www.pravo.gov.ru (accessed: 07.02.2020).

13. Radbil', T. B. Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts // Political Linguistics, Vol. 2 (74). 2019. Pp. 156-166. [Veroyatnostno-statisticheskie modeli v proizvodstve avtorovedcheskoj ekspertizy russkoyazychnyh tekstov // Politicheskaya lingvistika, Vyp. 2 (74). 2019. S. 156-166.] — (In Rus.)

14. Rodionova, E. S. Linguistic Methods of Attribution and Dating of Literary Works (to the Problem "Moliere / Corneille") [Electronic resource] : synopsis of doctoral thesis of Cand. Philol. Scinces, 2008. [Lingvisticheskie metody atributsii i datirovki literaturnykh proizvedeniy (k probleme «Mol'er — Kornel'») : avtoref. dis. … kand. filol. nauk, 2008]. URL: http://epir.ru/pragmat!/projects/corneille/files/autoreferat.pdf (accessed: 07.03. 2019). — (In Rus.)

15. Rubtsova I. I., Ermolaeva E. I., Bezrukova A. I. et al. Integrated methodology for the production forensic authorship examinations: Methodological recommendations — Moscow, Ministry of Internal Affairs of Russia, 2007. 192 p. [Kompleksnaya metodika proizvodstva avtorovedcheskih ekspertiz: Metodicheskie rekomendacii. — Moskva, Ministerstvo vnutrennih del Rossii, 2007. 192 s.] — (In Rus.)

16. Stepanenko A. A. Gender attribution in social network communication: the statistical analysis of pronouns frequency // Tomsk State University Journal, Vol. 415, 2017. Pp. 17—25 [Gendernaya atribuciya tekstov komp'yuternoj kommunikacii: statisticheskij analiz ispol'zovaniya mestoimenij // Vestnik Tomskogo gosudarstvennogo universiteta, N 415, 2017. Ss. 17—25], DOI: 10.17223/15617793/415/3. — (In Rus.)

17. Federal Law of May 31, 2001 N 73-ФЗ «On State Forensic Science Activities in the Russian Federation» [Electronic resource] // Russian newspaper, N 256 of December 31, 2001. [Federal'nyj zakon ot 31 maya 2001 g. N 73-FZ «O gosudarstvennoj sudebno-ekspertnoj deyatel'nosti v Rossijskoj Federacii» Rossiyskaya Gazeta, N 256 3.12. 2001], URL: https://base.garant.ru/12123142/ (accessed: 07.03.2019). — (In Rus.)

18. Campbell, L. The Sophisties and Polilicus of Plato / L. Campbell. — Oxford : Clarendon, 1867. — 170 p. — Text : unmediated.

19. Coulthard, M. Author identification, idiolect, and linguistic uniqueness / M. Coulthard. — Text : unmediated // Applied Linguistics. — 2004. — No 24 (4). — P. 431—447.

20. Juola, P. A Prototype for Authorship Attribution Studies / P. Juola, J. Sofko, P. Brennan. — Text : electronic // Literary and Linguistic Computing. — 2006. — Vol. 21. — Iss. 2. — 1 June. — P. 169—178. — URL: https://doi.org/10.1093/llc/fql0 (date of access: 05.07.2019).

21. Karlgren, J. Authorship Profiling Without Using Topical Information—Notebook for PAN at CLEF, 2018 / J. Karlgren, L. Esposito, Ch. Gratton, P. Kanerva. — URL: https://pdfs.semanticscholar.org/ee57/5920182cdc6de1337f71b07a25e830022459.pdf?_ga=2.139547835.909834531.1562339431-1809262388.1562339431 (date of access: 05.07.2019). — Text : electronic.

22. Koppel, M. Exploiting Stylistic Idiosyncrasies for Authorship Attribution / M. Koppel, J. Schler. — Text : unmediated // Proceedings of IJCAI'03 Workshop on Computational Approaches to Style Analysis and Synthesis. — 2003. — No 69. — P. 72—80.

23. Labbe, C. Inter-Textual Distance and Authorship Attribution / C. Labbe, D. Labbe. — Text : unmediated // Corneille and Molière. Journal of Quantitative Linguistics. — Taylor & Francis (Routledge), 2001. — No 8 (3). — P. 213—231.

24. Litvinova, T. Gender identification in Russian written texts / T. Litvinova, P. Seredin, O. Litvinova, O. Zagorovskaya. — Text : electronic // XLinguae. — 2017. — Vol. 10. — Iss. 3. — P. 176—183. — URL: http://xlinguae.eu/files/XLinguae3_2017_14.pdf (date of access: 05.07.2019).

25. Lutoslawski, W. The origin and growth of Plato's logic / W. Lutoslawski. — London, 1997. — 613 p. — Text : unmediated.

26. McMenamin, G. R. Forensic Linguistics: advances in forensic stylistics / G. R. McMenamin. — 2002. — 331 p. — Text : unmediated.

27. Mendenhall, T. The characteristic curves of composition / T. Mendenhall. — Text : unmediated // Science. — 1987. — No 9. — P. 237—249.

28. Merriam, T. An Application of Authorship Attribution by Intertextual Distance in English / T. Merriam. — Text : unmediated // Corpus. — 2003. — N 2. — P. 142—168.

29. Mosteller, F. Applied Bayesian and Classical Inference: The Case of the Federalist Papers / F. Mosteller, D. L. Wallace. — Addison-Wesley, Reading, MA, 1984. — Text : unmediated.

30. Wright, D. Using word n-grams to identify authors and idiolects: A corpus approach to a forensic linguistic problem / D. Wright. — Text : electronic // International Journal of Corpus Linguistics. — 2017. — 22 (2). — P. 212—241. — URL: https://core.ac.uk/download/pdf/84587040.pdf (date of access: 05.07.2019).