www.volsu.ru

DOI: https://doi.org/10.15688/jvolsu2.2021.5.4

UDC 81'322.2
LBC 81.112

Submitted: 14.12.2020
Accepted: 30.03.2021

RUSSIAN TEXT AUTHOR'S GENDER IDENTIFICATION IN FORENSIC EXAMINATION: PROBABILITY-AND-STATISTICS METHOD

Timur B. Radbil

Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

Marina V. Markina

Lobachevsky State University of Nizhny Novgorod, Nizhny Novgorod, Russia

Abstract. The article discusses intermediate research results in the development and improvement of a computerized model of Russian texts authorization, which is based on complex application of probabilistic-and-statistical methods. The study aims to describe the new capabilities of the created system in the aspect of its application to diagnostic examinations in text authorization for detection of the gender of the alleged author of the text. The work presents the next stage of fine-tuning and testing of the improved version of the computer program "CTA" (computerized text authorization), which at this stage was adapted for the task of determining and comparing stable relative frequencies of correlation coefficients (the ratio of specified linguistic phenomena of different levels of the language system) in the texts, the authors of which are men and women. The research material is the continuously updated primary bases of literary texts of the 19th and 21st centuries (4 bases, respectively). The work shows that for the texts written by men and women, significant differences can be noted in such correlation coefficients as average word length, average sentence length, objectivity coefficient, quality coefficient, activity coefficient, dynamism coefficient, connectivity coefficient, etc. Verification of the results obtained experimentally has demonstrated that the accuracy of gender determining at this stage of the study is approximately 65%. This indicator can be significantly exceeded with an increase in the volume and quality specification of databases and/or when using new models for calculating the correlation coefficients (Spearman's model, etc.).

Key words: text authorization, computer text authorization, gender, forensic studies in text authorization, automatic text processing, probability-and-statistics method, applied linguistics.

Citation. Radbil T.B., Markina M.V. Russian Text Author's Gender Identification in Forensic Examination: Probability-and-Statistics Method. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznanie* [Science Journal of Volgograd State University. Linguistics], 2021, vol. 20, no. 5, pp. 43-55. (in Russian). DOI: https://doi.org/10.15688/jvolsu2.2021.5.4

УДК 81'322.2
ББК 81.112

Дата поступления статьи: 14.12.2020
Дата принятия статьи: 30.03.2021

ВЕРОЯТНОСТНО-СТАТИСТИЧЕСКАЯ МЕТОДИКА УСТАНОВЛЕНИЯ ГЕНДЕРНОЙ ПРИНАДЛЕЖНОСТИ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ В СУДЕБНОМ АВТОРОВЕДЕНИИ

Тимур Беньюминович Радбиль

Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород, Россия

Марина Викторовна Маркина

Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, г. Нижний Новгород, Россия

Аннотация. В статье представлены промежуточные результаты разработки и усовершенствования компьютеризованной модели авторизации текстов на русском языке на основе вероятностно-статистической методики. Целью исследования является интерпретация экспериментального исследования расширенных возможностей компьютерной программы авторизации текста «КАТ» (компьютерная авторизация текста) при применении ее в диагностических автороведческих экспертизах по установлению гендерной принадлежности предполагаемого автора текста. Описаны результаты апробации усовершенствованной версии «КАТ», которая была адаптирована для определения и сопоставления стабильных относительных частот коэффициентов корреляции в текстах, авторами которых являются мужчины и женщины. Материалом исследования послужили созданные авторами и непрерывно пополняемые первичные базы художественных текстов XIX и XXI веков. Установлено, что в текстах, написанных мужчинами и женщинами, имеются значимые расхождения в таких коэффициентах корреляции, как средняя длина слов, средняя длина предложения, коэффициент предметности, коэффициент качественности, коэффициент активности, коэффициент динамизма, коэффициент связности. Проверка полученных результатов показала, что точность определения гендерной принадлежности на данном этапе исследования составляет приблизительно 65 %. Этот показатель может быть существенно превышен при увеличении объема и качественной спецификации баз данных с учетом типа дискурса и/или при использовании других моделей исчисления коэффициентов корреляции.

Ключевые слова: авторизация текста, компьютерная авторизация текста, гендер, судебное автороведение, автоматическая обработка текста, вероятностно-статистическая методика, прикладная лингвистика.

Цитирование. Радбиль Т. Б., Маркина М. В. Вероятностно-статистическая методика установления гендерной принадлежности текстов на русском языке в судебном автороведении // Вестник Волгоградского государственного университета. Серия 2, Языкознание. – 2021. – Т. 20, № 5. – С. 43–55. – DOI: https://doi.org/10.15688/jvolsu2.2021.5.4

Введение

В статье представлены результаты очередного этапа разработки компьютерной программы диагностики и авторизации текста «КАТ», описанной в наших предыдущих исследованиях применительно к собственно идентификации авторства [Радбиль, Маркина, 2019]. Он связан с решением новой экспериментальной задачи определения гендерной принадлежности авторов спорных текстов на русском языке. Еще на стадиях замысла, создания и первичной апробации программы (см.: [Юматов, Маркина, Ковалева, 2015; Юматов В.А., Маркина, Юматов С.В., 2016]) в нее была заложена возможность диверсифицировать не только проблемы идентификации текста, но и проблемы классификационного и диагностического типа, к каковым относится определение гендерной принадлежности автора.

Использование методик автоматической обработки текста востребовано в судебных автороведческих экспертизах. «В связи с развитием компьютерных технологий, с одной стороны, и с успехом применения математических вероятностно-статистических моделей в исследовании самых разнообразных аспектов бытия человека в мире – с другой, судебное автороведение сегодня получает новые импульсы» [Радбиль, Маркина, 2019, с. 157]. Это связано прежде всего с возрастанием объема и повышением качественного и структурного многообразия попадающих в сферу внимания правоприменительной практики текстов, которые часто могут распространяться через сеть Интернет без элементов атрибуции, анонимно или псевдонимно.

Актуальным для экспертов-авторoведов видится такой аспект проблемы, как выявление имплицитных компонентов спорного текста на разных уровнях их «залегания» [Галяшина, Ермолова, 2005; Радбиль, 2014; Радбиль, Юматов, 2014; и др.] посредством квантитативных методик, например подсчета частоты использования определенных элементов текста, которая может свидетельствовать об особенностях психологического состояния автора, его социальных, образовательных, гендерных характеристиках [Катышев, Осадчий, 2018; Хоменко, 2019; Литвинова, Громова, 2020].

Современные компьютерные методы авторизации и диагностики текста, отражающие инновационное развитие традиционных принципов стилеметрии, или стилостатистики, во многом восходящие к знаменитой работе Н.А. Морозова «Лингвистические спектры» [Морозов, 1916], отличаются разнообразием [Баранов, 2001; Верзохин, 2013; и др.]. Как от-

мечает С.С. Верзохин, «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, с. 24]. Сегодня большинство методов основано на применении разных версий вероятностно-статистического подхода к анализу текста [Головин, 1970; Хмелев, 2000; Кремер, 2007; Романов, Мещеряков, 2009; Хоменко, 2019]. С их помощью ставятся и решаются различные авторизационные задачи в области текстологии, в том числе обладающие культурной значимостью. Результаты применения указанного подхода в текстологии привлекли внимание и представителей лингвокриминалистики.

