РАЗДЕЛ 5. ЛИНГВИСТИЧЕСКАЯ ЭКСПЕРТИЗА: ЯЗЫК И ПРАВО

УДК 811.161.1'42
DOI 10.26170/pl19-01-13
ББК Ш141.12-51 *ГСНТИ 16.31.02* *Код ВАК 10.02.01; 10.02.19*

Т. А. Литвинова
Воронежский государственный педагогический университет, Воронеж, Россия
ORCID ID: 0000-0002-6019-3700 ☑

☑ ***E-mail:*** *centr_rus_yaz@mail.ru.*

Пунктуационные выборы как составляющая ортологического параметра идиолекта носителя современного русского языка в аспекте идентификационной автороведческой экспертизы

АННОТАЦИЯ. *В настоящее время в связи с развитием интернет-коммуникации и появлением массива текстовых данных, часть которого содержит вредоносный контент, проблема идентификации автора текста, в том числе как задача судебной экспертизы, стала особенно актуальной. Однако точных и объективных методик атрибуции текста, которые могли быть использованы при проведении автороведческой экспертизы, до сих пор не выработано, причем как исследователи, так и эксперты-практики отмечают особую сложность автороведческого анализа текстов интернет-коммуникации. Статья посвящена проблемам создания инструментария автороведческой экспертизы текста на основе квантифицируемых признаков идиолекта как индивидуальной реализации языковой системы. Предметом рассмотрения являются результаты пунктуационного выбора пишущего как одна из составляющих ортологического параметра идиолекта. Обосновывается правомерность использования данного признака для идентификации автора. Анализ научной литературы по атрибуции текста, а также наши собственные экспериментальные исследования, выполненные на материале текстов экстремистского форума, показывают, что выбор пишущим знаков пунктуации, представленный рядом квантифицируемых признаков, в том числе впервые предложенных нами, достаточно устойчив к изменению темы текста (топика), что позволяет использовать данный параметр в кросс-топиковом сценарии. Нами также сформулированы направления дальнейших исследований, связанные прежде всего с формированием специализированного корпуса текстов, содержащего различные типы речевых произведений авторов (тексты разных жанров, модусов и т. д.), а также с извлечением новых признаков, характеризующих результаты выбора автором текста пунктуации.*

КЛЮЧЕВЫЕ СЛОВА: *автороведческая экспертиза; русский язык; идиолекты; атрибуция текста; ортологические параметры идиолектов; пунктуационные знаки; лингвистика; корпус текстов.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Литвинова Татьяна Александровна, кандидат филологических наук, зав. лабораторией теоретической и прикладной идиолектологии, Воронежский государственный педагогический университет; 394043, Россия, г. Воронеж, ул. Ленина, 86; e-mail: centr_rus_yaz@mail.ru.*

ДЛЯ ЦИТИРОВАНИЯ: *Литвинова, Т. А.* Пунктуационные выборы как составляющая ортологического параметра идиолекта носителя современного русского языка в аспекте идентификационной автороведческой экспертизы / Т. А. Литвинова // Политическая лингвистика. — 2019. — № 1 (73). — С. 114—121.

БЛАГОДАРНОСТИ. Исследование выполнено при поддержке гранта Российского научного фонда: № 18-78-10081 «Моделирование идиолекта носителя современного русского языка в аспекте идентификации автора текста».

Переход общества к новым коммуникативным технологиям, появление новых форм и модулей существования языка, широкое распространение виртуального общения привели к резкому увеличению количества анонимных текстов, в том числе содержащих прямые и скрытые угрозы, призывы к противоправной деятельности, включая террористическую, героизирующих суицид, скрывающих криминальные намерения педофилов и т. п. Такие тексты являются орудием преступления, в связи с чем закономерно возросла потребность в развитии методик идентификации авторов подобных текстов. Установление факта авторства текста или его опровержение относится к одной из задач судебно-автороведческой экспертизы (САЭ) [Галяшина 2011: 14], которая «отпочковалась от почерковедческой в самостоятельный вид исследования» [Чулахов 2007: 23]. Наряду с лингвистическими и фоноскопическими, она относится к классу судебно-речеведческих экспертиз [Моштылева 2018: 133], что указывает на особую специфику автороведческого исследования текста в сравнении с другими видами лингвистического анализа [Соколова 2018: 125]. САЭ используется при расследовании уголовных дел, связанных с торговлей детьми и использованием рабского труда [Головко 2016], с доведением до самоубийства «группами смерти» [Панина 2016], экстремизмом и терроризмом [Кулешов 2016], а также с клеветой, оскорблением, нарушением авторских и

смежных прав, незаконным изготовлением и оборотом порнографических материалов или предметов и др. [Галяшина 2011: 14].

Задача идентификации автора текста решается отечественными лингвистами, юристами, а в последние годы — и специалистами по информационным технологиям, однако значительного прогресса в этой области сделано не было, что связано во многом с отсутствием интеграции методов указанных направлений, а также ориентированностью исследований лингвистов и специалистов по информационным технологиям на анализ текстов большого объема, преимущественно художественных, использованием лингвистами преимущественно неверифицируемых, субъективных методик анализа языкового материала.

