РАЗДЕЛ 6. РЕЦЕНЗИИ. ХРОНИКА

УДК 81'42
ББК Ш105.51 *ГСНТИ 16.21.27, 16.31.21* *Код ВАК 10.02.19, 10.02.21*
DOI 10.26170/1999-2629_2021_01_15

С. И. Красса
Ставрополь, Россия
ORCID ID: 0000-0002-6699-2159 ☑

☑ ***E-mail:*** *skrassa@yandex.ru.*

Первый круглый стол по практикам и стандартам судебного автороведческого анализа (обзор 2)

АННОТАЦИЯ. *Проводится обзор докладов Первого круглого стола по практикам и стандартам судебного автороведческого анализа, организованного Международной ассоциацией судебной лингвистики и Центром цифровых гуманитарных наук Манчестерского университета. Круглый стол проведен 15 мая 2019 года. В обзоре представлены совместный доклад Кшиштофа Креденса из Астонского университета и Петра Рензика «Крупномасштабная классификация автора: исследуя черный ящик»; доклад Джека Грива, профессора лингвистики из Бирмингемского университета, «Изменения стиля и автороведческий анализ». В конце каждого доклада приводятся вопросы к выступающему и ответы на них. В первом докладе установление авторства рассматривается на материале коротких сообщений популярных интернет-форумов (один из них — «Musmnet»). При обработке данных применен статистический подход k — ближайших соседей (k-nearest neighbors approach), использовались следующие маркеры стилей: N-граммы, POS-леммы со снятой неоднозначностью, коллокации, демонстрирующие типичные лексические реализации синтаксических конструкций (фразовые глаголы, прямые дополнения, подлежащие, выраженные местоимениями). Если нет универсально применимых маркеров, необходимо всякий раз отыскивать оптимальные параметры и воспринимать каждый случай как уникальный. Джек Грив в докладе, посвященном стилеметрии, утверждает, что в качестве основы для нее лучше, чем социолингвистика, подходит функциональная теория языкового варьирования: язык видоизменяется в разных контекстах потому, что различные языковые формы лучше подходят для этих контекстов. Для демонстрации преимуществ функционального подхода проводится стилеметрический анализ статей двух политических колумнистов издания «The Telegraph». Стиль аналитических колонок этих авторов различается в соответствии со стратегиями коммуникации: у одного автора стиль более нарративный, у другого — более аналитический. Докладчик делает вывод, что функциональная теория автороведения может стать более надежной, чем социолингвистическая теория идиолекта, базой для судебной лингвистики в целом.*

КЛЮЧЕВЫЕ СЛОВА: *круглый стол; судебный автороведческий анализ; профилирование автора; стилеметрия; компьютерная лингвистика; ядерные признаки стиля; судебная лингвистика; судебная экспертиза*.

ТИП ПУБЛИКАЦИИ: *рецензия*.

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Красса Сергей Иванович, кандидат филологических наук, доцент, Ставрополь, Россия; e-mail: skrassa@yandex.ru.*

ДЛЯ ЦИТИРОВАНИЯ: *Красса, С. И.* Первый круглый стол по практикам и стандартам судебного автороведческого анализа (обзор 2) / С. И. Красса // Политическая лингвистика. — 2021. — № 1 (85). — С. 159-166. — DOI 10.12345/1999-2629_2021_01_15.

ВВЕДЕНИЕ

15 мая 2019 г. Международная ассоциация судебной лингвистики (International Association of Forensic Linguistics) и Центр цифровых гуманитарных наук (Centre for Digital Humanities) Манчестерского университета провели Первый круглый стол по практикам и стандартам судебного автороведческого анализа [Forensic Authorship Analysis Roundtable]. Целью мероприятия являлось продвижение судебной автороведческой экспертизы в направлении стандартизации применения, улучшения диалога между учеными различных направлений. Одной из наиболее значимых целей круглого стола было создание площадки для обсуждения и развития междисциплинарного сотрудничества, которое могло бы привести к формированию строгих методических и теоретических достижений в данной области. Трансляция круглого стола велась в видеохостинге «YouTube» [Forensic Linguistics Roundtable Event]. Порядок проведения круглого стола предусматривал выступления докладчиков, ответы на вопросы, затем выступления и вопросы в режиме диалога. В целом круглый стол продлился более семи часов. Данный обзор является продолжением обзора, опубликованного ранее (Красса, С. И. Первый круглый стол по практикам и стандартам судебного автороведческого анализа (обзор 1) / С. И. Красса // Политическая лингвистика. — 2020. — № 6 (84). — С. 174-187. — DOI 10.26170/pl20-06-19).

