УДК 81'42
ББК Ш141.12-51 *ГСНТИ 16.31.21; 10.85.31* *Код ВАК 10.02.19; 05.25.05;12.00.12*
DOI 10.26170/pl19-02-18

Т. Б. Радбиль
Национальный исследовательский Нижегородский государственный ун-т им. Н. И. Лобачевского
ORCID ID: 0000-0002-7516-6705 ☑

М. В. Маркина
Национальный исследовательский Нижегородский государственный ун-т им. Н. И. Лобачевского
ORCID ID: — ☑

☑ ***E-mail:*** *timur@radbil.ru; markinamv6213@yandex.ru.*

Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов

АННОТАЦИЯ. *В статье представлен опыт разработки компьютеризованной модели авторизации текста и ее адаптации к нуждам идентификационной и диагностической автороведческой экспертизы русскоязычных текстов. Цель исследования —— продемонстрировать возможности идентификационного автороведческого экспертного исследования текстов посредством автоматической обработки текстов на основе комплексного применения вероятностно-статистических методов. Описан очередной этап апробации усовершенствованной версии программы «КАТ» (компьютерная авторизация текста) —— эксперимент по определению относительных частот соотношения тех или иных языковых элементов (вычислению коэффициентов корреляции) в нескольких выборках из сравниваемых текстов по комплексу разноуровневых параметров — коэффициентам Б. Н. Головина, дополненному существующими в научной практике и прикладной сфере индексами понятности текста («индекс Флеша — Кинкейда», «FOG-индекс») и рядом других параметров. Материалом исследования являются первичные базы данных текстов русской классики (произведений Л. Н. Толстого, Н. В. Гоголя, И. С. Тургенева). В результате экспериментального исследования были выработаны следующие критерии идентификации авторства: считать текст принадлежащим автору, если коэффициент корреляции текста с существующей базой больше 0,87, т. е. в качестве доверительного интервала принять интервал 0,87—1,0; считать текст не принадлежащим автору, если коэффициент корреляции текста с существующей базой меньше 0,82; учесть, что точность работы программы увеличивается с возрастанием объема текстов в базе. Под базой понимается средний показатель, исчисленный по всем текстам, с достоверностью атрибутированным как принадлежащие данному автору. В случае успешной доводки предложенной программы автоматической обработки текстов «КАТ» с ее помощью можно будет решать экспертные задачи по авторизации и диагностике спорных текстов, реализованных в пространстве медийного и политического дискурсов, в юридической, официально-деловой и коммерческой документации и пр.*

КЛЮЧЕВЫЕ СЛОВА: *судебное автороведение; автороведческая экспертиза; авторизация текстов; автоматическая обработка текстов; вероятностно-статистическая методика; русский язык.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Радбиль Тимур Беньюминович, доктор филологических наук, профессор, академик Российской академии естествознания, заведующий кафедрой теоретической и прикладной лингвистики, Институт филологии и журналистики, Национальный исследовательский Нижегородский государственный университет им. Н. И. Лобачевского; 603950, Россия, г. Нижний Новгород, пр-т Гагарина, 23; e-mail: timur@radbil.ru.*

ИНФОРМАЦИЯ ОБ АВТОРЕ: *Маркина Марина Викторовна, кандидат физико-математических наук, доцент, доцент кафедры теоретической, компьютерной и экспериментальной механики, Институт информационных технологий, математики и механики, Национальный исследовательский Нижегородский государственный университет им. Н. И. Лобачевского; 603950, Россия, г. Нижний Новгород, пр-т Гагарина, 23; e-mail: markinamv6213@yandex.ru.*

ДЛЯ ЦИТИРОВАНИЯ: *Радбиль, Т. Б.* Вероятностно-статистические модели в производства автороведческой экспертизы русскоязычных текстов / Т. Б. Радбиль, М. В. Маркина // Политическая лингвистика. — 2019. — № 2 (74). — С. 156-166. — DOI 10.26170/pl19-02-18.

0. Введение: к постановке проблемы

В статье представлен опыт разработки компьютерной программы диагностики и авторизации текста и ее адаптации к нуждам идентификационной и диагностической автороведческой экспертизы русскоязычных текстов.

Проблема отражения личности продуцента того или иного текста в результатах его речевой деятельности возникает еще задолго до появления лингвистики как науки. Как криминалистике, так и филологии, восходящей к герменевтике и экзегетике, хорошо известно, что активно и целенаправленно действующий субъект всегда оставляет значимые рефлексы, следы своей деятельности, по которым многое можно сказать о нем, о его внешних, внутренних и акциональных характеристиках.

Сегодня проблемы атрибуции текста находятся в самом эпицентре двух базовых взаимопересекающихся трендов современной гуманитаристики — это установка на

интердисциплинарность (все самое интересное в наши дни делается в гуманитарных исследовательских штудиях именно на «стыке наук») и тенденция к диверсификации, т. е. к расширению возможных сфер прикладного (общественно полезного) применения «высокого» теоретического знания в максимально большом количестве областей общественной практики.

Одной из такой социально значимых сфер приложения идей и принципов современной лингвистики является речеведческая экспертиза, особенно такая энергозатратная и трудоемкая ее разновидность, как **экспертиза автороведческая.** С одной стороны, в связи с развитием компьютерных технологий, с другой — с успехом применения математических вероятностно-статистических моделей в исследовании самых разнообразных аспектов бытия человека в мире судебное автороведение сегодня получает новые импульсы.

