РАЗДЕЛ 5. 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_02_19

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 3)

АННОТАЦИЯ. *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 состоялся 15 мая 2019 года. Представлены доклады Эрики Голд, преподавателя судебной речи в Университете Хаддерсфилда «Коэффициент вероятности в науке о судебной речи. Текущая ситуация» и Штефана Эверта, профессора Университета Эрлангена — Нюрнберга «Статистическая значимость в литературной атрибуции авторства». В конце каждого доклада приводятся вопросы к выступающему и ответы на них. В первом (пятом) докладе рассматривается ситуация, когда сравниваются образец речи известного говорящего из числа подозреваемых и образцы речи неизвестного говорящего. Обычно сочетаются два метода: аудитивный (перцептивный) и акустический фонетический анализ; автоматическое (компьютерное) распознавание речи. Предлагается при атрибуции текста использовать коэффициент вероятности, выражающий степень соответствия между рассматриваемыми речевыми образцами. К недостаткам компьютерного анализа речи относятся произвольность выбора референтных групп, нехватка статистической информации, необходимость учитывать разные аспекты текстов (стилистический, канал коммуникации, качество записи и т. д.). Следующий доклад посвящен стилеметрии и основан на анализе художественных произведений. Релевантные для таких исследований параметры: длина предложения, длина слова, класс частотности; богатство словаря; синтаксическая сложность; орфография; выбор синонимов. Рассматриваются разные методы статистических подсчетов вплоть до дельты Бёрроуза. Формулируются потенциальные ограничения этих методов (влияние жанра на стиль; применимость к текстам малого объема; устойчивость к «шуму», например ошибкам автоматического распознавания текста).*

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

ТИП ПУБЛИКАЦИИ: *обзор*.

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 3) / С. И. Красса // Политическая лингвистика. — 2021. — № 2 (86). — С. 196-203. — DOI 10.12345/1999-2629_2021_02_19.

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 применяемых методик, 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 обзоров, опубликованных ранее (Политическая лингвистика. — 2020. — № 6 (84). — С. 174-187. — DOI 10.26170/pl20-06-19; 2021. — № 1 (85). — С. 159-166. — DOI 10.12345/1999-2629_2021_01_15).

ОБЗОР ДОКЛАДОВ УЧАСТНИКОВ
Э. ГОЛД

Пятым было выступление Эрики Голд (Erica Gold) — преподавателя судебной речи в Университете Хаддерсфилда (Huddersfield University) на тему «Коэффициент вероятности в науке о судебной речи. Текущая ситуация».

Судебное речеведение — это использование речи как доказательства на следствии и в суде. В основном к экспертам обращаются с просьбой сравнить речь говорящих (рассматриваемая в докладе версия автороведческого анализа). Обычно имеется образец речи известного говорящего из числа подозреваемых (например, допросы PACE — Police and Criminal Evidence Act) и образцы речи неизвестного говорящего. Типичные вопросы, задаваемые лицами, проводящими допрос: на записях представлены одни и те же или различные говорящие?

В основном используются два подхода (или их комбинация):

1) аудитивный (перцептивный) и акустический фонетический анализ;

2) автоматическое (компьютерное) распознавание речи (ASR).

Важно отметить, что те, кто делает автоматический анализ, не ограничиваются компьютерным анализом — всегда есть элементы анализа, выполненного человеком.

Методы в значительной степени зависят от стран, правил, требований и индивидуальных преференций. Подход в Соединенном Королевстве основывается на примате фонетической стороны, в то же время проводятся современные исследования ASR, проходящие тестирование в полиции.

Кодексы практики и поведения для судебных экспертов и специалистов-практиков в системе уголовного правосудия в Соединенном Королевстве и приложения о судебной речи и аудиосервисах (2014) не выделяют этот подход в качестве рекомендуемого, они намекают на возможность его использования, но явно этого не утверждают.