ОБЗОР ДОКЛАДОВ УЧАСТНИКОВ

К. КРЕДЕНС, П. РЕНЗИК

Третьим было совместное выступление Кшиштофа Креденса (Krzysztof Kredens) из Астонского университета и Петра Рензика (Piotr Ręzik) на тему «Крупномасштабная классификация автора: исследуя черный ящик».

В докладе рассматриваются две проблемы: первая — общая классификация авторов. В значительной мере это исследования в области, в которой работает Е. Стамататос и другие специалисты по компьютерной лингвистике. К. Креденс и П. Рензик в своем исследовании тестировали показатели системы, используя различные измерения, чтобы увидеть, какие из них работают лучше других. Однако более важное направление сегодня — рассмотрение «черного ящика». Ранее К. Креденс и П. Рензик использовали термин «экспликация», сейчас — более простой термин, а именно «объяснение». Докладчики постарались продемонстрировать, можно ли объяснить что-либо, используя большое количество данных.

Контекст исследования включает следующие утверждения. Тексты Q и K могут существенно различаться по многим основаниям: жанр, тема, канал коммуникации, эффект аккомодации и так далее. В это очень длинное «так далее» включаются буквально сотни параметров, которые необходимо принимать во внимание в ходе судебной авторoведческой работы.

Вследствие отмеченной выше вариативности имеется множество параметров, о существовании некоторых из которых мы иногда не знаем, поэтому проверка достоверности анализа крайне трудна. Обычно в настоящее время при валидизации отдается предпочтение науке о судебной речи. Выступающие, по их признанию, хотели сделать то же самое, но на данной стадии для этого проще обратиться к объяснению. Однако объяснение не только возможно, но и необходимо.

Тим Грант ранее обращался к этому кейсу, в частности приводя следующее заявление судьи: «Я не в состоянии оценить достоверность данного анализа в качестве отрасли знания. Это нечто новое для меня. Я не знаю, каковы качество и надежность этого экспертного доказательства».

И судья не использовал данное заключение при вынесении вердикта. Он сказал, что не может оценить аргументированность, которая по определению связана с объяснимостью. Аргументированность (обоснованность, достоверность) имеет отношение к нашей уверенности в том, что система производит измерения того, что мы хотим измерять, а также к тому, как определенный параметр проявляет себя в различных контекстах.

Итак, это предпосылки исследования. Далее докладчик остановился на исследовательских вопросах. Какие маркеры стиля имеют наилучший индивидуализирующий потенциал и почему? Этот вопрос относится к совместным признакам, о которых докладчик спрашивал одного из предыдущих спикеров. Докладчик не верит в такие признаки, вместе с тем одни признаки проявляют себя лучше в одних жанрах, чем в других.

Какой параметр сравнения проявляет себя лучшим образом в атрибуции текстов в отношении того или иного автора и почему? Почему некоторые маркеры демонстрируют эффективность (это вопрос объяснения), и можем ли мы индексировать наши данные с парадигматически сходными для дальнейшего улучшения показателей системы? Если категория работает, если Ngram доказывает свою продуктивность, то давайте рассматривать парадигматически те категории, которые, возможно, будут работать на наши цели. И наконец, какие маркеры не проявляют себя наилучшим образом в идентификационных задачах? Как мы можем объяснить наличие ложных позитивных показателей в системе?

Рассмотрение позиций в докладе началось с маркеров стиля. В течение ряда лет были предложены сотни, если не тысячи идиолектных маркеров. Одна из фундаментальных работ в этом отношении — «Судебная стилистика» Джеральда Мак-Менамика (1993 г.) [McMenamic 1993], где приводится несколько сотен маркеров.

Описываются все уровни письменной речи — от сочетания букв (Ngram) до признаков дискурсивного уровня. Ngram достаточно эффективно достигает поставленных целей. Мы осознаем, что значительное число недавних исследований использует буквы и буквенные и словесные Ngram, но авторы не делают попыток объяснения. Если использование Ngram работает, а оно работает хорошо, то необходимо дать этому объяснение в структурных и языковых терминах.

