



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 4)

АННОТАЦИЯ. Проводится обзор заключительной дискуссии по итогам 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. Еще раз обсуждаются актуальные темы, поднимавшиеся во время круглого стола: важность учета частотности служебных слов по сравнению со знаменательными и влияние на частотность служебных слов различных факторов (стиль, жанр текста, авторский идиостиль, коммуникативная стратегия), регулирование судебными властями процедур судебной лингвистической экспертизы, валидизация практически используемых методов автороведческого анализа, роль формализованных методик и интуиции исследователя при анализе, научные и юридические аспекты использования массива доступных в Интернете данных (включая видеосообщения и видеопереговоры), обращение к наработкам корпусной лингвистики, составление узкоспециализированных языковых корпусов, выбор компьютерных программ для анализа текстовых массивов данных, применение автоматического распознавания речи, взаимодействие ручного и автоматизированного компьютерного анализа, приоритетность буквенных или словесных Ngram в зависимости от исследуемого материала. В ходе обсуждений вопросы задавали как докладчики и присутствовавшие на круглом столе, так и специалисты, наблюдавшие за мероприятием в онлайн-режиме. В заключительном слове ведущий выразил намерение организовывать круглый стол ежегодно для своевременного обсуждения новых идей и достижений в области судебного автороведческого анализа.

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

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 4) / С. И. Красса // Политическая лингвистика. — 2021. — № 3 (87). — С. 227-234. — DOI 10.26170/1999-2629_2021_03_24.

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 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 завершением 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; № 2 (86). — С. 196-203. — DOI 10.12345/1999-2629_2021_02_19).

ЗАКЛЮЧИТЕЛЬНАЯ ДИСКУССИЯ

Вопрос. Вопрос Штефану и, возможно, также Чаку о служебных словах. Судя по всему, для вас они обладают особым статусом. Они не подвержены осознанному контролю. Но я бы сказал, что служебные слова не существуют в вакууме, на них влияют многие параметры: тема, жанр, все характеристики, связанные с вариациями. И среди наиболее частотных слов большинство со-

ставляют именно служебные. Похожая проблема — отнесение слов к служебным. И близкий вопрос Штефану о выявленной вами закономерности: идиолектное варьирование в большей степени важно, чем тематическое, жанровое варьирование в романе, хотя некоторые из авторов пишут на определенные темы? Я просто хочу выяснить, в какой мере тематическое варьирование можно считать более «авторским», чем жанровым.

Ответ (Эверт). Это несколько вопросов. Прошу прощения за возвращение к прежнему слайду, поскольку краткий ответ на ваш вопрос будет таким: «Вы совершенно правы». Я не убежден, что это служебные (function) слова. Когда мы посмотрим на вершину вектора, то мы увидим, что в первых сотнях (200) наиболее частотных слов будут служебные. Если мы посмотрим на первые 2000 или 5000 слов, то, конечно же, среди них будет много знаменательных. Один из способов, который мы используем, — элиминация признака, выбор свойств. То есть мы выбираем много служебных, но также и много знаменательных слов, встречающихся в тексте. Мы даже замечаем, что служебные слова не совпадают с сигнальными, поскольку в основном служебные слова в данной части кривой показывают надежный результат, тогда как знаменательные создают шум. В диапазоне 200 слов служебные слова показывают надежный результат. Служебные слова более «шумные», чем лучшие параметры.

Ответ (Стаматос). Для меня ясно, что служебные слова также подвержены влиянию темы. Представленные эксперименты, возможно, показали, что некоторые кластеры (поджанры) документов в большей степени обуславливают частотность тех или иных слов, чем авторский стиль, а также тема. Определенные авторы предпочитают писать на определенные темы. Поэтому кластеризация документов — это в значительной мере кластеризация документов сходной, а не различной тематики. И с учетом авторского стиля, конечно же. С другой стороны, в ваших экспериментах используется CUSUM-сходство. Этот параметр подвержен влиянию больших цифр, и одно из значений — около нуля, если оно вообще поддается исчислению. Я думаю, CUSUM-сходство более подвержено влиянию служебных слов — или же этот параметр имеет более высокие значения? (Мы просчитывали на заданных значениях.) Понятно, вы правы. Таким образом, нет различий в шкале значений.