Пионером в области отечественного криминалистического исследования письменной речи по праву считается С.М. Вул. Именно он заложил основы современного судебного авторoведения и разработал его терминологический аппарат (см.: [Вул, 1977]). Предметом судебного авторoведения 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 [Галяшина, Ермолова, 2005].

Изначально судебное авторoведение сосредоточилось исключительно на определении авторства спорного текста [Литвинова, 2012], в том числе анонимного [Argamon et al., 2009], и добилось на этом пути значительных успехов. Однако с развитием компьютерных технологий круг проблем, требующих обсуждения, неуклонно расширяется. Специалисты обратили внимание на то, что в принципе стабильные относительные частоты встречаемости в тексте того или иного языкового элемента могут быть не только индивидуальным признаком автора, но и показателем общих черт людей, пребывающих в определенном психическом или психофизиологическом состоянии, а также маркером уровня образованности, профессии, возраста и гендера. Иными словами, в современном судебном авторoведении становятся легитимными проблемы диагностического плана, а также определение гендерной принадлежности автора спорного текста посредством методов автоматической обработки текстов.

Материал и методы

Теоретические основы исследования

Теоретической базой предлагаемого исследования стал сложившийся на современном уровне междисциплинарного лингвистического знания комплекс идей о наличии собственно языковых, коммуникативных и психологических различий между речью мужчин и женщин [Крючкова, 1976; Енгалычев и др., 2001; Литвинова и др., 2014]. Эти различия не эксплицитны, но между тем они пронизывают все уровни языковой системы и отражаются в речевой практике [Горошко, 1999], в разных типах дискурсов и речевых жанрах [Викторова, 2011; Сеченова, 2012].

Кроме того, эмпирически выяснено, что указанные различия имеют не столько качественный, сколько количественный характер, а значит, они в принципе могут быть подвергнуты вероятностно-статистической процедуре [Ионова, Огорелков, 2020]. Любой текст характеризуется определенными статистическими закономерностями, которые измеряемы и вычислимы с достаточной степенью объективности, что позволяет применять математические методы, например, модели А.А. Маркова, для достижения требуемых результатов [Хмелев, 2000]. Сегодня в науке о языке получены достоверные корреляции между параметрами текста и характеристиками личности, в частности гендером. Так, информативными для диагностирования личности по гендеру были признаны такие параметры текста, как «количество знаменательных слов / количество незнаменательных слов», «количество имен существительных / всего слов», «количество незнаменательных слов / число существительных», «отношение местоимений / общее число слов»; «личные местоимения / всего слов» и пр. [Литвинова и др., 2014]. При этом акцент делается на квантитативную интерпретацию не столько лексических, сколько формально-грамматических элементов текста (соотношение слов разных частей речи, разных синтаксических моделей), потому что они в меньшей степени контролируются автором, но при этом являются облигаторными для выражения.

Современное развитие компьютерных технологий позволяет в значительной степени формализовать и автоматизировать полученные результаты при наличии теоретически непротиворечивых и методологически оправданных параметров анализа. Цель работы – интерпретация экспериментального исследования расширенных возможностей компьютерной программы авторизации текста «КАТ» (компьютерная авторизация текста) применительно к новым задачам определения гендерной принадлежности спорного текста, то есть разработка научно-практической платформы диагностического гендерного анализа текста для экспертных авторoведческих исследований.

Нами принята методика вероятностно-статистического исчисления и оценки относительных частот соотношения тех или иных языковых элементов, в результате чего подсчитываются коэффициенты корреляции и колебания параметров в разных выборках из текстовых баз данных. Набор исчисляемых параметров основан на коэффициентах Б.Н. Головина [Головин, 1970] и дополнен некоторыми другими принятыми в стилостатистике параметрами.

Материалом исследования являются созданные нами на основе Национального корпуса русского языка непрерывно пополняемые первичные базы художественных текстов XIX и XXI вв. и тексты, достоверно атрибутированные как написанные мужчинами или женщинами (соответственно, 4 базы: XIX век – женщины, XIX век – мужчины; XXI век – женщины, XXI век – мужчины). На данном этапе в каждой базе примерно по 150 текстов.

Принципы построения автоматизированной компьютерной программы гендерной диагностики текста и выбор алгоритмов

Исходные данные. Имеется компьютерная программа идентификации авторства текста по определенным параметрам (о начальном этапе ее разработки см.: [Юматов, Маркина, Ковалева, 2015; Юматов В.А., Маркина, Юматов С.В., 2016], о корректировании и апробации применительно к идентификации автора см.: [Радбиль, Маркина, 2019]). В настоящей работе освещаются результаты усовершенствования этой программы для установления гендера предполагаемого автора текста, которое сводится к поиску параметров, отражающих гендерный инвариант. После его установления определение гендерной принадлежности автора текста существенно упрощается.

Выбор группы параметров. В программе идентификации автора ранее была выделена группа параметров:

– отношение знаков препинания к общему количеству слов в тексте – число знаков препинания / число всех слов в тексте (1);

– средняя длина слова – число букв в слове / число всех слов в тексте (2);

– средняя длина предложения – число слов в предложении / число предложений в тексте (3);

– коэффициент предметности (Pr) – отношение суммы существительных и местоимений к сумме прилагательных и глаголов (4);

– коэффициент качественности (Qu) – отношение суммы прилагательных и наречий к сумме глаголов и существительных (5);

– коэффициент активности (Ac) – отношение суммы глаголов и глагольных форм к количеству слов в тексте (6);

– коэффициент динамизма (Din) – отношение суммы глаголов и глагольных форм к сумме существительных, прилагательных и местоимений (7);

– коэффициент связности текста (Con) – отношение суммы предлогов и союзов к числу предложений (8).

Всего используется 8 параметров [Радбиль, Маркина, 2019].

Компьютеризованная программа опирается на предварительно заданный набор характеристик, что существенно ограничивает надежность полученных результатов.

В основу предлагаемого алгоритма компьютеризованной модели положен корреляционный анализ – статистический метод, посредством которого изучается связь между явлениями на основе установления связей между случайными величинами.

Согласно концепции Н.Ш. Кремера, для изучения корреляционной связи данные о статистической зависимости целесообразно задавать в виде корреляционной таблицы или в виде двумерной выборки X (X_1, X_2, ..., X_n), Y (Y_1, Y_2, ..., Y_n). Для наглядности каждую пару можно представить в виде точки на координатной плоскости.

По оси абсцисс откладываются значения одного вариационного ряда X_i, а по оси ординат – другого Y_i. Такое изображение статистической зависимости называется полем корреляции или корреляционным полем точек. Оно создает общую картину корреляции [Кремер, 2007].