Автоматическое распознавание речи в мире

В последние годы увеличивается число случаев применения систем автоматического распознавания речи: 17 % (2011) и 41 % (2019) применения систем ASR в комбинации с экспертным анализом — в роли эксперта может выступать инженер или фонетист [Gold, French 2019 in print]. Рост числа обращений к системам распознавания речи наблюдается как в коммерческом секторе, так и в банковской сфере, страховании и уголовном расследовании, в сфере национальной безопасности и судебной практике [Gold, French 2019 in print].

Фонетический метод оказывает влияние на выводы, которые делаются в заключении исследования. В докладе приводится обзор формулировок выводов на основании данных 2019 и 2011 г.

1. Бинарный ответ: да, это один человек / нет, это разные люди (5,3 % — 5,7 %).
2. Классическая вероятность: указывается, насколько вероятны две позиции с точки зрения статистики (28 % — 40 %).
3. Изложение позиции в Соединенном Королевстве: где-то посередине между байесовской моделью и свободным контекстом (13,2 % — 31,4 %).
4. Заявление в поддержку: вербальное выражение соотношения вероятности, очень близкое байесовской модели (27,2 % — 9 %).
5. Математическое соотношение вероятности: выражение соотношения в цифрах, что легко интерпретируется судьями и присяжными и используется в различных судебных экспертизах (13,2 % — 8,6 %).

В Соединенном Королевстве наблюдается значительное падение частотности изложения позиции. Наибольший рост демонстрируют формулировки, основанные на заявлении в поддержку. И это свидетельствует о движении в сторону мышления в рамках байесовской модели.

Коэффициент вероятности

В течение ряда лет наблюдается рост понимания и принятия байесовской логики в различных областях. Подтверждением этому могут служить рекомендации о том, чтобы коэффициент сходства был адаптирован для выводов по расследованию у криминалистов разного профиля [Sacks, Koehler 2005].

Высказываются мнения в пользу того, чтобы все выводы относительно сравнения говорящих формулировались с использованием LR. Эти представления широко и охотно защищаются с 2009 г. [Rose, Morrison 2009]. Многие другие исследователи также работают над LR, используя его. Докладчик считает это парадигмальным изменением. С 2009 г. наблюдается всплеск исследований по коэффициенту вероятности.

Что понимается под коэффициентом вероятности (Likelihood Ratios — LR)? Он выражает результаты поиска степени соответствия между образцами, принадлежащими, по мнению обвинения, одному и тому же говорящему, и образцами, которые, по предположению, принадлежат разным говорящим.

Докладчик приводит соотношения между данными исследования и данными по материалам уголовного дела.

Исследование	Изучение материалов уголовного дела
Данные одного параметра многих токенов из множества говорящих	В идеальном случае данные многих параметров и многих токенов из записей подозреваемых и преступников
Множество говорящих группируется по полу и акценту с говорящими в данных следствия	Понимание того, что параметры речи соотносятся друг с другом
В идеале — еще один набор говорящих из того же множества (для калибровки)	Референтная группа, соответствующая по параметрам акцента и пола говорящим в записях уголовного дела

Трудности

1. Ограничения в референтных группах (размер, состав и т. д.)
   1.1. Каким образом выбирается референтная группа?
   1.2. Кто выбирает референтную группу?
   1.3. Насколько большой и специфичной является референтная группа?

В целом подбираются соответствия по полу и акценту (см. работы Винсента Хьюза (Vincent Hughes) по этой проблеме).

2. Недостаточные статистические данные о населении.
   2.1. Сбор статистических данных о населении — занятие очень затратное по времени и дорогостоящее.
   2.2. Как результат, отсутствуют доступные данные по статистике населения для помощи в подсчетах LR.
   2.3. Предлагаются различные решения по использованию данных по референтным группам, собранных на базе индивидуальных случаев.
   2.4. Использование готовых баз данных [Rose 2007], например больших полицейских баз.
3. Естественная сложность речевых данных (что ведет к проблеме статистического моделирования / вычисления).

