https://doi.org/10.30764/1819-2785-2018-13-3-47-52

Новый подход к сравнительному анализу текстов в рамках комплексной компьютерно-технической и лингвистической экспертизы

А.А. Бойцов[1,2], А.А. Кулешова[1,3], А.М. Лизоркин[1]

[1] Федеральное бюджетное учреждение Российский федеральный центр судебной экспертизы при Министерстве юстиции Российской Федерации, Москва 109028, Россия

[2] ГАОУ ВО города Москвы «Московский городской педагогический университет», Москва 129226, Россия

[3] Национальный исследовательский университет «Высшая школа экономики», Москва 101000, Россия

Аннотация: Рассмотрен методический подход по исследованию текстов больших объемов на предмет установления их сходства/различия. Подход предполагает взаимодействие эксперта-лингвиста и эксперта, обладающего специальными знаниями в области информационных компьютерных средств, позволяет оптимизировать экспертную работу и сформулировать максимально объективные выводы.

Ключевые слова: *сравнение текстов, лингвистическая экспертиза, компьютерно-техническая экспертиза, плагиат, расстояние Левенштейна, оптимизация экспертной работы*

Для цитирования: Бойцов А.А., Кулешова А.А., Лизоркин А.М. Новый подход к сравнительному анализу текстов в рамках комплексной компьютерно-технической и лингвистической экспертизы // Теория и практика судебной экспертизы. 2018. Том 13. № 3. С. 47–52. https://doi.org/10.30764/1819-2785-2018-13-3-47-52

A New Approach to Forensic Text Comparison in the Framework of Integrated Computer Science and Linguistics Investigation

Aleksei A. Boitsov[1,2], Aleksandra A. Kuleshova[1,3], Aleksei M. Lizorkin[1]

[1] The Russian Federal Centre of Forensic Science of the Ministry of Justice of the Russian Federation, Moscow 109028, Russia

[2] Moscow City Pedagogical University, Moscow 129226, Russia

[3] National Research University Higher School of Economics, Moscow 101000, Russia

Abstract. The article presents a new methodological approach to forensic comparison of large amounts of text to establish their similarity/distinctiveness. The proposed approach assumes interaction between a forensic linguist and a forensic examiner with specialized knowledge in the field of forensic computer science, which the effect of optimizing forensic workflow and maximizing objectivity of the resulting expert conclusions.

Keywords: *text comparison, forensic linguistics, forensic computer science, plagiarism, Levenshtein distance, optimization of forensic workflow*

For citation: Boitsov A.A., Kuleshova A.A., Lizorkin A.M. A New Approach to Forensic Text Comparison in the Framework of Integrated Computer Science and Linguistics Investigation. *Theory and Practice of Forensic Science*. 2018. Vol. 13. No. 3. P. 47–52 (In Russ.). https://doi.org/10.30764/1819-2785-2018-13-3-47-52

В настоящее время в экспертной практике по уголовным и гражданским делам все чаще возникает необходимость сравнения больших объемов текстов на предмет установления наличия/отсутствия сходных фрагментов, в том числе по делам о мошенничестве, коррупции, плагиате и защите авторских прав [1]. Указанная задача является одной из частных задач судебной лингвистической экспертизы. Нередко объ-

ектами исследования становятся тексты объемом более 10 000 страниц (более 100 текстовых файлов). Исследование их невозможно произвести качественно и в разумные сроки в рамках одной лишь судебной лингвистической экспертизы. Одним из возможных путей решения этой проблемы является привлечение к производству экспертизы эксперта в области информационных компьютерных технологий. Возможности компьютерно-технической экспертизы позволяют качественно и в короткие сроки проанализировать фрагменты текста и предоставить эксперту-лингвисту материал для исследования в форме обобщенного статистического отчета по совпадениям, работа с которым значительно сократит временны*е* затраты эксперта-лингвиста [2, 3].

Для производства экспертизы на предмет установления наличия/отсутствия сходных фрагментов часто предоставляются файл с текстом, требующим проверки, и файлы-образцы с текстами, с которыми необходимо сравнить первый текст. В случае если на экспертизу предоставляются объекты в традиционном виде – на бумажном носителе, то для компьютерно-технической экспертизы их предварительно сканируют и распознают текст. Распознают и тексты, которые представлены в электронном виде, но не в текстовом формате (например, .doc, .txt), а в графическом (например, .jpg, .pdf) [3, 4]. Следует отметить, что необходима максимальная внимательность при проверке полученных результатов, поскольку нередки сбои и неточности.

В этой связи экспертами лаборатории судебной лингвистической экспертизы (ЛСЛЭ) и лаборатории судебной компьютерно-технической экспертизы (ЛСКТЭ) ФБУ РФЦСЭ при Минюсте России был совместно разработан следующий методический подход к решению подобных задач.