Сама идея применения математических, прежде всего статистических методов для атрибуции текста возникает в **стилеметрии** [Мартыненко 1988], которая, будучи прикладной филологической дисциплиной, первоначально не ставила себе криминалистических задач. Она пыталась «поверить алгеброй гармонию», т. е. объективировать и формализовать интуитивно ощущаемые любым грамотным читателем различия между стилем одного и другого автора, чтобы ответить на «вечные вопросы» литературоведения, кто является автором классических произведений — Гомер или не Гомер, Шекспир или не Шекспир, Шолохов или не Шолохов и пр.

В России истоки традиции выявления частот встречаемости тех или иных повторяющихся элементов текста для определения констант авторского стиля связывают с появлением работы Н. А. Морозова «Лингвистические спектры» (1916). В 70-е гг. XX в. выходит известное учебное пособие Б. Н. Головина «Язык и статистика», в котором рассмотрены возможности применения вероятностно-статистических методик не только для определения автора, но и для выявления особенностей того или иного стиля, для характеристики языковой эволюции тех или иных участков системы и пр. [Головин 1970]. Современное состояние вопроса обстоятельно охарактеризовано в книге А. Н. Баранова «Введение в прикладную лингвистику» [Баранов 2001].

Научный инструментарий современного лингвистического автороведения существенно обогащается за счет упрочения союза лингвистики и математики. Именно математики вносят существенный вклад в оснащение данной проблемы математическим научным аппаратом, прежде всего за счет повсеместного использования, например, вероятностных моделей А. А. Маркова [Хмелев 2000; Романов, Мещеряков 2009]. Совершенно справедливо отмечается многими авторами: «На сегодняшний день методы атрибуции текста отличаются большим разнообразием: одни направлены на изучение лексических показателей, другие на изучение синтаксических или грамматических характеристик. Существуют также некоторые другие подходы, авторы которых предлагают комплексный анализ текста на нескольких языковых уровнях» [Верзохин 2013: 24].

1. Цель, объект, методология и материал исследования

Одной из наиболее перспективных сфер применения вероятностных методов атрибуции текста является сегодня судебное автороведение и автороведческая экспертиза. Вероятностные модели используются для решения как идентификационных, связанных с установлением авторства спорных текстов, так и диагностических (определение необычного психофизиологического состояния, установление пола, возраста, области проживания, уровня образованности и пр.) задач [Галяшина, Ермолова 2005].

Термин «судебное автороведение» в криминалистической науке появился в начале 70-х гг. XX в. в работах известного автороведа С. М. Вула [Вул 1977]. Предметом судебного автороведения является установление фактических данных о личности автора при исследовании текста документа и иных материалов уголовного дела. Эти данные фиксируются в заключении эксперта и служат доказательством в процессе расследования и судебного разбирательства дел.

Между тем в судебном автороведении до сих пор не существует общепринятой теоретической платформы и единства методологических подходов в области идентификации и авторизации текста, абсолютно надежных методов и методик, дающих объективный, достоверный и не допускающий инотолкований результат экспертного автороведческого исследования, необходимый для доказательной базы в правоприменительной практике.

Таким образом, мы можем сформулировать цель настоящего исследования — продемонстрировать возможности разработанной нами компьютеризованной модели идентификационного автороведческого экспертного исследования текстов посредством

автоматической обработки текстов на основе комплексного применения вероятностно-статистических методов. Это очередной этап «прогона» усовершенствованной версии программы «КАТ» (компьютерная авторизация текста). Начальные стадии разработки описаны в работах [Юматов В., Маркина, Ковалева 2015; Юматов В., Маркина, Юматов С. 2016].

Иными словами, предполагается разработать концептуальные основы установления автора текста и на этой базе создать научно-практическую платформу для экспертного идентификационного анализа текстов на предмет их диагностики и авторизации. Поставленная задача требует междисциплинарного подхода. С одной стороны, исследование подобного рода базируется на данных лингвистики, в том числе фразеологии, грамматики и других областей знаний о языке и письменной речи, на системе знаний об условиях и закономерностях речевого поведения человека. С другой стороны, исследуемый текст характеризуется определенными статистическими закономерностями, которые измеряемы и вычислимы с достаточной степенью объективности, что позволяет применять математические методы для достижения требуемых результатов. Современное развитие компьютерных технологий позволяет в значительной степени формализовать и автоматизировать указанные исследования при наличии теоретически непротиворечивых и методологически оправданных параметров анализа.

Итак, в основу предлагаемой нами методики положено определение относительных частот соотношения тех или иных языковых элементов (вычисление коэффициентов корреляции и — на этой основе — колебания параметров) в нескольких выборках из сравниваемых текстов по комплексу разноуровневых параметров — коэффициентам Б. Н. Головина [Головин 1970], дополненному существующими в научной практике и прикладной сфере индексами понятности текста («индекс Флеша — Кинкейда», «FOG-индекс») и рядом других параметров. Материалом исследования являются первичные базы данных текстов русской классики (произведений Л. Н. Толстого, Н. В. Гоголя, И. С. Тургенева и др.).

2. Основной алгоритм и исходные принципы компьютеризованной модели авторизации текста

В настоящее время для теории и практики лингвистической экспертологии характерно существенное усложнение решаемых задач посредством постепенной выработки исследовательского инструментария, который сегодня позволяет выявить непрямо актуализованную информацию и дать экспертную квалификацию неявно выраженных, порою тщательно завуалированных и замаскированных диагностических языковых признаков, имеющих значение для выяснения обстоятельств дела (об этом подробнее см., например: [Радбиль 2014; Радбиль, Юматов 2014]). Все это актуально и для современного состояния авторoведческой экспертизы.

Задача идентификации авторства текстов и/или их диагностики, выступающая в качестве одной из основных в судебном авторoведении, в ряде специфических практических областей, например в правоприменительной практике или в экспертной деятельности по делам о плагиате, об экстремизме, в арбитражном судопроизводстве по делам о недобросовестной конкуренции, по документационным спорам и прочим, является в настоящее время актуальной проблемой. Это связано прежде всего с возрастанием объема текстов и, в связи с этим, с возможными спорами в области авторства, которые возникают при распространении ряда текстов через сеть Интернет без элементов атрибуции или под псевдонимами.