Реплика (Грив). Я не знаю, насколько обсуждалось сознательное/неосознанное ис-

пользование знаменательных слов. Я думаю, что преимуществом служебных слов является их высокая частотность и меньшее колебание в связи с тематикой текста, чем у знаменательных слов.

Вопрос (Грант). Мой вопрос направлен непосредственно на то, что говорил Джек. При автороведческом анализе мы в большей мере занимаемся стилевым анализом, чем социолингвистическим. Я думаю, Джек, вы продемонстрировали географическое варьирование в том, что могло быть стилевым измерением. Я думаю, что произношение на Глубоком Юге США в большей мере индивидуальное явление, чем групповое. О чем я думаю, так это о том, что в понятие языковой личности мы вкладываем коммуникативные стратегии, которые могут быть варьированием прагматических стратегий. В эту модель можно включить оба понятия, чтобы показать, как индивидуум использует коммуникативные стратегии.

Ответ (Грив). Это относится к социальной истории, социальной истории стиля, социальной истории бэкграунда. Мы находим один стиль более нарративным, более неформальным в твитах. Социолингвистический подход представляется более произвольным, чем, например, фонологический.

Вопрос (Кларк). Насколько коммуникативная стратегия функционально соответствует стилю?

Ответ (Эверт). Для сообщений на интернет-форумах, а также твитов, это соответствие релевантно. Несколько меньше — в романе, поскольку мы много работали с данными о романах и не видели этой связи в своем материале. Ремарки также важны в профилях «дельта». При анализе следует учитывать совокупность характеристик: черты индивидуального автора, стиля или жанра, более тщательного контроля вариаций, также черты стиля переводчика (если это перевод). Весьма интересная проблема, чьи сигналы более сильные — переводчика или автора. Имеются сигналы определенной сферы и так далее.

Ответ (Грант). Различные коммуникативные стратегии — это свидетельство стилевого варьирования и особенно социолингвистической идентичности. Мой исходный пункт: если в любом жанре или письменном тексте у вас есть выбор коммуникативных стратегий, и данные показывают, что можно применять более богатый набор коммуникативных стратегий — это предоставляет возможность для выражения идентичности в выборе того или иного варианта. Каждый пункт варьирования в использовании языка открывает доступ к выражению идентично-

сти. Поэтому если бы я захотел выразить свою идентичность как представителя Глубокого Юга США (сознательно или неосознанно), я бы стремился воспроизвести черты, типичные для стиля, я полагаю. Поэтому если говорить о газетном колумнисте, как я думаю, можно наблюдать вариации стилевого регистра — нарратива, которые также будут служить выражением идентичности.

Ответ (Грис). Это действительно сложный вопрос. Я думаю, вы говорите о чем-то типа мотивации, выбора структуры варьирования с целью сделать текст более нарративным. Мы могли бы подумать, каким образом конструирование делает текст более нарративным, ваш аргумент — более убедительным. Коммуникация — это выбор наиболее верных слов в данной ситуации.

Комментарий. Я хочу сделать небольшой комментарий — не как председатель, а от себя лично. Проблема регистра, взаимодействия между регистром и социолингвистикой имеет в качестве теоретического базиса так называемую «аксиому регистра», в функциональной лингвистике это варьирование кода и семантическое варьирование. И при любой дискуссии это нужно иметь в виду.

Комментарий (Райт). Я только хочу отметить, что необходимо опираться на текст: например, монологический текст, такой как редакционная или научная статья, возможно, более ограничен регистром, чем диалогические тексты или электронные письма. Также тип регистра, которым мы занимаемся, может быть более релевантен для социолингвистической истории. Особенно если он определяется при обращении к прошлым образцам, которые использовали ранее.

Комментарий. То, что Джек понимает под функциональным подходом, очень и очень интересно. В отношении того, что сказал Дэвид, я не думаю, что подходы, необходимые для анализа, в сущности, эксклюзивные. Это замечательная дискуссия.

Вопрос. У меня вопрос ко всем выступающим сегодня. Я не очень понял, кто определяет исходные положения дискуссии. Тим, вы четко говорите о том, что служебные слова коррелируют с содержанием, стиль и значение идут рука об руку. И когда мы рассматриваем первую тысячу слов, это очень важно. Джек говорит, что служебные слова связаны с регистром. Я просто хочу узнать, какова общая концепция: стиль, форма, содержание зависят друг от друга или нет? Какова ваша личная позиция по этому вопросу?