Следует отметить, что в последнее время лингвистами осознается необходимость широкого применения более объективных, количественных методов для решения задачи речеведческих экспертиз, и в частности, задачи атрибуции текста [Баранов 2006; Напреенко 2014], однако лингвистические работы, как правило, не учитывают всего многообразия и возможностей современных методов анализа данных. Усилиями специалистов по информационным технологиям создаются программные комплексы для атрибуции автора текста (см. обзор: [Романченко 2013]), однако, как справедливо отмечается в указанной работе, существующие программные решения ориентированы на тексты большого объема и не применимы в экспертной практике. Кроме того, как лингвисты, так и специалисты по информационным технологиям, как правило, не знакомы с основами общей теории судебной экспертизы и не учитывают требований, предъявляемых к экспертным заключениям. На отсутствие разработанных методик объективной авторoведческой экспертизы текстов и несовершенство используемых в названной экспертизе методов неоднократно указывали и ученые, и эксперты-практики. Ср., например, мнение профессора Института судебной экспертизы Московской государственной юридической академии имени О. Е. Кутафина Е. И. Галяшиной [Галяшина 2006]; ср. также критический анализ экспертных заключений, выполненных учеными-филологами без учета требований, предъявляемых к такого рода исследованиям, в работе [Соколова 2018].

В рамках названной научной области зарубежными исследователями (прежде всего специалистами по информационным технологиям) активно проводятся работы, ориентированные на решение задачи атрибуции текста как одной из задач классификации с использованием инструментария информационного поиска (*information retrieval)* и добычи данных (data mining), проводятся хакатоны по выявлению самых точных классификаторов [Overview of the author identification task at PAN-2018… 2018], однако, как показано в обзорной работе [Authorship Attribution for Social Media Forensics 2017], специально посвященной анализу текстов социальных сетей в идентификационном аспекте, в этой области требуется разработка новых методов, связанных с малым размером текстов, сложностью их автоматической обработки, обусловленной языковыми особенностями естественных письменных текстов. Отметим также нерешенность многих теоретико-прикладных вопросов, связанных с выбором параметров текста при кросс-жанровой атрибуции (типичная ситуация, с которой сталкивается эксперт-авторовед), определением минимального объема текста, необходимого для проведения авторoведческого исследования; возможным отсутствием автора в тестовой выборке и многих других.

Очевидно, что без теоретической основы, как и без использования больших корпусов текстов вкупе с современными методами добычи данных невозможно создание обоснованных и доказательных методик САЭ, однако, как показывает анализ научной литературы, до настоящего времени в науке не сложилось междисциплинарного направления, в которым бы сочетались указанные подходы. На наш взгляд, таким направлением может стать развиваемая нами междисциплинарная область — **корпусная идиолектология**, объектом которой является идиолект как индивидуальная реализация национальной языковой системы. Теоретические проблемы, разрабатываемые в рамках указанного направления, связаны прежде с построением комплексной многофакторной параметрической модели идиолекта, определением степени интериндивидуальной и интраиндивидуальной вариативности идиолектных признаков, определением вклада разных факторов в варьирование идиолектных признаков и т. д. Без решения этих и многих других вопросов невозможно решить и прикладные задачи, такие как идентификация и моделирование личности автора текста, анализ текста на заимствования, выявление в тексте намеренно искаженной информации и т. д.

Следует отметить прежде всего отсутствие общепринятого подхода к определению самого термина «идиолект». В российской науке нет четкой дифференциации между терминами «идиолект» и «идиостиль», при-

чем оба термина преимущественно используются в контексте исследования языка писателя, ученого и других лиц, профессионально владеющих языком. На наш взгляд, идиостиль, т. е. идиолект лица, профессионально владеющего языком, имеющего уникальный авторский стиль, является объектом идиостилистики, тогда как идиолект, т. е. индивидуальный вариант языка, присущий каждому его носителю, должен являться отдельным объектом исследования. Нами предлагается исследовать идиолект в рамках корпусной идиолектологии. Развитие Интернета привело к появления уникального по объему массива непрофессиональных текстов, и их исследование в аспекте авторства логично проводить в рамках названного направления. Именно исследование «естественных» письменных текстов разных жанров [Лебедева 2001] и — шире — идиолектов рядовых носителей языка, которому до последнего времени уделялось мало внимания в сравнении с текстами, созданными мастерами слова, является особенно актуальным для судебной лингвистики [Соколова 2018: 128].

В современной зарубежной лингвистике, в том числе судебной, идиолект понимается прежде всего как совокупность языковых привычек индивида (паттернов), который по-своему использует языковую систему, общую для многих людей, как автоматическое и бессознательное поведение (см., например, [Chaski 2001: 8]), однако общепринятого определения, пригодного для решения задач судебного авторoведения, также не выработано [Crankshaw 2012]. Заметим, что исследователи, занимающиеся проблемой идентификации автора текста, исходят из идеи о стабильности и уникальности идиолекта, однако специальных исследований по интра- и интериндивидуальной стабильности признаков идиолекта крайне мало [On the Stability of Some Idiolectal Features 2018; Litvinova et al. 2018]).