В разрабатываемой компьютеризованной модели авторизации и диагностики осуществляется комплексное применение вероятностно-статистических методов в судебной авторoведческой экспертизе, в результате чего производится характеристика разных уровней текста и объективируются константные и дифференциальные признаки спорных текстов на основе сопоставления относительных частот встречаемости повторяющихся элементов в нескольких репрезентативных выборках из сопоставляемых текстов по отношению друг к другу: это повышает степень достоверности и верифицируемости, так как относительные величины являются более индивидуализированными признаками речи.

Исследовательская модель базируется на следующих принципах.

Имеется несколько баз текстов. Каждая база — это тексты, написанные одним человеком. Имеется новый текст, авторство которого неизвестно. Необходимо определить, написан ли текст одним из авторов, тексты которых хранятся в базах, либо же новым автором.

Представим базы текстов в виде матриц, где количество столбцов m — количество обработанных текстов, количество строк n — количество параметров этих текстов.

Пусть у нас, например, три базы текстов с матрицами A, B, C (рис. 1).

$$\begin{matrix} a_{1_1} & a_{2_1} & \dots & a_{m1} \\ a_{1_2} & a_{2_2} & \dots & a_{m2} \\ \dots & \dots & \dots & \dots \\ a_{1n} & a_{2n} & \dots & a_{mn} \end{matrix} , \begin{matrix} b_{1_1} & b_{2_1} & \dots & b_{m1} \\ b_{1_2} & b_{2_2} & \dots & b_{m2} \\ \dots & \dots & \dots & \dots \\ b_{1n} & b_{2n} & \dots & b_{mn} \end{matrix} , \begin{matrix} c_{1_1} & c_{2_1} & \dots & c_{m1} \\ c_{1_2} & c_{2_2} & \dots & c_{m2} \\ \dots & \dots & \dots & \dots \\ c_{1n} & c_{2n} & \dots & c_{mn} \end{matrix}$$

Рис. 1

Для нового текста определены те же n параметров, как и для текстов из баз. Тогда новый текст можно охарактеризовать следующим вектор-столбцом:
new (*new*$_1$, *new*$_2$, … *new*$_n$)^T.

Для каждой матрицы посчитаем коэффициент корреляции Пирсона каждого столбца с вектором *new* и получим (с учетом наличия 3 баз) три вектора s, p, r длины m [Кремер 2007].

Далее найдем среднее значение элементов каждого вектора. Получим усредненные *s̃, p̃, r̃,* из этих значений составим вектор *k(s̃, p̃, r̃).*

Новый текст можно добавить в уже существующую базу текстов в том случае, если компонент вектора *k(s̃, p̃, r̃)*, соответствующий коэффициенту корреляции именно с этой базой, близок к единице в пределах доверительного интервала. Если это условие не выполняется, то текст добавляется в новую базу.

***Выбор параметров анализа текста*.** Наиболее важным моментом создания алгоритма является правильный подбор параметров корреляции. Параметры должны отражать сознательный выбор автором грамматических моделей и синтаксических конструкций и специфику отбора слов.

В качестве базовых параметров, которые позволяют только исключить определенное авторство, но не диагностировать его, предлагается включить показатели понятности текста: индекс Флеша — Кинкейда и индекс «туманности» текста — так называемый FOG-индекс. В данном исследовании учет этих параметров не проводится.

Еще раз отметим, что это — вспомогательные параметры, позволяющие исключить чье-либо авторство, но не идентифицировать его; иными словами, существенные расхождения в значениях этих индексов означают, что текст создан разными авторами, но одинаковое значение индексов может быть характерно и для разных авторов, и для одного автора. Из являющихся традиционными для методик данного рода вспомогательных параметров, которые исключают, но не идентифицируют авторство, мы включили значения по таким параметрам, как **средняя длина слова** и **средняя длина предложения**, и один формальный показатель — **отношение количества знаков препинания к количеству слов.**

Собственно идентификационное значение для авторизации имеют параметры, представленные в работе Б. Н. Головина «Язык и статистика» (1970).

- **Коэффициент предметности (Pr)** — отношение суммы существительных и местоимений к сумме прилагательных и глаголов.
- **Коэффициент качественности (Qu)** — отношение суммы прилагательных и наречий к сумме глаголов и существительных.
- **Коэффициент активности (Ac)** — отношение суммы глаголов и глагольных форм к количеству слов в тексте.
- **Коэффициент динамизма (Din)** — отношение суммы глаголов и глагольных форм к сумме существительных, прилагательных и местоимений.
- **Коэффициент связности текста (Con)** — отношение суммы предлогов и союзов к числу предложений [Головин 1970].

Вычисление указанных параметров для большого количества текстов позволяет сделать вывод о возможности применения коэффициента корреляции Пирсона, так как выборки подчинены нормальному закону распределения.

Следует отметить, что при вычислении этих восьми параметров корреляции используются в целом двенадцать параметров текста:

1) количество знаков препинания;
2) количество слов;
3) количество предложений;
4) средняя длина слова;
5) средняя длина предложения;
6) количество существительных;
7) количество прилагательных;
8) количество глаголов;
9) количество местоимений;
10) количество наречий;
11) количество предлогов
12) количество союзов.

На этой основе в программе автоматизированы следующие алгоритмы:

1. Алгоритм формирования массива слов текста и знаков препинания после каждого слова. Список возможных знаков препинания задается.
2. Алгоритм определения частоты встречаемости слов и знаков препинания.
3. Алгоритм нахождения наиболее употребляемых слов и знаков препинания.
4. Алгоритм нахождения прилагательных в тексте и подсчета их количества. Прилагательные идентифицируются по окончанию слова.

