ISSN 2658-5014 (Print) СИИТ ISSN 2686-7044 (Online)
http://siit.ugatu.su 2026. Т. 8, № 2(26)
СИСТЕМНАЯ ИНЖЕНЕРИЯ И ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
Научная статья

УДК 004.8 DOI: 10.54708/SIIT-2026-no2-p75 EDN: UOLWFG

Мультипараметрический метод установления авторства текста: интеграция стилометрического, тематического и прагматического анализа

Д. С. Алексеева

Уфимский университет науки и технологий

В условиях тотальной цифровизации коммуникаций текст стал основным носителем криминалистически значимой информации, что требует развития новых методов его автоматизированного анализа для нужд расследования. Традиционные подходы, такие как стилометрия, семантический или прагматический анализ, применяются изолированно, что ограничивает полноту и надежность лингвистической экспертизы. В данной статье предлагается комплексный мультипараметрический метод атрибуции анонимных или спорных текстов, интегрирующий три независимых лингвистических уровня: формально-статистический (стилометрия на основе дельты Бёрроуза), смысловой (тематическое моделирование с использованием Latent Dirichlet Allocation) и прагматический (оценка достоверности нарратива по критериям CBCA – Criteria-Based Content Analysis). Метод предполагает извлечение соответствующих векторов признаков, вычисление на их основе метрик стилистического, тематического и прагматического сходства и их последующую взвешенную агрегацию в итоговый показатель соответствия (Compliance Indicator), максимизирующий вероятность корректной атрибуции авторства. Теоретическая значимость работы заключается в разработке формальной модели комплексного лингвистического «цифрового профиля», а практическая – в создании методологического базиса для инструментов поддержки принятия решений при расследовании киберпреступлений, экстремизма и других сложно структурированных деяний, опосредованных цифровым текстом.

Лингвокриминалистика; атрибуция текста; установление авторства; мультипараметрический анализ; стилометрия; тематическое моделирование (LDA); анализ достоверности (CBCA); цифровая криминалистика.

Введение

Исторически применение лингвистических знаний в правоприменительной практике ограничивалось преимущественно судебно-почерковедческой экспертизой, объектом которой выступали индивидуально-устойчивые особенности двигательного навыка исполнения рукописного текста [Кры89]. Однако тотальная цифровизация социальных коммуникаций привела к парадигмальному сдвигу в природе криминалистически релевантных данных. Доминирующим медиумом фиксации преступных намерений, сговоров и действий стал текст в его электронной форме: коммуникация в мессенджерах, социальных сетях, тематических форумах и зашифрованных каналах [Tur10]. Данная трансформация потребовала пересмотра методологического аппарата и интеграции методов компьютерной лингвистики и анализа больших данных в криминалистический инструментарий.

Ключевым объектом исследования современной лингвокриминалистики выступает не материальный носитель, а лингвистическая структура текста, рассматриваемая как проекция идиолекта – устойчивой совокупности языковых привычек индивида, проявляющихся

Алексеева Д. С. Мультипараметрический метод установления авторства текста: интеграция стилометрического, тематического и прагматического анализа // СИИТ. 2026. Т. 8, № 2(26). С. 75-83. DOI: 10.54708/SIIT-2026-no2-p75. EDN: UOLWFG.

Alekseeva D. S. “Multiparametric method of text authorship attribution: integration of stylometric, thematic, and pragmatic analysis” // SIIT. 2026. Vol. 8, no. 2(26), pp. 75-83. DOI: 10.54708/SIIT-2026-no2-p75. EDN: UOLWFG. (In Russian).

на лексическом, синтаксическом, морфологическом и прагматическом уровнях [Cou07]. Идиолект обладает свойствами устойчивости и непроизвольности формирования, что позволяет рассматривать его в качестве аналога биометрического идентификатора личности в цифровой среде и криминалистике [Gra13].