Ответ (Эверт). Это большая проблема в гуманитарных науках с акцентом на языко-

вую сложность — синтаксическую и семантическую, лексическую в романах, требующая проведения множества различных измерений, в том числе с использованием серьезного математического аппарата, но нас вовсе не пугает эта сложность квантификации, и в то же время не ясно, как количественные показатели связаны с интуицией. У нас есть интуитивное представление о том, что некоторые романы более сложные, чем другие, но мы не располагаем ясной и операциональной дефиницией этого явления. То же относится к сопряжению количественной составляющей метода с качественной — интуицией, которой мы располагаем. Это составляет большую часть еще не выполненной работы. Вопрос состоит в том, что такое стиль, как может быть определен авторский стиль таким образом, чтобы выявлялась его связь с количественными данными.

Ответ (Стаматос). Да, главный вопрос заключается в том, как измерить стиль. Служебные слова представляют хороший, легкий путь получить набор измерений. Вы знаете, что эти измерения коррелируют с синтаксическими структурами, которые автор предпочитает использовать. Ввиду того что синтаксический анализ весьма проблематичен, вы предпочитаете измерять употребление служебных слов в качестве непосредственного шага в продвижении к информативному синтаксическому анализу. Как я отмечал ранее, эти измерения позволяют получить определенные ценные данные, мы знаем, что они также коррелируют с темой и жанром. Но это путь работы с документами, более легкий по сравнению с другими, которые, конечно, представляют собой более сложные проекты. Служебные слова не подвержены влиянию других факторов, таких как тема.

Вопрос (Креденс). Один факт о служебных словах. Мне в значительной мере нравится положение о том, что служебные слова — это суперпозиции смешанных сигналов. И мне хорошо известно использование Диккенсом фразы *from heads to feet* 70 раз с различными целями для описания чье-либо эмоционального состояния, как делают повествователи в нарративных контекстах. Это служит примером, как подобные выражения могут вносить свой вклад в частотность, и подталкивает дифференцированно рассматривать контексты с частотными служебными словами. Полагаю, при таком описании служебные слова — суперпозиция для многих типов сигналов.

Вопрос. Я думаю, это вопрос для всей аудитории. Как судебный регулятор в Со-

единенном Королевстве оказывает влияние в области судебной лингвистики? И как судебная лингвистика справляется с этими проблемами — требованиями для стандартов, валидизацией методов?

Ответ (Грант). В настоящее время судебный регулятор не оказывает никакого влияния на судебную лингвистику. И напротив, мы наблюдаем большой интерес к судебной речи. У нас была конференция, где также выступала Эрика. И мы были в значительной мере проигнорированы. Мы составляем небольшое сообщество, и мы больше похожи на судебно-экспертных исследователей. Мы хотели донести до участников мысль о том, что работы по авторской атрибуции должны быть отнесены к нашей области — в плане валидизации методов и т. д. Но сейчас эта концепция не очень влиятельна. Я бы хотел продолжать исследования в этом направлении, и некоторая работа ведется. Хотелось бы узнать, кто со мной согласен, а кто нет.

Реплика (Креденс). Вы употребили слово «валидизация». За ним стоит большая проблема, перед лицом которой мы находимся. Начиная с того, с чем мы имеем дело. Текст, письменный текст — это комплексное явление, существующее в условиях сложного социального взаимодействия. Как было отмечено сегодня, непросто проследить, насколько много переменных проявляется в этом взаимодействии. Да, давайте валидизировать, но что именно подвергать валидизации? Иными словами, если я выступаю в суде, какой корпус нам необходим? В традиционной социолингвистической парадигме нам нужен корпус из более чем 40 оригинальных писем. Это корпус, достаточный для валидизации, не так ли? Но очевидно, что мы знаем много больше. Нам не нужен этот вид корпуса, нам определенно нужен корпус для кого-то, отвечающего искомым характеристикам, но взаимодействующего с людьми, для такого, как вы, другие в этой комнате, в Манчестере, с некоторыми переменными. У нас есть такой корпус? Можем ли мы получить корпус, подобный этому? Можем ли мы смоделировать эту ситуацию? Это сложно. Можем ли аппроксимировать ситуацию, приближая, насколько это возможно, к проблемной? Я работал над сотней дел авторской атрибуции, и они очень отличаются друг от друга — тем, кто написал тексты, когда, с какой целью. Подобно всем исследователям, мы используем большие корпуса общего английского, например COCA (Corpus of Contemporary American English). Некоторые из корпусов были созданы в исследовательских целях. Но в последнее

