DOI: https://doi.org/10.15688/jvolsu2.2020.1.7

UDC 81'322:340.6
LBC 81.11

Submitted: 03.10.2019
Accepted: 27.12.2019

CURRENT PROBLEMS OF FORENSIC AUTHORSHIP ANALYSIS AND THE POSSIBILITY OF THEIR SOLUTION WITH THE USE OF COMPUTER METHODS: PROBLEMS AND PROSPECTS [1]

Tatyana A. Litvinova

Voronezh State Pedagogical University, Voronezh, Russia

Anastasiya V. Gromova

Forensic Centre of the the Ministry of Internal Affairs of the Russian Federation, Moscow, Russia

Abstract. Active development of Internet communication in recent years caused an increase in the number of forensic text examinations aimed at identifying and profiling (i.e. inferring gender, age, personality, etc. of the author from textual analysis) the author of written texts. Despite the availability of proven methodological recommendations for the production of such examinations, in this area there are many unresolved problems associated mainly with the emergence of new research objects. In addition, the existing expert practice does not fully utilize the achievements of corpus, computer, and quantitative linguistics. In this situation, there is a gap between the "qualitative" and "quantitative" methods of textual authorship analysis, which hinders further development of both theoretical research in the area of authorship attribution and profiling and an increase in the level of objectivity and reproducibility of forensic authorship analysis. The paper represents some typical tasks solved by a forensic expert; describes the characteristics of the objects of forensic authorship analysis, and determines the main difficulties forensic experts face in the course of this analysis. The possibilities of using existing computer methods to solve these tasks are analyzed. It is shown that not all the existing computer methods are useful for forensic authorship analysis. We also highlight the ways of development of forensic authorship analysis related to further theoretical research in the field of idiolect using corpus data and natural language processing techniques.

Key words: forensic authorship analysis, authorship attribution, authorship profiling, computational linguistics, text corpora, stylometry.

Citation. Litvinova T.A., Gromova A.V. Current Problems of Forensic Authorship Analysis and the Possibility of Their Solution with the Use of Computer Methods: Problems and Prospects. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2020, vol. 19, no. 1, pp. 77-88. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2020.1.7

УДК 81'322:340.6
ББК 81.11

Дата поступления статьи: 03.10.2019
Дата принятия статьи: 27.12.2019

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В СУДЕБНОЙ АВТОРОВЕДЧЕСКОЙ ЭКСПЕРТИЗЕ: ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ ИСПОЛЬЗОВАНИЯ [1]

Татьяна Александровна Литвинова

Воронежский государственный педагогический университет, г. Воронеж, Россия

Анастасия Викторовна Громова

Экспертно-криминалистический центр МВД России, г. Москва, Россия

Аннотация. В связи с активным развитием интернет-коммуникации в последние годы наблюдается рост количества назначаемых судебных автороведческих экспертиз, направленных на идентификацию и

диагностирование личности автора текста. Несмотря на наличие апробированных методических рекомендаций по производству таких экспертиз, в данной области существуют нерешенные задачи, связанные преимущественно с появлением новых объектов для изучения. В экспертной практике не в полной мере используются достижения корпусной, компьютерной и квантитативной лингвистики. Авторами констатируется наличие разрыва между «качественными» и «количественными» методами автороведческого анализа текста, который препятствует как дальнейшему развитию теоретических исследований в данной области, так и повышению уровня объективности и воспроизводимости экспертиз. В статье сформулированы типовые задачи, решаемые экспертом-автороведом, охарактеризованы объекты автороведческой экспертизы, определены основные трудности, с которыми сталкиваются эксперты-авторoведы. Проанализированы возможности применения существующих компьютерных методов для решения указанных задач. Намечены пути развития методов судебной автороведческой экспертизы, связанные с дальнейшим теоретическим осмыслением идиолектов с использованием корпусных данных и технологий автоматической обработки текстов.

Ключевые слова: автороведческая экспертиза, идентификация автора текста, диагностика характеристик автора текста, компьютерная лингвистика, корпусы текстов, стилеметрия.

Цитирование. Литвинова Т. А., Громова А. В. Компьютерные технологии в судебной автороведческой экспертизе: проблемы и перспективы использования // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2020. – Т. 19, № 1. – С. 77–88. – DOI: https://doi.org/10.15688/jvolsu2.2020.1.7

Введение

Развитие виртуальной коммуникации неизбежно влечет за собой рост преступлений, совершаемых с использованием возможностей сети Интернет. Анонимность и дистанционный характер общения способствуют росту преступлений в отношении несовершеннолетних (развратные действия, склонение к самоубийству, шантаж), вовлечению граждан в экстремистские, террористические сообщества, осуществлению угроз жизни и здоровью, мошенничества. При расследовании таких преступлений у правоохранительных органов может возникать необходимость в установлении автора сообщений, имеющих криминалистическую значимость. Данная задача решается в рамках судебной автороведческой экспертизы (далее – САЭ), которая производится в экспертно-криминалистических подразделениях системы МВД России уже более 25 лет. Для ее решения разработаны и успешно используются методические рекомендации по идентификации [Методические рекомендации..., 2010] и диагностированию личности [Диагностика половозрастных..., 2014] автора нерукописного текста. Однако, как показывает практика, динамичное развитие функциональных возможностей создания и обработки текстов, связанное прежде всего с развитием технологий интернет-коммуникации, требует совершенствования инструментария автороведческих экспертиз на постоянной основе (см. подробнее: [Назарова, Громова, 2016]).