5. Алгоритм нахождения глаголов в тексте и подсчета их количества. Глаголы идентифицируются по окончанию слова.
6. Алгоритм формирования массива предложений, из которых состоит исходный текст.
7. Алгоритм подсчета среднего количества слов в предложениях текста.
8. Алгоритм поиска слов в словаре, использующий бинарный поиск элемента в массиве. Для поиска замены для ошибочно написанного в тексте слова используется понятие «дистанция Левенштейна». Эта «дистанция» — минимальное количество правок первого слова (под правками подразумеваются три возможные операции: стирание символа, замена символа и вставка символа), чтобы превратить его во второе слово.
9. Алгоритм вычисления индекса Флеша — Кинкейда (FRE-индекс).
10. Алгоритм вычисления индекса Роберта Ганнинга (FOG-индекс).

Рекомендуемый размер выборки текста — не менее 500 слов и 30 предложений [Головин 1970].

Практическая реализация предложенной модели нашла свое выражение в создании компьютерной программы криминалистической диагностики и авторизации текста «КАТ», предназначенной **д**ля автоматизированной обработки текстовых данных.

3. Характеристики компьютерной программы криминалистической диагностики и авторизации текста «КАТ»

Компьютерная программа криминалистической диагностики и авторизации текста «КАТ» предназначена для автоматизированной обработки текстовых данных. В программе «КАТ» для получения интересующих нас данных об исполнителе текстов заложена возможность комплексного исследования: создаются заданные параметры (они могут быть расширены в зависимости от целей исследователя). Для реализации этих целей используется программный комплекс обработки текстовой информации с учетом особенностей русского языка. В этих целях исследуемый текст, предназначенный для обработки, должен быть подготовлен в виде текстового файла или введен в онлайн-режиме пользователем программы. Для проверки правильности написания слов применяется словарь, который также является текстовым файлом.

Основные функции программы. Selection(list<Item> res, long&k) — выборка слов из исходной строки и вывод таблицы. Используется для нахождения количества глаголов и прилагательных; IsEndSens(Item c) — поиск конца предложений; GetCountAdjective() — вычисление количества прилагательных в тексте; GetCountVerb() — вычисление количества глаголов; GetCountSentence() — вычисление количества предложений текста; GetMeanCountWordinSentence() — поиск среднего количества слов в предложении; Convert(char&symb) — перевод символа из верхнего регистра в нижний; DistLevinstane (const std::string& from, const std::string& to) — возвращает количество действий для добавления/удаления/замены символов (используется для проверки грамотности); findWord(string in) — поиск слов в словаре; getIndexFOG() — подсчет индекса туманности; getIndexFRE() — подсчет индекса удобочитаемости; IsVerb() — поиск всех глаголов; GetWord() — поиск всех слов в тексте; GetPrep() — поиск знаков препинания; Get CountSyllableInWord() — подсчет слогов; Is Adjective() — поиск всех прилагательных в тексте.

Для работы с программой создан интерфейс, позволяющий пользователю выбрать одну из опций:

- создать новую базу текстов;
- сделать просмотр или редакцию существующих баз текстов;
- сделать проверку нового текста.

При создании новой базы запрашивается имя автора, добавляется текст, осуществляется расчет всех коэффициентов.

При просмотре выдается список всех существующих баз текстов. При редактировании предусмотрена возможность удаления текста из базы.

При проверке нового текста выдается один из двух запросов — либо запрос на разрешение поместить проверенный текст в уже существующую базу данных (если вычисленный коэффициент корреляции имеет соответствующую этой базе величину), либо запрос на создание новой базы текстов.

4. Описание работы программы «КАТ»

Приведем пример вычисления всех параметров и коэффициентов корреляции на предварительно созданных в целях экспериментальной проверки модели базах данных с текстами Н. В. Гоголя, Л. Н. Толстого и И. С. Тургенева. Ниже приводятся графики колебания каждого параметра в пределах существующей базы текстов (соответственно **рис. 2, 3** и **4**).