В современных исследованиях (преимущественно англоязычных) идиолект, наряду с голосом, походкой и другими уникальными формами человеческого поведения, рассматривается в рамках поведенческой биометрии [Rozz 2018]), основной задачей которой является идентификация личности. На наш взгляд, такой подход является более обоснованным, чем сравнение идиолекта с ДНК или отпечатками пальцев, как это делается в ряде работ [New Machine Learning Methods… 2005], поскольку идиолект является формой поведения, а не физиологической характеристикой.

В настоящее время научные работы в области изучения идиолекта носителей русского языка активно ведутся в Лаборатории теоретической и прикладной идиолектологии (ранее — Лаборатория корпусной социолингвистики и авторoведческих исследований) (*RusProfilingLab)*, созданной на базе Воронежского государственного педагогического университета под руководством автора статьи (*http://rusprofilinglab.ru*).

Исследования названной лаборатории ориентированы на изучение идиолекта носителя русского языка как «структуры стабильных и вариативных его параметров, репрезентируемых в тексте», или — иначе — «как совокупности устойчивых и вариативных квантифицируемых языковых признаков, обладающих неодинаковыми различительными способностями в аспекте идентификации личности» [Litvinova 2018]) с использованием корпусных данных и методов компьютерной лингвистики.

Как показывают наши исследования, перспективным является рассмотрение идиолекта как набора параметров разного уровня. Одним из важных параметров идиолекта является ортологический (от греч. ortos ‘правильный’) параметр, связанный с отношением продуцента текста к языковой норме и выбору ее вариантов [ср.: Загоровская 2018б]. Правомерность включения названного параметра в структуру идиолекта носителя русского языка подтверждается и реальной практикой специалистов в области судебного авторoведения, учитывающих однотипность ошибок, связанных с нарушением языковых норм, при атрибуции текстов (см. об этом, например: [Маркова 1956]), и достижениями современной теоретической лингвистики, теории языковой нормы и русской ортологии, доказавших особую значимость нормативных/ненормативных выборов в организации языковой личности и языкового сознания носителя русского языка [Загоровская 2016а; Загоровская 2016б; Загоровская 2017]. Наши исследования позволяют также утверждать, что в зависимости от вида языковых норм и норм русского литературного словоупотребления (как известно, виды норм могут разграничиваться на разных основаниях, но для исследований в области лингвистической экспертизы текста наиболее значимым является их типология в соответствии с уровнями языковой системы и формой реализации речи, что предполагает выделение прежде всего норм орфоэпических, лексических, стилистических, грамматических и норм правописания, включающих орфографические и пунктуационные нормы) ортологический параметр идиолекта может репрезентироваться в различных составляющих и предполагать в том числе пунк-

туационный выбор продуцента текста [Загоровская 2018а; Загоровская 2018б].

В настоящее время в ряде работ (преимущественно на материале англоязычных текстов) достаточно определенно доказано, что носителю языка могут быть свойственны определенные пунктуационные привычки, которые проявляются в частотностях знаков препинания (как отдельных, так и в целом), а также в выборе названных знаков в определенной синтаксической позиции, что позволяет говорить о возможности использования пунктуационных выборов в качестве одного из признаков ортологического параметра идиолекта.

Выводы о стабильности пунктуационных привычек носителей языка содержатся, в частности, в работе [Baayen 2002], доказывающей, что частотность пунктуационных знаков является одним из эффективных лингвистических признаков, используемых в кросс-топиковом и кросс-жанровом сценариях (то есть в тех случаях, когда тестовые и контрольные образцы принадлежат разным темам и/или жанрам).

Вывод об устойчивости пунктуационных выборов пишущего представлен в работе известного американского судебного лингвиста К. Часки [Chaski 2001], в которой показано, что синтаксически обоснованные пунктуационные выборы пишущего (syntactically-classified punctuation) в качестве параметров классификатора дают бóльшую точность, чем просто частóты знаков препинания, демонстрирующие различия разных авторов (inter-author identification), а также позволяют установить авторство текстов, созданных одним и тем же автором (intra-author identification). Разные пишущие могут использовать одни и те же знаки препинания с одинаковой частотой, но при этом в разных позициях. Важно отметить, что в работах К. Часки использован весьма узкий круг пунктуационных знаков и синтаксических позиций: анализируются знаки конца предложения, знаки в словосочетаниях (фразах) и знаки в словах (дефис). Также в работе используется ограниченный корпус текстов (5 авторов).

В исследовании [Sapkota 2015] был использован обширный корпусный материал, а также современные алгоритмы машинного обучения для идентификации продуцента текста. Авторы экспериментировали с разными типами n-грамм символов (т. е. последовательностью символов) с учетом позиции символов в слове и выявили, что наивысшая точность моделей достигается для комбинации affix + punct n-grams, т. е. n-грамм (последовательностей), содержащих префиксы и суффиксы, и n-грамм, содержащих, кроме прочих символов, знаки препинания, при этом n-граммы, содержащие знаки пунктуации, лучше всего работают в кросс-топиковом сценарии, что со всей очевидностью доказывает устойчивость пунктуационных привычек носителей языка.

Таким образом, предыдущие работы использовали в качестве признака частóты пунктуационных знаков, в том числе с учетом их синтаксической позиции и контекста в широком понимании (n-граммы символов).