Далее доклад продолжил Петр Рензик. Он привел примеры по исследованию метафоры черного ящика. Есть еще одна метафора, которая, возможно, не настолько распространена, но также может быть использована, — метафора песочницы. Идея заключается в том, что авторoведческие исследования анализа уровня точности и дру-

гих параметров проводятся в рамках компьютерной песочницы.

Одна из идей, которая может быть использована, — несколько песочниц, с которыми мы можем играть или действовать иным образом, чтобы получить наиболее возможные положительные результаты.

Другая идея — это то, что мы оперируем очень большим количеством данных. Одна из наиболее крупных коллекций сообщений — *Mumsnet*, онлайновый дискуссионный форум, содержащий полилоговые тексты, в основном по проблемам родителей. В итоговой версии там содержится более 50 млн постов, более 2 млрд слов, более 600 тыс. авторов.

Имеются и другие дискуссионные форумы, которые мы анализируем, они обозначены как X, Y, Z. Одна из их основных характеристик — они содержат полилоговые тексты: авторы вступают в интеракцию друг с другом, в них мы можем найти все, что находим в других разговорах (участники проявляют любопытство, вмешиваются в чужие дела, шутят с использованием лексического значения слова, выбирая те или иные фразы для достижения прайминг-эффекта, используют фразы, которые рассматриваются в качестве полезных в этом разговоре и больше не используются).

Также есть проблема, кажущаяся наивной: каждый аккаунт, каждый ник соответствует отдельному автору. В то же время существует множественное, разделяемое авторство. Это не такая серьезная проблема для традиционной дискуссии, как для материалов социальных сетей и троллинговых аккаунтов и подобных явлений. Но это должно быть принято во внимание и учтено в результатах.

Выбор методов мотивирован двумя основными факторами.

1. Относительно малый размер тестовых образцов (1—50 постов у автора). Показатели являются результатом обработки большого числа авторов.
2. Контролируемая классификация текстов.

Быстрый базовый тест — это тест с внешним словом (словами), заключенным во входной слой. Идея заключается в том, что если обучение проводится на внешних коллекциях, ваша модель более надежна в применении к другим коллекциям. Теоретически это может распространяться на межжанровое применение, но только теоретически.

Также проводился продолжающийся эксперимент с конволюционными сетями, однако пока нет положительных результатов, поскольку обучение идет очень медленно. Это связано с большим числом классов, которые требуется предсказать — несколько десятков тысяч.

Однако также при IR (Information retrieval) был применен подход k — ближайших соседей (K-Nearest Neighbors approach, k-NN, KNN), перекочевавший из более ранних работ П. Рензика по информационному поиску: 10 лет назад ученый должен был разработать классификаторы по отношению к таксонам приблизительно из 20 тыс. категорий — требовалась бинарная классификация, применимая к такой огромной таксономии. Этот подход также позволил получить некоторые результаты.

Маркеры стилей, используемые в классификаторах KNN

- Словесные Ngram (моно-, би-, три- и тетраграммы).
- POS-леммы со снятой неоднозначностью.
- Лексические реализации синтаксических конструкций (дерево зависимостей): фразовые глаголы, прямые дополнения, определения, определения к прямым дополнениям, подлежащие, выраженные местоимениями. Идея заключается в том, чтобы фиксировать сочетания (коллокации), которые могли быть использованы автором: типичное прямое дополнение, которое может быть использовано, но не зафиксировано посредством Ngram; поскольку это прилегающие друг к другу последовательности слов, исследователи вынуждены использовать 6-gram, чтобы определить повторяющиеся паттерны прямого дополнения.

Подход KNN в определении автора

Исследователи пытались определить маркеры для каждого документа. Внимание фокусировалось на ключевых маркерах, что бы это ни означало, на методе их выделения. Набор ключевых маркеров репрезентирует запросы и представлен в индексе. В материал исследования вошла тысяча постов, которые агрегировались с установлением автора. Основная идея состоит в поиске подобных элементов в представленных документах, рассматривались наиболее часто повторяемые единицы. Это простой метод, но у него есть реальные преимущества.

KNN (за и против)

Плюсы:

– обновляемость в режиме реального времени;

– теоретически неограниченное число авторов, классов, как в случае с использованием глубинных нейросетей, для чего требуются серьезные компьютерные ресурсы;

– прозрачный механизм классификации (маркеры прослеживаемые и интерпретируемые);

– межжанровая применимость;

– возможность использования для атрибуции небольших текстовых образцов с невысоким соотношением токенов.

