КРИМИНАЛИСТИКА

УДК 343.98

Применение отношения правдоподобия для разрешения избыточного многообразия в задачах криминалистического автороведения

М. А. Марусенко, В. В. Петров

Санкт-Петербургский государственный университет,
Российская Федерация, 199034, Санкт-Петербург, Университетская наб., 7–9

Для цитирования: Марусенко, М. А., В. В. Петров. 2024. «Применение отношения правдоподобия для разрешения избыточного многообразия в задачах криминалистического автороведения». *Вестник Санкт-Петербургского университета. Право* 4: 1086–1097.
https://doi.org/10.21638/spbu14.2024.411

В статье рассматривается проблема избыточного многообразия алгоритмов, применяемых для решения задач криминалистического автороведения. Доступность текстов в электронной форме и банализация использования программных средств привели к неконтролируемому многообразию алгоритмов, используемых для решения таких задач. Современный этап развития теории и практики авторской идентификации характеризуется широким распространением и легкой доступностью программных средств, позволяющих обрабатывать тексты, существующие в электронной форме, без дополнительной подготовки. Если на начальном этапе развития этой предметной области объектами исследования были в основном литературные тексты, то сегодня в центре внимания оказались электронные тексты (e-mail, SMS, чаты). Это создает у пользователей подобных текстов иллюзию отсутствия необходимости хотя бы общего ознакомления с основными положениями стилеметрии (как минимум таких, как требование жанрово-стилевой однородности исследуемых текстов, исключение речи персонажей и обработка только авторской речи, приоритет синтаксиса над лексикой), возникшей уже довольно давно. Нередко за стилистический анализ выдаются формальные процедуры, например разбиение текста на n-граммы. Отсутствие инструментов валидации результатов криминалистических экспертиз в этой сфере затрудняет их использование в юридической практике. В англосаксонских юрисдикциях обязательно применение метода оценки отношения правдоподобия (likelihood ratio, LR) при проверке статистических гипотез. Данное отношение показывает вероятность сходства между текстом известного происхождения и атрибутируемым текстом для нулевой гипотезы (Но), согласно которой оба тек-

ста имеют одно происхождение, против альтернативной гипотезы (Ha), в соответствии с которой у них разное происхождение. Применение этого метода позволяет сократить избыточное многообразие алгоритмов идентификации за счет отказа от использования алгоритмов, не включающих проверку статистических гипотез и оценку LR.

Ключевые слова: криминалистическое авторoведение, отношение правдоподобия, проверка статистических гипотез, авторская идентификация, индивидуальный авторский стиль, атрибуция, нулевая гипотеза, альтернативная гипотеза.

1. Введение

Огромный объем текстовой информации породил большой спрос на методы классификации текстов, среди которых выделяются задачи авторской идентификации, когда истинный автор некоего документа определяется на основе выборок из текстов, написанных известными авторами. Эта проблема может решаться разными способами, одним из которых является атрибуция с известным числом априорных классов, когда реальный автор находится в числе нескольких известных кандидатов (минимальное число возможных авторов равно двум, и тогда проблема сводится к решению задачи бинарной классификации) (Aitken, Stoney 1991).

В общем случае проблема авторской идентификации оказывается типичной проблемой классификации, решение которой зависит от дискриминирующих признаков, определяющих авторский стиль. Стилеметрические количественные признаки, представляющие измерения различных уровней языковой структуры, играют значимую роль. Исторически можно выделить четыре этапа развития таких исследований.

На начальном этапе атрибуции проводились на основе единственного инвариантного признака, характеризующего индивидуальный авторский стиль и меняющегося от одного автора к другому.

На втором этапе стали использоваться различные статистики, основанные на доле разных слов, коэффициенте лексического богатства, доле слов с низкой частотой (1 или 2), средней длине слова, средней длине предложения, мере Юла, характеристиках распределения Ципфа и т. д. Однако эти методы оказались недостаточно эффективными, частично из-за того, что в них использовались элементы с низкой вероятностью встречаемости в тексте (Holmes 1998).

Позднее, в результате снятия ограничения на использование единственного признака, стали разрабатываться методы многомерного анализа, направленные на выделение наборов индивидуальных дискриминирующих признаков для каждого автора либо на построение методами факторного и компонентного анализа (метод главных компонент) новых композитных признаков, являющихся линейными преобразованиями известных признаков. В этом случае атрибутируемый документ представлял собой точку в новом пространстве признаков, а определение возможного автора проводилось по образцу, который был ближе всего по своим стилистическим, хронологическим и тематическим аспектам и автор которого считался автором атрибутируемого текста. Автор такого «ближайшего» документа и считался автором атрибутируемого текста (Марусенко 1990).