Автороведческое исследование является продолжительным и трудоемким. Практически полное отсутствие в поступающих на экспертизу текстах рукописных элементов, а также возможность использования технологий синтеза речи (и, как следствие, невозможность идентификации говорящего по голосу и речи) приводят к тому, что анализ лингвистических признаков приобретает первостепенное значение для решения задачи идентификации и диагностирования личности автора текста. Разработка техник, позволяющих автоматизировать отдельные этапы автороведческого анализа, представляет собой актуальную задачу, поскольку подобные техники дадут возможность оптимизировать сроки производства экспертиз. Кроме того, ученые и практикующие эксперты все чаще говорят о необходимости использования статистических методов при производстве САЭ с целью увеличения объективности экспертных заключений (см. об этом: [Литвинова, 2019]).

На протяжении последних 15 лет специалисты в области информационных технологий (computer scientists) активно разрабатывают методы идентификации и диагностирования личности по тексту (преимущественно на материале английского языка). Ежегодно проводятся хакатоны PAN (https://pan.webis.de), направленные на выявление наиболее точных моделей идентификации и диагностирования

личности по тексту (как правило, исследователи ставят задачу диагностировать пол и возраст, реже – некоторые психологические характеристики автора), однако такие работы не направлены специально на решение задач экспертной практики.

В работах зарубежных специалистов идет активное обсуждение вопросов использования «компьютерных»[2] методов при производстве САЭ [Argamon, 2018; Chaski, 2013], и единых подходов к использованию количественных методов для решения задач САЭ до настоящего времени не выработано. Существующий разрыв между преимущественно качественными методами анализа текста, применяемыми в экспертной практике, и многочисленными наработками последних лет в области анализа текста с применением технологий интеллектуального анализа данных может быть сокращен, по нашему мнению, путем выделения основных задач, которые решаются экспертами при производстве САЭ, выявления особенностей современных объектов САЭ, а также обзора «компьютерных» работ в свете возможностей использования представленных в них методов для решения задач, стоящих перед экспертом-автороведом.

Результаты и обсуждение

Типовые задачи, стоящие перед экспертом-автороведом

Задачи, которые решает эксперт, можно разделить на две большие группы:

1) идентификационные задачи, связанные с определением автора текста путем проведения раздельного и сравнительного анализа признаков, проявившихся в спорном тексте и текстах – образцах письменной речи подозреваемого лица;

2) диагностические задачи, связанные с определением половозрастных, индивидуально-личностных характеристик, уровня коммуникативной компетенции, речевой культуры, сферы профессиональной деятельности автора текста (см. подробнее: [Назарова, Громова, 2016]).

Проблема идентификации автора текста может иметь разные аспекты, обусловленные задачами, решение которых актуально в конкретной криминалистической ситуации:

1. Кто из замкнутого круга лиц (небольшого, как правило 2–3 человека) является автором криминалистически значимого текста (в зарубежной литературе – closed-set problem)?

2. Является ли данное лицо, сравнительные образцы текстов которого представлены эксперту, автором криминалистически значимого текста (verification problem)?

Практика производства автороведческих экспертиз в Экспертно-криминалистическом центре (ЭКЦ) МВД России свидетельствует о том, что идентификационные задачи ставятся чаще, чем диагностические. Решение диагностических задач наиболее актуально при необходимости сужения круга лиц и, как правило, значимо в рамках предварительных исследований, проводимых по заданиям сотрудников оперативных подразделений.

Когда круг сужен, у подозреваемых изымаются образцы письменной речи для проведения идентификационной автороведческой экспертизы, при этом в большинстве случаев эксперт решает вопрос о том, является ли данное лицо, сравнительные образцы текстов которого представлены эксперту, автором криминалистически значимого текста, то есть круг подозреваемых сужается до одного человека.

Характеристики объектов, направляемых для производства САЭ

Объект САЭ – текст, являющийся воплощением идиолекта автора. Под идиолектом мы понимаем воплощенную в каждом речевом произведении (тексте) любого носителя языка уникальную реализацию языковой системы, состоящую как из устойчивых (зона стабильности), так и вариабельных (зона вариативности) сознательных и бессознательных выборов, производимых процудентом речи под влиянием совокупности факторов внутриидиолектного и межидиолектного варьирования, обеспечивающих уникальность и в то же время возможность типизации каждого конкретного идиолекта [Литвинова, 2019]. С точки зрения криминалистики текст рассматривается как некий интеллектуальный

след, продукт целенаправленной деятельности, выступающий носителем информации, в том числе и о личности автора. Объекты, направляемые на автороведческую экспертизу, могут быть различными по форме и условиям речевого представления: письменные тексты (рукописные и электронные, в том числе созданные в условиях интернет-коммуникации); устные тексты, в том числе содержащие признаки использования программ для синтеза речи.

В условиях интернет-коммуникации основной формой взаимодействия и самопрезентации пользователей являются письменные или устные тексты, создаваемые в рамках личной или публичной коммуникации. Данные тексты условно можно разделить на монологические и диалогические. Монологические формы реализуются в виде публичных сообщений, размещаемых пользователями под своим ником (пост, статус, заметка, статья, открытое письмо и др.), или сообщений, направляемых в адрес какой-либо инстанции. Диалогические формы и их разновидности представлены в виде переписки с другими пользователями, которая может протекать как личная или публичная беседа, обмен сообщениями в микрогруппе.