Минусы:

– в большей мере зависит от уникальных слов, чем от дистрибуции наиболее частотных слов;

– может быть квазиоптимальным (теоретически в большей мере непосредственно контролируемые подходы могут работать лучше для небольших наборов авторов/классов);

– несмотря на мгновенное обучение, рост показателей производительности не очень быстрый.

Оценка

50 постов, созданных 100—135 авторами, были извлечены из обучающего набора, вручную (поверхностно) верифицированы как анонимные (удалены все ссылки и другие атрибуты) и включены в исследовательский набор.

Полученные результаты весьма интересны.

Mumsnet. 19,2 % — это бизнес-предложения. Эффективность поиска в первых 30 предположениях: средний высокий балл — 5,46, медиана — 3. Вероятно увеличение с возрастанием времени.

Форум X. 135 авторов в тестовом наборе, число одиночных постов — 7392, средний высокий балл — 7,2, медиана — 4.

Объединив 50 случайным образом перемешанных постов каждого автора, получаем:

Mumsnet: 45,1 % (средний высокий балл — 5,46, медиана — 3).

Форум X: 89,7 % в первых 30 предположениях (средний высокий балл — 2,7, медиана — 1).

Это были результаты «песочницы».

Эффективность поиска в KNN составила 0,7; ранг 30.

Объяснение полученных результатов

Для контролируемого классификатора мы используем «метод случайного перемешивания данных» (data perturbation) (Ribeiro et al., 2016).

Тестовые образцы видоизменены (чтобы одно слово встречалось один раз) и введены в классификатор. Слова, которые помогают или мешают классификаторам, выделены посредством многократной переквалификации: *sense*, *anyhow* (помогают); *was*, *university* (мешают).

Для KNN-подхода идентифицировались объекты коллекции, которые доказали возврат в релевантные документы. Любой тип маркера может быть проанализирован с позиции его различительного потенциала (пропорции релевантных документов), возможности внести вклад в верную идентификацию автора.

К продолжению доклада вернулся Кшиштоф Креденс. Как он заметил, «вот что происходит, когда мы заглядываем в „черный ящик“»: мы ищем объяснения, и нет очевидных типологических паттернов в наборах совместных признаков, которые представляют собой результат. С одной стороны, это разочаровывает, а с другой — это замечательно.

Наборы созданы из элементов, принадлежащих — в разном количестве и непредсказуемо — к различным лексико-грамматическим категориям. Некоторые категории имеют нулевую реализацию.

Объяснения обнаруживаются для относительно небольшого числа авторов, демонстрирующих: нестандартные формы / ошибки; опечатки, описки, творческие окказионализмы, совместный отбор тематически связанных слов.

Импликации по теории идиолекта

Некоторые авторы демонстрируют узнаваемый стиль идиолекта (ключевое слово — «некоторые»). Некоторые авторы проявляют тенденцию соблюдать жанровые конвенции (или то, чем является стандарт). Ни одна из лексико-грамматических категорий не имеет сильного индивидуализирующего потенциала. Совместный набор признаков часто является ключом к идентификации автора.

Информация об индивидуальных идиолектах обнаруживается только частично каждый раз, когда публикуется индивидуальный пост в режиме онлайн, и чем более его автор активен, тем больше информации проявляется (когда продолжает наблюдаться эффект плато).

Если нет универсально применимых маркеров, тогда мы вынуждены работать от случая к случаю, всякий раз отыскивая оптимальные параметры, и каждый случай уникален. Универсально применимые маркеры, вероятно, просто миф.

Вопросы по докладу

Вопрос. Что вы понимаете под индексом характеристик в парадигматическом аспекте?

Ответ. Если у вас Ngram (N-грамма), которая, как подтвердилось, работает, и данная Ngram представляет лексико-грамматическую характеристику, например определительное наречие + прилагательное, то

это очень интересно. Если вы полагаете, что сочетание такого вида различительное, то это может быть весьма интересным не только для данного, но и для ряда других случаев.

Вопрос. Другой вопрос — о слове *that*. Исследовали ли вы контексты, в которых оно употребляется?

Ответ. Нет.