На третьем этапе стали использоваться стилистически независимые признаки, считавшиеся связанными с индивидуальным авторским стилем, среди которых можно выделить три группы: 1) на лексическом уровне — частоты встречаемости некоторых слов (или n-грамм-последовательностей из n элементов произвольной

длины; это могут быть последовательности звуков, слогов, слов или букв с числом элементов более двух), знаков пунктуации, вводных слов, предлогов, союзов, местоимений, некоторых глагольных форм и т. д.; 2) на синтаксическом уровне — распределение различных частей речи в тексте, устойчивых выражений и словосочетаний; 3) использование таких структурных и типографских признаков, как общее число строк, число строк в предложении и абзаце, абзацные отступы, число токенов в абзаце и т. д. (Aitken, Taroni 2004; Amelin et al. 2018; Glaudes et al. 2022; Johnson, Wright 2014; Pavelec, Justino, Oliveira 2007; Savoy 2012). Сюда же относится использование различных дополнительных признаков, таких как орфографические отличия (британский и американский варианты написания), анализ орфографических ошибок и т. д. Число таких дополнительных признаков достигает 270 (Alred, Brusaw, Oliu 2008; Baldwin 1979; Chaski 2005; Savoy 2012; Zheng et al. 2006). Наивысшим достижением в этой области можно признать дельта-классификатор Берроуза, основанный на средней абсолютной разности между z-оценками от 40 до 150 (в последних версиях 800) наиболее частотных слов в контрольном и атрибутируемом текстах (Марусенко и др. 2019; Burrows 2002; Glaudes et al. 2022).

Сегодня мы находимся на четвертом этапе развития, который характеризуется широким распространением и легкой доступностью программных средств, позволяющих обрабатывать тексты, существующие в электронной форме, без дополнительной подготовки. Если на начальном этапе развития этой предметной области объектами исследования были в основном литературные тексты, то сегодня в центре внимания оказались электронные тексты (e-mail, SMS, чаты). Неподготовленных пользователей это может привести к мысли о том, что нет необходимости более глубоко знакомиться с основными положениями стилеметрии, включающими соблюдение требования жанрово-стилевой однородности текстов, удаление прямой речи персонажей и учет только авторской речи, понимание стиля как преимущественно структурно-синтаксической категории.

В качестве экзотического примера можно привести использование стандартных архиваторов при решении задач авторской атрибуции (Malyutov, Wickramasinghe, Li 2007; Oliveira, Justino, Oliveira 2013). Этот метод основан на предположении, что тексты, принадлежащие одному автору, при сжатии показывают больший процент компрессии.

Доктор экономических (!) наук Е. В. Луценко из Кубанского государственного аграрного университета полагает, что «задача идентификации текстов на основе анализа предложений является тривиальной из-за практически абсолютной уникальности предложений. Поэтому больший интерес представляет задача идентификация текстов на основе анализа слов» (Луценко 2004).

В подобных работах, как отмечает французский социолог Д. Лаббе, очень часто отсутствуют стандартные процедуры валидации (оценки пригодности) методов идентификации: во многих случаях они применяются непосредственно к спорным текстам без предварительного тестирования на бесспорных образцах (Basson, Labbé 2020). Кроме того, нередко исследователи выбирают определенные слова, например служебные, создавая возможность бесконечного манипулирования данными до тех пор, пока не добьются «хорошего» (т. е. нужного им) ответа. Наконец, они не оценивают вероятность ошибки при принятии или отклонении статистической гипотезы. Многие игнорируют то, что, перед тем как попасть в руки читателя,

текст может подвергаться многим операциям (слияниям и разделениям на абзацы, главы, части, тома и т. д., литературному и техническому редактированию, стандартизации, верстке типа ликвидации «висячих строк» и т. д.), которые могут изменять его параметры (Alred, Brusaw, Oliu 2008; Zheng et. al. 2006), а также то, что индивидуальный авторский стиль эволюционирует и эта эволюция вызывается изменениями авторских интенций, влиянием коллег, изменениями в социальном положении и т. д. (Robertson, Vignaux, Berger 1995). В криминалистике подобные явления считаются неустойчивостью признаков и их изменяемостью с течением времени (Седова, Кушниренко, Пристансков 2021). Неудивительно, что публикации по самым известным случаям авторской атрибуции часто содержат противоположные результаты на основе одних и тех же исходных данных.