Однако современные инструменты лингвокриминалистики часто развиваются и применяются изолированно. Стилометрические методы (такие как дельта Бёрроуза) эффективны для измерения формального сходства, но игнорируют смысловое содержание и коммуникативные интенции. Семантические технологии (тематическое моделирование, LDA) позволяют выявлять скрытые смысловые структуры, но не чувствительны к индивидуальным авторским особенностям формы. Прагматические подходы (лингвистический анализ содержания, CBCA) фокусируются на достоверности и интенциях нарратива, но не решают задачу атрибуции напрямую. Таким образом, возникает научно-практическая проблема: отсутствие формализованного метода, который интегрировал бы разноуровневые лингвистические признаки в единую модель для повышения точности, надежности и обоснованности криминалистической атрибуции текста, особенно в условиях больших массивов неструктурированных цифровых данных.

Целью настоящего исследования является разработка мультипараметрического метода атрибуции текста, основанного на интеграции стилометрического, тематического и прагматического анализа. Для достижения поставленной цели решаются следующие задачи:

1. Систематизировать требования к комплексной лингвистической модели авторского профиля («цифрового отпечатка»).
2. Предложить архитектуру метода, объединяющего три независимых компонента: вычисление стилистического сходства на основе модифицированной дельты Бёрроуза, тематического сходства с применением Latent Dirichlet Allocation (LDA) и прагматического сходства на основе критериев оценки достоверности утверждений (CBCA).
3. Определить формальные процедуры извлечения соответствующих векторов признаков, вычисления метрик сходства и их взвешенной агрегации в итоговый показатель соответствия (Compliance Indicator).
4. Обосновать теоретическую и практическую значимость предлагаемого подхода для задач судебно-экспертной деятельности и оперативно-розыскного анализа.

Научная новизна исследования заключается в следующем:

- Впервые предлагается формальная интегративная модель атрибуции текста, одновременно учитывающая формальные (стилометрические), семантические (тематические) и прагматические (контент-аналитические) лингвистические признаки.
- Предлагается новый алгоритмический конвейер (pipeline), включающий этапы многомерного признакового описания, вычисления трех специализированных метрик сходства и их линейной агрегации с весовыми коэффициентами.
- Вводится композитный показатель соответствия (CI), количественно оценивающий вероятность авторства на основе комплексного лингвистического профиля.

Методы атрибуции текста и стилометрия

Ключевой задачей судебной лингвистики в контексте расследования преступлений является установление авторства анонимного или спорного текста. Данная задача, известная как атрибуция текста, заключается в определении, принадлежит ли анализируемый документ конкретному подозреваемому или является продуктом деятельности одного и того же анонимного источника [Kop11]. Современная методология решения этой задачи отошла от интуитивно-стилистического анализа в сторону формализованных, статистически верифицируемых процедур, основанных на концепции идиолекта.

Основу методологии составляет стилометрический анализ, направленный на выявление и количественное измерение подсознательно воспроизводимых, устойчивых лингвистических

признаков индивидуума. Эти признаки обладают низкой осознаваемостью для автора и, следовательно, высокой диагностической ценностью. К числу наиболее репрезентативных стилометрических маркеров относятся: лексико-статистические параметры (частотность употребления служебных слов, местоимений, показатель лексического разнообразия), синтаксические характеристики (средняя длина и глубина вложенности предложений, преференции в использовании определенных грамматических конструкций) и морфологические паттерны (соотношение частей речи) [Juo08]. Ключевым допущением является гипотеза о том, что совокупность таких сублексических признаков образует уникальный и относительно стабильный профиль, аналогичный поведенческому биометрическому шаблону.

Наряду со статистическими методами, углубленный анализ фокусируется на выявлении идиолектных особенностей, составляющих уникальную лингвистическую «подпись» лица. Сюда входят индивидуально-авторские неологизмы, устойчивые сочетания слов, специфические ошибки (орфографические, пунктуационные), регулярное использование определенных дискурсивных маркеров и прагматических стратегий [Gra13]. В отличие от стилометрии, анализирующей распределение абстрактных признаков, идиолектный анализ требует герменевтического погружения в семантику и прагматику текста.

Операционализация данных методов в условиях больших объемов цифровой информации стала возможной благодаря применению технологий машинного обучения и анализа больших данных. Алгоритмы контролируемого обучения (такие как метод опорных векторов, случайный лес, нейронные сети) тренируются на референсных корпусах текстов, аутентифицированных как принадлежащие конкретному лицу, и в дальнейшем классифицируют анонимные тексты по степени сходства выявленных лингвистических паттернов [Sta09]. Данный подход позволяет не только проводить бинарное сравнение, но и решать задачи кластеризации, выявляя группы текстов, вероятно созданных одним автором, в массивах неразмеченных данных, например, при мониторинге коммуникаций в даркнете с целью идентификации ключевых фигур преступного сообщества.