Необходимо учитывать такие факторы, как жанр, вариативность данных, возраст, пол — все эти параметры оказывают влияние на LR. Эту трудность можно обозначить как многоуровневый характер материала, подвергаемого анализу: сравниваемые тексты различаются по стилю, различается характер данных (протяженный или дискретный), их многоуровневая природа, не совпадают каналы коммуникации, собеседники, тексты записываются в разное время, различается качество записи и др.

Заявление о поддержке

В настоящее время используется всеми (или их большинством) практическими работниками в Соединенном Королевстве. Введено Французской ассоциацией в январе 2015 г. Заимствовано из Европейской сети института криминалистики (ENFSI).

Заключение будет иметь краткую часть, которая похожа на вербальное представление коэффициента вероятности, но конечное утверждение в дальнейшем упрощено.

Прогресс

Хотя в Соединенном Королевстве (и в большинстве стран) не используют цифровое выражение LR, имеет место продвижение от индивидуального анализа в сторону интерпретации данных по делу.

Пока процентное соотношение с общими статистическими данными по населению в попытке принятия решения остается относительно стабильным (около 70 %), хотя количество данных и возросло, а анализ показывает большее разнообразие в корпусе данных и статистике. Докладчик предположил, что со временем решение будет становиться все более надежным. Полностью цифровое решение LR относится к весьма и весьма отдаленному будущему, поскольку исследователи постепенно осознают проблемы, которые ощутимо влияют на решение поставленной задачи.

Современные исследования

Имеет значение общая стабильность референтной группы населения (пример: WYRED — West Yorkshire Regional English Database) [Gold, Ross, Earnshaw 2018]. Как много токенов нужно для определения конкретной референтной группы населения в терминах различных фонетических параметров?

Токены / население

Количественные параметры, определяющие группу населения, даются в следующих работах: V. Hughes et al. (2013), V. Hughes (2017), Hughes & P. Foulkes (2014, 2015), Wang et al (2019) [Hughes, Haddican, Foulkes, Richards 2013; Hughes 2017; Hughes, Foulkes 2014, 2015]. Рассматриваются эффекты образцов применения предложенных параметров. Описывается соотношение факторов «собеседник / стиль / эффекты». Это разновидность парадигмального сдвига (оценка параметров, поиск «золотого» параметра, лучших параметров).

Криминалистический анализ авторства

Это не является специализацией докладчика, но он предполагает, что имеются сходства в развитии описанных процедур с ситуацией в криминалистическом сопоставлении образцов речи: «Мы собираемся отделить морфологический подход, приспо-

собление морфологического подхода. Мы стоим перед теми же сходствами, возможностями, я бы хотела сказать, препятствиями. И я считаю, для нас то, что изменилось за последние десять лет в лучшую сторону, послужило прогрессу».

Обзор делался с целью понять, на чем специалисты фокусируют свое внимание и что они на самом деле делают. По мнению готовивших доклад, должно быть меньше секретности в отношении того, что происходит, и того, на чем построены исследования. Наибольшее значение имеет время, поскольку новые люди приходят в эту область, стараются сделать что-то новое и просто развивают ее. Будущее сотрудничества между атрибуцией автора и сравнением речи состоит в использовании одних и тех же баз данных. Конечно, жанровые различия делают сопоставляемые тексты разными, но научный коллектив, на разработках которого основан доклад, собрал данные о сотнях говорящих, транскрибировал их, некоторые из них сделал доступными онлайн. Есть часы и часы записанных данных. Конечно, следует принимать во внимание языковые трудности и все, что с ними связано.

Вопросы по докладу

Вопрос. Я заметил, что вы упомянули «Morrison's Interpol Survey for Forensic Speaker Comparison». Они используют полностью автоматизированное распознавание речи, устраняя таким образом эксперта-человека из процесса анализа?

Ответ. Да, мы упоминали эту работу в кратком обзоре, но ее выводы зависят от данных, на которых покров секретности. Я думаю, некоторые так делают в части Америки и в Азии. Да, мы даем отсылку к этой статье.