Математической мерой корреляции двух случайных величин служит ***коэффициент корреляции*** *r*. Линейный коэффициент корреляции (или коэффициент корреляции Пирсона), разработанный К. Пирсоном, Р. Уэлдоном и Ф. Эджуортом в 90-х гг. XIX в. (см.: [Кремер, 2007]), рассчитывается по формуле

$$r_{XY} = \frac{\mathrm{cov}_{XY}}{\sigma_X \sigma_Y} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sqrt{\sum (X - \overline{X})^2 \sum (Y - \overline{Y})^2}},$$

где $\overline{X} = \frac{1}{n}\sum_{t=1}^{n} X_t, Y = \frac{1}{n}\sum_{t=1}^{n} Y_t$ – средние значения выборок.

Алгоритм действия данной программы

Пусть имеются несколько баз данных художественных произведений, авторство которых известно (*A*, *B*, *C*), и текст *new* (new_1, new_2, ..., new_m)^T, авторство которого не известно. Необходимо определить, написан ли текст одним из известных писателей либо новым автором.

Представим базы произведений, авторство которых известно, в виде матриц, где *m* – количество обработанных текстов, *n* – количество их параметров:

$$\begin{matrix} a_{11}\, a_{21} \ldots a_{m1} \\ a_{12}\, a_{22} \ldots a_{m2} \\ \ldots \ldots \ldots \ldots \\ a_{1n}\, a_{2n} \ldots a_{mn} \end{matrix}$$

$$\begin{matrix} b_{11}\, b_{21} \ldots b_{m1} \\ b_{12}\, b_{22} \ldots b_{m2} \\ \ldots \ldots \ldots \ldots \\ b_{1n}\, b_{2n} \ldots b_{mn} \end{matrix}$$

$$\begin{matrix} c_{11}\, c_{21} \ldots c_{m1} \\ c_{12}\, c_{22} \ldots c_{m2} \\ \ldots \ldots \ldots \ldots \\ c_{1n}\, c_{2n} \ldots c_{mn} \end{matrix}$$

Новый текст представим в виде вектор-столбца *new* (new_1, new_2, ..., new_m)^T.

Для каждой матрицы посчитаем коэффициент корреляции каждого столбца с вектором *new* и получим три вектора *s* (s_1, s_2, ..., s_m); *p* (p_1, p_2, ..., p_m); *r* (r_1, r_2, ..., r_m).

Найдем среднее значение элементов каждого вектора. Получим $\widetilde{s}, \widetilde{p}, \widetilde{r}$, из этих значений составим вектор $k(\widetilde{s}, \widetilde{p}, \widetilde{r})$.

Конечным результатом работы алгоритма по определению авторства текста выступают вектор, длина которого равна количеству писателей, и значения, позволяющие установить авторство спорного текста.

Далее предстоит выяснить, подходят ли рассмотренные выше параметры, коэффициенты и методы их исчисления для определения гендера предполагаемого автора текста.

Результаты и обсуждение

Предварительная стадия реализации поставленной цели диагностики гендерной принадлежности автора текста

На предварительной стадии исследования мы, используя данные нашей работы [Радбиль, Маркина, 2019], установили 8 релевантных параметров, которые предполагалось проверить посредством программы:

– отношение всех знаков препинания к числу слов;
– средняя длина слов;
– средняя длина предложения;
– коэффициент предметности;
– коэффициент качественности;
– коэффициент активности;
– коэффициент динамизма;
– коэффициент связности.

Затем на основе баз Национального корпуса русского языка методом сплошной выборки были созданы две первичных базы художественных текстов (в одной – написанные мужчинами, в другой – женщинами). Все тексты были обработаны программой «КАТ» по 8 параметрам. Были определены усредненные коэффициенты для всех баз (см. таблицу).

Для каждого нового текста вычисляются 8 параметров и устанавливаются корреляции с усредненными значениями по каждой базе. Затем проверяются уже атрибутированные тексты авторов – мужчин и женщин – для выяснения возможности разграничить их по набору указанных параметров (см. рис. 1, 2).

Значения усредненных коэффициентов для баз текстов

Values of averaged coefficients for text bases

Параметры	Авторы	
	Женщины	Мужчины
Отношение всех знаков препинания к числу слов	0,222	0,254
Средняя длина слов	5,292	4,988
Средняя длина предложения	13,540	11,014
Коэффициент предметности	0,994	1,023
Коэффициент качественности	0,364	0,317
Коэффициент активности	0,157	0,165
Коэффициент динамизма	0,315	0,359
Коэффициент связности	4,535	3,974

Текст 1. Сергей Самаров. «Операция “Антитеррор”»

Название	Корелляция	Отношение всех знаков препинания к числу слов	Средняя длина слова	Средняя длина предложения	Коэффициент предметности	Коэффициент качественности	Коэффициент активности	Коэффициент динамизма	Коэффициент связности
ТЕКСТ	**1**	**0,256**	**5,513**	**7,977**	**1,073**	**0,416**	**0,123**	**0,232**	**2,795**
Мужчины	0,803	0,254	4,985	11,034	1,023	0,317	0,166	0,36	3,983
Женщины	0,794	0,222	5,292	13,541	0,995	0,364	0,158	0,315	4,536

Программа рекомендует данную базу как максимально близкую: **Мужчины**

Рис. 1. Значения коэффициентов для текста, написанного мужчиной

Fig. 1. Values of coefficients for a text written by a man

Текст 2. Галина Чернецкая. «Невеста Дракона»

Название	Корелляция	Отношение всех знаков препинания к числу слов	Средняя длина слова	Средняя длина предложения	Коэффициент предметности	Коэффициент качественности	Коэффициент активности	Коэффициент динамизма	Коэффициент связности
ТЕКСТ	**1**	**0,224**	**5,371**	**14,3**	**0,824**	**0,428**	**0,171**	**0,331**	**4,45**
Женщины	0,933	0,222	5,292	13,541	0,995	0,364	0,158	0,315	4,536
Мужчины	0,864	0,254	4,989	11,014	1,024	0,318	0,166	0,359	3,975

Программа рекомендует данную базу как максимально близкую: **Женщины**

Рис. 2. Значения коэффициентов для текста, написанного женщиной

Fig. 2. Values of coefficients for a text written by a woman

Далее были проведены эксперименты по анализу новых текстов и определению их отношения к усредненному показателю в базе. Это позволило уточнить значения параметров и выявить три наиболее релевантных из них для определения различий между текстами, написанными мужчинами и женщинами: (1) средняя длина слов: у женщин – 5,292, у мужчин – 4,988; (2) средняя длина предложения: у женщин – 13,540, у мужчин – 11,014; (3) коэффициент связности: у женщин – 4,535, у мужчин – 3,974.