Семантический и дискурс-анализ в криминалистических задачах

Задача автороведческой экспертизы по установлению индивидуального авторства дополняется и обогащается задачами семантического и дискурс-анализа, направленными на реконструкцию имплицитных смыслов, идеологических установок, интенций и структур социального взаимодействия, закодированных в текстовой коммуникации преступных групп [Cou07]. В отличие от стилометрии, фокусирующейся на форме, данный подход исследует содержание и прагматику высказываний, что позволяет перейти от идентификации автора к пониманию логики, организации и динамики преступной деятельности.

Методологический аппарат данного направления образует конвергенция компьютерной лингвистики и критического дискурс-анализа. Одним из ключевых инструментов является тематическое моделирование, в частности, метод латентного размещения Дирихле (Latent Dirichlet Allocation, LDA). Данный алгоритм вероятностного моделирования позволяет автоматически выявлять скрытые (латентные) тематические кластеры в больших корпусах неразмеченных текстов, таких как архивы форумов или переписка в мессенджерах [Ble03]. В криминалистическом контексте это позволяет объективно категоризировать контент, отслеживать эволюцию дискуссий и идентифицировать ключевых идеологических агентов, наиболее активно генерирующих сообщения по целевым темам.

Параллельно анализ тональности и эмоциональной окраски обеспечивает идентификацию речевых актов, несущих агрессивную, угрожающую или манипулятивную нагрузку. Современные методы, основанные на глубоком обучении, способны детектировать не только явную вербальную агрессию, но и косвенные формы запугивания, иронию или психологическое давление, что имеет критическое значение при оценке степени общественной опасности высказываний или доказательстве факта угрозы [Liu12].

Наиболее комплексный уровень анализа представляет собой сетевой анализ дискурса, интегрирующий лингвистические данные с методами социометрии. Данный подход рассматривает участников коммуникации как узлы сети, а их лингвистические взаимодействия (цитирование, обращение по имени, использование общей специфической лексики, ответные реплики) – как связи. Последующий анализ центральности позволяет с высокой степенью достоверности реконструировать неформальную иерархию внутри группы, выделяя лидеров мнений, координаторов, исполнителей и периферийных участников [Die14]. Так, при расследовании деятельности организованной преступной группы частотность и паттерны директивных речевых актов, распределение ролей в диалоге и лексическая конвергенция между участниками могут служить не менее весомым доказательством ролевой дифференциации и наличия сговора, чем прямое упоминание преступных планов.

Лингвистический анализ содержания (LCA) и оценка достоверности сообщений

Задача оценки достоверности устных и письменных показаний представляет собой отдельное направление судебной лингвистики, ориентированное на выявление лингвистических коррелятов когнитивных процессов, связанных с производством правдивых или ложных сообщений [Vri14]. В отличие от полиграфа, фиксирующего физиологические реакции, лингвистический анализ сосредоточен на вербальных и просодических особенностях речи, возникающих в результате повышенной когнитивной нагрузки, стратегического управления информацией или эмоционального стресса, сопровождающих деконтекстуализацию события.

Основным методологическим подходом в данной области является лингвистический анализ содержания (Linguistic Content Analysis, LCA), разработанный на основе критериев оценки достоверности утверждений (Criteria-Based Content Analysis, CBCA). LCA фокусируется на систематической оценке качественных параметров повествования, среди которых ключевыми являются: логическая последовательность и внутренняя непротиворечивость описания события; специфичность и контекстуальная встроенность деталей; особенности лексического выбора, такие как избегание местоимения первого лица единственного числа, повышенная частота глаголов действия в ущерб описанию психических состояний, а также наличие аномальных пауз и оговорок [Spo16]. Предполагается, что описание реально пережитого события, извлекаемого из эпизодической памяти, будет демонстрировать более высокие показатели по параметрам сенсорной детализации и логической связности, в то время как сконструированный нарратив будет характеризоваться обобщенностью, логическими разрывами и стратегическим избеганием самоидентификации с сообщаемыми фактами.