Для сравнения больших объемов текстов предлагается использовать локальный метод проверки текста по сегментам – абзацам. Абзац – отрезок письменного или печатного текста от одной красной строки до другой, обычно заключающий в себе сверхфразовое единство или его часть [5, с. 10]. Абзацы анализируются на особенности совпадения фрагментов текста – *X* и более последовательно расположенных одинаковых слов без каких-либо дополнительных или отсутствующих слов в одном из сравниваемых текстов. Где *X* – средняя длина абзацев в представленных на исследование массивах текстов. В целях повышения представительности выборки и уменьшения потенциальной погрешности проверяемый массив абзацев можно расширить за счет включения абзацев меньшей, чем *X* – 1 длины. При этом такое условное округление предлагается делать всегда в меньшую сторону, например с 11,6 до 10.

Под совпадающими фрагментами текста предлагаем понимать:

1) полностью совпадающие фрагменты текста (100%-ное совпадение) – 10 и более последовательно расположенных одинаковых словоформ, занимающих одинаковую синтаксическую позицию, без каких-либо дополнительных или отсутствующих слов в одном из сравниваемых текстов;

2) частично (процент совпадения не менее Y %[1]) совпадающие фрагменты текста – *X* – 1 и более последовательно расположенных слов, со следующими изменениями в одном из текстов:

– слова заменены на синонимы[2] (например: «от 1 *декабря* 2014 г.» – «от 01.*12*.2014 г.», «система *электропитания* должна обеспечивать» – «система *электроснабжения* должна обеспечивать», «по созданию *системы*» – «по созданию *комплекса*»; «влияние эффекта масштаба регионов со *сверхпоказателями* нивелировано» – «влияние эффекта масштаба регионов со *сверхрасходами* нивелировано»);

– удалены/внесены знаки препинания, нумерация (цифр) (например: «**-** участвует в разработке проектов» – «*4.7.* участвует в разработке проектов»);

– изменено написание слов (например: «*научнотехнической*» – «*научно-технической*»; «*ковариацион-ной* матрицы» – «*ковариационной* матрицы»);

– глаголы заменены на отглагольные существительные (например: «*проводит* работу по вычислению» – «*проведение* работы по вычислению»);

– изменена грамматическая форма (например: «*применяют* следующие обозначения» – «*применялись* следующие обозначения», «печать результатов опросов в табличной и графической *форме*» – «печать

[1] Данная величина также является условной и рассчитывается эмпирически исходя из лингвистического анализа выборки, полученной в ходе компьютерной обработки. По результатам лингвистического анализа в число частично совпадающих фрагментов текста предлагается не включать те фрагменты, в которых имеются значительные изменения, затрагивающие семантику предложений [6].

[2] В том числе замены равнозначных обозначений.

результатов опросов в табличной и графической *формах*»);

– изменен порядок слов (например: «организациями в области обеспечения *экономической и финансовой* безопасности» – «организациями в области обеспечения *финансовой и экономической* безопасности»).

Непосредственно анализ текстов предлагается проводить в два этапа.

Первый этап – компьютерно-техническое исследование, в рамках которого осуществляется автоматизированный поиск совпадающих или частично совпадающих фрагментов. Автоматизированный поиск может быть произведен с помощью освоенного экспертом языка программирования; эксперты ФБУ РФЦСЭ при Минюсте России при решении подобных задач используют язык программирования C# в среде разработки Microsoft Visual Studio Community [2].

При автоматизированном поиске выполняется следующий алгоритм.

– Из сравниваемых файлов извлекается текст. Текст, представленный в виде файлов в графическом формате (например, сканированные документы), распознается. Остальной текст приводится к единой кодировке[3]. В результате вся имеющаяся в сравниваемых документах текстовая информация приводится к двум единообразным текстам.

– В полученных текстах находятся все абзацы[4] с длиной слов *X* – 1 и более.

– Такие абзацы сравниваются по принципу «каждый с каждым». Результат сравнения – числовой параметр, определяющий степень совпадения сравниваемых абзацев. Алгоритм расчета числового параметра приведен ниже.

– Формируются отчеты по выходным формам, в том числе с расчетом статистических показателей, необходимых для второго этапа (см. ниже).

Расчет числового параметра степени совпадения сравниваемых абзацев производится на основании вычисления расстояния Левенштейна[5]. Так как расстояние Левенштейна не может быть больше, чем количество слов в наибольшем (по количеству слов) из сравниваемых абзацев, то степень совпадения абзацев может быть вычислена по формуле:

$$D(a,b) = \left(1 - Lev(a,b) \Big/ \max\left(L(a), L(b)\right)\right),$$

где *Lev(a, b)* – расстояние Левенштейна между строками *a* и *b*,

L(a) и *L(b)* – длины (в количестве слов) строк *a* и *b* соответственно.