Вопрос. Потому что это очень интересный признак?

Ответ. Да, но очень сложно отстроить для этих целей программу. Возможно, следует проанализировать конкорданс с *that*. Это следующий шаг. Мы уже заинтересованы в объяснении. Одно из главных отличий в подходе заключается в том, что *that* может быть выбрано в соответствии со строго контролируемым подходом, но не KNN-подходом.

Вопрос. У вас есть данные по *Mumsnet* относительно девайсов и интерфейсов, с которых заходили авторы постов?

Ответ. Нет, сбор материала занял длительное время, около 20 лет. Это хороший вопрос, такая информация может быть еще одним параметром, используемым в рассмотрении постов. Мы предполагаем продление проекта на 10—11 месяцев, и это определенно будет принято во внимание, поскольку есть очень много путей рассмотрения данного материала, как и параметров его исследования.

ДЖ. ГРИВ

Четвертым выступал Джек Грив (Jack Grieve), профессор лингвистики из Бирмингемского университета, с докладом на тему «Изменения стиля и автороведческий анализ».

Докладчик начал выступление с объяснения того, что такое стилеметрия.

Мы можем различать написанное разными авторами посредством количественного анализа. Мы особенно успешны в атрибуции авторства, когда у нас протяженные спорные тексты и большое количество данных для сопоставления (например, использование служебных слов). Но мы не располагаем проверенной и общепризнанной теорией автороведения, чтобы основывать на ней стилеметрический анализ, как отметили Е. Стамататос, Т. Грант и К. Креденс.

Среди специалистов отсутствует единство во мнениях о том, почему люди имеют уникальный письменный стиль (и даже о том, существуют ли уникальные стили) и почему специальные языковые переменные следовало бы анализировать как маркеры этого стиля. Эта проблема очень важна не только для судебной лингвистики, но и для когнитивной лингвистики, овладения языком и других областей.

Далее докладчик остановился на стандарте Даубера (США; Daubert standart), упомянув необходимость теории в судебном контексте (правило 702); допрос экспертов. Если научные, технические или иные специальные знания эксперта помогают лицу, проводящему следствие, понять или установить факты, относящиеся к делу, свидетель квалифицируется как эксперт на основе знаний, умений, образования, подготовки и может давать показания в форме мнения или другой, если показания основываются на существенных фактах или данных, свидетельство является произведением, базирующимся на принципах и методах, надежных для данных фактов и дела.

Идиолект

Автороведческий анализ, особенно в правовом контексте, часто для подтверждения выводов апеллирует к социолингвистической концепции идиолекта — уникальной индивидуальной разновидности языка.

Лингвисты подходят к проблеме спорного авторства с теоретических позиций, согласно которым каждый носитель языка имеет свою собственную отличительную и индивидуальную версию языка, на котором он говорит и пишет — свой собственный идиолект [Coulthard. 2004: 431].

Социолингвистика

Форма языка, используемая людьми с различными социальным бэкграундом и идентификаторами, системно и постоянно изменяется. Существует предположение о том, что в автороведческом анализе этот тип социолингвистического варьирования расширяется до уровня индивидуума, поскольку каждый имеет социально уникальную историю (социальную идентичность, как утверждал в своем докладе Тим Грант — более подробно об этом написано в его совместной с Никки Маклеод статье).

Критично, что социолингвистическая теория и исследования основываются на чередовании переменных — альтернативных способах подтвердить одно и то же, а это часто является основой для анализа в судебной стилистике.

Социолингвистика и стилеметрия

В то время как судебная стилистика может быть основана на социолингвистических принципах, это неверно в отношении стилеметрии, поскольку она основывается не на анализе чередования переменных, а главным образом на относительной частотности индивидуальных форм (набор служебных слов, буквенных Ngram), которые не являются ва-

лидными переменными в социолингвистической теории: не *the* в чередовании с чем-то, а просто *the*.

Например, Деннис Престон утверждает, что частотность индивидуальных форм «do not meet the basic requirement for the study of variation — the choice of more than one semantically equivalent element in environments where all have a privilege of occurrence» [Preston 2001: 291] — «не согласуется с основными требованиями для изучения вариативности — выбора более одного семантически эквивалентного элемента в окружении, где все имеют преимущество в употреблении» (перевод наш. — *С. К.*). То есть если нет чередования, то нет и параметра.