Терминальная стадия реализации поставленной цели диагностики гендерной принадлежности автора текста

На основании результатов эмпирического анализа значительного массива текстов было сделано предположение, что указанные коэффициенты не являются абсолютными, но зависят от хронологического периода. В соответствии с ним необходимо было проверить следующее наблюдение: при переходе от текстов XIX в. к текстам XXI в. коэффициенты средней длины предложения приближаются друг к другу (у женщин уменьшаются, а у мужчин увеличиваются), то есть формально-грамматические различия между текстами авторов разной гендерной принадлежности сокращаются. Проверка этого наблюдения осуществлялась на материале 4 баз текстов: XIX век – женщины, XIX век – мужчины; XXI век – женщины, XXI век – мужчины (см. рис. 3–6 соответственно).

Значения коэффициентов для текстов, написанных женщиной, XIX в. (см. рис. 3):

– отношение всех знаков препинания к числу слов: 0,253;
– средняя длина слов: 5,472;
– средняя длина предложения: 13,744;
– коэффициент предметности: 1,602;
– коэффициент качественности: 0,286;
– коэффициент активности: 0,138;
– коэффициент динамизма: 0,261;
– коэффициент связности: 4,200.

Значения коэффициентов для текстов, написанных мужчиной, XIX в. (см. рис. 4):

– отношение всех знаков препинания к числу слов: 0,239;
– средняя длина слов: 5,161;
– средняя длина предложения: 15,062;
– коэффициент предметности: 0,954;
– коэффициент качественности: 0,377;
– коэффициент активности: 0,156;
– коэффициент динамизма: 0,324;
– коэффициент связности: 5,245.

Значения коэффициентов для текстов, написанных женщиной, XXI в. (см. рис. 5):

– отношение всех знаков препинания к числу слов: 0,243;
– средняя длина слов: 5,209;
– средняя длина предложения: 11,070;
– коэффициент предметности: 1,049;
– коэффициент качественности: 0,329;
– коэффициент активности: 0,160;
– коэффициент динамизма: 0,332;
– коэффициент связности: 3,789.

Значения коэффициентов для текстов, написанных мужчиной, XXI в. (см. рис. 6):

– отношение всех знаков препинания к числу слов: 0,231;
– средняя длина слов: 5,380;
– средняя длина предложения: 12,013;
– коэффициент предметности: 1,019;
– коэффициент качественности: 0,346;
– коэффициент активности: 0,162;
– коэффициент динамизма: 0,321;
– коэффициент связности: 3,851.

Данные результаты получены на первичных базах текстов, и их следует оценивать как промежуточные. Однако уже и на этом, весьма схематичном, уровне диагностики можно выдвинуть некоторые предположения в области их качественной интерпретации. Так, при переходе от текстов XIX к текстам XXI в. уменьшается средняя длина предложения, что свидетельствует об общей тенденции мужской и женской речи к упрощению синтаксиса в целом, к динамизму, к ускорению коммуникации в условиях временных ограничений на речевой акт. Кроме того, можно говорить о стирании различий между мужской и женской речью в современном мире, что, вероятно, отражает тенденции к стандартизации и унификации всех форм мыслительной, психической и вербальной активности людей в меняющихся условиях коммуникации.