Нами [Litvinova 2019] был расширен список признаков, основанных на автоматическом анализе пунктуационных выборов пишущего. Помимо признаков, использованных в работе [Sapkota 2015], были апробированы три группы признаков:

– триграммы токенов и пунктуационных знаков (как минимум 1 пунктуационный знак в триграмме), при этом слова заменяются обозначением их грамматических категорий, например триграмма *Ты где?* представляется в следующем виде: PRON ADV? (группа признаков PunctPOS);

– n-граммы (n = {3, 4, 5}) токенов и пунктуационных знаков (как минимум 1 пунктуационный знак в н-грамме), при этом слова заменяются знаком * (StarMark) (группа признаков StarMark);

– n-граммы (n = {3, 4, 5}) токенов и пунктуационных знаков (как минимум 1 пунктуационный знак в н-грамме), при этом слова заменяются знаком *, пунктуационный знак заменяется на ‘PNCT’ (группа признаков StarPunct).

Эксперименты по идентификации автора проводились нами на материале текстов форума «Кавказчат» (внесен в Федеральный список экстремистских материалов). Нами были проведены эксперименты по идентификации автора текста в рамках одной темы, а также в кросс-топиковом сценарии. Была выявлена эффективность всех групп «пунктуационных» признаков, при этом их эффективность не падала даже в кросс-жанровом сценарии, что позволяет говорить о стабильности пунктуационных признаков идиолекта, причем выделенные нами группы признаков учитывают не только частотность знаков препинания и контекст, как n-граммы из работы [Sapkota 2015], но и расстояния между ними (в словах), т. е. дополняют существующие в современной науке группы признаков идиолекта, используемые в идентификационных исследованиях [Литвинова 2015].

В дальнейшем мы планируем расширить наши исследования пунктуационных выборов пишущего в двух направлениях.

Во-первых, мы предполагаем извлечь новые группы признаков в зависимости от тех синтаксических позиций, которые в отечественной традиции связываются с понятием «пунктограмма» и предполагают использование определенных нормативных пунктуационных знаков, в том числе как обязательных, так и факультативных, в следующих позициях: 1) конец предложения; 2) границы предикативных частей сложного предложения; 3) границы обособленных членов предложения; 4) границы слов, грамматически не связанных с предложением; 5) границы пунктуационно не разделяемых и не выделяемых членов предложения (главные члены, второстепенные члены, однородные члены, актанты разных типов).

Мы предполагаем, что именно в таких позициях выбор того или иного пунктуационного знака, его замена и даже нарушение пунктуационной нормы чаще всего являются следствием индивидуальных предпочтений автора текста.

Разметка в таком случае может выполняться только вручную специалистом-лингвистом, что, безусловно, требует трудовых и временных затрат, однако сочетание ручных и автоматизированных методов анализа языкового материала является, как показывает наш опыт работы по моделированию личности автора письменного текста, обязательным методологическим принципом корпусной идиолектологии.

Вторым направлением работ является расширение корпусного материала. В настоящее время мы работаем с материалами корпуса «Кавказчат» как реальным языковым материалом, который, в силу своего содержания, нуждается в идентификационном анализе, а также мы используем материалы и возможности созданного в Лаборатории корпуса естественных письменных текстов *Ruspersonality* с метаданными о социально-демографических и личностных характеристиках авторов текстов (пол, возраст, уровень образования, данные психологического тестирования, профессия и т. д.) (подробнее о корпусе см.: [“RusPersonality”: A Russian corpus… 2016]).

В настоящее время нами проводится сбор корпуса текстов *RusIdioStyle* одних и тех же авторов в различных языковых модусах (формах существования национального русского языка и его функциональных разновидностях) и различных жанрах, продуцированных в условиях специального эксперимента, а также в «реальных» условиях. Представляется, что материалы названного корпуса окажутся наиболее значимыми для проведения статистического анализа, ориентированного на определение уровня стабильности и вариативности пунктуационных выборов автора русского письменного текста, и существенно дополнят представления о зонах квантифицируемых признаков идиолекта, на основании которых могут быть построены надежные методики идентификационной авторoведческой экспертизы.

Следует отметить, что изучение стабильности и вариативности пунктуационных выборов автора русского письменного текста как компонентов ортологического параметра идиолекта и использование соответствующего параметра для идентификации продуцента текста представляется особенно перспективным ввиду типологических особенностей системы пунктуации в русском языке, а также специфики самой пунктуационной нормы, которая, по мнению многих современных специалистов в области русского синтаксиса и русской пунктуации, по самой своей сущности является коммуникативно-прагматической, т. е. регулирующей употребление пунктуационных знаков во многих случаях не в соответствии с предписаниями, а в соответствии в теми условиями, которые устанавливаются в конкретной коммуникативной ситуации. Вместе с тем широкие возможности пунктуационных выборов, заложенные в самой пунктуационной норме русского языка, требуют от исследователей, работающих в рассматриваемой области, не только глубокого осмысления синтаксических явлений, передаваемых теми или иными пунктуационными знаками, но и поиска новых подходов к предварительной обработке языковых материалов, разметке текстов и их анализу с помощью современных методов корпусной лингвистики и математической статистики.