Приводится пример определения авторства 15-й книги о стране Оз, в частности, с помощью анализа использования служебных слов. Этот пример трудно объяснить с позиций социолингвистики, что вступает в противоречие с положением Д. Престона. Приводится 50 служебных слов, и нет чередования: сообщается процент использования слов *and*, *the*, *to*. Такой способ работает весьма эффективно: когда много данных, два автора и нет контроля регистра.

Докладчик полагает, что функциональная теория языкового варьирования представляет собой лучшую основу для стилеметрии, чем социолингвистическая теория. Если вы посмотрите на телефонный разговор и обычную беседу, вы сразу обнаружите отличия. То же касается статьи в газете и рассказа. Это делается намного легче, чем в случае авторoведческого анализа. Так происходит за счет распознавания языковых моделей. Реципиента при этом нельзя обвинить в субъективности: узнавание происходит, поскольку опосредовано функцией данного текста.

Язык видоизменяется в разных контекстах потому, что различные языковые формы лучше подходят для этих контекстов.

Язык видоизменяется от автора к автору, потому что они используют какие-то иные коммуникативные стратегии. Особенно, полагает Дж. Грив, релевантны для стилеметрии корпусные методы и теории, основанные на анализе регистров [Biber 1988]. В стилеметрии измерения стилистического варьирования имеют целью исследование регистра, основанного на многомерном анализе относительной частотности грамматических форм.

Докладчик демонстрирует измерения, различающие исследуемые тексты.

Первое измерение — разговорные, неформальные (телефонный разговор, разговор «лицом к лицу», личное письмо, интервью) и более формальные, литературные (официальные документы, научные тексты, репортажи в СМИ, обзоры прессы, биографии) тексты [Biber 1988]:

«приватные» глаголы	существительные
опущение *that*	длинные слова
стяженные формы	предлоги
настоящее время	прилагательные в роли определений
местоимения 2-го лица	
эмфаза	
местоимения 1-го лица	
дискурсивные маркеры	

«Приватные» (*private*) глаголы обозначают ненаблюдаемые интеллектуальные акты и состояния.

Второе измерение: нарративные (любовные истории, детективы, научная фантастика, художественные тексты в целом) — ненарративные (радио и телепередачи, описания хобби, научные тексты, деловые письма, телефонные переговоры) тексты:

прошедшее время	настоящее время
местоимения 3-го лица	прилагательные в роли определений
перфект	
«публичные» (public) глаголы (say)	

Стилеметрия

Case study. Чтобы проиллюстрировать, как функциональный подход может служить базой стилеметрического анализа, докладчик привел стилеметрический case study. Он сопоставил стиль двух авторов, которые являются постоянными политическими колумнистами в «The Telegraph», Уильяма Хейга (William Hague) и Чарльза Мора (Charles Moore). Целью Дж. Грива было определить, можем ли мы разграничивать этих авторов (конечно, такая возможность подтвердилась). Но если можем, то как различаются их стили?

Докладчик проанализировал корпус из 130 статей с применением стандартного стилеметрического подхода. Была определена относительная частотность верхних 50 служебных слов в корпусе из 260 текстов. Эти данные были подвергнуты анализу методом главных компонент, чтобы выделить два измерения вариативности. Исследователь выстроил тексты относительно двух измерений, чтобы идентифицировать стилистические различия между двумя авторами. Затем были продемонстрированы кривые распределения для *and*, *it*, верхних 50 служебных слов, и была построена их корреляционная матрица. Эти служебные слова заметно разграничивают стили.

Анализ регистров

Для понимания, чтó стилеметрический анализ говорит нам о различии стилей авто-

ров, докладчик провел повторный анализ данных, используя регистровый анализ. Также он провел разметку данных и обработку 66 грамматических характеристик с использованием *Andrea Nin's MAT Tagger* [Nini 2019]. Сократил набор данных до двух измерений, используя факторный анализ. Интерпретировал эти два фактора на основе переменных и текстов, в наибольшей мере ассоциированных друг с другом. Составил диаграмму текстов двух авторов в разрезе этих факторов. Сравнил эти факторы с анализом регистров до измерений предыдущего стилеметрического анализа, для чего использовал набор характеристик Д. Байбера [Biber 1988].