2. Основное исследование

Всего около 30–40 лет назад традиционные методы криминалистической идентификации в большей степени были основаны на вере, а не на эмпирических доказательствах (Saks 2010). Криминалистическая идентификация начинается с признания того, что физические признаки объектов и следы, которые они оставляют, вариабельны. Однако вместо того, чтобы измерять эту вариабельность и ее влияние на доказательность принимаемых решений, каждая из традиционных криминалистических наук (дактилоскопия, почерковедение, распознавание голоса, ДНК, осколков стекла, отпечатков ушных раковин, следов обуви и т. д.) стала делать акцент на идее уникальности: два объекта не могут оставлять неотличимые следы, и если два следа неотличимы друг от друга, они должны принадлежать одному объекту, отличному от всех других. По непонятным причинам многие эксперты-криминалисты верили, что в этом случае вероятность ошибки близка или равна нулю (Johnson, Wright 2014).

Идея уникальности восходит к Адольфу Кетле (1796–1874), основателю социальной статистики, который заявлял, что «природа никогда не повторяется» (Рейхесберг 1894, 38). После него полицейский чиновник Альфонс Бертильон (1853–1914) использовал эту идею для идентификации преступников и изобрел антропометрию (систему идентификации преступников по данным антропометрических измерений, известную как бертильонаж) (Марусенко 1990, 85–86). Она никогда не была доказана эмпирически и не может быть доказана статистически, однако повторялась тремя поколениями экспертов-криминалистов. В результате англосаксонская наука о криминалистической идентификации основывалась на фундаментальной идее, согласно которой вариабельность объясняется случайными причинами, и поддерживала веру в уникальность и свободу от ошибок.

Только целый ряд громких трагических ошибок в конце XX в. (например, в США за последние 50 лет более 100 чел., приговоренных к смертной казни, были оправданы) показал необходимость поиска новой криминалистической теории. Толчок к этому поиску дало, в частности, экспериментальное доказательство того, что кожа может сжиматься или растягиваться под влиянием внешних обстоятельств, что приводит к изменениям дактилоскопического отпечатка.

Направлением лингвистических исследований, подвергающимся наиболее жестким и структурированным требованиям, является криминалистическая линг-

вистика, входящая в круг таких наук, как криминалистические медицина, стоматология, химия и т. д. Она имеет три основных направления: почерковедение, фонетика/фонология и дискурсивный анализ, базирующийся на стилистическом анализе письменных текстов и устной речи. В отечественной криминалистике исследования почерка и авторского стиля также разделены (Седова, Кушниренко, Пристансков 2021, 176).

В области авторской идентификации это привело к выделению в самостоятельную науку криминалистической лингвистики (forensic linguistics) (Varney 1977), базовыми в которой являются два положения (Aitken, Stoney 1991; Holmes 1998; Johnson, Wright 2014):

— два писателя (с одним материнским языком) не пишут одинаково;

— один и тот же писатель не пишет одинаково все время.

К сожалению, в криминалистике стилеметрический подход в компьютерной реализации долгое время ограничивался машиночитаемыми параметрами (длина слова и предложения, частотность, распределение слов разной длины) и стандартным статистическим анализом, а важность стандартного синтаксического анализа и прочие достижения доминирующей парадигмы теоретической лингвистики недооценивались.

При формулировании заключений эксперты-криминалисты оперируют несколькими категориями (табл. 1).

Таблица 1. **Оценочные категории, используемые экспертами-криминалистами**

Положительная идентификация	Отрицательная идентификация
Sure beyond reasonable doubt (безусловно, без всяких сомнений)	–
There can be very little doubt (может быть очень небольшое сомнение)	–
Highly likely (высокая вероятность)	Highly likely (высокая вероятность)
Likely (имеется вероятность)	Likely (имеется вероятность)
Very probable (весьма вероятно)	Quite probable (вполне вероятно)
Probable (вероятно)	Probable (вероятно)
Quite possible (вполне возможно)	–
Possible (возможно)	–
…что это один и тот же человек	

Источник: (Baldwin 1979, 231–232).

Одна из этих оценочных категорий (highly likely) стала в России мемом из-за частого употребления экс-премьер-министром Великобритании Терезой Мэй по поводу так называемого отравления Скрипалей.

Как отмечают сами англосаксонские криминалисты, категории likely (имеется вероятность) и probable (вероятно) практически являются синонимами.