ЛИТЕРАТУРА

1. Баранов А. Н. Теория лингвистических экспертиз как направление прикладной лингвистики // Компьютерная лингвистика и интеллектуальные технологии : материалы ежегод. конф. «Диалог». — М. : Наука, 2004. С. 27—31.

2. Галяшина Е. И. Речеведческие экспертизы в судопроизводстве // Законы России: опыт, анализ, практика. 2011. № 12. С. 12—29.

3. Галяшина Е. И., Приводнова Е. В. Авторoведческая экспертиза в российском судопроизводстве // Lex Russica. 2006. № 4. С. 55—61.

4. Головко Н. В. Значение судебных экспертиз для успешного расследования уголовных дел о торговле детьми и использовании их рабского труда // Вестн. Акад. 2016. № 2. С. 101—104.

5. Загоровская О. В. Нормы русского литературного языка: типология и основания для классификации // Изв. Воронеж. гос. пед. ун-та. 2016а. № 3 (272). С. 129—134.

6. Загоровская О. В. Языковая норма в современной русской визуально-письменной речи, функционирующей в интернет-коммуникации: к постановке проблемы // Изв. Воронеж. гос. пед. ун-та. 2017. № 4 (277). С. 168—172.

7. Загоровская О. В. Языковая норма и норма литературного языка как лингвистические понятия // Изв. Воронеж. гос. пед. ун-та. 2016б. № 2 (271). С. 161—165.

8. Загоровская О. В., Литвинова Т. А. Корпус текстов RusPersonality как основа исследований «реальной» языковой нормы в современной русской письменной речи // Современные проблемы лингвистики и методики преподавания русского языка в вузе и школе / под ред. О. В. Загоровской. Вып. 28. — Воронеж : ИПЦ «Научная книга», 2018а. С. 51—57.

9. Загоровская О. В., Литвинова Т. А. Электронная база данных о языковой норме и ее вариантности как основа научных исследований ортологического параметра идиолекта // Изв. Воронеж. гос. пед. ун-та. 2018б. № 3 (280). С. 138—143.

10. Кулешов Р. В. Роль судебно-авторoведческой экспертизы в расследовании преступлений экстремистской и террористической направленности: типичные задачи, особенности назначения, соотношение со смежными видами экспертиз // Юридическая наука и правоохранительная практика. 2016. № 3 (37). С. 147—152.

11. Лебедева Н. Б. Естественная письменная русская речь: проблемы изучения // Русский язык: исторические судьбы и современность : Междунар. конгр. исследователей русского языка : труды и материалы. — М., 2001. С. 260—261.

12. Литвинова Т. А., Литвинова О. А. Идентификация и моделирование личности автора письменного текста. — Воронеж : Изд-во ВГПУ, 2015. 322 с.

13. Маркова Г. Д. Идентификационные признаки письма в советской криминалистической экспертизе : автореф. дис. … канд. юрид. наук. — Харьков, 1956. 24 с.

14. Моштылева Е. С. Классификационное место речеведческих экспертиз в теории и практике судебной экспертизы // Вестн. ННГУ. 2018. № 4. С. 131—135.

15. Напреенко Г. В. Идентификация текста по его авторской принадлежности на лексическом уровне (формально-количественная модель) // Вестн. Томск. гос. ун-та. 2014. № 379. С. 17—23.

16. Панина Н. А. О роли судебной авторoведческой экспертизы при расследовании преступлений, связанных с доведением до самоубийства «группами смерти» // Традиции и новации в системе современного российского права : сб. тезисов 17-й Междунар. науч.-практ. конф. молодых ученых. — М. : ООО «Проспект». 2018. С. 848—850.

17. Романченко Т. Н. Методы атрибуции в авторoведческой экспертизе // Вестн. СГЮА. 2013. № 2 (91). С. 228—233.

18. Соколова Т. П. Роль специальных знаний в судебной авторoведческой экспертизе // Вестн. Ун-та им. О. Е. Кутафина. 2018. № 7 (47). С. 123—131.

19. Чулахов В. Н. Криминалистическое учение о навыках и привычках человека / под ред. Е. Р. Россинской. — М. : Юрлитинформ, 2007. 285 с.

20. "RusPersonality": A Russian corpus for authorship profiling and deception detection / T. Litvinova [et. al.] // Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT 2016). IEEE. С. 1-7.

21. Authorship Attribution for Social Media Forensics / A. Rocha [et al.] // IEEE Transactions on Information Forensics and Security. 2017. Vol. 12, Iss. 1. P. 5-33.

22. Baayen H., Halteren van H., Neijt A., Tweedie F. An experiment in authorship attribution // Proc. of 6th JADT. 2002. P. 29—37.

23. Chaski C. Empirical evaluations of language-based author identification techniques // Forensic Linguistics. 2001. Vol. 8. P. 1-65.

24. Crankshaw R. The validity of the Linguistic Fingerprint in forensic investigation. Diffusion: the UCLan Journal of Undergraduate Research. 2012. Vol., 5 Iss. 2. URL: http://bcur.org/journals/index.php/Diffusion/article/view/92 (last accessed: 17.01.2019).