Ответ. Я считаю, что три года назад была большая дискуссия о том, как объединить эти два вида доказательств. Теперь есть люди, которые используют множественный подход, некоторые используют эксперта, другие — подход, основанный на группах населения. И мнение складывается в основном по отношению к эксперту, который применяет тот или иной подход. Такие страны, как Швеция, используют множественный подход и затем предлагают систему того, как это сделать. В Соединенном Королевстве нет совершенного пути того, как этого достичь. Мое мнение — предложить многоуровневый подход. Если вы делаете какую-то часть машиной, если вы делаете другую часть более целостно, вы осуществляет оба подхода. В идеале вы двигаетесь в одном и том же направлении, если нет — суды воспринимают больше тот или иной подход, в любом случае вы делаете одну и тут же работу.

Ш. ЭВЕРТ

Шестым было выступление Штефана Эверта (Stefan Evert) — профессора Университета Эрглангена — Нюрнберга (Erlangen-Nuremberg) на тему «Статистическая значимость в литературной атрибуции авторства».

Докладчик рассказал, что занимается литературной стилеметрией, совместно с коллегами из Центра гуманитарных исследований (Humanity Research Center) проводя исследования в отношении ряда художественных произведений. Данный подход заметно отличается от доминирующего в судебном автороведении. Основная цель литературного автороведческого анализа обычно лежит в области стилеметрии, идентификации авторского стиля — «отпечатков пальцев» автора, описание которых основано на количественных параметрах. Это традиционные черты, характеризующие автора:

– длина предложения, длина слова, класс частотности;
– богатство словаря (тип — токен);
– синтаксическая сложность;
– частотность служебных слов, синтаксических структур;
– орфография;
– выбор синонимов.

Многие из этих характеристик пересекаются (по крайней мере эмпирически) с манчестерскими вариациями.

Докладчик представил заслуживающие внимание, по его мнению, примеры применения литературной атрибуции автора:

1. Публикация в «Федералистских документах» [Mosteller, Wallace 1963] — один из наиболее ранних объектов исследования в авторской атрибуции текста.
2. Кто написал письмо Биксби (Bixby)? [Grieve et al 2018].
3. Существовал ли на самом деле Шекспир? (Thisteal and Efton, 1987) — попытка установить, кто был автором произведений, подписанных именем Шекспира.

Метод определения автора может быть применим к разным языкам, например в авторской атрибуции поэзии на средневерхненемецком:

1) поэзия на Middle High German (Dimpel 2018);
2) роман, изданный под псевдонимом Роберт Гелбрейт (Robert Galbrieth), — «The Cuckoo's Calling», — написанный Д. Роулинг (J. K. Rowling).

Названные случаи интересны не только сами по себе, как примеры выявления ав-

торства, но как представляющие «отпечатки пальцев» автора в авторoведческом анализе. Мы можем обучать модели или тестировать их, выделяя параметры в задачах определения авторства. И, предположительно, эти характеристики хорошо зарекомендовали себя в решении типичных задач по установлению авторства, т. е. отличают стиль автора от стиля других авторов.

Стилеметрия и атрибуция авторства

Если рассматривать атрибуцию авторства в более широкой перспективе, можно увидеть два основных подхода, которые позволяют осуществлять ее различными способами. В криминалистике классификационной подход — это естественный выбор.

Атрибуция авторства как классификационная задача включает в себя:

– закрытый список кандидатур в авторы неизвестного текста;

– учебный набор текстов с известным авторством;

– контролируемый алгоритм машинного обучения (глубокий, если необходимо);

– оценку классификационной точности.

Некоторые из методов являются вероятностными, тогда есть указания на то, насколько они точны в решении поставленной задачи.

Атрибуция авторства как задача кластеризации включает:

– данный набор неизвестных текстов;

– группировку текстов, написанных одним и тем же автором, в кластер;

– измерение текстового сходства и алгоритм кластеризации;

– оценку скорректированного индекса Рэнда (Adjusted Rand Index — ARI), который основан на том, сколько текстов верно приписано в ходе рандомной кластеризации.