время возникает все больше корпусов с использованием некоторых австралийских данных. Петр обращался к данным о преподавателях австралийского варианта английского как иностранного при анализе случая, когда подозреваемым был австралийский преподаватель. И очень важно, насколько полно удастся разработать нужный корпус. Проблема валидизации всегда будет острой. Мы имеем голосовой тракт, и у нас есть основания полагать, что голосовой тракт в целом различается. Голосовой тракт — биологическое явление, так ведь? Но язык — общественное явление.

Реплика. Есть два типа валидизации. Первый — вы должны валидизировать эксперта, второй — валидизировать метод. Я думаю, валидизация метода — более простое, прямое решение. И, валидизируя метод, мы валидизируем эксперта, способного применять этот метод. Это проблема в некотором роде состязательных, конкурирующих видов анализа, с помощью которого мы можем определять автора. Но наша дисциплина, следует предполагать, заметно отличается от других, использующих автороведческий анализ. Нам нужен практический подход к проблеме. Мы не отказываемся от трудностей, на самом деле важно преодолевать их — мы должны стремиться к столь же четким стандартам, как в исследовании ДНК, в области анализа речи.

Реплика (Грант). Я думаю, на самом деле это важный вопрос разработки методов, которые мы пытаемся валидизировать в потенциальном аспекте судебной лингвистической экспертизы. Я думаю, в ранних работах по АА (атрибуции авторства) выбирались характеристики, в частности синтаксические, такие как цепочки прилагательных или группы прилагательного и существительного. И на ранних этапах это были хорошие, поддающиеся подсчету маркеры. Эта стадия валидизации была пройдена. Особенно важно, что были проведены исследования с широким привлечением населения. Я полагаю, что валидизация не сводится к какой-то разновидности переменных или даже к измерению переменных в случае небольшого числа авторов. Каким образом можно устанавливать переменные, уже существующие в базе данных? Поэтому следует говорить о валидизации общего метода, осуществляемой от одного дела к другому. И в определенном смысле валидизацией будет использование Ngram как метода поиска характеристик. Ngram-анализ — источник характеристик, в основном генерализованный метод для поиска сочетаний низкого уровня, структуры в тексте. В любом случае вы найдете

некоторые характеристики, если вам повезет. И человек-эксперт найдет строгий метод, подходящий к любому случаю, может, нечто похожее для различения. Но если есть различия, которые я обнаружил, я демонстрирую их в качестве надежных. Когда я думаю о валидации, я валидирую метод. И некоторые переменные изменяются от случая к случаю. Я думаю, стилистический автороведческий анализ, какой проводит Креденс, более подходит в качестве модели, чем валидизация на основе набора переменных.

Реплика (Стаматос). С технической точки зрения это зависит от ситуации в целом — закрытый ли у вас список подозреваемых, открытый или кейс верификации, при котором следует учитывать характеристики населения в целом. Для меня постановка вопроса о валидации не имеет смысла, если речь идет о закрытом списке. В таком случае я пытаюсь определить наиболее вероятного автора из установленного круга лиц. Если же список открытый, то проблема валидации, конечно же, более важна, нужно каким-либо образом установить, какое в целом население, что оно из себя представляет. Я имею в виду выделение людей с определенными характеристиками, похожими на черты подозреваемых. Иначе не имеет смысла сопоставлять имеющийся текст со случайно выбранным (принадлежащим случайной личности). Поэтому я полагаю, что алгоритм верификации — один из вопросов, на которых нам следует фокусировать внимание. Все характеристики невозможно учитывать при верификации. Это очень похоже на фундаментальную проблему. Если мы ее не решим, то мы (наша система) не сможем справиться с задачей верификации. Между тем она определяет наши ожидания от атрибуции.

Реплика (Грант). Имеется в виду верификация большой группы или малой? С малой группой относительно легко иметь дело. В этом вопросе вам нужны данные по некой референтной группе населения, потому что требуется сопоставлять пару подходящих кандидатов, черты каждого из которых могут быть дистинктивны на фоне некоторой релевантной референтной группы. А насколько могут различаться сравниваемые романы?