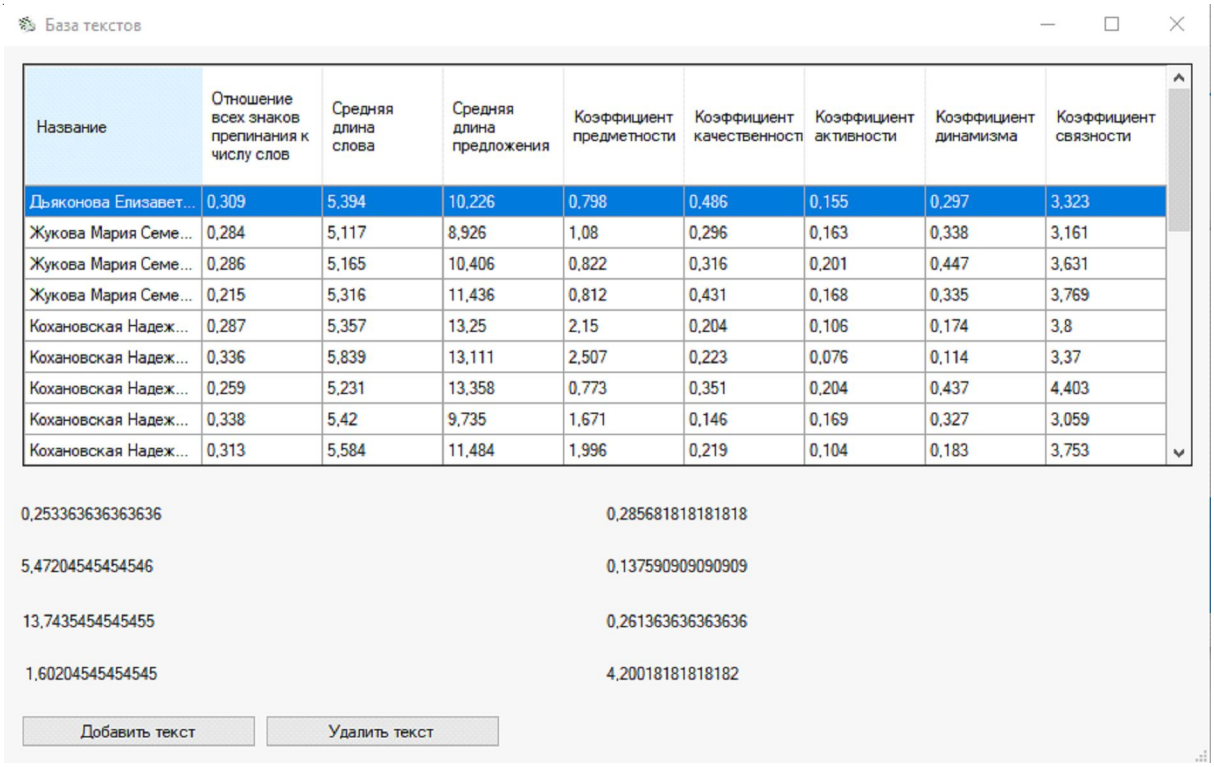

Рис. 3. Фрагмент базы текстов: XIX век – женщины

Fig. 3. Base of texts: the 19th century – women

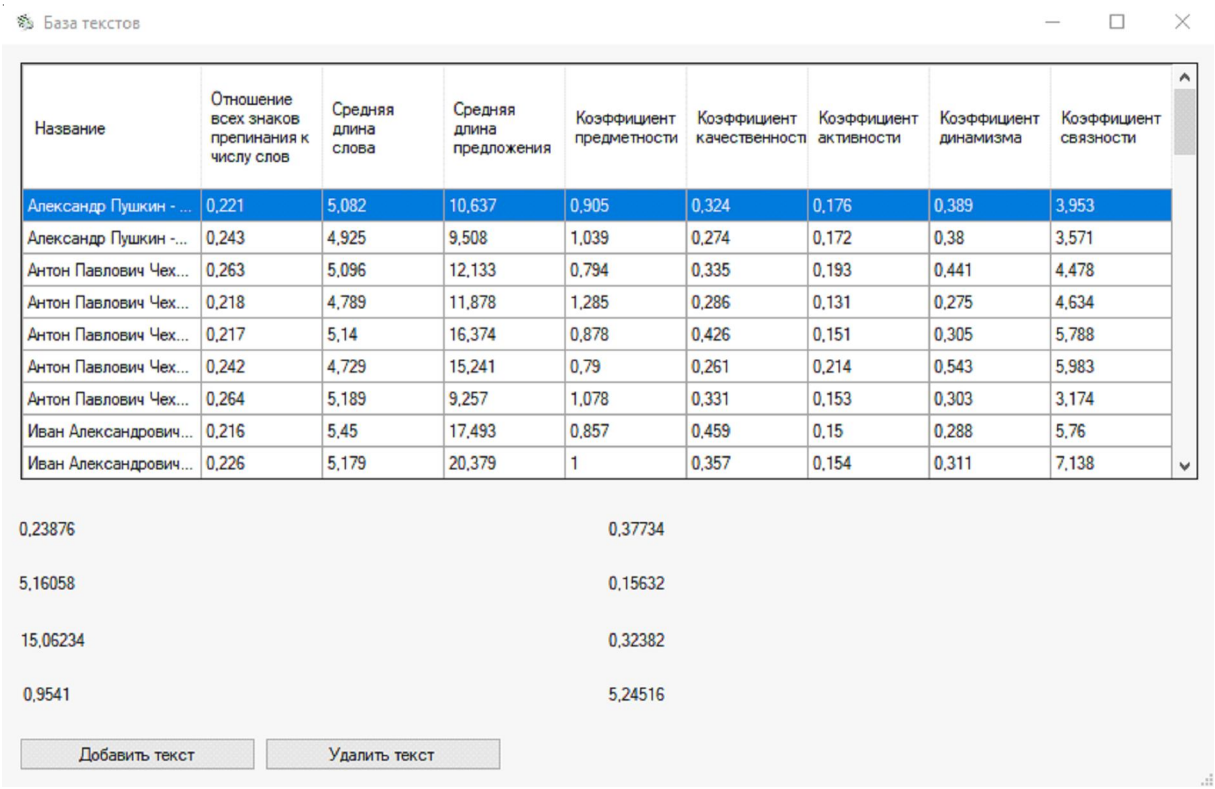

Рис. 4. Фрагмент базы текстов: XIX век – мужчины

Fig. 4. Base of texts: the 19th century – men

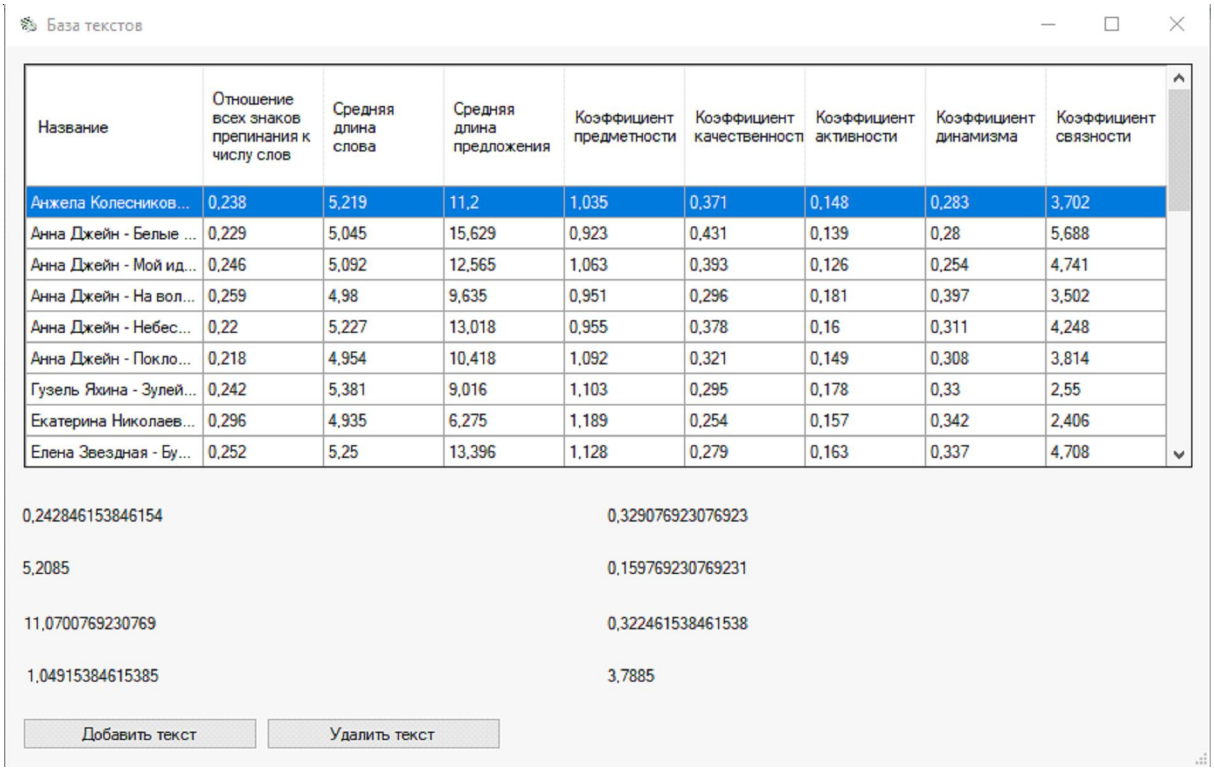

Рис. 5. Фрагмент базы текстов: XXI век – женщины

Fig. 5. Base of texts: the 21st century – women

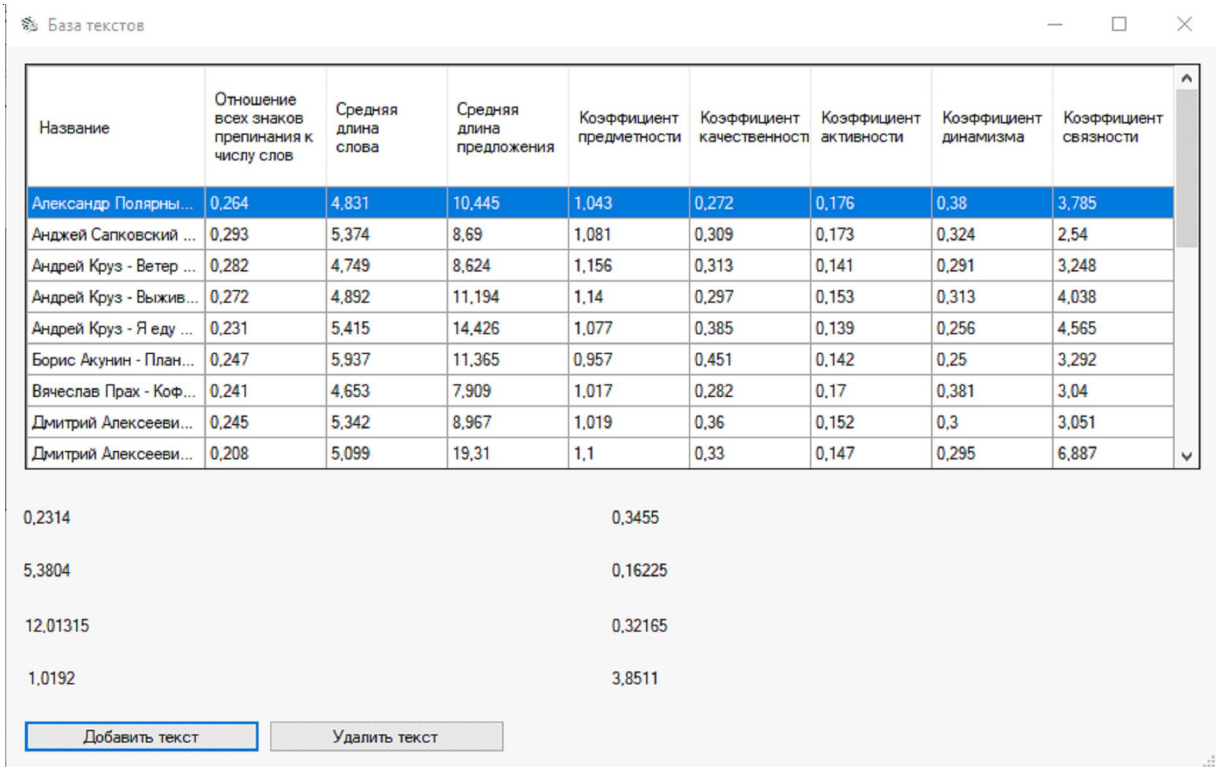

Рис. 6. Фрагмент базы текстов: XXI век – мужчины

Fig. 6. Base of texts: the 21st century – men

Выводы

Проведенное исследование выявило ряд проблем, которые требуют дальнейшей корректировки программы компьютерной диагностики текста по нескольким направлениям. Прежде всего необходимо предусмотреть дифференцирование пороговых значений коэффициентов применительно к текстам разных хронологических периодов с учетом уменьшения различий по параметрам между женской и мужской речью (уменьшить диапазон значимых расхождений для более современного периода).

Применительно к общим принципам работы системы отметим ее преимущества и недостатки. Преимущества заключаются в относительной простоте использования, интуитивной понятности параметров авторизации и логики их исчисления, а также в прозрачности качественной интерпретации результатов. Недостатком программы на данной стадии разработки является приблизительность в вычислениях коэффициентов корреляции. Кроме того, нуждается в дополнительной проверке на релевантность состав параметров – какие из параметров действительно необходимы, а какие имеют случайный характер. Остаются пока не решенными задачи установления необходимого и достаточного количества выборок текстов (какое количество минимально допустимое?), объема данных выборок (достаточен ли, например, объем 10 000 единиц?) и пр.

Так, на данном этапе программа правильно определяет гендерную принадлежность автора с точностью примерно 65 %. Для начальной стадии исследования это приемлемый результат, но для возможностей дальнейшего применения «КАТ» в судебно-авторовед-ческих экспертизах этого недостаточно. Необходимы более тонкие и точные расчеты, что напрямую зависит от качественного состава и объема баз данных текстов. Чем больше текстов в базах, тем точнее диагностика. Возможно, следует проверить и другие методы вычисления корреляций, например метод Спирмена.

В качестве перспектив усовершенствования программы предполагается осуществить спецификацию процесса определения коэффициентов по разным типам дискурса. Тогда с помощью программы «КАТ» можно будет полноценно решать экспертные задачи по диагностике гендерной принадлежности авторов спорных текстов художественного, медийного и политического дискурсов, юридической, официально-деловой и коммерческой документации.