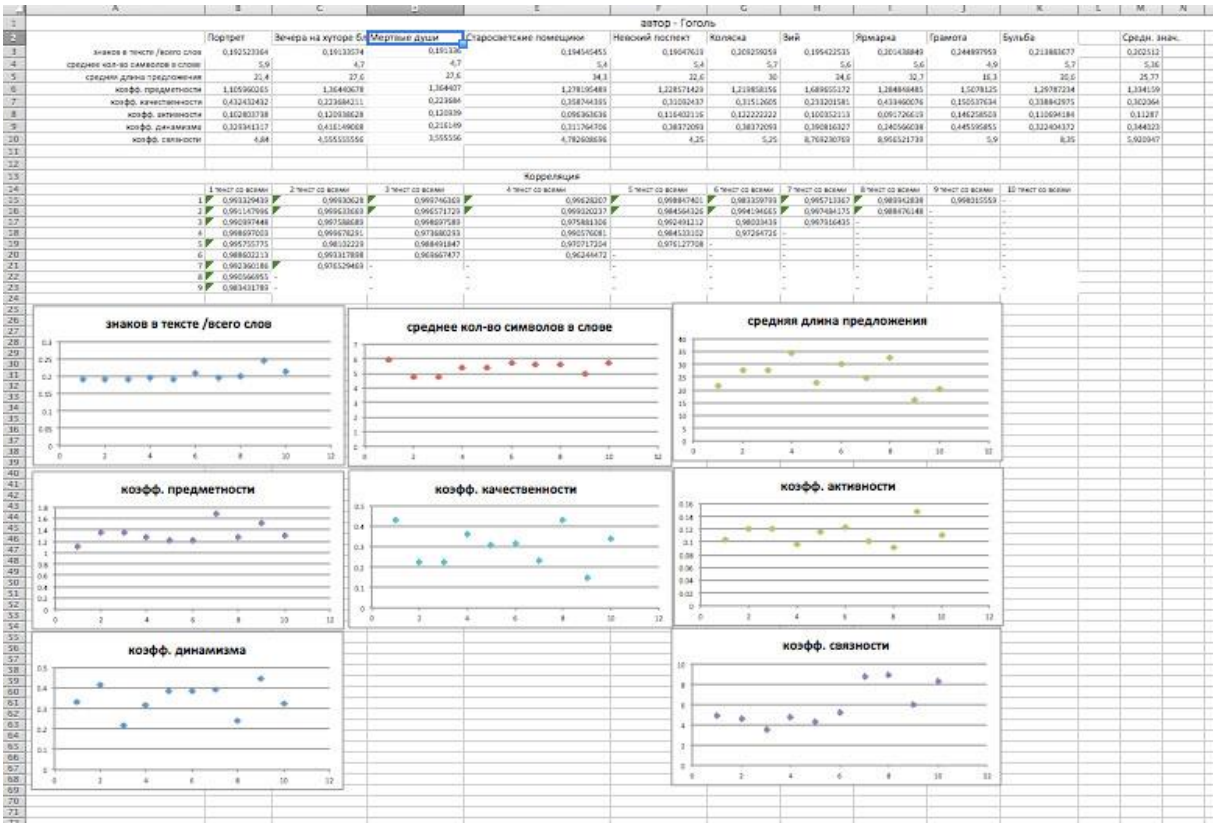

Рис. 2. Н. В. Гоголь

Рис. 3. Л. Н. Толстой

Рис. 4. И. С. Тургенев

Таблица 1. Параметры корреляции по выборке

знаков в тексте/всего слов	0,228417266
среднее кол-во символов в слове	5,2
средняя длина предложения	22,2
коэфф. предметности	0,96
коэфф. качественности	0,300884956
коэфф. активности	0,147482014
коэфф. динамизма	0,569444444
коэфф. связности	8,2

Таблица 2. Сопоставительный анализ параметров корреляции

Параметры	Новый текст	1 база — Гоголь	2 база — Толстой	3 база — Тургенев
отношение кол-ва знаков в тексте к числу слов	0,191336	0,20622	0,251259	0,21622
среднее кол-во символов в слове	4,7	5,37	7,66	5,53
средняя длина предложения	27,6	25,83	20,19	18,93
коэфф. предметности	1,364407	1,293718	1,275741	2,924918
коэфф. качественности	0,223684	0,309784	0,284997	0,185786
коэфф. активности	0,120939	0,115524	0,12536	0,104544
коэфф. динамизма	0,216149	0,369352	0,501173	0,173962
коэфф. связности	4,555556	5,385392	18,76997	16,09842

Таблица 3. Сопоставительный анализ средних значений коэффициентов корреляции

Коэфф. корреляции с базой «Гоголь»	Коэфф. корреляции с базой «Толстой»	Коэфф. корреляции с базой «Тургенев»
0,998595	0,760158	0,813944

Для демонстрации работы программы возьмем новый отрывок текста. Пусть это будет отрывок из поэмы Н. В. Гоголя «Мертвые души», начинающийся со слов: «*Лицо Ноздрева, верно, уже сколько-нибудь знакомо читателю. Таких людей приходилось всякому встречать немало*…» — и заканчивающийся словами: «*…напротив, если случай приводил его опять встретиться с вами, он обходился вновь по-дружески и даже говорил: „Ведь ты такой подлец, никогда ко мне не заедешь“*».

Вычислим все параметры корреляции текста (табл. 1).

Теперь сравним эти параметры с усредненными параметрами по каждой из трех существующих баз (которые играют роль сравнительных образцов для традиционной автороведческой экспертизы) — табл. 2.

Вычислим средние значения всех коэффициентов корреляции в сопоставлении нового текста с каждой базой (табл. 3).

Качественная интерпретация полученных коэффициентов. Возникает резонный вопрос, какая реальность стоит за сопоставлением средних значений коэффициентов корреляции нового текста в сравнении с базами 1, 2 и 3? Самая сложная проблема для всех без исключения количественных методик авторизации текста всегда заключалась в том, какие числовые показатели маркируют границу между отнесением принадлежности анализируемого текста к тому же автору, к которому относятся сравнительные образцы, и исключением его принадлежности данному автору. Иными словами, какие различия в числовых показателях сопоставляемых коэффициентов следует считать допустимыми, обусловленными, например, разницей в жанре, теме или времени написания текста одним и тем же автором, а какие нужно считать принципиальными с точки зрения исключения авторства?

Возвращаясь к полученным нами результатам (анализируемый «новый» текст имеет сходство с базой 1 «Гоголь» — 0,998595 при максимуме сходства 1 = тождество, что существенно выше, чем сходство, соответственно, с базой 2 «Толстой» — 0,760158 и базой 3 «Тургенев» — 0,813944), отметим, что на данном этапе исследования эмпирическим и экспериментальным путем были выработаны следующие критерии идентификации авторства:

- считать текст принадлежащим автору, если коэффициент корреляции текста с существующей базой больше 0,87, т. е. в качестве доверительного интервала принять интервал 0,87—1,0;
- в диапазоне коэффициента корреляции 0,82—0,87 требуется дополнительное мнение эксперта (встречались единичные случаи авторства при таких значениях коэффициента);
- считать текст не принадлежащим автору, если коэффициент корреляции текста с существующей базой меньше 0,82;
- учесть, что точность работы программы увеличивается с увеличением текстов в базе. Под базой понимается средний показатель, исчисленный по всем текстам, с достоверностью известным как принадлежащие данному автору.