Вопрос. В каких пределах, вы полагаете, судебная лингвистика использует данные корпусов? Естественный язык, транслируемый сегодня с помощью Интернета, упрощает решение вопросов судебного речевого действия.

Ответ (Голд). Я думаю, существует определенная проблема в обработке большого

количества данных, когда требуется много записей и транскриптов, помогающих в выполнении анализа.

Реплика. Многие проблемы, поднятые Эрикой (зашумленность, компрессия), определенно могут быть решены в будущих публикациях. Да, записи во многих случаях сделаны под контролем, и кодеки сжатия оказывают влияние. Тем не менее определенно может быть использован телефонный трафик, мобильная и IP-телефония. Можно обеспечить доступ к банку голосов пользователей «Скайп». Материал для исследования есть, но он представлен неоднородными, «загрязненными» данными. Доступно много цифровой аудиоинформации, в перспективе можно привлекать и новые ресурсы со звуковыми данными. Однако проводить сопоставление образцов звучащей речи достаточно сложно.

Реплика (Эверт). Под данными, полученными с помощью Сети, чаще всего подразумеваются речевые корпуса. Да, есть «шумовая» проблематика, необходимо держать под контролем, кто говорящие, откуда информация, носители ли они языка. Я думаю, использовать эти большие корпуса очень сложно.

Вопрос (Голд). Некоторые автоматизированные системы анализа, имеющиеся на рынке, поддерживают использование видео. Мой вопрос: как люди делятся данными внутри научного сообщества? Например, Кристофер использует специализированный корпус для достаточно узких целей. Мы используем разнообразные данные для широкого круга целей.

Ответ (Эверт). Мы используем специально отобранные данные для определенных целей: существуют законные, лицензированные способы обмениваться данными. Это часто случается, когда специалисты обращаются к коллегам, нередко тайно, но допуская возможность публичного разглашения данного факта. Это одна из причин, заставляющая обращаться к ресурсам Сети, чтобы получить законный доступ к материалам. Один из потенциальных путей обмена данными — возможность делиться результатами анализа. Если интерес вызывают конкретные характеристики, их набор может быть квантифицирован и сопровожден данными, включая идентификацию информации о говорящих. Такое осуществимо для многих кейсов.

Вопрос (Райт). В соответствии с наиболее общими подходами в ваших исследованиях, восприятие речи переходит в акустический анализ. Как это соотносится со стилистическим анализом?

Ответ (Голд). Прежде всего соединяются данные исследования акустических и аудиторных параметров. Это требует комбинации автоматизированных приемов акустического анализа и фонетических методов. Много работы в данном направлении было сделано в рамках Йоркского проекта идентификации — автоматизированного метода в комбинации с изучением качества голоса — с привлечением подозреваемых в Германии, Швеции. Данному проекту принадлежит лидерство в комбинации автоматизированных и фонетических методов. У его организаторов разные правила относительно того, как можно вносить вклад в исследования. Сейчас в данном направлении ведется активная работа. Мне кажется, мы пробили не одну стену в области регулирования данного вида деятельности. Возможно, определяющим фактором в регулировании являются правила конкретной страны. Это один аспект. Для подобных методов одно дело теоретические исследования и совсем другое — практика их применения. Суть в том, что нужно найти доказательство и представить его в суде, чтобы помочь вынесению решения по делу.

Реплика. Эрика говорит о том, какие методы мы должны использовать при разбирательстве дел в суде, какие методики. Разрабатываются новые судебные экспертные технологии. Например, автоматическое распознавание голоса по записи.

Вопросы из Сети. В какой мере мы можем сделать автороведческий анализ без применения компьютеров?

Стамататос. Имеется в виду, как быстро?

Ведущий. Я полагаю, вопрос о точности.

Ответ (Стамататос). Может, этот анализ получится сделать с помощью смартфона? Я думаю, вы спрашиваете не у аудиторию.

Ответ (Грант). Я думаю, мы часто делаем это с очень короткими текстами. В анализе профиля, например, единственное письмо из шести строк. И полиция спрашивает нас, что мы можем сказать о написавшем. Причем это, на самом деле, обычный случай. У меня около месяца назад был запрос от людей, расследовавших дело о вымогательстве. Нас просили высказать свое мнение о написавшем. И ответ был очень коротким. Письмо составил не носитель языка, т. е. мы выяснили происхождение автора. Но обычно из-за узких временных рамок мы делаем компьютерный анализ. Раньше общение со следователем выстраивалось иначе, и мы проводили основательный анализ, причем не компьютерный.