Длительное время монологические тексты выступали основным объектом автороведческих исследований, однако развитие сетевой коммуникации привело к появлению большого числа диалогических текстов (переписка в мессенджерах, чаты, форумы), задача изучения которых все чаще ставится перед экспертом-автороведом [Назарова, Громова, 2016].

Основные трудности, с которыми сталкивается эксперт-автороведа

Среди таких трудностей выделяются две группы:

1) связанные с криминалистически значимым текстом:

– *малый объем текста и/или текстов автора.* Большинство текстов, по которым востребовано проведение автороведческого исследования, содержат менее 1 000 слов. Существующие экспертные методики позволяют работать с текстами длиной от 200 словоупотреблений, однако в таком случае является критичной высокая степень сопоставимости анализируемого текста со сравнительными образцами;

– *возможность коллективного авторства.* Существуют различные варианты участия двух и более лиц в подготовке текста: фрагменты текста принадлежат разным авторам; текст написан одним автором, но редактируется другим; два (и более) автора принимали одинаковое участие в создании текста и т. д.

– *возможность имитации и маскировки признаков авторского текста.* Автор текста может имитировать социолект, характерный, по его мнению, для лиц той или иной группы (половой, возрастной, социальной), к которой он не принадлежит, либо же идиолект какого-либо конкретного лица. Кроме того, возможны случаи маскировки идиолекта без ориентации на какой-либо стандарт с целью сокрытия своей идентичности (например, путем снижения уровня грамотности). Сюда же входит проблема синтезированной письменной речи (боты), однако отметим, что в настоящее время уровень развития таких систем позволяет допустить возможность их использования для продуцирования лишь определенных текстов (например, диалогов в сфере обслуживания);

2) связанные со сравнительными образцами. Основная трудность заключается в получении и отборе сопоставимых с исследуемым криминалистически значимым текстом образцов письменной речи проверяемого лица. Значимыми характеристиками текстов являются: жанровая сопоставимость, адресат текстов, их тематика, объем, временной период создания (согласно сложившейся экспертной практике, разница между сравниваемыми текстами не должна превышать 5 лет).

Пути решения задач САЭ методами computer science

Зарубежные специалисты по судебной лингвистике (см., например: [Chaski, 2013]) выделяют две основные группы методов автороведческого исследования:

1) стилистические, то есть основанные на проводимом лингвистами анализе отдель-

ных, уникальных для каждого текста его элементов (forensic stylistics);

2) статистические, то есть основанные на использовании «компьютерных» методов, которые, как считают многие авторы (см. обзор: [Литвинова Т.А., Литвинова О.А., 2015]), являются более объективными, поскольку базируются не на интуиции эксперта, а результаты, полученные на их основе, – более воспроизводимыми.

В общем виде «компьютерный» подход к идентификации автора текста заключается в построении классификатора, на входе которого – численные значения различных квантифицируемых параметров текста, извлеченных, как правило, автоматически (униграммы и *n*-граммы, то есть последовательности из *n* элементов – символов (букв, знаков препинания, цифр и т. п.), слов, частей речи и т. д.); реже – синтаксические и семантические параметры; на выходе – класс объекта (текста), то есть принадлежность тому или иному автору.

В опубликованных исследованиях говорится о достижении точности свыше 80 % для текстов 100 авторов, свыше 30 % – для 10 000 авторов (см. подробнее: [Argamon, 2018]), однако проведенный нами критический анализ таких работ показал, что подавляющее большинство из них обладает рядом особенностей, делающих их результаты непригодными для использования в условиях авторoведческого исследования в силу следующих обстоятельств:

– в таких работах исследуются далекие от задач авторoведения проблемы, например задача идентификации автора из большого круга лиц (несколько сотен и даже тысяч);

– используются тексты большого объема (несколько тысяч и даже десятков тысяч слов), либо же анализируется большое число текстов от каждого автора;

– мало внимания уделяется собственно лингвистическим признакам и их различающей способности, поскольку ученые сфокусированы преимущественно на точности создаваемых ими моделей. Ряд методов, дающих наибольшую точность (например, методы глубокого обучения), крайне сложны для интерпретации, которая важна для решения задач судебного экспертного исследования.

Тем не менее в работах последних лет, выполненных, как правило, лингвистами, в том числе судебными, приводятся описания методов и результатов, которые могут быть использованы при проведении судебного авторoведческого исследования и учитывают те трудности, с которыми сталкиваются эксперты:

– *малый объем текста и/или текстов, предоставляемых для производства САЭ*, не позволяет использовать методы машинного обучения, однако в таких условиях применимы методы, основанные на измерении расстояния между текстами (см. подробнее: [Argamon, 2018; Stamatatos, 2009]);

– *возможность коллективного авторства.* Задача исследования соавторства компьютерными методами в различных ее постановках начинает привлекать внимание ученых. Ей были специально посвящены хакатоны PAN, в том числе 2019 г. [Overview of PAN..., 2019]. Она является очень сложной для решения, в связи с чем в 2019 г. организаторами соревнований она была упрощена: участникам нужно было не выделить фрагменты текста, созданные разными авторами, а определить, сколько авторов участвовало в создании текста.