Таким образом, предполагаемое авторство Гоголя доказывается значением 0,998595, что существенно выше рекомендуемого «порогового» значения для установления авторства (0,82—0,87), на фоне значений предполагаемого авторства Толстого 0,760158 или Тургенева 0,813944, которые, соответственно, ниже «порогового» значения 0,82—0,87.

В целях верификации установленных критериев идентификации авторства в проводимом эксперименте для автоматической обработки был взят текст нового автора, не входящего в существующие базы — отрывок текста Д. Донцовой из романа «Огнетушитель Прометея».

Вычислим все параметры корреляции текста (табл. 4).

Сравним эти параметры с усредненными параметрами по каждой из трех существующих баз (табл. 5).

Вычислим средние значения всех коэффициентов корреляции в сопоставлении нового текста с каждой базой (табл. 6).

Таблица 4. Параметры корреляции по выборке

знаков в тексте/всего слов	0,65
среднее кол-во символов в слове	4,5
средняя длина предложения	9,3
коэфф. предметности	0,3329
коэфф. качественности	0,161551
коэфф. активности	1,763359
коэфф. динамизма	3,53333
коэфф. связности	3,215556

Таблица 5. Сопоставительный анализ параметров корреляции

Параметры	Текст для сравнения	1 база — Гоголь	2 база — Толстой	3 база — Тургенев
отношение кол-ва знаков в тексте к числу слов	0,65	0,20622	0,251259	0,21622
среднее кол-во символов в слове	4,5	5,37	7,66	5,53
средняя длина предложения	9,3	25,83	20,19	18,93
коэфф. предметности	0,3329	1,293718	1,275741	2,924918
коэфф. качественности	0,161551	0,309784	0,284997	0,185786
коэфф. активности	1,763359	0,115524	0,12536	0,104544
коэфф. динамизма	3,533333	0,369352	0,501173	0,173962
коэфф. связности	3,215556	5,385392	18,76997	16,09842

Таблица 6. Сопоставительный анализ средних значений коэффициентов корреляции

Коэфф. корреляции с базой «Гоголь»	Коэфф. корреляции с базой «Толстой»	Коэфф. корреляции с базой «Тургенев»
0,718595	0,75235	0,747093

Нетрудно видеть, что в данном случае коэффициент идентификации совпадения авторства «нового» текста с Гоголем, Толстым или Тургеневым ожидаемо ниже «порогового» значения 0,82—0,87: соответственно 0,718595 при сопоставлении с Гоголем, 0,75235 при сопоставлении с Толстым и 0,747093 при сопоставлении с Тургеневым. Данный анализ дает еще и ряд дополнительных результатов, например, позволяет установить, что текст Донцовой по стилю несколько «ближе» к Толстому, чем к Тургеневу и тем более к Гоголю, но это уже представляет чисто умозрительный интерес и не имеет идентификационного значения.

5. Выводы и перспективы

Предлагаемая в работе модель автоматической обработки текста для его дальнейшей авторизации и диагностики, разумеется, может найти свое применение не только в судебном автороведении, но и в решении разнообразных задач атрибуции текста в гуманитарных исследованиях разной направленности.

«Пилотная» версия экспериментальной компьютерной программы «КАТ» прошла апробацию в реальной автороведческой экспертной практике. В 2018 г. одним из соавторов была произведена идентификационная судебная автороведческая экспертиза трех документов официально-делового стиля в рамках арбитражного судебного дела по иску налоговой инспекции о признании недействительной сделки. Необходимо было авторизовать данные документы на предмет возможной фальсификации авторства какого-либо из них. В силу отсутствия базы данных по официально-деловой документации в сфере финансово-договорных отношений тексты исследовались попарно, в сопоставлении 1 → 2, 1 → 3 и 2 → 3, и коэффициенты корреляции сравнивались также попарно. Использование в ходе экспертного исследования программы «КАТ» подтвердило выводы истца о том, что документ 1 и 2 написаны одним лицом, а документ 3 написан другим лицом, т. е. помогло установить факт фальсификации авторства.

Однако при этом следует отметить, что представленная в настоящей работе компьютерная программа идентификации и диагностики авторства текста все же находится еще в стадии разработки. Уже на этом этапе видны преимущества и недостатки данной модели. Преимущества заключаются в относительной простоте использования, интуитивной понятности параметров авторизации и логики их исчисления, а также прозрачности качественной интерпретации результатов, что имеет порою решающее значение в представлении результатов автороведческого экспертного исследования в суде, перед лицами, не являющимися специалистами в данной области.

Определенным недостатком программы на данной стадии разработки является некоторая «грубость», «приблизительность». Относительно вычисления коэффициентов корреляции имеются вопросы и к самому составу параметров, и к необходимому и достаточному количеству выборок текстов (также и к объему самих выборок) и пр. Необходимы более тонкие и точные расчеты, что напрямую зависит от качественного состава и объема баз данных текстов.

Поэтому в перспективе необходимо как количественное расширение существующих баз данных, так и создание новых баз данных по текстам разной стилевой принадлежности (тексты СМИ, политические тексты, юридические документы, официально-деловые документы и пр.). Также желательно осуществить дифференциацию баз данных

текстов по хронологическому принципу, по тематической направленности, по жанровой принадлежности, по авторам и т. п. Именно в этом случае можно будет решать экспертные задачи по авторизации и диагностике спорных текстов в наиболее юридизированных областях бытования текстов в современном мире — в медийном и политическом дискурсе, в юридической, официально-деловой и коммерческой документации и пр.