Во многих юрисдикциях зарубежных стран, в основном использующих англосаксонское право (common law), с целью повышения доказательности представляемых в суды заключений экспертов-криминалистов используется отношение прав-

доподобия (likelihood ratio, LR), которое считается самым подходящим инструментом, помогающим суду при определении значения, которое должно придаваться экспертным заключениям (Morrison 2011).

Среди других отраслей научной криминалистики, где отношение правдоподобия уже стало стандартной парадигмой оценки доказательности экспертных оценок, авторская идентификация находится на одном из последних мест. Использование LR-парадигмы вошло в основные учебники по доказательной криминалистике и криминалистической статистике. Методы авторской идентификации сегодня применяются на практике в уголовном праве (установление авторов требований о выкупе, писем с угрозами), гражданском праве (авторские права и имущественные споры), компьютерной безопасности (исследование содержания электронной переписки). Если установлен конкретный компьютер, на котором выполнен документ, легитимной задачей является идентификация автора этого документа, т. е. того, кто находился за клавиатурой во время написания документа (Robertson, Vignaux, Berger 1995; Savoy 2012; Zheng et al. 2006).

По мнению англосаксонских криминалистов, их задачей является оценка достоверности экспертного заключения, которая выражается при помощи LR, представляющего собой отношение вероятности того, что достоверность достигается, если одна гипотеза (нулевая гипотеза H0) подтверждается, к вероятности того, что достоверность достигается, если подтверждается альтернативная гипотеза Ha:

$$LR = \frac{P\left(E \mid H_0\right)}{P\left(E \mid H_a\right)},$$

где P — вероятность; E — реализация той или иной гипотезы; H_0 — гипотеза, согласно которой два текста написаны одним автором; H_a — гипотеза, согласно которой два текста написаны разными авторами.

Числитель уравнения содержит вероятность решения, основанного на нулевой гипотезе H_0, а знаменатель — на альтернативной гипотезе *Ha*.

Отношение правдоподобия представляет собой утверждение, которое оценивает вероятность решений, связанных с каждой из гипотез. Так, *LR* показывает вероятность сходства между текстом известного происхождения и атрибутируемым текстом для нулевой гипотезы (H_0), согласно которой оба текста имеют одно происхождение, против альтернативной гипотезы (H_a), в соответствии с которой у них разное происхождение.

Таким образом, *LR* представляет собой отношение вероятностей реализации двух конкурирующих гипотез. Если вероятность реализации нулевой гипотезы больше, чем вероятность реализации альтернативной гипотезы, *LR* имеет величину больше единицы. В противном случае *LR* меньше единицы. Другими словами, относительная достоверность решения, основанного на конкурирующих гипотезах, связана с величиной *LR*. Чем больше *LR* отличается от единицы, тем бóльшую достоверность получает одна или другая гипотеза. Значения *LR* интерпретируются следующим образом:

$LR > 1$ — результаты склоняются в пользу нулевой гипотезы;

$LR = 1$ — результаты в равной степени поддерживают обе гипотезы»;

$LR < 1$ — результаты поддерживают альтернативную гипотезу.

Таблица 2. **Стандарты для цифрового и вербального выражения отношения правдоподобия**

Рекомендуемая терминология для отношения правдоподобия	
Цифровое выражение отношения правдоподобия	**Вербальное выражение (вспомогательное)**
> 1–10	Weak or limited (слабое или ограниченное)
10–100	Moderate (умеренное)
100–1000	Moderately strong (умеренно сильное)
1000–10 000	Strong (сильное)
10 000–1 000 000	Very strong (очень сильное)
> 1 000 000	Extremely strong (чрезвычайно сильное)

Источник: Association of Forensic Science Providers. 2009. "Standards for the formulation of evaluative forensic science expert opinion". *Scientific Justice* 3: 161–164.

Такие утверждения могут формулироваться как в цифровом виде, так и вербально. Американская ассоциация провайдеров научной криминалистики разработала шкалу для перехода от цифрового формата к вербальному (табл. 2).

Таким образом, эксперт может заявить, что, по его мнению, совпадения между атрибутируемым текстом и текстом известного автора представляют сильное подтверждение гипотезы, согласно которой оба текста принадлежат одному автору, либо сказать, что совпадения между двумя текстами показывают, что вероятность принадлежности одному автору в n раз больше, чем для альтернативной гипотезы.