Крайне важно подчеркнуть, что лингвистические маркеры не являются диагностическими признаками недостоверной информации, а интерпретируются как индикаторы речевого поведения, отклоняющегося от базовой модели правдивого повествования, сформированной для данного коммуникативного контекста [Mas17]. Следовательно, их выявление указывает не на установление факта лжи, а на зоны нарратива, требующие дополнительного процессуального внимания и проверки иными доказательствами. Надежность выводов напрямую зависит от учета индивидуальных речевых особенностей, ситуационного контекста и кросс-культурных различий в коммуникативных стилях.

Интегрированный метод атрибуции текста на основе стилометрических, тематических и прагматических признаков

Предлагаемый метод интегрирует три независимых лингвистических подхода:

- стилометрический анализ на основе дельты Бёрроуза для измерения формально-статистического сходства с текстами-эталонами;
- моделирование с применение распределения Дирихле для выявления и сравнения смысловых структур;

- лингвистический анализ содержания на основе критериев оценки достоверности для оценки правдоподобия.

На рисунке представлен алгоритм предлагаемого метода, который позволит идентифицировать авторство контента.

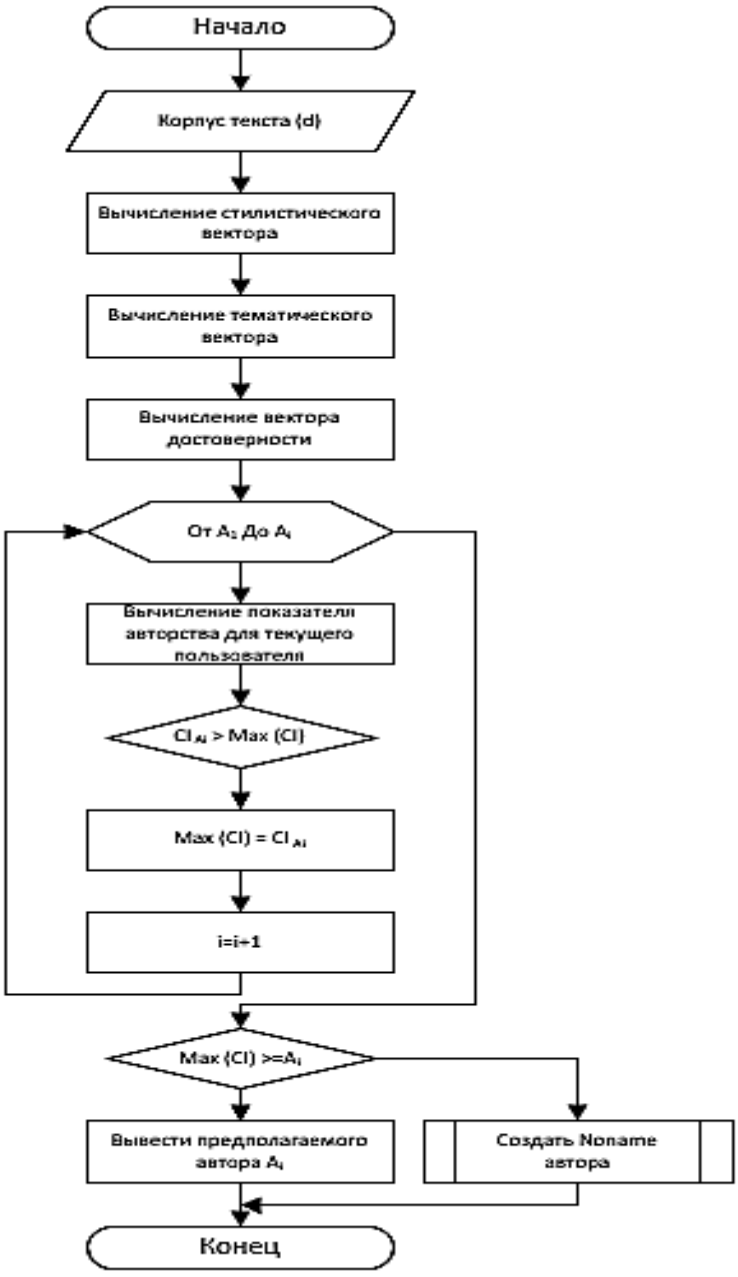

Рис. Алгоритм предлагаемого метода определения авторства

Каждый блок алгоритма предполагает работу над корпусом входного текста и эталонного. Для начала вычисляется дельта Бёрроуза:

$$\Delta_B(d, A) = \frac{1}{m}\sum_{i=1}^{m} |z_i(d) - z_i(A)|, \quad (1)$$

где d – анонимный корпус текста; m – количество ключей; $z_i(A)$ – средний показатель упоминания ключевого слова по массиву текста пользователя A. Чем меньше $\Delta_B(d, A)$, тем выше вероятность авторства пользователя.

Вычисление стилистического вектора α – нормированных частот употребления ключевых слов:

$$\alpha = \frac{\left(\frac{\sum m_i \in d}{N_d}\right) - \omega_i}{\sigma_i}, \quad (2)$$

где d – анонимный корпус текста; m_i – ключевое слово; N_d – количество слов в анонимном корпусе текста; ω_i – средняя частота ключевого слова; σ_i – стандартное отклонение частоты ключевого слова.

Вычисление тематического вектора β – распределения по категориям, выявленным с помощью распределения Дирихле:

$$\beta = \frac{n_{dk} + \tau_k}{N_d + \sum_{j=1}^{K} \tau_j}, \quad (3)$$

где β – вероятность категории K для анонимного корпуса текста d; n_{dk} – количество слов в анонимном корпусе текста d, принадлежащих категории K; N_d – количество слов в анонимном корпусе текста; τ — гиперпараметр для категории текста.

Вычисление вектора достоверности γ – путём применения системы критериев оценки достоверности утверждений (CBCA) к тексту генерируется числовой вектор, каждый элемент которого соответствует оценке выполнения конкретного лингвистического критерия:

$$\gamma = \frac{q_l(d) - \mu_l}{\sigma_l}, \quad (4)$$

где $q_l(d)$ – значение оценки критерия относительно анонимного корпуса текста d; μ_l – среднее значение оценки критерия; σ_l – стандартное отклонение оценки критерия.

Расчет компонент сходства – это процедура расчета множества связанных метрик: стилистической S, тематической T, метрики достоверности C:

$$S = P_S(A|d) = \frac{e(-\lambda \Delta_B(d,A))}{\sum_{A'' \in A} e(-\lambda \Delta_B(d,A''))}, \quad (5)$$

где $S = P_S(A|d)$ – стилистическая метрика; λ – параметр статистической дисперсии распределения; $\Delta_B(d, A)$ – дельта Бёрроуза; A – пользователь как предполагаемый автор;

$$T = P_T(d|A) = \frac{e(-\rho\, KL(\beta_d || \overline{\beta}_A))}{\sum_{A' \in A} e(-\rho\, KL(\beta_d || \overline{\beta}_{A'}))}, \quad (6)$$

где ρ – параметр веса тематического компонента; β_d – тематический вектор анонимного корпуса текста d; $\overline{\beta_A}$ – тематический вектор профиля пользователя A; KL – параметр расхождения вероятностей стандартных тем пользователя от темы корпуса текста d:

$$C = P_C(d|A) = \frac{(F_{tr}(d) - \tau_A)}{1 + e^{-x}}, \quad (7)$$

где τ_A – порог правдоподобия для пользователя A; $F_{tr}(d)$ – оценка достоверности документа.

Вычисление показателя соответствия авторства – это оценка возможности авторства каждого пользователя вычисляется как взвешенная линейная комбинация трех компонент с учетом каждого ранее рассчитанного вектора (стилистического, тематического, вектора достоверности). В итоге формула дает значение показателя

$$\mathrm{CI} = \alpha\, \mathrm{S} + \beta\, T + \gamma\, \mathrm{C},$$

при котором пользователь с его максимальным значением будет наиболее вероятным автором текста, а сама оценка дает комплексную характеристику документа, объединяющую авторский стиль, смысловое наполнение и прагматическую валидность.