25. Litvinova T.A., Panicheva P.V., Litvinova O.A. Authorship Attribution of Russian Extremist Forum Texts with Different Types of N-gram Features // Submitted for CICLING 2019.

26. Litvinova T.A., Seredin P.V., Litvinova O.A. Assessing the Level of Stability of Idiolectal Features across Modes, Topics and Time of Text Production // S. Balandin, T. Cinotti, F. Viola, T. Tyutina (eds). Proceedings of the 23rd Conference of Open Innovations Association FRUCT. — IEEE, 2018. P. 223-230.

27. New Machine Learning Methods Demonstrate the Existence of a Human Stylome / H.V. Halteren [et al.] // Journal of Quantitative Linguistics. 2005. № 12. P. 65-77.

28. On the Stability of Some Idiolectal Features / T. Litvinova [et. al.] // Lecture Notes in Computer Science. 2018. Vol. 11096. С. 331—336.

29. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection / M. Kestemont [et al.] // Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018 / L. Cappellato [edit.]; et al. 2018. С. 1-25.

30. Rozz Y., Menezes R. Author Attribution Using Network Motifs // Cornelius S. et al. (eds). Complex Networks IX: Proceedings of the 9th Conference on Complex Networks. — Springer, 2018. P. 199-207.

31. Sapkota U., Bethard S., Montes M., Solorio T. Not all character n-grams are created equal: A study in authorship attribution // Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies. P. 93-102.

T. A. Litvinova
Voronezh State Pedagogical University, Voronezh, Russia
ORCID ID: 0000-0002-6019-3700 ☑

☑ ***E-mail:*** *centr_rus_yaz@mail.ru.*

Punctuation Choice as a Component of Orthological Parameter of the Modern Russian Speaker's Idiolect in Forensic Authorship Analysis

ABSTRACT. *Currently, due to the development of Internet communication and the emergence of an array of text data, part of which contain malicious content, the problem of identifying the author of the text in forensic settings has become particularly urgent. However, exact and objective methods of text attribution, which could be used in forensic authorship analysis, have not yet been worked out, and both researchers and experts emphasize the particular complexity of the authorship analysis of the texts of Internet communication. The paper deals with the issues of creation of the tools of forensic authorship analysis based on quantitative markers of an idiolect as individual realization of the language system. The author analyzes punctuation choice of the writer as one of the components of the orthological parameter of an idiolect and justifies the relevance of the analysis of punctuation choice for authorship identification. Analysis of the scientific literature on text attribution, as well as the author of the paper's own experimental studies carried out on the material of extremist forum texts, shows that the punctuation choice of the writer represented by a number of quantifiable idiolect features, including those first suggested by the author of this article, are quite resistant to topic change, which allows them to be used for assessment of cross-topic scenarios. The author also formulates areas for further research, primarily related to the development of a specialized corpus containing texts of different genres, modes etc. by the same authors, as well as to the detection of new features characterizing the punctuation choice of the author of the text.*

KEYWORDS: *forensic authorship analysis, Russian language; idiolects, text attribution; orthological parameters of idiolects, punctuation marks, linguistics, text corpus.*

AUTHOR'S INFORMATION: *Litvinova Tat'yana Aleksandrovna, Candidate of Philology, Head of Laboratory of Theoretical and Applied Idiolectology, Voronezh State Pedagogical University, Voronezh, Russia.*

FOR CITATION: *Litvinova, T. A.* Punctuation Choice as a Component of Orthological Parameter of the Modern Russian Speaker's Idiolect in Forensic Authorship Analysis / T. A. Litvinova // Political Linguistics. — 2019. — No 1 (73). — P. 114—121.

ACKNOWLEDGMENTS: Research is accomplished with financial support of the Russian Foundation for Basic Research grant within the scientific project №18-78-10081 "Modeling of the Modern Russian Speaker's Idiolect in Forensic Authorship Analysis".

REFERENCES

1. Baranov A. N. The Theory of Linguistic Expertise as a Direction of Applied Linguistics // Computational Linguistics and Intellectual Technologies : materials of annually conf. "Dialogue". — Moscow : Science, 2004. P. 27—31. [Teoriya lingvisticheskikh ekspertiz kak napravlenie prikladnoy lingvistiki // Komp'yuternaya lingvistika i intellektual'nye tekhnologii : materialy ezhegod. konf. «Dialog». — M. : Nauka, 2004. S. 27—31]. — (In Rus.)

2. Galyashina E. I. Speech Expertise in Legal Proceedings // Laws of Russia: Experience, Analysis, Practice. 2011. No. 12. P. 12— 29. [Rechevedcheskie ekspertizy v sudoproizvodstve // Zakony Rossii: opyt, analiz, praktika. 2011. № 12. S. 12—29]. — (In Rus.)

3. Galyashina E. I., Privodnova E. V. Autorological Examination in Russian Legal Proceedings // Lex Russica. 2006. № 4. P. 55—61. [Avtorovedcheskaya ekspertiza v rossiyskom sudoproizvodstve // Lex Russica. 2006. № 4. S. 55—61]. — (In Rus.)