Хотя вербальное представление чаще используется в ходе судебных прений, цифровое представление должно использоваться в доказательной криминалистике. Специалисты считают, что двойная форма представления (вербальная и цифровая) не только содержит больше информации, но и является более убедительной для участников судебного процесса, облегчая им выбор одной из конкурирующих гипотез (Aitken et al. 2011; Robertson, Vignaux, Berger 1995). В то же время лишь вербальное представление, по заключению Американской ассоциации провайдеров научной криминалистики, может приносить только слабую или ограниченную пользу и имеет малую доказательную силу. В криминалистике использование отношения правдоподобия считается оптимальным способом выражения неопределенности в экспертных заключениях, хотя в некоторой степени затрудняет понимание между участниками судебного процесса.

Как известно, анализ статистических данных в криминалистике опирается на идею, согласно которой результаты исследований могут быть объективными с известной вероятностью ошибки. Таким образом, ключевым компонентом формирования решений о выборе из конкурирующих гипотез являются процедуры проверки гипотез, при которых каждой из них приписывается определенная вероятность. С этой точки зрения оптимальны алгоритмы распознавания образов (Марусенко 1990, 126–156).

В итоге авторская идентификация должна сводиться к проверке гипотез о том, является ли конкретный автор настоящим автором атрибутируемого документа. Первым этапом такой процедуры должно быть представление этого документа

в виде многомерной математической модели, построенной на релевантных признаках, пригодных для различения разных авторов. Данная процедура требует отбора наиболее информативных признаков или построения новых признаков, представляющих собой комбинацию уже существующих признаков и полезных для определения различий между индивидуальными авторскими стилями. На втором этапе проводится взвешивание этих признаков с целью определения их значимости для решения данной задачи и их относительной дискриминирующей силы. В результате система должна указать на наиболее вероятного автора из созданного алфавита априорных классов.

Описанная процедура была реализована при идентификации автора «Записки Юровского» о расстреле императорской семьи (Петров, Марусенко 2017). Нулевая гипотеза была сформулирована следующим образом:

H_0 — атрибутируемый текст представляет собой единоличное произведение Я. М. Юровского.

Альтернативная гипотеза имеет два варианта:

H_a^1 — атрибутируемый текст представляет собой единоличное произведение М. Н. Покровского;

H_a^2 — атрибутируемый текст представляет собой совместное произведение Я. М. Юровского и М. Н. Покровского.

Вероятностный алгоритм распознавания определил принадлежность «Записки Юровского» классу Ω (Юровский) с вероятностью 0,6, а классу Ω (Покровский) — с вероятностью 0,4. Таким образом, отношение правдоподобия LR = 0,6/0,4 = 1,5. В соответствии с табл. 2, эта величина означает слабую или ограниченную надежность данной гипотезы, поэтому решение принимается в пользу второго варианта альтернативной гипотезы — совместная работа Я. М. Юровского и М. Н. Покровского над текстом «Записки», о форме которой мы можем только догадываться.

3. Выводы

С учетом того, для в криминалистическом авторoведении разрешено применение избыточного количества разнообразных алгоритмов, следует прийти к выводу, что алгоритмы, не использующие аппарат проверки статистических гипотез и не позволяющие оценку отношения правдоподобия, не могут быть рекомендованы к использованию. Вычисление отношения правдоподобия и его вербальное выражение являются полезными инструментами для валидации алгоритмов авторской идентификации, используемых в криминалистическом авторoведении.

Библиография

Луценко, Е. В. 2004. «Атрибуция анонимных и псевдонимных текстов в системно-когнитивном анализе». *Политематический сетевой электронный научный журнал Кубанского государственного аграрного университета* 5: 36–56. Дата обращения 11 ноября, 2024. http://ej.kubagro.ru/2004/03/03.

Марусенко, М. А. 1990. *Атрибуция анонимных и псевдонимных литературных произведений методами теории распознавания образов*. Л.: Изд-во Ленингр. гос. ун-та.

Марусенко, М. А., В. В. Петров, К. Р. Пиотровская, И. Н. Маньяс, Н. К. Мамаев. 2019. «Об авторстве "писем Берии из заточения"». *Вестник Санкт-Петербургского университета. Право* 3: 568–605.

Петров, В. В., М. А. Марусенко. 2017. «Об истинном авторе "Записки Юровского"». *Вестник Санкт-Петербургского университета. Право* 8 (1): 76–107.

Рейхесберг, Н. М. 1894. *Адольф Кетле. Его жизнь, и научная деятельность*. СПб.: Тип. Ю. Н. Эрлих.

Седова, Т. А., С. П. Кушниренко, В. Д. Пристансков, ред. 2021. *Криминалистика*. М.: Юстиция.