Одним из перспективных методов решения задачи определения наличия у текста нескольких авторов (шире – выявления случаев стилистической неоднородности) представляется метод, получивший название «скользящая стилеметрия» (rolling stylometry) [Rybicki, Eder, Hoover, 2016]. В указанной работе анализ одинаковых по длине последовательностей тех или иных языковых единиц (символов, самых частотных слов корпуса) с последующей визуализацией стилистических изменений был успешно применен для изучения литературных произведений с целью выявления фрагментов текста, созданных разными авторами, однако необходимы дальнейшие экспериментальные исследования с использованием данного подхода на материале текстов меньшей длины.

Другим методом, который может быть использован для определения наличия у текста нескольких авторов, является кластерный анализ. Он часто применяется в исследованиях авторства [Argamon, 2018; Panicheva, Litvinova O.,

Litvinova T., 2019]. Преимущество данной техники состоит в том, что она позволяет изучать группы схожих объектов (текстов) и визуализировать полученные результаты.

Следует отметить, что современные методы анализа предлагают множество возможностей для визуализации полученных результатов, которые используются в научных работах по исследованию авторства (метод главных компонент, многомерное шкалирование и т. д.), но, как справедливо указано в работе Ш. Аргамона, нужно понимать, что графики могут иметь разный вид и, соответственно, давать разные результаты для разных текстов и разных параметров, следовательно, нужно проводить несколько экспериментов и хорошо понимать особенности применяемых методов [Argamon, 2018]. То же относится и к кластеризации, однако роль визуализации в автороведческом анализе трудно переоценить;

– *возможность имитации и маскировки признаков.* Как показывают специальные исследования, имитация идиолекта существенно снижает точность классификационных моделей [The Case for Being Average..., 2017]. При этом очевидно, что первым этапом автороведческого анализа должно быть определение самого намерения исказить идиолект, но исследований, посвященных данной проблеме, крайне мало. В одной из немногочисленных работ по данной теме была показана принципиальная возможность такого анализа [Juola, 2012], однако не дано ответа на очень важный как для лингвиста-теоретика, так и для эксперта-автороведа вопрос: какие именно языковые элементы вносят наибольший вклад в разделение текстов по признаку наличия / отсутствия искажения идиолекта?

В работе [Afroz, Brennan, Greenstadt, 2012] был представлен классификатор, с точностью 96,6 % определяющий наличие в тексте признаков сокрытия / имитации идиолекта, при этом в указанном исследовании был проведен детальный анализ языковых признаков. Авторы рассматривали две ситуации: имитации идиостиля известного писателя людьми, профессионально владеющими языком, и маскировки идиолекта рядовыми носителями языка. Очевидно, для САЭ более актуальной является вторая задача, поэтому остановимся на ней подробнее. Авторы определили, что наиболее информативными параметрами, различающими две группы текстов (с искажением и без), оказались частоты строевых слов (function words). Так, в текстах с искажением употреблено больше наречий, частиц и личных местоимений, но меньше существительных. Отмечены попытки исказить свой идиолект путем использования более коротких предложений, простых слов с меньшим числом слогов – такие тексты в целом характеризовались более низкими показателями сложности. В целом в текстах с искажениями зафиксирована повышенная частотность строевых слов *I* «я», *my* «мой», *there* «здесь», *you* «ты», но реже встречались слова *as* «как», *his* «его», *her* «ее», предлог *by*, глагол-связка *to be* «быть». Авторы пришли к важному выводу о том, что обнаружение попытки искажения идиолекта в принципе возможно и важную роль в решении данной задачи играет частотный анализ строевых слов.

Как говорилось выше, важной проблемой современного автороведения, которая, по нашему мнению, будет все более острой по мере развития технологий синтезированной письменной речи, является проблема определения синтезированной речи. В настоящее время активно рассматривается проблема обнаружения ботов в социальных сетях, но для ее решения используются, как правило, неязыковые признаки [Badawy, Ferrara, Lerman, 2018] либо же применяются методы глубокого обучения [Kudugunta, Ferrara, 2018], не позволяющие проводить лингвистический анализ признаков. Безусловно, обнаружение бота в ситуации диалогической синхронной коммуникации является актуальной, однако пока мало исследованной проблемой классификации текста, как и проблема обнаружения намерения искажения идиолекта в целом.

Как уже было сказано, для повышения надежности результатов САЭ сравнительные образцы должны быть аналогичны криминалистически значимому тексту по целому ряду факторов. Однако такие тексты не всегда можно получить в силу объективных обстоятельств, как в случае с предсмертной запиской, а также в силу того, что подозреваемые вправе отказаться от написания текстов требуемых характеристик, и тогда в распоряже-

ние эксперта поступают те тексты, которые автор создавал ранее в различных ситуациях. Это могут быть тексты, кардинально отличающиеся от криминалистически значимого по жанру, теме и т. д. Таким образом, возникает проблема кросс-жанровой, кросс-топиковой и даже кросс-модальной (в случае с текстами устной и письменной речи) атрибуции. Данная задача все чаще привлекает внимание ученых в последние годы (см. подробнее: [Литвинова, 2019]), при этом абсолютное большинство исследователей отмечает резкое падение точности классификационных моделей в кросс-жанровом и/или кросс-топиковом сценарии в сравнении с одножанровым и/или однотопиковым сценарием (см. обзор новейших исследований и результатов хакатона: [Overview of PAN..., 2019]). Ведутся разработки, направленные на построение моделей, эффективных и в таком, крайне сложном для анализа, сценарии, при этом одними из эффективных оказываются слабо контролируемые автором параметры – последовательности символов, пунктуационные привычки и т. д. [Литвинова, 2019], однако очевидно, что для разработки методов САЭ в условиях недостаточной сопоставимости образцов должны быть проведены масштабные экспериментальные исследования на обширном языковом материале.