Выводы

Стиль, которым Хейг и Мор пишут аналитические колонки, различается потому, что у Мора стиль более нарративный, тогда как Хейг использует более аналитический стиль. Результаты продемонстрировали, как авторы используют слегка различающиеся стратегии для коммуникации в данном контексте. Именно поэтому стилеметрические методы работают: они определяют эти типы риторических различий.

Стилеметрия не нуждается в подтверждении положения, что идиолект существует: только стилистические вариации меняются и отражаются в соответствии с коммуникативной целью, и эта цель может различать авторов в измеряемом и интерпретируемом аспектах. Функциональная теория автороведения также потенциально обеспечивает в дальнейшем более надежное основание для судебной лингвистики в целом, чем социолингвистическая теория идиолекта.

Вопросы по докладу

Вопрос. Вы основывались на встречаемости (употребительности) или частотности?

Ответ. На частотности, да. В коротких текстах, объемом в сотни слов, мы основываемся на употреблении, встречаемости (есть или нет). Но в текстах размером тысячи слов это, конечно, частотность.

Вопрос. Выбор параметров — это строго контролируемая ситуация, или вы проводили отбор случайным образом?

Ответ. Скорее, это осознанный отбор, поскольку я лингвист. Тем более что необходимо объяснение того или иного выбора. Корпусный лингвист, специалист по анализу дискурса это делает сознательно.

Вопрос. Я согласен с точкой зрения, согласно которой множество характеристик может быть объяснено с функциональной точки зрения. Однако я должен согласиться и с тем, что многие характеристики, которые вы назвали, представляют собой функциональные черты.

Ответ. Лично я считаю, что все характеристики функциональные. Ngram, коллокации, знаки препинания имеют функциональную нагрузку.

Вопрос. Вопрос о риторической теории жанров. Мы относим текст к тому или иному жанру. Они представляют собой разновидность текстов, мы выделяем коммуникативные цели, и эти цели относятся ко всем текстам этого жанра или лишь к некоторым текстам?

Ответ. Я считаю, что в автороведческом анализе рассматривается не социальное, а регистровое варьирование — в 95 % случаев. Например, мы анализируем блоги или троллей. Если вы рассматриваете регистр — пристально изучаете его характеристики, индивидуальный язык в рамках регистра. Если вы посмотрите мои научные статьи, то в них я подробнее характеризую указанные подходы.

S. I. Krassa
Stavropol, Russia
ORCID ID: 0000-0002-6699-2159 ☑

☑ ***E-mail:*** *skrassa@yandex.ru.*

First Roundtable on Practices and Standards in Forensic Authorship Analysis (overview 2)

ABSTRACT. *The article provides an overview of the reports of the 1st Roundtable on Practices and Standards in Forensic Authorship. The roundtable was held by the International Association of Forensic Linguistics and the Centre for Digital Humanities at the University of Manchester on May 15, 2019. The review presents a co-authored report by Krzysztof Kredens from the University of Aston and Piotr Ręzik "Large-scale author classification — looking into the black box", Jack Grieve, Professor of Linguistics at the University of Birmingham "Register variation and authorship analysis" and the report by Jack Grieve, Professor of Linguistics at the University of Birmingham, "Register variation and authorship analysis". Each report is supplemented with questions to the speaker and their answers to them. The first report looks at the issue of authorship identification on the material of short messages from popular Internet forums («Musmnet» is one of them). Data processing involves the k-nearest neighbors algorithm and uses the following style markers: n-grams, POS-lemmas with neutralized polysemanticity, collocations demonstrating typical lexical realizations of syntactic constructions*

(phrasal verbs, direct objects and subjects expressed by pronouns). In case there are no universally accepted markers, it is necessary each time to find optimal parameters and treat each case as unique. In his report on stylometry, Jack Grieve argues that the functional theory of language variation is more suitable as a foundation for it than sociolinguistics: language varies in different contexts because some linguistic forms suit these contexts better than others. To demonstrate the advantages of the functional approach, a stylometric analysis of the articles of two political columnists of "The Telegraph" is performed. The style of analytical columns of these authors differs depending on the communication strategy: the style of one author is more narrative, and that of the other is more analytical. The reporter concludes that the functional theory of authorship identification studies can become more reliable than the sociolinguistic theory of individual style and can be considered a foundation of forensic linguistics on the whole.