Aitken, C. G. G., F. Taroni. 2004. *Statistics and the evaluation of evidence for forensic scientists.* Chichester: John Wiley & Sons.

Aitken, C. G., D. A. Stoney. 1991. *The use of statistics in forensic science*. New York; London: Ellis Horwood.

Aitken, C., C. E. Y. Berger, J. S. Buckleton, C. Champod, J. Curran, A. P. Dawid, I. W. Evett, P. Gill, J. Gonzalez-Rodriguez, G. Jackson, A. Kloosterman, T. Lovelock, D. Lucy, P. Margot, L. McKenn, D. Meuwly, C. Neumann, N. N. Daéid, A. Nordgaard, R. Puch-Solis, B. Rasmusson, M. Redmayne, P. Roberts, B. Robertson, C. P. Roux, M. Sjerps, F. Taroni, T. Tjin-A-Tsoi, G. Vignaux, S. Willis, G. Zadora. 2011. "Expressing evaluative opinions: A position statement". *Scientific Justice* 51: 1–2. https://doi.org/10.1515/9783110228069.1

Alred, G. J., C. T. Brusaw, W. E. Oliu. 2008. *Handbook of technical writing.* 9th ed. Bedford: St. Martin's Press.

Amelin, K., O. Granichin, N. Kizhaeva, Z. Volkovich. 2018. "Patterning of writing style evolution by means of dynamic similarity". *Pattern Recognition* 77: 45–64.

Baldwin, J. 1979. "Phonetics and speaker identification". *Medicine, Science and the Law* 9: 231–232.

Basson, J.-C., D. Labbé. 2020. "Les précieux manuscrits". *Proceedings of the 15th International Conference on Statistical Analysis of Textual Data (16–19 june 2020). Toulouse.* Дата обращения 11 ноября, 2024. http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/BASSON_LABBE_JADT2020.pdf.

Burrows, J. F. 2002. "Delta: A measure of stylistic difference and a guide to likely authorship". *Literary and Linguistic Computing* 17 (3): 267–287.

Chaski, C. E. 2005. "Who is at the keyboard. Authorship attribution in digital evidence investigations". *International Journal of Digital Evidence* 4 (1): 1–14.

Glaudes, P., A. Cervoni, F. Guglielmi, C. Mayaux, M. Marusenko, Ye. Kuralesina, M. Miretina, Y. Nikitina, M. Solovyeva, O. Khutoretskaya. 2022. "Jules Barbey d'Aurevilly et corpus journalistique. Problèmes d'attribution". *Observer la vie littéraire. Études littéraires et numériques* 2: 261–304.

Holmes, D. I. 1998. "The evolution of stylometry in humanities scholarship". *Literary and Linguistic Computing* 13 (3): 111–117.

Johnson, A., D. Wright. 2014. "Identifying idiolect in forensic authorship attribution". *Language and Law = Linguagem e Direito* 1 (1): 37–69.

Malyutov, M. B., C. I. Wickramasinghe, S. Li. 2007. "Conditional complexity of compression for authorship attribution". *SFB 649 Discussion Paper*. Дата обращения 11 ноября, 2024. https://scholar.google.co.id/citations?view_op=view_citation&hl=en&user=auS8PHEAAAAJ&cstart=20&pagesize=80&citation_for_view=auS8PHEAAAAJ:LkGwnXOMwfcC.

Morrison, G. S. 2011. "Measuring the validity and reliability of forensic likelihood-ratio systems". *Scientific Justice* 51: 91–98.

Oliveira J., W. E. Justino, L. S. Oliveira. 2013. "Comparing compression models for authorship attribution". *Forensic Science International* 228: 100–104.

Pavelec, D., E. Justino, L. S. Oliveira. 2007. "Author identification using stylometric features". *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* 11 (36): 59–65.

Robertson, B., T. Vignaux, C. Berger. 1995. *Interpreting evidence: Evaluating forensic science in the courtroom*. Chichester: John Wiley & Sons.

Saks, M. J. 2010. "Forensic identification: From a faith-based 'Science' to a scientific science". *Forensic Science International* 201: 14–17.

Savoy, J. 2012. "Authorship attribution: A comparative study of three text corpora and three languages". *Journal of Quantitative Linguistics* 19 (2): 132–161.

Varney, M. H. 1977. "Forensic linguistics". *English Today* 13 (4): 42–47.

Zheng, R., J. Li, H. Chen, Z. Huang. 2006. "A framework for authorship identification of online messages: Writing-style features and classification techniques". *Journal of the American Society for Information Science & Technology* 57 (3): 378–393.