ЛИТЕРАТУРА

1. Баранов А. Н. Введение в прикладную лингвистику : учеб. пособие. — М. : Эдиториал УРСС, 2001. 347 с.

2. Верзохин С. С. К вопросу о лингвотеоретических основах методик авторизации текста // Учен. зап. ЗабГГПУ. 2013. № 2 (49). С. 22—27.

3. Вул С. М. Теоретические и методические вопросы криминалистического исследования письменной речи. — М. : ВНИИСЭ, 1977. 109 с.

4. Галяшина Е. И., Ермолова Е. И. Перспективы развития автороведческой экспертизы в России // Судебная экспертиза : науч.-практ. журн. 2005. № 3. С. 5—11.

5. Головин Б. Н. Язык и статистика. — М. : Просвещение, 1970. 190 с.

6. Кремер Н. Ш. Теория вероятностей и математическая статистика. Изд. 3-е, перераб. и доп. — М. : ЮНИТИ—ДАНА, 2007. 543 с.

7. Мартыненко Г. Я. Основы стилеметрии. — Л. : Изд-во Ленингр. ун-та, 1988. 173 с.

8. Морозов Н. А. Лингвистические спектры: средство для отличения плагиатов от истин. произведений того или др. известного авт. [Электронный ресурс]. — Петроград : тип. Имп. Акад. наук, 1916. 42 с. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3.

9. Радбиль Т. Б. Выявление содержательных и речевых признаков недобросовестной информации в экспертной деятельности лингвиста // Вестн. Нижегор. ун-та им. Н. И. Лобачевского. 2014. № 6. С. 146—149.

10. Радбиль Т. Б., Юматов В. А. Способы выявления имплицитной информации в лингвистической экспертизе // Вестн. Нижегор. ун-та им. Н. И. Лобачевского. 2014. № 3 (2). С. 18—21.

11. Романов А. С., Мещеряков Р. В. Идентификация автора текста с помощью аппарата опорных векторов // Компьютерная лингвистика и интеллектуальные технологии : по материалам ежегодной Междунар. конф. «Диалог-2009» (Бекасово, 27—31 мая 2009 г.). — М. : РГГУ, 2009. Вып. 8 (15). С. 432—437.

12. Хмелев Д. В. Распознавание автора текста с использованием цепей А. А. Маркова // Вестн. МГУ. Сер. 9, Филология. 2000. № 2. С. 115—126.

13. Юматов В. А., Маркина М. В., Ковалева А. С. Программа криминалистической диагностики и авторизации текста «КАТ» // Вестн. Костром. ун-та. 2015. Т. 21. № 3. С. 199—202.

14. Юматов В. А., Маркина М. В., Юматов С. В. Математические методы криминалистической диагностики и авторизации текста в речеведческой экспертизе // Вестн. Нижегор. ун-та им. Н. И. Лобачевского. 2016. № 5. С. 227—232.

T. B. Radbil
National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia
ORCID ID: 0000-0002-7516-6705 ☑

M. V. Markina
National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia
ORCID ID: — ☑

☑ ***E-mail:*** *timur@radbil.ru; markinamv6213@yandex.ru.*

Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts

ABSTRACT. *The article presents the experience of developing a computerized text authorization model and its adaptation to the needs of identification and diagnostic authoring expertise of Russian texts. The purpose of the study is to demonstrate the possibilities of identification authoring expert examination of texts through automatic text processing based on the integrated application of probabilistic-statistical methods. The article describes one more stage of testing an improved version of the CAT program (computerized text authorization) – an experiment to determine the relative frequencies of the ratio of certain linguistic elements (calculation of correlation coefficients) in several samples of compared texts using a set of different-level parameters —— B.N. Golovin's coefficients, supplemented by the text clarity indexes existing in scientific practice and applied field ("Flesch-Kincaid Index", "FOG-Index") and a number of other parameters. The research materials consist of primary databases of texts of Russian classics (works by L.N. Tolstoy, N.V. Gogol, I.S. Turgenev). As a result of the experimental study, the following criteria for identifying authorship were developed: the text is considered to belong to the author, if the correlation coefficient of the text with the existing base is greater than 0.87, i.e. the interval 0.87-1 should be taken as a confidence interval; the text is assumed not to belong to the author, if the correlation coefficient of the text with the existing base is less than 0.82; we should note that the accuracy of the program increases with longer texts in the database. The author defines a base as the average indicator, calculated for all texts, certainty known as belonging to this author. In case of successful refinement of the proposed CAT automatic text processing program, it will be possible to solve expert problems of authorization and diagnostics of contentious texts produced in the space of media and political discourses, in legal and official business, commercial documentation, etc.*

KEYWORDS: *forensic authoring; authoring expertise; text authoring; automatic text procession, probabilistic-statistical method; Russian language.*

AUTHOR'S INFORMATION: *Radbil' Timur Ben'yuminovich, Doctor of Philology, Professor, Head of Department of Theoretical and Applied Linguistics, Institute of Philology and Journalism, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia.*

AUTHOR'S INFORMATION: *Markina Marina Viktorovna, Candidate of Physics and Mathematics, Associate Professor of Department of Theoretical, Computer and Experimental Mechanics, Institute of Information Technologies, Mathematics and Mechanics, National Research Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia.*

FOR CITATION: *Radbil', T. B.* Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts / T. B. Radbil', M. V. Markina // Political Linguistics. — 2019. — No 2 (74). — P. 156—166. — DOI 10.26170/pl19-02-18.

REFERENCES

1. Baranov A. N. Introduction to Applied Linguistics : teaching aid. — Moscow : Editorial URSS, 2001. 347 p. [Vvedenie v prikladnuyu lingvistiku : ucheb. posobie. — M. : Editorial URSS, 2001. 347 s.]. — (In Rus.)