Заключение

Современный этап развития лингвокриминалистики характеризуется процессом глубокой трансформации, движимой конвергенцией методов обработки естественного языка, анализа

больших данных и искусственного интеллекта. Из дисциплины, выполняющей преимущественно вспомогательную экспертно-консультативную функцию, она эволюционирует в один из ключевых инструментариев расследования сложно структурированных преступлений, таких как киберпреступления, экстремизм, коррупционные схемы и организованные угрозы информационной безопасности. Данная трансформация порождает потребность в формировании нового типа специалиста – гибридного эксперта, который совмещает фундаментальное лингвистическое образование, практические навыки программирования и работы с данными, а также глубокое понимание криминалистической методологии и процессуальных норм.

Прогнозируемые векторы развития данной области связаны с тремя взаимосвязанными трендами. Во-первых, это переход к анализу потоковых данных в режиме, приближенном к реальному времени. Технологии позволяют осуществлять автоматизированный мониторинг и скрининг коммуникационных потоков в мессенджерах и социальных сетях для превентивного выявления маркеров планирования преступной деятельности или вербовки, что смещает фокус с ретроспективного расследования на прогнозную аналитику. Во-вторых, усиливается тренд на мультимодальный анализ, при котором лингвистические данные интегрируются с экстралингвистическими: акустическими параметрами голоса, визуальным контентом и метаданными коммуникации. Такой холлистический подход позволяет реконструировать событие с существенно более высокой полнотой и достоверностью. В-третьих, внедрение архитектур глубокого обучения, в частности трансформерных моделей, открывает путь к глубинному семантическому и прагматическому пониманию текста. Это подразумевает переход от анализа поверхностных статистических корреляций к моделированию контекстуальных значений, распознаванию речевых актов, иронии и истинных интонаций пользователей.

Таким образом, текст в современной криминалистической парадигме перестает быть пассивной «уликой» и приобретает статус комплексного цифрового отпечатка личности и социального взаимодействия. Его декодирование требует симбиоза классического гуманитарного герменевтического подхода, основанного на понимании природы языка как социального явления, и передового инструментария data science, способного выявлять в больших данных скрытые, статистически значимые паттерны.

Благодарности и поддержка

Работа поддержана Министерством науки и высшего образования Российской Федерации в рамках базовой части Государственного задания для высших учебных заведений # FRRR2026-0006. В контексте данного исследования автор считает целесообразным отметить работы [Кор24, Ишк26].

Список литературы | References

[Ble03] Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Advances in Neural Information Processing Systems 14 (NIPS 2001). 2001. Pp. 601–608. [[Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet Allocation // Advances in Neural Information Processing Systems 14 (NIPS 2001). 2001. P. 601-608. (In English)].]

[Cou07] Coulthard M., Johnson A. An Introduction to Forensic Linguistics: Language in Evidence. 1st ed. London: Routledge, 2007. 237 p. DOI: 10.4324/9780203969717. [[Coulthard M., Johnson A. An Introduction to Forensic Linguistics: Language in Evidence. 1st ed. London: Routledge, 2007. 237 p. (In English)].]

[Die14] Diesner J. ConText: Network-based analysis of unstructured text data // Encyclopedia of Social Network Analysis and Mining / eds. R. Alhajj, J. Rokne. New York: Springer, 2014. DOI: 10.1007/978-1-4614-6170-8_357. [[Diesner J. ConText: Network-based analysis of unstructured text data // Encyclopedia of Social Network Analysis and Mining / eds. R. Alhajj, J. Rokne. New York: Springer, 2014. (In English)].]

[Gra13] Grant T. TXTBK 2NVR: The idiolect of a terrorist? // The Routledge Handbook of Forensic Linguistics / eds. M. Coulthard, A. Johnson. London: Routledge, 2013. Pp. 493–508. [[Grant T. TXTBK 2NVR: The idiolect of a terrorist? // The Routledge Handbook of Forensic Linguistics / eds. M. Coulthard, A. Johnson. London: Routledge, 2013. P. 493-508. (In English)].]

[Juo08] Juola P. Authorship Attribution // Foundations and Trends in Information Retrieval. 2008. Vol. 1, No. 3. Pp. 233–334. DOI: 10.1561/1500000005. [[Juola P. Authorship Attribution // Foundations and Trends in Information Retrieval. 2008. Vol. 1, No. 3. P. 233-334. (In English)].]