4. Golovko N. V. The Value of Forensic Examinations for Successful Investigation of Criminal Cases on the Sale of Children and the Use of Their Slave Labor // Herald of Academy. 2016. No. 2. P. 101—104. [Znachenie sudebnykh ekspertiz dlya uspeshnogo rassledovaniya ugolovnykh del o torgovle det'mi i ispol'zovanii ikh rabskogo truda // Vestn. Akad. 2016. № 2. S. 101—104]. — (In Rus.)

5. Zagorovskaya O. V. Norms of the Russian Literary Language: Typology and Grounds for Classification // News of Voronezh State Ped. Univ. 2016a. Number 3 (272). P. 129—134. [Normy russkogo literaturnogo yazyka: tipologiya i osnovaniya dlya klassifikatsii // Izv. Voronezh. gos. ped. un-ta. 2016a. № 3 (272). S. 129—134]. — (In Rus.)

6. Zagorovskaya O. V. Language Norm in Modern Russian Visual-written Speech, Functioning in Internet Communications: to the Formulation of the Problem // News of Voronezh State Ped. Univ. 2017. No. 4 (277). P. 168—172. [Yazykovaya norma v sovremennoy russkoy vizual'no-pis'mennoy rechi, funktsioniruyushchey v internet-kommunikatsii: k postanovke problemy // Izv. Voronezh. gos. ped. un-ta. 2017. № 4 (277). S. 168—172]. — (In Rus.)

7. Zagorovskaya O. V. Language Norm and Norm of Literary Language as Linguistic Concepts // News of Voronezh State Ped. Univ. 2016b. Num. 2 (271). P. 161—165. [Yazykovaya norma i norma literaturnogo yazyka kak lingvisticheskie ponyatiya // Izv. Voronezh. gos. ped. un-ta. 2016b. № 2 (271). S. 161—165]. — (In Rus.)

8. Zagorovskaya O. V., Litvinova T. A. The Corpus of Texts RusPersonality as the basis for Research on the "Real" Language Standard in Modern Russian Written Speech // Modern Problems of Linguistics and Methods of Teaching Russian in High School and School / ed. O.V. Zagorovskoy. Iss. 28. — Voronezh: Scientific Research Educational Center, 2018a. P. 51—57. [Korpus tekstov RusPersonality kak osnova issledovaniy «real'noy» yazykovoy normy v sovremennoy russkoy pis'mennoy rechi // Sovremennye problemy lingvistiki i metodiki prepodavaniya russkogo yazyka v vuze i shkole / pod red. O. V. Zagorovskoy. Vyp. 28. — Voronezh : IPTs «Nauchnaya kniga», 2018a. S. 51—57]. — (In Rus.)

9. Zagorovskaya O. V., Litvinova T. A. Electronic Database on the Language Norm and Its Variance as the Basis for Scientific Research on the Orthological Parameter of the Idiolect // News of Voronezh State Ped. Univ. 2018b. Num. 3 (280). P. 138—143. [Elektronnaya baza dannykh o yazykovoy norme i ee variantnosti kak osnova nauchnykh issledovaniy ortologicheskogo parametra idiolekta // Izv. Voronezh. gos. ped. un-ta. 2018b. № 3 (280). S. 138—143]. — (In Rus.)

10. Kuleshov R. V. The Role of Forensic-Authorship Expert Examination in Investigating Crimes of Extremist and Terrorist Orientation: Typical Tasks, Specifics of Assignment, Relationship with Related Types of Examinations. 2016. № 3 (37). P. 147—152. [Rol' sudebno-avtorovedcheskoy ekspertizy v rassledovanii prestupleniy ekstremistskoy i terroristicheskoy napravlennosti: tipichnye zadachi, osobennosti naznacheniya, sootnoshenie so smezhnymi vidami ekspertiz // Yuridicheskaya nauka i pravookhranitel'naya praktika. 2016. № 3 (37). S. 147—152]. — (In Rus.)

11. Lebedeva N. B. Natural Written Russian Speech: Problems of Studying // Russian Language: Historical Destinies and Modernity: Intern. Congr. of Russian Language Researchers : works and materials. — Moscow, 2001. P. 260—261. [Estestvennaya pis'mennaya russkaya rech': problemy izucheniya // Russkiy yazyk: istoricheskie sud'by i sovremennost' : Mezhdunar. kongr. issledovateley russkogo yazyka : trudy i materialy. — M., 2001. S. 260—261]. — (In Rus.)

12. Litvinova T. A., Litvinova O. A. Identification and Modeling of the Personality of the Author of the Written Text. — Voronezh : VGPU Publ. House, 2015. 322 p. [Identifikatsiya i modelirovanie lichnosti avtora pis'mennogo teksta. — Voronezh : Izd-vo VGPU, 2015. 322 s.]. — (In Rus.)

13. Markova G. D. Identification Signs of the Letter in the Soviet Forensic Examination: abstract of doctoral thesis ... of Cand. in Law. — Kharkov, 1956. 24 p. [Identifikatsionnye priznaki pis'ma v sovetskoy kriminalisticheskoy ekspertize : avtoref. dis. … kand. yurid. nauk. — Khar'kov, 1956. 24 s.]. — (In Rus.)