СПИСОК ЛИТЕРАТУРЫ

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

Верзохин С. С., 2013. К вопросу о лингвотеоретических основах методик авторизации текста // Ученые записки Забайкальского государственного гуманитарно-педагогического университета им. Н.Г. Чернышевского. № 2 (49). С. 22–27.

Викторова Е. Ю., 2011. Влияет ли гендер на использование дискурсивов? : (На материале письменного научного дискурса) // Известия Саратовского университета. Новая серия. Серия: Филология. Журналистика. Вып. 3. С. 8–14.

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

Галяшина Е. И., Ермолова Е. И., 2005. Перспективы развития авторoведческой экспертизы в России // Судебная экспертиза. № 3. С. 5–11.

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

Горошко Е. И., 1999. Особенности мужского и женского стиля письма // Гендерный фактор в коммуникации : сб. науч. тр. Иваново : Иван. гос. ун-т. С. 28–41.

Енгалычев В. Ф., Белянин В. П., Константинова Е. С., Ощепкова Е. С., 2001. Психолингвистические особенности «мужского» и «женского» языков // Труды регионального конкурса научных проектов в области гуманитарных наук. Калуга : Эйдос. Вып. 2. С. 177–187.

Ионова С. В., Огорелков И. В., 2020. Речевая диагностика личности по гендерному признаку в авторoведении: квантитативный подход // Вестник Волгоградского государственного университета. Серия 2, Языкознание. Т. 19, № 1. С. 115–127. DOI: https://doi.org/10.15688/jvolsu2.2020.1.10.

Катышев П. А., Осадчий М. А., 2018. Метод параметрического моделирования в судебной лингвистике // Вестник Волгоградского государственного университета. Серия 2, Языкознание. Т. 17, № 3. С. 24–34. DOI: https://doi.org/10.15688/jvolsu2.2018.3.3.

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

Крючкова Т. Б., 1976. К вопросу о дифференциации языка по полу говорящего // Восточное языкознание : сб. тр. / отв. ред. В. П. Старинин. М. : Наука. С. 152–158.

Литвинова Т. А., 2012. Языковые корреляты личностных особенностей автора письменного текста: алгоритм исследования // В мире научных открытий. Серия: Проблемы науки и образования. № 9.3 (33). С. 236–255.

Литвинова Т. А., Загоровская О. В., Черванева В. А., Литвинова О. А., 2014. Проблема диагностирования пола автора письменного текста: фактор жанра // Современные исследования социальных проблем : электрон. науч. журн. № 1 (33). DOI: 10.12731/2218-7405-2014-1-4. URL: https://cyberleninka.ru/article/n/problema-diagnostirovaniya-pola-avtora-pismennogo-teksta-faktor-zhanra/viewer (дата обращения: 14.08.2020).