Кластеризация в более общем подходе

Если вы проводите успешную кластеризацию, это, конечно, дает возможность взглянуть на кластер, которому принадлежит спорный текст, и приписать его большинством голосов к этому кластеру. Преимущество заключается в том, что имеется возможность его автоматического использования.

Далее автор попытался очень кратко представить историю авторской атрибуции. Одна из наиболее ранних работ, опубликованных в начале 60-х годов, «Сопоставление частотных слов» [Mosteller, Wallace 1963]. Работа основана на простом сопоставлении частоты слова из спорного текста и частоты этого же слова в тексте возможного автора. Mosteller и Wallace предложили основанный на усовершенствовании байесовской модели подход; коэффициент подобия известен, поскольку они исходили из статистического основания.

Много позже появилась работа «Машинное обучение на основе широкого использования стилеметрических характеристик» (Juola, Stamatatos, 2009). В ней рассматривается большое количество лексических, синтаксических, семантических, лексико-статистических и буквенных характеристик. Для выбора наиболее информативных характеристик используется ML-алгоритм.

Метод Delta [Burrows 2002] — простое сравнение частотности нескольких сотен общеупотребительных слов, по существу метод, который предложили Mosteller и Wallace, без сопоставления. Поскольку никто реально не рассматривал дельта-метод в деталях, остановимся вкратце на том, как работает «дельта». Идея, на которой основывается «дельта», заключается в том, что наиболее частотные слова — MFW (100—5000), формируют «отпечатки пальцев» авторского стиля и таким образом позволяют идентифицировать автора. Как мы определяем наиболее частотные слова? В полной коллекции текстов, подлежащих анализу. Исследователи идут от наиболее частотных служебных слов к полнозначным словам средней частотности.

Основная единица измерения — это вектор относительной частотности данных слов в английских романах. Это очень простой метод, и он может быть применен к любым видам языковых данных. Наиболее частотен определенный артикль (частотность 5,1 %), неопределенный артикль (2,7 %), *was* (2,7 %) и т. д. Как можно заметить, метод даже не требует применения каких-либо лингвистических процессоров к данным, которые представлены в необработанном виде (слова в оригинальном написании) с указанием частоты, с которой они употребляются в тексте. Сопоставление этих векторов показывает, насколько сходны сравниваемые тексты.

Одна из ключевых проблем авторoведческого анализа — это применение дельта-метода к характеристикам, которые сложно интерпретировать, например употребление неопределенных артиклей в романе Томаса Гарди «Вдали от обезумевшей толпы» («Far from the Madding Crowd», Thomas Hardy). Что значат выявленные характеристики? Является ли это частью стиля Томаса Гарди или стиля этой конкретной книги?

Конечно, это не касается случаев, когда мы видим указанные с высокой точностью употребления специфические конструкции. Для «Вдали от обезумевшей толпы» это, согласно конкордансу слова *few*, выражения

for a few days, *for a day or two*, а также *a man (a woman) of: a man of misty views, of spirit, a woman of gardening*. Все эти конструкции вносят вклад в употребление неопределенного артикля.

Основание данной работы, вероятно, многовариантное. Мы должны рассматривать лежащие в основе характеристики, которые неоднозначны, но в любом многоаспектном исследовании есть возможность реконструировать данные на основе коррелирующих данных в дистрибуции единиц. Конечно, в атрибуции авторства с помощью 12 слов не достичь успеха. Докладчик представил визуализации 150 наиболее частотных слов в романе «Вдали от обезумевшей толпы». Этот график исключает наиболее частотные и наименее частотные слова. Основной интерес представляют не они, а каждая отдельная единица, которая значительно превосходит по частоте употребления в романе другие или, наоборот, частотность которой значительно ниже средней.

Рассматривается средняя частотность каждого слова во всем тексте и затем сравнивается с относительной частотой. Сравниваются отклонения от средних значений — выше или ниже средней частоты. Затем можно построить относительно четкий график с пиками выше или ниже средней частоты. На изображении хорошего качества и масштаба можно заметить различия между двумя рассматриваемыми романами в пиках частот употребления слов, представленных зеленоватыми или красноватыми фрагментами графика.