KEYWORDS: *roundtable discussions; forensic authorship analysis; author profiling; stylometry; computer linguistics; core features of a style; forensic linguistics; forensic expertise.*

TYPE OF PUCBLICATION: *review.*

AUTHOR'S INFORMATION: *Krassa Sergey Ivanovich, Candidate of Philology, Associate Professor, Stavropol, Russia.*

FOR CITATION: *Krassa, S. I.* First Roundtable on Practices and Standards in Forensic Authorship Analysis (overview 2) / S. I. Krassa // Political Linguistics. — 2021. — No 1 (85). — P. 159—166. — DOI 10.12345/1999-2629_2021_01_15.

REFERENCES

1. Biber, D. Variation across Speech and Writing / Douglas Biber. — Cambridge : Cambridge Univ. Pr., 1988. — 13, 299 p.

2. Coulthard, M. Author identification, idiolect, and linguistic uniqueness / Malcolm Coulthard. — Text : unmediated // Applied Linguistics. — 2004. — Vol. 25. — Iss. 4. — P. 431—447.

3. Forensic Authorship Analysis Roundtable. — URL: https://www.eventbrite.co.uk/e/forensic-authorship-analysis-roundtable-tickets-59772040783# (date of access: 27.10.2020). — Text : electronic.

4. Forensic Linguistics Roundtable Event / International Association of Forensic Linguists ; Centre for Digital Humanities, University of Manchester // YouTube. — Duration: 7:29:40. — URL: https //www.youtube.com/watch?v=ZUfxdLstIOc (date of access: 27.10.2020). — Image (moving; 2D) : electronic.

5. Grant, T. Assuming identities online: experimental linguistics applied to the policing of online paedophile activity / Tim Grant, Nicci Macleod. — Text : unmediated // Applied Linguistics. — 2016. — Vol. 37. — Iss. 1. — P. 50—70.

6. Grant, T. Resources and constraints in linguistic identity performance: a theory of authorship / T. Grant, N. Macleod. — Text : unmediated // Language and Law. — 2018. — Vol. 5. — Iss. 1. — P. 80—96.

7. Grieve, J. Quantitative authorship attribution: an evaluation of techniques / J. Grieve. — Text : unmediated // Literary and Linguistic Computing. — 2007. — Vol. 22. — P. 251—270.

8. Grieve, J. Regional Variation in Written American English / J. Grieve. — Cambridge Univ. Pr., 2016. — Text : unmediated.

9. Grieve, J. Sociolinguistics: Quantitative Methods / J. Grieve. — Text : unmediated // The Encyclopedia of Applied Linguistics / C. A. Chapelle (ed.). — Hoboken, NJ : Wiley-Blackwell, 2012.

10. MacLeod, N. Whose Tweet? Authorship analysis of microblogs and other short-form messages / Nicci MacLeod, Tim Grant. — Text : unmediated // Proceedings of The International Association of Forensic Linguists' Tenth Biennial Conference / Centre for Forensic Linguistics, Aston Univ. — 2012. — P. 210—224.

11. MAT Tagger = Multidimensional Analysis Tagger / Dr Andrea Nini. — Program : electronic. — URL: https://sites.google.com/site/multidimensionaltagger/.

12. McMenamic, G. R. Forensic Stylistics / G. R. McMenamic. — Amsterdam : Elsevier Science Publisher, 1993. — XV, 249 p. — Text : unmediated.

13. Nini, A. The Multi-Dimensional Analysis Tagger / A. Nini. — Text : unmediated // Multi-Dimensional Analysis: Research Methods and Current Issues / Berber Sardinha, T. & Veirano Pinto M. (eds.). — London ; New York : Bloomsbury Academic. — 2019. — P. 67—94.

14. Preston, D. Style and the psycholinguistics of sociolinguistics: the logical problem of language variation / Dennis R. Preston. — Text : unmediated // Style and Sociolinguistic Variation / P. Eckert, J. R. Rickford (Eds.). — Cambridge : Cambridge Univ. Pr., 2001. — P. 279—304.

15. Ribeiro, M. T. Why Should I Trust You?: Explaining the Predisctions of Any Classifier / Marco Tulio Ribeiro, Sameer Singh, Carlos Guestrin. — DOI 10.1145/2939672.2939778. — Text : electronic // Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining / ACM, 2016. — P. 1135—1144. — URL: http://www.kdd.org/kdd2016/papers/files/rfp0573-ribeiroA.pdf.