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

Морозов Н. А., 1916. Лингвистические спектры : Средство для отличения плагиатов от истинных произведений того или другого известного автора. Пг. : Тип. Императ. Акад. наук. 42 с. URL: http://www.textology.ru/library/book.aspx?bookId=1&textId=3 (дата обращения: 12.05.2020).

Радбиль Т. Б., 2014. Выявление содержательных и речевых признаков недобросовестной информации в экспертной деятельности лингвиста // Вестник Нижегородского университета им. Н.И. Лобачевского. № 6. С. 146–149.

Радбиль Т. Б. Маркина М. В., 2019. Вероятностно-статистические модели в производстве автороведческой экспертизы русскоязычных текстов // Политическая лингвистика. № 2 (74). С. 156–166.

Радбиль Т. Б., Юматов В. А., 2014. Способы выявления имплицитной информации в лингвистической экспертизе // Вестник Нижегородского университета им. Н.И. Лобачевского. № 3 (2). С. 18–21.

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

Сеченова Е. Г., 2012. Гендерная идентичность в фокусе современного научного дискурса // Вестник Тюменского государственного университета. Гуманитарные исследования. № 1. С. 86–91.

Хмелев Д. В., 2000. Распознавание автора текста с использованием цепей А.А. Маркова // Вестник Московского университета. Серия 9, Филология. № 2. С. 115–126.

Хоменко А. Ю., 2019. Лингвистическое атрибуционное исследование коротких письменных текстов: качественные и количественные методы // Политическая лингвистика. № 2 (74). С. 177–187. DOI: 10.26170/pl19-02-20.

Юматов В. А., Маркина М. В., Ковалева А. С., 2015. Программа криминалистической диагностики и авторизации текста «КАТ» // Вестник Костромского государственного университета им. Н.А. Некрасова. Т. 21, № 3. С. 199–202.

Юматов В. А., Маркина М. В., Юматов С. В., 2016. Математические методы криминалистической диагностики и авторизации текста в речеведческой экспертизе // Вестник Нижегородского университета им. Н.И. Лобачевского. № 5. С. 227–232.

Argamon Sh., Koppel M., Pennebaker J. W., Schler J., 2009. Profiling the Author of an Anonymous Text // Communication of the ACM. Vol. 52 (2). P. 119–123.

REFERENCES

Baranov A.N., 2001. *Vvedeniye v prikladnuyu lingvistiku* [Introduction to Applied Linguistics]. Moscow, Editorial URSS Publ. 347 p.

Verzokhin S.S., 2013. K voprosu o lingvoteoreticheskikh osnovakh metodik avtorizatsii teksta [On Issue of Linguistic and Theoretical Foundations of Authorship Attribution Methods]. *Uchonyye zapiski Zabaykal'skogo gosudarstvennogo gumanitarno-pedagogicheskogo universiteta im. N.G. Chernyshevskogo*, no. 2 (49), pp. 22-27.

Viktorova Ye.Yu., 2011. Vliyayet li gender na ispolzovaniye diskursivov?: (Na materiale pismennogo nauchnogo diskursa) [Does Gender Influence the Use of Discursive Words? (On the Material of the Written Scientific Discourse)]. *Izvestiya Saratovskogo universiteta. Novaya seriya. Seriya: Filologiya. Zhurnalistika*, iss. 3, pp. 8-14.

Vul S.M., 1977. *Teoreticheskiye i metodicheskiye voprosy kriminalisticheskogo issledovaniya pis'mennoy rechi* [Theoretical and Methodological Problems of Forensic Research of Written Speech]. Moscow, VNIISE. 109 p.

Galyashina Ye.I., Yermolova Ye.I., 2005. Perspektivy razvitiya avtorovedcheskoy ekspertizy v Rossii [Prospects for the Development of Authoring Expertise in Russia]. *Sudebnaya ekspertiza* [Forensic Examination], no 3, pp. 5-11.

Golovin B.N., 1970. *Yazyk i statistika* [Language and Statistics]. Moscow, Prosveshcheniye Publ. 190 p.

Goroshko Ye.I., 1999. Osobennosti muzhskogo i zhenskogo stilya pis'ma [Features of Masculine and Feminine Writing Styles]. *Gendernyy faktor v kommunikatsii: sb. nauch. tr.* [Gender Factor in Communication. Collection of Scientific Papers]. Ivanovo, Ivanovskiy gosudarstvennyy univeritet, pp. 28-41.

Yengalychev V.F., Belyanin V.P., Konstantinova Ye.S., Oshchepkova Ye.S., 2001. Psikholingvisticheskiye osobennosti «muzhskogo» i «zhenskogo» yazykov [Psycholinguistic Features of "Male" and "Female" Languages]. *Trudy regional'nogo konkursa nauchnykh proyektov v oblasti gumanitarnykh nauk* [Proceedings of the Regional Competition of Scientific Projects in the Humanities]. Kaluga, Eydos Publ. Iss. 2, pp. 177-187.

Ionova S.V., Ogorelkov I.V., 2020. Rechevaya diagnostika lichnosti po gendernomu priznaku v avtorovedenii: kvantitativnyy podkhod [Personality Speech Diagnostics in Author Identification Based on Gender Parameter: Quantitative Approach]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznaniye* [Science Journal of Volgograd State University. Linguistics], vol. 19, no. 1, pp. 115-127. DOI: https://doi.org/10.15688/jvolsu2.2020.1.10.

Katyshev P.A., Osadchiy M.A., 2018. Metod parametricheskogo modelirovaniya v sudebnoy lingvistike [Method of Parametric Modeling in Forensic Linguistics]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznaniye* [Science Journal of Volgograd State University. Linguistics], vol. 17, no. 3, pp. 24-34. DOI: https://doi.org/10.15688/jvolsu2.2018.3.3.

Kremer N.Sh., 2007. *Teoriya veroyatnostey i matematicheskaya statistika* [Theory of Probability and Mathematical Statistics]. Moscow, YUNITI-DANA Publ. 543 p.

Kryuchkova T.B., 1976. K voprosu o differentsiatsii yazyka po polu govoryashchego [On the Issue of Language Differentiation by Gender of the Speaker]. Starinin V.P., ed. *Vostochnoye yazykoznaniye* [Eastern Linguistics]. Moscow, Nauka Publ., pp. 152-158.

Litvinova T.A. 2012. Yazykovyye korrelyaty lichnostnykh osobennostey avtora pismennogo teksta: algoritm issledovaniya [Language Correlates of Personal Characteristics of the Author of a Written Text: Research Algorithm]. *V mire nauchnykh otkrytiy. Seriya: Problemy nauki i obrazovaniya* [In the World of Scientific Discoveries. Series: Problems of Science and Education], no. 9.3 (33), pp. 236-255.