Заметим, что если проблема судебного авторoведческого идентификационного исследования активно изучается, хотя в этой области, безусловно, очень много вопросов, на которые пока нет ответов, то проблема диагностирования автора криминалистически значимого текста остается малоразвитым направлением (схожее мнение высказано известным английским специалистом по судебной лингвистике Андреа Нини [Nini, 2018]). Несмотря на большое количество работ, посвященных диагностированию (преимущественно определению пола) автора текста, выполненных специалистами в области computer science, следует отметить, что большинство из них основано на анализе значительного по объему языкового материала, что релевантно для решения задач выявления трендов, актуальных, например, для анализа поведения покупателя и разработки таргетированной рекламы, но не подходит для целей САЭ.

Перспективы развития методического и технического обеспечения авторoведческих экспертиз

Как показывает анализ имеющихся работ, задача идентификации и диагностирования автора текста привлекает все большее внимание специалистов по computer science. Однако, несмотря на интерес исследователей к данной теме, следует констатировать разрыв между задачами, которые ставятся исследователями, и реальными потребностями САЭ. Отметим также недостаточную разработанность ряда теоретических проблем языкознания, прежде всего отсутствие общей теории идиолекта, а также корпусной поддержки экспертиз и – шире – низкий уровень интеграции усилий лингвистов, авторoведов, специалистов в области computer science в решении сложной междисциплинарной задачи идентификации и диагностирования автора криминалистически значимого текста.

Кратко сформулируем основные направления дальнейшей работы.

1. При постановке задачи исследования следует ориентироваться не на эффективность того или иного метода машинного обучения в обработке больших баз данных, а на задачи, наиболее часто встречающиеся в практике эксперта-авторoведа. В качестве материала для анализа необходимо выбирать тексты, схожие по объему и характеристикам с объектами, которые имеют реальную криминалистическую значимость и наиболее часто направляются для производства авторoведческих экспертиз. Кроме того, особое внимание нужно уделять интуитивно понятным интерпретируемым статистическим методам с обязательной визуализацией полученных результатов и анализу языковых признаков.

2. Необходимо помнить о том, что «экспертная деятельность не тождественна научной деятельности как процессу познания объективной действительности, а основывается на его достижениях – новых знаниях, полученных в результате всестороннего и достоверного изучения объектов, процессов или явлений при помощи имеющихся в науке принципов и методов познания» [Ионова, 2017, с. 29], в связи с чем считаем важным для совершенствования методик САЭ проводить

дальнейшие теоретические изыскания в области изучения идиолекта, выделять и исследовать факторы внутриидиолектного и межидиолектного варьирования (рис. 1), оценивая их взаимосвязи, их влияние на те или иные параметры текста.

Подобные работы обязательно должны опираться на эмпирическую базу в виде специальным образом построенных корпусов текстов, в связи с чем нами было предложено исследовать идиолект в рамках отдельного направления – корпусной идиолектологии [Литвинова, 2019].

3. Необходимо создавать корпусы текстов как эмпирическую базу теоретических исследований идиолекта и САЭ. Нам представляется значимой работа по обеспечению корпусной поддержки авторoведческих исследований как минимум в трех направлениях:

– корпусы как источник для изучения идиолекта. Специальным образом построенные корпусы, размеченные с точки зрения факторов идиолектного варьирования, позволят оценить роль тех иных факторов в функционировании различных языковых единиц, выявить устойчивые к смене темы, жанра и т. д. языковые единицы, наиболее типичные языковые признаки текстов лиц с теми или иными характеристиками и т. д. Создание таких корпусов является трудоемкой, но актуальной задачей современной корпусной лингвистики. В настоящее время в лаборатории корпусной идиолектологии ВГПУ ведется работа над расширением базы данных RusIdiolect (см. рис. 2), специально предназначенной для изучения влияния факторов идиолектного варьирования на параметры текста;

– корпусы как источник информации о частотности тех или иных языковых единиц в речи различных социальных, возрастных и других групп, в текстах тех или иных стилей и т. д. (base-rate knowledge). Имеющийся в настоящий момент самый крупный корпус текстов на русском языке НКРЯ на 40 % состоит из художественных текстов и поэтому не может служить надежной базой таких исследований, поскольку основной массив криминалистически значимых текстов – это тексты так называемой непрофессиональной письменной речи (термин О.В. Загоровской [Загоровская, 2019]);

– для совершенствования методов социопсихолингвистического моделирования автора текста необходимо создание закрытой базы данных криминалистически значимых текстов, авторство которых и, соответственно, характеристики автора установлены в ходе судебного исследования.