Статья поступила в редакцию 7 февраля 2023 г.;
рекомендована к печати 30 июля 2024 г.

Контактная информация:

Марусенко Михаил Александрович — д-р филол. наук, проф.;
https://orcid.org/0000-0002-0441-7845; m.marusenko@spbu.ru
Петров Вадим Вадимович — канд. мед. наук, доц.;
https://orcid.org/0000-0001-7753-0083; vadim.petrov@spbu.ru

The application of the likelihood ratio to the resolution of redundant diversity in forensic authoring

M. A. Marusenko, V. V. Petrov

St. Petersburg State University,
7–9, Universitetskaya nab., St. Petersburg, 199034, Russian Federation

For citation: Marusenko, M. A., V. V. Petrov. 2024. "The application of the likelihood ratio to the resolution of redundant diversity in forensic authoring". *Vestnik of Saint Petersburg University. Law* 4: 1086–1097. https://doi.org/10.21638/spbu14.2024.411 (In Russian)

The article deals with the problem of resolving the redundant diversity of algorithms used to solve the problems of forensic authoring. The availability of texts in electronic form and the banalization of the use of software tools have led to an uncontrollable variety of algorithms used to solve such problems. The modern stage in the development of the theory and practice of author identification is characterized by the widespread and easy availability of software tools that allow to process texts existing in electronic form without additional training. If at the initial stage of development of this subject domain the objects of research were mainly literary texts, today the center of interest are electronic texts (e-mails, SMS, chat rooms). This creates for their users an illusion of lack of necessity at least a general familiarity with the basic principles of stylometry (at least, such as the requirement of genre-stylistic homogeneity of the texts studied, exclusion of personages' speech and processing only the author's speech, the priority of syntax over vocabulary), which already has a rather respectable history. The lack of tools for validating the results of forensic analysis in this area makes it difficult to use them in legal practice. In the Anglo-Saxon jurisdictions, it is mandatory to use the method of likelihood ratio (*LR*) assessment when testing statistical hypotheses. *LR* shows the probability of similarity between a text of known origin and an attributed text for the null hypothesis that both texts have the same origin, versus the alternative hypothesis claiming that they have different origins. The application of this method reduces the redundant diversity of identification algorithms by eliminating the use of algorithms that do not include statistical hypothesis checking and LR estimation.

Keywords: forensic authorship, likelihood ratio, statistical hypothesis testing, author identification, individual author style, attribution, null hypothesis, alternative hypothesis.

References

Aitken, C. G. G., D. A. Stoney. 1991. *The Use of Statistics in Forensic Science*. New York; London, Ellis Horwood.
Aitken, C. G. G., F. Taroni. 2004. *Statistics and the evaluation of evidence for forensic scientists.* Chichester, John Wiley & Sons.

Aitken, C., C. E. Y. Berger, J. S. Buckleton, C. Champod, J. Curran, A. P. Dawid, I. W. Evett, P. Gill, J. Gonzalez-Rodriguez, G. Jackson, A. Kloosterman, T. Lovelock, D. Lucy, P. Margot, L. McKenn, D. Meuwly, C. Neumann, N. N. Daéid, A. Nordgaard, R. Puch-Solis, B. Rasmusson, M. Redmayne, P. Roberts, B. Robertson, C. P. Roux, M. Sjerps, F. Taroni, T. Tjin-A-Tsoi, G. Vignaux, S. Willis, G. Zadora. 2011. Expressing evaluative opinions: a position statement. *Scientific Justice* 51: 1–2. https://doi.org/10.1515/9783110228069.1

Alred, G. J., C. T. Brusaw, W. Oliu. E. 2008. *Handbook of technical writing.* 9th ed. Bedford, St. Martin's Press.

Amelin, K., O. Granichin, N. Kizhaeva, Z. Volkovich. 2018. "Patterning of writing style evolution by means of dynamic similarity". *Pattern Recognition* 77: 45–64.

Baldwin, J. 1979. "Phonetics and speaker identification". *Medicine, Science and the Law* 9: 231–232.

Basson, J.-C., D. Labbé. 2020. "Les précieux manuscrits". *Proceedings of the 15th International Conference on Statistical Analysis of Textual Data (16–19 june 2020). Toulouse.* Accessed November 11, 2024. http://lexicometrica.univ-paris3.fr/jadt/JADT2020/jadt2020_pdf/BASSON_LABBE_JADT2020.pdf.