Если посмотреть на график доминирующих в употреблении нескольких слов, то мы видим значительное варьирование в частоте. Далее проводится нормализация употребления для выявления стандартного отклонения, и затем разделяются характеристики по стандартному отклонению z-scores.

$$z_i(\Delta) = \frac{f1(\Delta) - Mi}{\delta i}$$

Мы видим, что каждая черта, каждое свойство, выбранные на математических основаниях, вносит вклад в информацию относительно различия профилей автора данных текстов.

Если теперь сопоставить данные относительно «Оливера Твиста» и «Вдали от обезумевшей толпы», зеленые и красные пики становятся более явными показателями, демонстрируя бóльшие различия.

Последний компонент, который необходим, чтобы завершить исследование методом Δ, — математически сопоставимые величины этих векторов, не только на уровне визуализации, поскольку математическое сопоставление дает возможность сравнить цифровые дистрибуции.

Используется геометрический подход для измерения расстояния между этими двумя пунктами. Это манхэттенская метрика, и докладчик применяет манхэттенское расстояние для измерения.

Несколькими годами позже после публикации дельта-метода Шлю Агемон опубликовал математическое исследование дельты Бёрроуза (Burrows) и в основном объяснил несоответствия с математической точки зрения, поскольку, если мы определяем манхэттенское расстояние в качестве вероятного распределения, то компьютерные оценки не соответствуют друг другу. Другая формула позволяет использовать метод расстояний, совместимый со стандартизацией.

Финальная версия метода содержит множество вариаций «дельты», позволяет провести исследование вариаций и содержит новую опцию — угол расстояния, поскольку это весьма популярный способ в извлечении информации, который полезно применять. Здесь расстояние — это угол между двумя точками.

Иерархическая кластеризация на основе Δ проводится в малых множествах. На больших множествах ситуация несколько более сложная.

Некоторые типичные исследовательские вопросы, которые задают в связи с использованием Δ:

1. Может ли Δ точно идентифицировать автора текста?
2. Как нужно исследовать MFW (наиболее частотные слова)?
3. В чем отличие Δ от других параметров?
4. Какие слова формируют характерные «отпечатки пальцев» автора?
5. Выступают ли эти «отпечатки» вместе с другими сигналами (например, жанром)?
6. Можем ли мы применить Δ к коротким текстам?
7. Насколько устойчива Δ к «шуму» в массиве данных (например, ошибки OCR)?
8. Почему Δ работает хорошо?

9 (вопрос, задающийся намного реже). Являются ли результаты статистически значимыми?

Эмпирическая оценка

«Теоретически теория и практика одно и то же. Практически — нет».

База данных исследования — романы XIX в. [Jannidis et al. 2018].

Также использовались три литературных корпуса (немецкий, английский, французский). В каждый корпус включалось 25 авторов, от каждого из которых бралось по 3 романа (всего 75 романов). Временной диапазон — от начала XIX в. до середины XX сто-

летия.

Получено 10 млн токенов для каждого языка. Конкретный текст — 25 тыс. токенов.

Докладчик провел сопоставление разных видов Δ, сопоставил метод Δ и Ngram в отношении длинных и коротких текстов.

Вопросы по докладу

Вопрос. Существует ряд исследовательских вопросов, поставленных вами эксплицитно. Число авторов, рассмотренных вами, фиксировано. Я хочу спросить: влияет ли объем данных на разрешающую способность метода? Если вы выйдете за рамки этих авторов, будете ли вы иметь ряд дистрибуций одного и того же вида?

Ответ. Я полагаю, что следующим вопросом будет, станут ли результаты такого эксперимента лучшими? 25 авторов показывают хороший результат. Если авторы, не входящие в этот список, не покажут худший результат, то измерения «дельта» полностью надежны.