Рис. 1. Взаимодействие факторов внутриидиолектного и межидиолектного варьирования

Fig. 1. Correlation of intra- and inter-idiolect variability factors

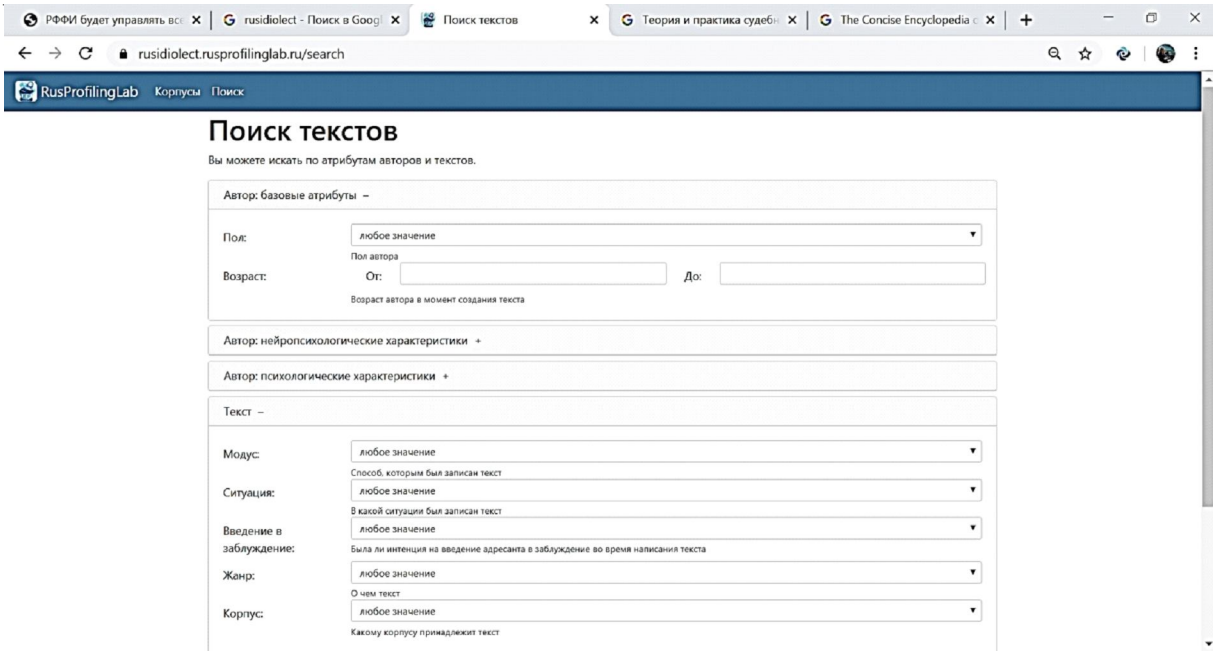

Рис. 2. Возможности поиска по факторам идиолектного варьирования в БД RusIdiolect

Fig. 2. RusIdiolect database search capability with regard to idiolect variability factors

Выводы

Применение компьютерных методов при решении задач САЭ ориентировано на повышение объективности полученных результатов, однако полностью исключать элемент субъективности исследователя (эксперта) даже при использовании исключительно «компьютерных» методов невозможно. Выбор самого метода, языковых параметров (их количества и типа) остается на усмотрение эксперта, поэтому ни один метод не лишен субъективизма. Таким образом, компьютерные методы являются не панацеей, но дополнительным инструментом, расширяющим возможности эксперта. По нашему глубокому убеждению, не следует противопоставлять традиционные методы, основанные на анализе отдельных языковых элементов текста, и методы компьютерные. Для построения эффективных методик авторoведческого анализа текста необходимо использовать преимущества каждого из них, а не выбирать какой-либо один. Лингвистический анализ текста, основанный на корпусных данных и дополненный анализом языковых параметров с обязательной визуализацией полученных результатов и статистическими методами, позволит, как нам представляется, значительно повысить точность полученных выводов. При этом нельзя забывать, что в САЭ анализируются языковые признаки текста, а статистические методы являются вспомогательными. Очевидно, что развитие авторoведческих методик невозможно без глубоких теоретических изысканий в области идиолекта с использованием достижений корпусной и компьютерной лингвистики.

ПРИМЕЧАНИЯ

[1] Т.А. Литвинова осуществила работу над статьей в рамках гранта РНФ № 18-78-10081 «Моделирование идиолекта носителя современного русского языка в аспекте идентификации автора текста».

T.A. Litvinova carried out the work on the reported article in the framework of the Russian Science Foundation grant no. 18-78-10081 “Modeling the Modern Russian Speaker’s Idiolect in Forensic Authorship Analysis”.

[2] Мы условно называем техники лингвистического анализа и последующей категоризации текста, основанные на использовании различных статистических методов и методов интеллектуального анализа данных, «компьютерными».

СПИСОК ЛИТЕРАТУРЫ

Диагностика половозрастных и индивидуально-личностных характеристик автора нерукописного

текста : метод. рекомендации, 2014 / Т. В. Назарова [и др.]. М. : ЭКЦ МВД России. 112 с.

Загоровская О. В., 2019. Естественная (непрофессиональная) письменная речь как модус существования современного русского языка и «зеркало» идиолекта его носителя // Известия ВГПУ. Т. 283, № 2. С. 202–206.