14. Moshtyleva E. S. Classification of Speech Expertise in the Theory and Practice of Forensic Examination // Herald of UNN. 2018. No. 4. P. 131—135. [Klassifikatsionnoe mesto rechevedcheskikh ekspertiz v teorii i praktike sudebnoy ekspertizy // Vestn. NNGU. 2018. № 4. S. 131—135]. — (In Rus.)

15. Napreenko G. V. Text Identification by Its Author's Identity at the Lexical Level (formal-quantitative model) // Herald of Tomsk State Univ. 2014. No. 379. P. 17—23. [Identifikatsiya teksta po ego avtorskoy prinadlezhnosti na leksicheskom urovne (formal'no-kolichestvennaya model') // Vestn. Tomsk. gos. un-ta. 2014. № 379. S. 17—23]. — (In Rus.)

16. Panina N. A. On the Role of Judicial Authoring Expertise in Investigating Crimes Related to "Bringing Death Groups To Suicide" // Traditions and Innovations in the System of Modern Russian Law: Coll. Abstracts of the 17th International scientific-practical conf. for young scientists. — Moscow : Prospect LLC, 2018. P. 848—850. [O roli sudebnoy avtorovedcheskoy ekspertizy pri rassledovanii prestupleniy, svyazannykh s dovedeniem do samoubiystva «gruppami smerti» // Traditsii i novatsii v sisteme sovremennogo rossiyskogo prava : sb. tezisov 17-y Mezhdunar. nauch.-prakt. konf. molodykh uchenykh. — M. : OOO «Prospekt». 2018. S. 848—850]. — (In Rus.)

17. Romanchenko T. N. Methods of Attribution in the Author's Expert Examination // Herald of SGUA. 2013. № 2 (91). P. 228—233. [Metody atributsii v avtorovedcheskoy ekspertize // Vestn. SGYuA. 2013. № 2 (91). S. 228—233]. — (In Rus.)

18. Sokolova T. P. The Role of Special Knowledge in the Judicial Authors Theory Expertise // Herald of Univ. named after O. E. Kutafina. 2018. No. 7 (47). P. 123—131. [Rol' spetsial'nykh znaniy v sudebnoy avtorovedcheskoy ekspertize // Vestn. Un-ta im. O. E. Kutafina. 2018. № 7 (47). S. 123—131]. — (In Rus.)

19. Chulakhov V. N. Forensic Theory of Human Habits and Behavior / ed. E. R. Rossinskaya. — M.: Yurlitinform, 2007. 285 p.

[Kriminalisticheskoe uchenie o navykakh i privychkakh cheloveka / pod red. E. R. Rossinskoy. — M. : Yurlitinform, 2007. 285 s.]. — (In Rus.)

20. RusPersonality": A Russian corpus for authorship profiling and deception detection / T. Litvinova [et. al.] // Proceedings of the International FRUCT Conference on Intelligence, Social Media and Web (ISMW FRUCT 2016). IEEE. C. 1-7.

21. Authorship Attribution for Social Media Forensics / A. Rocha [et al.] // IEEE Transactions on Information Forensics and Security. 2017. Vol. 12, Iss. 1. P. 5-33.

22. Baayen H., Halteren van H., Neijt A., Tweedie F. An experiment in authorship attribution // Proc. of 6th JADT. 2002. P. 29—37.

23. Chaski C. Empirical evaluations of language-based author identification techniques // Forensic Linguistics. 2001. Vol. 8. P. 1—65.

24. Crankshaw R. The validity of the Linguistic Fingerprint in forensic investigation. Diffusion: the UCLan Journal of Undergraduate Research. 2012. Vol., 5 Iss. 2. URL: http://bcur.org/journals/index.php/Diffusion/article/view/92 (last accessed: 17.01.2019).

25. Litvinova T.A., Panicheva P.V., Litvinova O.A. Authorship Attribution of Russian Extremist Forum Texts with Different Types of N-gram Features // Submitted for CICLING 2019.

26. Litvinova T.A., Seredin P.V., Litvinova O.A. Assessing the Level of Stability of Idiolectal Features across Modes, Topics and Time of Text Production // S. Balandin, T. Cinotti, F. Viola, T. Tyutina (eds). Proceedings of the 23rd Conference of Open Innovations Association FRUCT. — IEEE, 2018. P. 223-230.

27. New Machine Learning Methods Demonstrate the Existence of a Human Stylome / H.V. Halteren [et al.] // Journal of Quantitative Linguistics. 2005. № 12. P. 65-77.

28. On the Stability of Some Idiolectal Features / T. Litvinova [et. al.] // Lecture Notes in Computer Science. 2018. Vol. 11096. P. 331—336.

29. Overview of the author identification task at PAN-2018: cross-domain authorship attribution and style change detection / M. Kestemont [et al.] // Working Notes Papers of the CLEF 2018 Evaluation Labs. Avignon, France, September 10-14, 2018 / L. Cappellato [edit.]; et al. 2018. C. 1-25.

30. Rozz Y., Menezes R. Author Attribution Using Network Motifs // Cornelius S. et al. (eds). Complex Networks IX: Proceedings of the 9th Conference on Complex Networks. — Springer, 2018. P. 199-207.

31. Sapkota U., Bethard S., Montes M., Solorio T. Not all character n-grams are created equal: A study in authorship attribution // Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics: Human language technologies. P. 93-102.