2. Verzokhin S. S. On the Issue of Linguistic-theoretical Fundamentals of text Authorization Methods // Proceedings of ZabGGPU. 2013. No 2 (49). P. 22—27. [K voprosu o lingvoteoreticheskikh osnovakh metodik avtorizatsii teksta // Uchen. zap. ZabGGPU. 2013. № 2 (49). S. 22—27]. — (In Rus.)

3. Vul S. M. Theoretical and Methodological Issues of Forensic Research in Writing. — Moscow : VNIISE, 1977. 109 p. [Teoreticheskie i metodicheskie voprosy kriminalisticheskogo issledovaniya pis'mennoy rechi. — M. : VNIISE, 1977. 109 s.]. — (In Rus.)

4. Galyashina E. I., Ermolova E. I. Prospects for the Development of an Author's Expert Examination in RUSSIA // Forensic Examination : scientific and practical journal. 2005. No. 3. P. 5—11. [Perspektivy razvitiya avtorovedcheskoy ekspertizy v Rossii // Sudebnaya ekspertiza : nauch.-prakt. zhurn. 2005. № 3. S. 5—11]. — (In Rus.)

5. Golovin B. N. Language and Statistics. — Moscow : Enlightenment, 1970. 190 p. [Yazyk i statistika. — M. : Prosveshchenie, 1970. 190 s.]. — (In Rus.)

6. Kremer N. Sh. Probability Theory and Mathematical Statistics. Ed. 3rd, rev. and add. — Moscow : UNITY — DANA, 2007. 543 p. [Teoriya veroyatnostey i matematicheskaya statistika. Izd. 3-e, pererab. i dop. — M. : YuNITI—DANA, 2007. 543 s.]. — (In Rus.)

7. Martynenko G. Ya. Basics of style-metrics. — Leningrad : Publishing House of Leningrad Univ., 1988. 173 p. [Osnovy stilemetrii. — L. : Izd-vo Leningr. un-ta, 1988. 173 s.]. — (In Rus.)

8. Morozov N. A. Linguistic Spectrs: a means for distinguishing of plagiarism and original works for famous authots. [Electronic resource]. — Petrograd : Type of Imp. Acad. Sciences, 1916. 42 p. [Lingvisticheskie spektry: sredstvo dlya otlicheniya plagiatov ot istin. proizvedeniy togo ili dr. izvestnogo avt. — Petrograd : tip. Imp. Akad. nauk, 1916. 42 s.]. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3. — (In Rus.)

9. Radbil' T. B. Identification of content and speech signs of unfair information in expert activities of a linguist // Nizhniy Novgorod Univ. n. a. N. I. Lobachevsky Journ. 2014. No. 6. P. 146—149. [Vyyavlenie soderzhatel'nykh i rechevykh priznakov nedobrosovestnoy informatsii v ekspertnoy deyatel'nosti lingvista // Vestn. Nizhegor. un-ta im. N. I. Lobachevskogo. 2014. № 6. S. 146—149]. — (In Rus.)

10. Radbil' T. B., Yumatov V. A. Ways to Identify Implicit Information in Linguistic Expertise // Nizhniy Novgorod Univ. n. a. N. I. Lobachevsky Journ. 2014. № 3 (2). P. 18—21. [Sposoby vyyavleniya implitsitnoy informatsii v lingvisticheskoy ekspertize // Vestn. Nizhegor. un-ta im. N. I. Lobachevskogo. 2014. № 3 (2). S. 18—21]. — (In Rus.)

11. Romanov A. S., Meshcheryakov R. V. Identification of the Author of the Text Using the Support Vector Machine // Computational linguistics and intellectual technologies: based on the materials of the annual International conf. "Dialogue-2009" (Bekasovo, May 27–31, 2009). — Moscow : RGGU, 2009. Vol. 8 (15). P. 432—437. [Identifikatsiya avtora teksta s pomoshch'yu apparata opornykh vektorov // Komp'yuternaya lingvistika i intellektual'nye tekhnologii : po materialam ezhegodnoy Mezhdunar. konf. «Dialog-2009» (Bekasovo, 27—31 maya 2009 g.). — M. : RGGU, 2009. Vyp. 8 (15). S. 432—437]. — (In Rus.)

12. Khmelev D. V. Recognition of the Author of the Text Using the Chains of A. Markov // Moscow State Univ. Journ. Ser. 9, Philology. 2000. № 2. P. 115—126. [Raspoznavanie avtora teksta s ispol'zovaniem tsepey A. A. Markova // Vestn. MGU. Ser. 9, Filologiya. 2000. № 2. S. 115—126]. — (In Rus.)

13. Yumatov V. A., Markina M. V., Kovaleva A. S. The Program of Forensic Diagnostics and Authorization of the Text "KAT" // Kostroma Univ. Journ. 2015. Vol. 21. No. 3. P. 199—202. [Programma kriminalisticheskoy diagnostiki i avtorizatsii teksta «KAT» // Vestn. Kostrom. un-ta. 2015. T. 21. № 3. S. 199—202]. — (In Rus.)

14. Yumatov V. A., Markina M. V., Yumatov S. V. Mathematical Methods of Forensic Diagnostics and Text Authorization in Speech Expertise // Nizhniy Novgorod Univ. n. a. N. I. Lobachevsky Journ. 2016. No. 5. P. 227—232. [Matematicheskie metody kriminalisticheskoy diagnostiki i avtorizatsii teksta v rechevedcheskoy ekspertize // Vestn. Nizhegor. un-ta im. N. I. Lobachevskogo. 2016. № 5. S. 227—232]. — (In Rus.)