Вопрос. Рассматриваете ли вы «дельту» в качестве потенциально надежного инструмента при большом числе измерений без потери надежности?

Ответ. Я не уверен, насколько это транслируется на авторов. Чем больше текст, тем большее число идиосинкретичных слов, которые будут специфичны для романов, и это более «топикальная» информация.

Вопрос. Судя по вашим объяснениям, некоторая «дельта» более устойчива.

Ответ. Да, 90 %, и поэтому мы полагаем, что это совершенно надежный метод. Если я смогу его расширить, то интересно будет представить финальную надежную «дельту». Да, мы имеем надежный метод. Мы также применяем модификации, которые увеличат его надежность. Он валидизирован этими экспериментами. Это не такой мягкий метод, каким он может показаться.

Вопрос. Спасибо за выступление. Я недавно прочитал статью о дисперсии. Я анализировал перевод, переводной текст. Я анализировал его главы. Мне интересно, как вы планировали делать образцы. Потому что каждая глава демонстрирует вариации в отношении лексической плотности в значительной мере. Вы разрезали текст или как-то иначе формировали образцы? Это «мешок слов»?

Ответ. Я думаю, это «мешок слов», поскольку другие подходы нереалистичны. Это хорошо, поскольку отображаются основные вариации образцов романа, систематических изменений романа. Нам нужна тонко настроенная «дельта». Она должна работать, несмотря на стилистические различия. Что интересно, так это возможность соединения «дельты», Ngram и наиболее частотных слов. Вы всегда можете сказать, что большинство слов с данной характеристикой представляет лучшие результаты, чем один набор характеристик.

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 3)

ABSTRACT. *The article provides an overview of the reports of the 1st Roundtable on Practices and Standards in Forensic Authorship Analysis held by the International Association of Forensic Linguistics and the Centre for Digital Humanities at the University of Manchester on May 15, 2019. It includes the reports by Erica Gold, Lecturer on forensic speech at the University of Huddersfield "Probability coefficient in the Forensic Speech Science: Current Situation" and Stefan Evert, Professor of the University of Erlangen-Nuremberg "Statistical Significance in the Literary Authorship Attribution". At the end of each report, there are questions to the speaker and answers to them. The first (fifth) report examines a situation where a sample of the speech of a known speaker from among the suspects and samples of the speech of an unknown speaker are compared. Usually two methods are combined: auditory (perceptual) and acoustic phonetic analysis; automatic (computer) speech recognition. It is proposed to use a probability coefficient when attributing the text, which expresses the degree of correspondence between the speech samples under consideration. The disadvantages of computer speech analysis include the arbitrary choice of reference groups, the lack of statistical information, and the need to take into account different aspects of texts (stylistic, communication channel, recording quality, etc.). The next report is devoted to stylometry and is based on the analysis of works of art. Relevant parameters for such studies include: sentence length, word length, frequency class; vocabulary richness; syntactic complexity; spelling; choice of synonyms. Different methods of statistical calculations including the Burrows's delta are considered. The potential limitations of these methods are formulated (the influence of genre on style; applicability to small-volume texts; resistance to "noise", for example, errors in automatic text recognition).*

KEYWORDS: *round table; forensic authorship analysis; author profiling; stylometry; computer linguistics; core features of the style; forensic linguistics; forensic expertise; automatic text recognition.*

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 3) / S. I. Krassa // Political Linguistics. — 2021. — No 2 (86). — P. 196—203. — DOI 10.12345/1999-2629_2021_02_19.

REFERENCES

1. Burrows, J. F. 'Delta': a Measure of Stylistic Difference and a Guide to Likely Autorship / J. F. Burrows. — Text : unmediated // Literary and Linguistic Computing. — 2002. — Vol. 17 (3). — P. 267—287.

2. 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.

3. Forensic Linguistics Roundtable Event / International Asso¬cia¬tion 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.