Litvinova T.A., Zagorovskaya O.V., Chervaneva V.A., Litvinova O.A., 2014. Problema diagnostirovaniya pola avtora pismennogo teksta: faktor zhanra [The Problem of Diagnosing the Gender of the Author of a Written Text: The Genre Factor]. *Sovremennyye issledovaniya sotsialnykh problem: elektron. nauch. zhurn.* [Modern Studies of Social Problems. Internet Scientific Journal], no. 1 (33). DOI: 10.12731/2218-7405-2014-1-4. URL: https://cyberleninka.ru/article/n/problema-diagnostirovaniya-pola-avtora-pismennogo-teksta-faktor-zhanra/viewer (accessed 14 August 2020).

Litvinova T.A., Gromova A.V., 2020. Kompyuternyye tekhnologii v sudebnoy avtorovedcheskoy ekspertize: problemy i perspektivy ispol'zovaniya [Current Problems of Forensic Authorship Analysis and the Possibility of Their Solution with the Use of Computer Methods: Problems and Prospects]. *Vestnik Volgogradskogo gosudarstvennogo universiteta. Seriya 2. Yazykoznaniye* [Science Journal of Volgograd State University. Linguistics], vol. 19, no. 1. pp. 77-88. DOI: https://doi.org/10.15688/jvolsu2.2020.1.7.

Morozov N.A., 1916. *Lingvisticheskiye spektry: Sredstvo dlya otlicheniya plagiatov ot istin. proizvedeniy togo ili drugogo izvestnogo avtora* [Linguistic Spectra: A Tool for Distinguishing Plagiarism from Authentic Works of One or Another Famous Author]. Petrograd, Typographiya Imperatorskoy Akademii nauk Publ. 42 p. URL: http://www.textology.ru/library/ (accessed 12 May 2020).

Radbil T.B., 2014. Vyyavleniye soderzhatel'nykh i rechevykh priznakov nedobrosovestnoy informatsii v ekspertnoy deyatel'nosti lingvista [Identification of Content and Speech Signs of Unfair Information in the Expert Activity of a Linguist]. *Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo* [Vestnik of Lobachevsky State University of Nizhni Novgorod], no. 6, pp. 146-149.

Radbil T.B., Markina M.V, 2019. Veroyatnostno-statisticheskiye modeli v proizvodstve avtorovedcheskoy ekspertizy russkoyazychnykh tekstov [Probabilistic-Statistical Models in Conducting Authoring Expertise of Russian Texts]. *Politicheskaya lingvistika* [Political Linguistics], no. 2 (74), pp. 156-166.

Radbil T.B., Yumatov V.A., 2014. Sposoby vyyavleniya implitsitnoy informatsii v

lingvisticheskoy ekspertize [Methods for Identifying Implicit Information in Linguistic Expertise]. *Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo* [Vestnik of Lobachevsky State University of Nizhni Novgorod], no. 3 (2), pp. 18-21.

Romanov A.S., Meshcheryakov R.V., 2009. Identifikatsiya avtora teksta s pomoshch'yu apparata opornykh vektorov [Identification of the Author of the Text Using the Support Vector Apparatus]. *Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: po materialam yezhegod. Mezhdunar. konf. «Dialog 2009» (Bekasovo, 27–31 maya 2009 g.)* [Computational Linguistics and Intellectual Technologies. On the Materials of the Annual Intern. Conf. "Dialogue 2009" (Bekasovo, May 27–31, 2009)]. Moscow, RGGU. Iss. 8 (15), pp. 432-437.

Sechenova Ye.G., 2012. Gendernaya identichnost v fokuse sovremennogo nauchnogo diskursa [Gender Identity in the Focus of Modern Scientific Discourse]. *Vestnik Tyumenskogo gosudarstvennogo universiteta. Gumanitarnye issledovaniya* [Tyumen State University Herald. Humanities Research. Humanitates], no. 1, pp. 86-91.

Khmelev D.V., 2000. Raspoznavaniye avtora teksta s ispolzovaniyem tsepey A.A. Markova [Recognition of the Author of the Text Using A.A. Markov's Chains]. *Vestnik Moskovskogo universiteta. Seriya 9. Filologiya* [Moscow State University Bulletin. Series 9. Philology], no. 2, pp. 115-126.

Khomenko A.Yu., 2019. Lingvisticheskoye atributsionnoye issledovaniye korotkikh pismennykh tekstov: kachestvennyye i kolichestvennyye metody [Linguistic Attribution of Research Short Written Texts: Qualitative and Quantitative Methods]. *Politicheskaya lingvistika* [Political Linguistics], no. 2 (74), pp. 177-187. DOI: 10.26170/pl19-02-20.

Yumatov V.A., Markina M.V., Kovaleva A.S., 2015. Programma kriminalisticheskoy diagnostiki i avtorizatsii teksta «KAT» [The Program of Forensic Diagnostics and Text Authorization "CTA"]. *Vestnik Kostromskogo gosudarstvennogo universiteta im. N.A. Nekrasova* [Vestnik of Kostroma State University], vol. 21, no. 3, pp. 199-202.

Yumatov V.A., Markina M.V., Yumatov S.A., 2016. Matematicheskiye metody kriminalisticheskoy diagnostiki i avtorizatsii teksta v rechevedcheskoy ekspertize [Mathematical Methods of Forensic Diagnostics and Text Authorization in Speech Expertise]. *Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo* [Vestnik of Lobachevsky State University of Nizhni Novgorod], no. 5, pp. 227-232.

Argamon Sh., Koppel M., Pennebaker J.W., Schler J., 2009. Profiling the Author of an Anonymous Text. *Communication of the ACM*, vol. 52 (2), pp. 119-123.

Information About the Authors

Timur B. Radbil, Doctor of Sciences (Philology), Professor, Head of the Department of Theoretical and Applied Linguistics, Lobachevsky State University of Nizhny Novgorod, Prosp. Gagarina, 23, 603950 Nizhny Novgorod, Russia, timur@radbil.ru, ResearcherID: AAO-6983-2020, ScopusID: 57210390493, https://orcid.org/0000-0002-7516-6705

Marina V. Markina, Candidate of Sciences (Physics and Mathematics), Associated Professor, Department of Theoretical, Computer and Experimental Mechanics, Lobachevsky State University of Nizhny Novgorod, Prosp. Gagarina, 23, 603950 Nizhny Novgorod, Russia, markinamv6213@yandex.ru, https://orcid.org/0000-0002-1042-8006

Информация об авторах

Тимур Беньюминович Радбиль, доктор филологических наук, профессор, заведующий кафедрой теоретической и прикладной лингвистики, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, просп. Гагарина, 23, 603950 г. Нижний Новгород, Россия, timur@radbil.ru, ResearcherID: AAO-6983-2020, ScopusID: 57210390493, https://orcid.org/0000-0002-7516-6705

Марина Викторовна Маркина, кандидат физико-математических наук, доцент кафедры теоретической, компьютерной и экспериментальной механики, Национальный исследовательский Нижегородский государственный университет им. Н.И. Лобачевского, просп. Гагарина, 23, 603950 г. Нижний Новгород, Россия, markinamv6213@yandex.ru, https://orcid.org/0000-0002-1042-8006