Ионова С. В., 2017. Аспекты исследования письменного текста как объекта лингвистической экспертизы // Вестник Волгоградского государственного университета. Серия 2, Языкознание. Т. 16, № 2. С. 28–38. DOI: 10.15688/jvolsu2.2017.2.3.

Литвинова Т. А., 2019. Пунктуационные выборы как составляющая ортологического параметра идиолекта носителя современного русского языка в аспекте идентификационной автороведческой экспертизы // Политическая лингвистика. № 1 (73). С. 114–121. DOI: 10.26170/pl19-01-13.

Литвинова Т. А., Литвинова О. А., 2015. Идентификация и диагностирование автора письменного текста. Воронеж : Изд-во ВГПУ. 332 с.

Методические рекомендации по производству судебной автороведческой экспертизы нерукописных текстов, 2010. М. : ЭКЦ МВД России. 55 с.

Назарова Т. В., Громова А. В., 2016. Объекты и задачи лингвистических и автороведческих экспертиз, проводимых в экспертно-криминалистических подразделениях органов внутренних дел Российской Федерации // Судебная экспертиза Беларуси. № 1 (2). С. 43–46.

Afroz S., Brennan M., Greenstadt R., 2012. Detecting Hoaxes, Frauds, and Deception in Writing Style Online // 2012 IEEE Symposium on Security and Privacy. San Francisco : IEEE. P. 461–475. DOI: 10.1109/SP.2012.34.

Argamon S., 2018. Computational Forensic Authorship Analysis Language and Law // Linguagem e Direito. Vol. 5 (2). P. 7–37.

Badawy A., Ferrara E., Lerman K., 2018. Analyzing the Digital Traces of Political Manipulation : The 2016 Russian Interference Twitter Campaign // Proceedings of 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM). IEEE. P. 258–265. DOI: 10.1109/ASONAM.2018.8508646.

Chaski C. E., 2013. Best Practices and Admissibility of Forensic Author Identification // Journal of Law and Policy. Vol. 21 (2). P. 333–376.

Juola P., 2012. Detecting Stylistic Deception // Proceeding EACL 2012 Workshop on Computational Approaches to Deception Detection. ACL. P. 91–96.

Kudugunta S., Ferrara E., 2018. Deep Neural Networks for Bot Detection // Information Sciences. Vol. 467. P. 312–322. DOI: 10.1016/j.ins.2018.08.019.

Nini A., 2018. Developing Forensic Authorship Profiling // Language and Law. Vol. 5 (2). P. 38–58.

Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection, 2019 / W. Daelemans [et al.] // Lecture Notes in Computer Science. Vol. 11696. P. 402–416. DOI: 10.1007/978-3-030-28577-7_30.

Panicheva P., Litvinova O., Litvinova T., 2019. Author Clustering with and Without Topical Features // Lecture Notes in Computer Science. Vol. 11658. P. 348–358. DOI: 10.1007/978-3-030-26061-3_36.

Rybicki J., Eder M., Hoover D., 2016. Computational Stylistics and Text Analysis // Doing Digital Humanities / ed. by C. Crompton, R. L. Lane, R. Siemens. L. ; N. Y. : Routledge. P. 123–144. DOI: 10.4324/9781315707860-19.

Stamatatos E., 2009. A Survey of Modern Authorship Attribution Methods // Journal of the American Society for Information Science and Technology. Vol. 60 (3). P. 538–556. DOI: 10.1002/asi.21001.

The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation, 2017 / G. Karadzhov [et al.] // Lecture Notes in Computer Science. Vol. 10456. P. 173–185. DOI: 10.1007/978-3-319-65813-1_18.

REFERENCES

Nazarova T.V. et al., 2014. *Diagnostika polovozrastnykh i individualno-lichnostnykh kharakteristik avtora nerukopisnogo teksta: metod. rekomendatsii* [Predicting Gender, Age and Personality Traits of the Author of the Printed Text: Guidelines]. Moscow, EKTs MVD Rossii. 112 p.

Zagorovskaya O.V., 2019. Estestvennaya (neprofessionalnaya) pismennaya rech kak modus sushchestvovaniya sovremennogo russkogo yazyka i «zerkalo» idiolekta ego nositelya [Natural (Unprofessional) Written Speech as a Mode of the Modern Russian Language Existence and the Idiolect “Mirror” of Its Speaker]. *Izvestiya VGPU* [Izvestia VSPU], vol. 283, no. 2, pp. 202-206.

Ionova S.V., 2017. Aspekty issledovaniya pismennogo teksta kak obyekta lingvisticheskoy ekspertizy [Aspects of Research Written Text as an Object of Forencis Linguistic Expertise]. *Vestnik Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2, Yazykoznanie* [Science Journal of

Volgograd State University. Linguistics], vol. 16, no. 2, pp. 28-38. DOI: 10.15688/jvolsu2.2017.2.3.

Litvinova T.A., 2019. Punktuatsionnye vybory kak sostavlyayushchaya ortologicheskogo parametra idiolekta nositelya sovremennogo russkogo yazyka v aspekte identifikatsionnoy avtorovedcheskoy ekspertizy [Punctuation Choice as a Component of Orthological Parameter of the Modern Russian Speaker's Idiolect in Forensic Authorship Analysis]. *Politicheskaja lingvistika* [Political Linguistics], no. 1 (73), pp. 114-121. DOI: 10.26170/pl19-01-13.