Burrows, J. F. 2002. "Delta: A measure of stylistic difference and a guide to likely authorship". *Literary and Linguistic Computing* 17 (3): 267–287.

Chaski, C. E. 2005. "Who is at the keyboard. Authorship attribution in digital evidence investigations". *International Journal of Digital Evidence* 4 (1): 1–14.

Glaudes, P., A. Cervoni, F. Guglielmi, C. Mayaux, M. Marusenko, Y. Kuralesina, M. Miretina, Y. Nikitina, M. Solovyeva, O. Khutoretskaya. 2022. "Jules Barbey d'Aurevilly et corpus journalistique. Problèmes d'attribution". *Observer la vie littéraire. Études littéraires et numériques* 2: 261–304.

Holmes, D. I. 1998. "The evolution of stylometry in humanities scholarship". *Literary and Linguistic Computing* 13 (3): 111–117.

Johnson, A., D. Wright. 2014. "Identifying idiolect in forensic authorship attribution". *Language and Law = Linguagem e Direito* 1 (1): 37–69.

Lutsenko, E. V. "Attribution of anonymous and pseudonymous texts in system-cognitive analysis". *Politematicheskii setevoi elektronnyi nauchnyi zhurnal Kubanskogo gosudarstvennogo agrarnogo universiteta* 5: 36–56. Accessed November 11, 2024. http://ej.kubagro.ru/2004/03/03.

Malyutov, M. B., C. I. Wickramasinghe, S. Li. 2007. "Conditional complexity of compression for authorship attribution". *SFB 649 Discussion Paper*. Accessed November 11, 2024. https://scholar.google.co.id/citations?view_op=view_citation&hl=en&user=auS8PHEAAAAJ&cstart=20&pagesize=80&citation_for_view=auS8PHEAAAAJ:LkGwnXOMwfcC.

Marusenko, M. A. 1990. *Attribution of anonymous and pseudonymous literary works by methods of pattern recognition theory*. Leningrad, Leningradskii gosudarstvennyi universitet Publ. (In Russian)

Marusenko, M. A., V. V. Petrov, K. R. Piotrovskaya, I. N. Manyas, N. K. Mamaev. 2019. "On the authorship of 'Beria's letters from imprisonment'". *Vestnik of Saint Petersburg University. Law* 3: 568–605. (In Russian)

Morrison, G. S. 2011. "Measuring the validity and reliability of forensic likelihood-ratio systems". *Scientific Justice* 51: 91–98.

Oliveira J., W. E. Justino, L. S. Oliveira. 2013. "Comparing compression models for authorship attribution". *Forensic Science International* 228: 100–104.

Pavelec, D., E. Justino, L. S. Oliveira. 2007. "Author identification using stylometric features". *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial* 11 (36): 59–65.

Petrov, V. V., M. A. Marusenko. 2017. "On the true author of Yurovsky's Notes". *Vestnik of Saint Petersburg University. Law* 8 (1): 76–107. (In Russian)

Reichesberg, N. M. 1894. *Adolph Quetelet. His life and scientific work*. St. Petersburg, Tipografiia Yu. N. Ehrlich Publ. (In Russian)

Robertson, B., T. Vignaux, C. Berger. 1995. *Interpreting evidence: Evaluating forensic science in the courtroom*. Chichester, John Wiley & Sons.

Saks, M. J. 2010. "Forensic identification: From a faith-based 'Science' to a scientific science". *Forensic Science International* 201: 14–17.

Savoy, J. 2012. "Authorship attribution: A comparative study of three text corpora and three languages". *Journal of Quantitative Linguistics* 19 (2): 132–161.

Sedova, T. A., S. P. Kushnirenko, V. D. Pristanskov, eds. 2021. *Forensic science*. Moscow, Iustitsiia Publ. (In Russian)
Varney, M. H. 1977. "Forensic linguistics". *English Today* 13 (4): 42–47.
Zheng, R., J. Li, H. Chen, Z. Huang. 2006. "A framework for authorship identification of online messages: Writing-style features and classification techniques". *Journal of the American Society for Information Science & Technology* 57 (3): 378–393.

Received: February 7, 2023
Accepted: July 30, 2024

Authors' information:

Mikhail A. Marusenko — Dr. Sci. in Philology, Professor;
https://orcid.org/0000-0002-0441-7845; m.marusenko@spbu.ru
Vadim V. Petrov — PhD in Medicine, Associate Professor;
https://orcid.org/0000-0001-7753-0083; vadim.petrov@spbu.ru