Ответ (Грив). Многое зависит от данных: если их массив большой, то необходим компьютер. В других случаях можно обойтись и без него. Мой ответ — да, конечно.

Ответ (Эверет). Судебная лингвистика — весьма интригующая меня область. Я думаю, это вопрос о квантитативном анализе — оптимальном выборе характеристик, или переменных. В специфическом кейсе мы выбираем индикативные характеристики и затем проводим количественный анализ.

Ответ (Грант). Мы проводим ручной анализ для выявления «корзины характеристик», чтобы понять, какие из них постоянны, а какие различают данных авторов в кейсе. И можно выписать некоторые отбракованные выражения, чтобы проверить результаты, полученные неавтоматизированно, сделать на этой стадии небольшие подсчеты — очень простые, чтобы подтвердить ваши предположения и правильность восприятия материала, проверить устойчивость и различительность характеристик. И еще проследить, какие характеристики будут варьироваться от кейса к кейсу.

Вопрос (Эверет). Насколько велик риск предвзятого отношения при подобном подходе?

Ответ (Грант). Конечно, мы осознаем возможность необъективности, когнитивного склонения к той или иной позиции. Но для проведения качественного анализа подобный путь возможен. Один из вариантов — если это кейс классификации с двумя авторами и сначала нужно изучить уже известный материал, прежде чем обращаться к спорному.

Реплика (Креденс). Мы осознаем опасность необъективного рассмотрения. В последнее время мы все больше и больше рассматривали использование Ngram, на самом деле. Я был вдохновлен работой с Дэвидом с применением Ngram. Мы рассматриваем их в качестве весьма удобной точки доступа, которая позволяет выявить очень интересные характеристики при анализе материала большей протяженности и с большей глубиной проникновения в него. Фактически мы покупаем много пакетов программного обеспечения, которое позволяет составлять Ngram по различным текстам и искать Ngram, имеющие много общего в различных текстах. Этим мы защищаем себя от необъективного рассмотрения материала. При этом важно определиться: считать приоритетными буквы или слова? Мы предпочитаем слова, поскольку результаты по этому параметру легче интерпретировать. Но буквенные Ngram лучше при работе с данными, в которых много «шума», особенно с тексто-

выми и ватсап-сообщениями, где много неправильных написаний. В таком случае мы обращаемся к буквенным Ngram.

Вопрос (Эверт). Вопрос Эрике. Мне интересно, когда вы опираетесь на чат-лист, чтобы корректно интерпретировать коэффициент вероятности? Многие дают поддерживающие утверждения — демонстрируют умеренную поддержку для данных целей и защиты этого текста. Следует ли принимать это во внимание при выявлении коэффициента вероятности сходства, его увеличения или снижения?

Ответ (Голд). Соответствующие исследования есть, но их сложно бегло охарактеризовать в удобной форме. Я представляла некоторые доказательства в суде, которые не вызвали возражений. Но также есть и поборники заключений, основанных исключительно на автоматическом распознавании речи. Некоторые аспекты автороведческого анализа выходят за границы данного поля или могут быть отнесены к нему только формально. Жюри и судьи по-разному интерпретируют ухудшение показателей. У нас есть актуальная терминология, разные трактовки терминов, взаимосвязи между ними, которые мы выстраиваем. Таким образом, имеется большое поле для работы, и это не узкая проблема.

S. I. Krassa
Stavropol, Russia
ORCID ID: —

E-mail: skrassa@yandex.ru.