[Kop11] Koppel M., Schler J., Argamon S. Authorship attribution in the wild // Language Resources and Evaluation. 2011. Vol. 45, No. 1. Pp. 83–94. EDN: GQJMZB. [[Koppel M., Schler J., Argamon S. Authorship attribution in the wild // Language Resources and Evaluation. 2011. Vol. 45, No. 1. P. 83-94. (In English)].]

[Liu12] Liu B. Sentiment Analysis and Opinion Mining // Synthesis Lectures on Human Language Technologies. 2012. Vol. 5, No. 1. Pp. 1–167. DOI: 10.2200/S00416ED1V01Y201204HLT016. [[Liu B. Sentiment Analysis and Opinion Mining // Synthesis Lectures on Human Language Technologies. 2012. Vol. 5, No. 1. P. 1-167. (In English)].]

[Mas17] Masip J. Deception Detection: State of the Art and Future Prospects // Psicothema. 2017. Vol. 29, No. 2. Pp. 149–159. EDN: YGJPWY. [[Masip J. Deception Detection: State of the Art and Future Prospects // Psicothema. 2017. Vol. 29, No. 2. P. 149-159. (In English)].]

[Spo16] Sporer S. L. Deception and cognitive load: Expanding our horizon with a working memory model // Frontiers in Psychology. 2016. Vol. 7, Article 420. DOI: 10.3389/fpsyg.2016.00420. [[Sporer S. L. Deception and cognitive load: Expanding our horizon with a working memory model // Frontiers in Psychology. 2016. Vol. 7, Article 420. (In English)].]

[Sta09] Stamatatos E. A survey of modern authorship attribution methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60, No. 3. Pp. 538–556. EDN: MFXHYX. [[Stamatatos E. A survey of modern authorship attribution methods // Journal of the American Society for Information Science and Technology. 2009. Vol. 60, No. 3. P. 538-556. (In English)].]

[Tur10] Turell M. T. The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison // International Journal of Speech, Language and the Law. 2010. Vol. 17, No. 2. Pp. 211–250. DOI: 10.1558/ijsll.v17i2.211. [[Turell M. T. The Use of Textual, Grammatical and Sociolinguistic Evidence in Forensic Text Comparison // International Journal of Speech, Language and the Law. 2010. Vol. 17, No. 2. P. 211-250. (In English)].

[Vil19] Villafana T. The Prosody of Deception: A Forensic Phonetic Approach // The Oxford Handbook of Language and Law / eds. L. M. Solan, P. M. Tiersma. Oxford: Oxford University Press, 2019. Pp. 347–362. [[Villafana T. The Prosody of Deception: A Forensic Phonetic Approach // The Oxford Handbook of Language and Law / eds. L. M. Solan, P. M. Tiersma. Oxford: Oxford University Press, 2019. P. 347-362. (In English)].]

[Vri14] Vrij A. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection // Handbook of Communication in the Legal Sphere / eds. C. A. Hafner, P. Tiersma. Berlin; Boston: De Gruyter Mouton, 2014. Pp. 321–344. DOI: 10.1515/9783110276794.321. [[Vrij A. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection // Handbook of Communication in the Legal Sphere / eds. C. A. Hafner, P. Tiersma. Berlin; Boston: De Gruyter Mouton, 2014. P. 321-344. (In English)].]

[Ишк26] Ишкинин Р. А., Ризванов Д. А. Классификация текстов на основе семантической близости с использованием встраиваемых моделей // СИИТ. 2026. Т. 8, № 1(25). С. 127–133. EDN: DIIKOG. [[Ishkinin R. A., Rizvanov D. A. “Text classification based on semantic similarity using embedding models” // SIIT. 2026. Vol. 8, No. 1(25). P. 127-133. (In Russian).]]

[Кор24] Коровин Е. А., Чиглинцева С. А. и др. Медицинская рекомендательная система на основе автоматического извлечения знаний из текстов // СИИТ. 2024. Т. 6, № 4(19). С. 111–121. EDN: OTVTXR. [[Korovin E. A., Chiglintseva S. A., et al. Medical recommender system based on automatic knowledge extraction from texts // SIIT. 2024. Vol. 6, No. 4(19). P. 111-121. (In Russian).]]