4. French, P. Comparing apples with apples, apples with oranges and apples with apples and oranges: the effects of (mis)matching reference population accents in ASR speaker comparisons : presentation / P. French, P. Harrison, V. Hughes, D. Watt, C. Llamas, A. Braun. — Text : unmediated // 27th Annual Conference of the International Association of Forensic Phonetics and Acoustics, University of Huddersfield, UK, 29 July — 1 August 2018.

5. Gold, E. International practices in forensic speaker comparisons: second survey / Erica Gold, Peter French. — DOI 10.1558/ijsll.38028. — Text : unmediated // International Journal of Speech Language and the Law. — 2019. — Vol. 26, No 1. — P. 1—20.

6. Gold, E. The 'West Yorkshire Regional English Database': investigations into the generalizability of reference populations for forensic speaker comparison casework / E. Gold, S. Ross, K. Earnshaw. — DOI 10.21437/interspeech.2018-65. — Text : unmediated // Proceedings of Interspeech. — Hyderabad, India, 2018. — P. 2748—2752.

7. Grieve, J. Attributing the Bixby Letter using N-gram Tracing / Jack Grieve, Isobelle Clarke, Emily Chiang, Hannah Gideon, Annina Heini, Andrea Nini, Emily Waibel. — DOI 10.1093/llc/fqy042. — Text : unmediated // Digital Scholarship in the Humanities. — 2018. — Vol. 34 (3). — P. 493—512.

8. Hughes, V. Interaction of social and linguistic constraints on two vowel changes in northern England / V. Hughes, B. Haddican, P. Foulkes, H. Richards. — Text : unmediated // Language Variation and Change. — 2013. — No 25 (3). — P. 371—403.

9. Hughes, V. Sample size and the multivariate kernel density likelihood ratio: how many speakers are enough? / V. Hughes. — Text : unmediated // Speech Communication. — 2013. — Vol. 94. — P. 15—29.

10. Hughes, V. The relevant population in forensic voice comparison: Effects of varying delimitations of social class and age / V. Hughes, P. Foulkes. — Text : unmediated // Speech Communication. — 2015. — Vol. 66. — P. 218—230.

11. Hughes, V. Variability in analyst decisions during the computation of numerical likelihood ratios / V. Hughes, P. Foulkes. — Text : unmediated // Int. J. Speech Lang. Law. — 2014. — Vol. 21 (2). — P. 279—315.

12. Jannidis, F. The Shape of Data in the Digital Humanities: Modeling Texts and Text-based Resources / F. Jannidis, J. Flanders (eds.). — London : Routledge, 2018. — 360 p. — Text : unmediated.

13. Juola, P. Authorship attribution / P. Juola. — Text : unmediated // Foundations and Trends in Information Retrieval. — 2006. — Vol. 1 (3). — P. 233—334.

14. Mosteller, F. Inference in an Autorship Problem. A Comparative Study of Discrimination Methods Applied to the Autorship of the Disputed / F. Mosteller, D. L. Wallace. — Text : unmediated // Federal Papers. Journal of the American Statistical Association. — 1963. — Vol. 58 (302). — P. 275—309.

15. Rose, P. A response to the UK Position Statement on forensic speaker comparison / P. Rose, G. S. Morrison. — DOI 10.1558/ijsll.v16i1.139. — Text : unmediated // International Journal of Speech, Language and the Law. — 2009. — Vol. 16 (1). — P. 139—163.

16. Rose, P. Forensic speaker discrimination with Australian English vowel acoustics / P. Rose. — Text : electronic // ICPhs. XVI (Saarbrücken, Germany, 6—10 August 2007). — ID 1339. — P. 1817—1820.

17. Sacks, M. J. The Coming Paradigm Shift in Forensic Science / Michael J. Sacks, Jonathan J. Koehler. — DOI 10.1126/science.1111565. — Text : unmediated // Science. — 2005. — Vol. 309 (5736). — P. 892—895.

18. Stamatatos, E. A survey of modern authorship attribution methods / E. Stamatatos. — Text : unmediated // Journal of the American Society for Information Science and Technology. — 2009. — Vol. 60 (3). — P. 538—556.