First Roundtable on Practices and Standards in Forensic Authorship Analysis (overview 4)

ABSTRACT. *The article provides an overview of the reports of the final discussion 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. Once again, the current topics raised during the round table are discussed: the importance of taking into account the frequency of function words in comparison with notional ones and the impact of various factors (style, genre of the text, author's idiosyncrasy, communication strategy) on the frequency of function words, the regulation of the procedures of forensic linguistic expertise by judicial authorities, the validation of the methods of authorship analysis used in practice, the role of formalized methods and the researcher's intuition in the analysis, the scientific and legal aspects of using the data array available on the Internet (including video messages and video conversations), the reference to the best practices of corpus linguistics, the compilation of highly specialized language corpora, the selection of computer programs for analyzing text data arrays, the use of automated speech recognition, the interaction between manual and automated computer analysis, the prioritizing of alphabetic or verbal Ngrams depending on the material under study. During the discussions, questions were asked both by the speakers and those present at the round table, as well as by experts who watched the event online. In his closing remarks, the moderator expressed his intention to organize this round table annually to discuss new ideas and achievements in the field of forensic analysis in a timely manner.*

KEYWORDS: *round table; forensic authorship analysis; author profiling; stylometry; computer linguistics; core features of a style; forensic linguistics; forensic expertise; automatic text recognition; forensic linguistics; corpus linguistics; referent groups; automatic speech recognition.*

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 4) / S. I. Krassa // Political Linguistics. — 2021. — No 3 (87). — P. 227-234. — DOI 10.26170/1999-2629_2021_03_24.*

Вопрос (Кшиштофу). О каком пакете программного обеспечения вы говорите?

Ответ. Я говорю о «Jungle» — это пакет программ, разработанный и произведенный Дэвидом Вулфсом.

Реплика (Креденс). Я хочу поблагодарить организаторов и выступающих. Мой вопрос: когда состоится следующий круглый стол?

Ведущий. Ответ на этот вопрос входит в мою заключительную речь. Время закрывать мероприятие. Я бы хотел в первую очередь поблагодарить выступавших с докладами, которые пришли сюда с замечательными сообщениями, а также других выступавших. Если говорить о частоте, то мы предполагаем организовывать ежегодные встречи с вами для того, чтобы иметь обратную связь. Для осуществления этих планов мы сделаем все от нас зависящее в рамках реальных возможностей. Я бы хотел высказать благодарность Центру цифровых гуманитарных наук Манчестерского университета и в особенности Джозефу Льюису, который все координировал. И я бы хотел сказать спасибо Международной ассоциации судебной лингвистики, и в частности ее президенту и секретарю. Также большое спасибо всем, кто наблюдал за нами онлайн. Спасибо Барселоне — они смотрели нас.

REFERENCES

1. Coulthard, M. Author identification, idiolect, and linguistic uniqueness / Malcolm Coulthard. — Text : unmediated // *Applied Linguistics*. — 2004. — Vol. 25. — Iss. 4. — P. 431–447.
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 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.
4. Gold, E. International practices in forensic speaker comparisons: second survey / Erica Gold, Peter French. — DOI 10.1558/ijsl.38028. — Text : unmediated // *International Journal of Speech Language and the Law*. — 2019. — Vol. 26, No 1. — P. 1–20.
5. 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.
6. 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.
7. 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.
8. Grant, T. Txt 4n6: Idiolect free authorship analysis? / T. Grant. — Text : unmediated // *The Routledge Handbook of Forensic Linguistics* / M. Coulthard, A. Johnson (eds.). — Abingdon : Routledge, 2010. — P. 508.—522.
9. 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.
10. Hardcastle, R. A. CUSUM: a credible method for the determination of authorship? / R. A. Hardcastle. — DOI 10.1016/S1355-0306(97)72158-0. — Text : unmediated // *Science & Justice*. — 1997. — Vol. 37, Iss. 2. — P. 129–138.
11. Hardcastle, R. A. Forensic linguistics: an assessment of the CUSUM method for the determination of authorship / R. A. Hardcastle. — DOI 10.1016/S0015-7368(93)72987-4. — Text : unmediated // *Journ. of the Forensic Science Society*. — 1993. — Vol. 33, Iss. 2. — P. 95–106.
12. 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.
13. 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.
14. Sidorov, G. Syntactic n-grams as machine learning features for natural language processing / G. Sidorov, F. Velasquez, E. Stamatatos, A. Gelbukh, L. Chanona-Hernández. — Text : unmediated // *Expert Systems with Applications*. — 2014. — Vol. 41. — Iss. 3. — P. 853–860.
15. Stamatatos, E. Text Genre Detection Using Common Word Frequencies / E. Stamatatos, N. Fakotakis, G. Kokkinakis. — Text : electronic // *COLING 2000. Vol. 2. The 18th International Conference on Computational Linguistics*. — P. 808–814. — URL: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.14.7621&rep=rep1&type=pdf>.