[Кры89] Крылов И. Ф. Криминалистическое учение о следах. Л.: Изд-во Ленингр. ун-та, 1976. 197 с. [[Krylov I. F. Forensic Science on Traces. Leningrad: Leningrad University Press, 1976. (In Russian).]]

ОБ АВТОРАХ | ABOUT THE AUTHORS

АЛЕКСЕЕВА Дарья Сергеевна
Уфимский университет науки и технологий, Россия.
ads.stat@mail.ru ORCID: 0009-0002-8955-9849.
Аспирантка, каф. автоматизированных систем управления.

ALEKSEEVA Daria Sergeyevna
Ufa University of Science and Technology, Russia.
ads.stat@mail.ru ORCID: 0009-0002-8955-9849.
Postgraduate student, Dept. of Automated Control Systems.

МЕТАДАННЫЕ | METADATA

Заглавие: Мультипараметрический метод установления авторства текста: интеграция стилометрического, тематического и прагматического анализа.

Авторы: Алексеева Д. С.

Аннотация: В условиях тотальной цифровизации коммуникаций текст стал основным носителем криминалистически значимой информации, что требует развития новых методов его автоматизированного анализа для нужд расследования. Традиционные подходы, такие как стилометрия, семантический или прагматический анализ, применяются изолированно, что ограничивает полноту и надежность лингвистической экспертизы. В данной статье предлагается комплексный мультипараметрический метод атрибуции анонимных или спорных текстов, интегрирующий три независимых лингвистических уровня: формально-статистический (стилометрия

Title: Multiparametric method of text authorship attribution: integration of stylometric, thematic, and pragmatic analysis.

Authors: Alekseeva D. S.

Abstract: In the context of total digitalization of communications, text has become the main carrier of criminally significant information, which requires the development of new methods of automated analysis for investigative purposes. Traditional approaches such as stylometry, semantic or pragmatic analysis are applied in isolation, which limits the completeness and reliability of linguistic expertise. This article proposes a comprehensive multiparametric attribution method for anonymous or controversial texts that integrates three independent linguistic levels: formal statistical (stylometry based on Burroughs delta), semantic (thematic modeling using Latent Dirichlet Allocation) and pragmatic (assessment of narrative reliability according to CBCA

на основе дельты Бёрроуза), смысловой (тематическое моделирование с использованием Latent Dirichlet Allocation) и прагматический (оценка достоверности нарратива по критериям CBCA – Criteria-Based Content Analysis). Метод предполагает извлечение соответствующих векторов признаков, вычисление на их основе метрик стилистического, тематического и прагматического сходства и их последующую взвешенную агрегацию в итоговый показатель соответствия (Compliance Indicator), максимизирующий вероятность корректной атрибуции авторства. Теоретическая значимость работы заключается в разработке формальной модели комплексного лингвистического «цифрового профиля», а практическая – в создании методологического базиса для инструментов поддержки принятия решений при расследовании киберпреступлений, экстремизма и других сложно структурированных деяний, опосредованных цифровым текстом.

Ключевые слова: Лингвокриминалистика; атрибуция текста; установление авторства; мультипараметрический анализ; стилометрия; тематическое моделирование (LDA); анализ достоверности (CBCA); цифровая криминалистика.

Язык: Русский.

Статья поступила в редакцию 2 февраля 2026 г.

criteria– Criteria-Based Content Analysis). The method involves extracting appropriate feature vectors, calculating stylistic, thematic, and pragmatic similarity metrics based on them, and then aggregating them into a final Compliance Indicator that maximizes the likelihood of correct attribution of authorship. The theoretical significance of the work lies in the development of a formal model of a complex linguistic "digital profile", and the practical one is to create a methodological basis for decision support tools in the investigation of cybercrime, extremism and other complex structured acts mediated by digital text.

Key words: Linguocriminalism; attribution of text; establishment of authorship; multiparametric analysis; stylometry; thematic modeling (LDA); reliability analysis (CBCA); digital forensics.

Language: Russian.

The article was received by the editors on 2 February 2026.