Litvinova T.A., Litvinova O.A., 2015. *Identifikatsiya i diagnostirovanie avtora pismennogo teksta* [Authorshop Identification and Profiling]. Voronezh, Izd-vo VGPU, 2015. 332 p.

Metodicheskie rekomendatsii po proizvodstvu sudebnoy avtorovedcheskoy ekspertizy nerukopisnykh tekstov, 2010 [Guidelines for the Production of Forensic Authorship Examination of Printed Texts]. Moscow, EKTs MVD Rossii. 55 p.

Nazarova T.V., Gromova A.V., 2016. Obyekty i zadachi lingvisticheskikh i avtorovedcheskikh ekspertiz, provodimykh v ekspertno-kriminalisticheskikh podrazdeleniyakh organov vnutrennikh del Rossiyskoy Federatsii [The Objects and the Tasks of Linguistic and Author's Right Examinations Conducted at Forensic Science Units in the System of the Ministry of the Interior of Russian Federation]. *Sudebnaya ekspertiza Belarusi* [Forensic Examination of Belarus], no. 1 (2), pp. 43-46.

Afroz S., Brennan M., Greenstadt R., 2012. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. *2012 IEEE Symposium on Security and Privacy.* San Francisco, IEEE, pp. 461-475. DOI: 10.1109/SP.2012.34.

Argamon S., 2018. Computational Forensic Authorship Analysis Language and Law. *Linguagem e Direito*, vol. 5 (2), pp. 7-37.

Badawy A., Ferrara E., Lerman K., 2018. Analyzing the Digital Traces of Political Manipulation: The 2016 Russian Interference Twitter Campaign. *Proceedings of 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM).* IEEE, pp. 258-265. DOI: 0.1109/ASONAM. 2018.8508646.

Chaski C.E., 2013. Best Practices and Admissibility of Forensic Author Identification. *Journal of Law and Policy*, vol. 21 (2), pp. 333-376.

Juola P., 2012. Detecting Stylistic Deception. *Proc. EACL 2012 Workshop on Computational Approaches to Deception Detection. ACL*, pp. 91-96.

Kudugunta S., Ferrara E., 2018. Deep Neural Networks for Bot Detection. *Information Sciences*, vol. 467, pp. 312-322. DOI: 10.1016/j.ins.2018.08.019.

Nini A., 2018. Developing Forensic Authorship Profiling. *Language and Law*, vol. 5 (2), pp. 38-58.

Daelemans W. et al., 2019. Overview of PAN 2019: Bots and Gender Profiling, Celebrity Profiling, Cross-Domain Authorship Attribution and Style Change Detection. *Lecture Notes in Computer Science*, vol. 11696, pp. 402-416. DOI: 10.1007/978-3-030-28577-7_30.

Panicheva P., Litvinova O., Litvinova T., 2019. Author Clustering with and Without Topical Features. *Lecture Notes in Computer Science*, vol. 11658, pp. 348-358. DOI: 10.1007/978-3-030-26061-3_36.

Rybicki J., Eder M., Hoover D., 2016. Computational Stylistics and Text Analysis. Crompton C., Lane R.L., Siemens R., eds. *Doing Digital Humanities.* London, New York, Routledge, pp. 123-144. DOI: 10.4324/9781315707860-19.

Stamatatos E., 2009. A Survey of Modern Authorship Attribution Methods. *Journal of the American Society for Information Science and Technology*, vol. 60 (3), pp. 538-556. DOI: 10.1002/asi.21001.

Karadzhov G. et al., 2017. The Case for Being Average: A Mediocrity Approach to Style Masking and Author Obfuscation. *Lecture Notes in Computer Science*, vol. 10456, pp. 173-185. DOI: 10.1007/978-3-319-65813-1_18.

Information About the Authors

Tatyana A. Litvinova, Candidate of Sciences (Philology), Head of the Research Laboratory of Corpus Idiolectology, Voronezh State Pedagogical University, Lenina St., 86, 394043 Voronezh, Russia, centr_rus_yaz@mail.ru, https://orcid.org/0000-0002-6019-3700

Anastasiya V. Gromova, Candidate of Sciences (Philology), Deputy Head of the Department of Forensic Linguistic and Authorship Examination of Spoken and Written Text, Head of the Subdepartment of Forensic Linguistic and Authorship Text Examination, Forensic Centre of the Ministry of Internal Affairs of the Russian Federation, Zoi i Aleksandra Kosmodemyanskikh St., 5, 125130 Moscow, Russia, Gromova_85@mail.ru, https://orcid.org/0000-0001-8255-5680

Информация об авторах

Татьяна Александровна Литвинова, кандидат филологических наук, заведующая научно-исследовательской лабораторией корпусной идиолектологии, Воронежский государственный педагогический университет, ул. Ленина, 86, 394043 г. Воронеж, Россия, centr_rus_yaz@mail.ru, https://orcid.org/0000-0002-6019-3700

Анастасия Викторовна Громова, кандидат филологических наук, заместитель начальника отдела фоноскопических, лингвистических и автороведческих экспертиз, начальник отделения лингвистических и автороведческих экспертиз, Экспертно-криминалистический центр МВД России, ул. Зои и Александра Космодемьянских, 5, 125130 г. Москва, Россия, Gromova_85@mail.ru, https://orcid.org/0000-0001-8255-5680