Такой подход позволяет определить степень схожести строк, отличия которых выражены заменой слов по месту (например, подстановкой синонимов), однако не учитывает возможность изменения порядка слов без изменения смысла. Чтобы учесть порядок слов, в расчет числового параметра степени совпадения абзацев включена степень схожести строк *a'* и *b'*, состоящих из набора слов строк *a* и *b* соответственно, выстроенных в алфавитном порядке с исключенными дубликатами. Поскольку вес расстояния Левенштейна от оригинальных строк должен быть выше, используется квадрат степени схожести строк *a'* и *b'*. Итоговая формула расчета процента совпадения:

$$D'(a,b) = \frac{k_1 \cdot \left(D(a',b')\right)^2 + k_2 \cdot D(a,b)}{k_1 + k_2} \times 100\,\%$$

Коэффициенты k_1 и k_2 подобраны эмпирически (на основании репрезентативных выборок из результатов сравнительных анализов текстов небольших объемов по ранее проведенным экспертным исследованиям на базе ЛСЛЭ и ЛСКТЭ) и равны 4 и 2 соответственно.

Второй этап – лингвистическое исследование, в рамках которого осуществляется непосредственно анализ найденных в ходе первого этапа совпадающих абзацев, а именно категоризация найденных абзацев по семантическим и формальным характеристикам [6]. Пример такой категоризации и его наглядное оформление приводим в таблице 1).

Полученные результаты ввиду их большого количества предлагается сводить

[3] На носителях информации содержатся байты. Для передачи текстовой информации имеется ряд международных договоренностей о соответствии каждого байта (или нескольких байт) определенной букве определенного алфавита. Такие договоренности называются кодировками (каждая буква кодируется байтом или набором байт) [2].

[4] Здесь и далее абзацы преобразуются в применимый для компьютерного анализа вид, при котором заглавные буквы и знаки препинания не учитывались.

[5] Расстояние Левенштейна – это минимальное количество операций вставки, удаления и замены одного символа на другой, необходимых для превращения одной строки в другую. Под «символом» понимается любой символ некоего «алфавита». В используемом методе сравнения в качестве «алфавита» выступает весь набор слов, имеющихся в обоих сравниваемых абзацах. Соответственно, каждое слово представляет собой сравниваемый символ.

Таблица 1. Категории абзацев
Table 1. *Categories of paragraphs*

Категория	Формат ячейки в соответствующих таблицах
Нумерация	текст
Названия организаций, учреждений и т. п.	текст
Названия документов, нормативных актов и т. п.	текст
Описание полномочий	текст
Ссылки на использованные источники: нормативные акты, научные труды и т. п.	текст
Описание реквизитов документов	текст
Описание правил, процедур, порядка действий	текст
Готовые речевые формулы (речевые клише), характерные для нормативно-правового дискурса и текстов данной тематики	текст

в таблицы MS Excel и прикладывать к заключению в качестве приложения (на оптическом диске или USB-накопителе). В таблицах предлагается указывать общее количество абзацев и процент покрытия для каждого файла (например, таблица 2 «Общее»), покрытия каждого проверяемого файла каждым проверочным файлом (например, таблица 3 «По файлам»), а также фразы (абзацы), размеченные соответствующими форматами ячеек, с указанием количества соответствующих фраз в проверяемом и проверочном массивах текстов, максимального и среднего[6] процента совпадения (например, таблица 4 «Фразы»). Некатегоризируемые фразы для наглядности предлагается включать в указанные таблицы №№ 2–4 без форматирования – без выделения цветом.

Далее предлагается проиллюстрировать проведенный сравнительный анализ текстов и дать оценку их уникальности следующим образом:

– Уникальность одного текста по отношению к другому оценивается и указывается в процентном соотношении, например: *уникальность текста XXX составляет 99,59%. Общий средний процент совпадений текста XXX с предоставленными на исследование текстами УУУ составляет 0,439 %.*

– Привести фрагменты текста, имеющие дословное совпадение.

– Привести фрагменты текста, имеющие частичное совпадение.

– Проанализировать и категоризацию совпадений (в случае большого количества совпадений в целях оптимизации объема текста заключения приводятся примеры таких совпадений с указанием суммарного количества фраз, относящихся к той или иной категории, в проверяемых файлах, а полностью все совпадения указываются в приложении на оптическом диске или флеш-накопителе), например: *в сравниваемых текстах совпадают фрагменты, которые можно отнести к следующим категориям:*

[6] Например, если фраза из проверяемого текста встречается 2 раза в проверочном тексте и в одном случае совпадает с проверяемой на 96 %, а во втором – на 94 %, то максимальный процент совпадения – 96 %, средний – 95 % (вычисляется как среднее арифметическое).

– нумерация – 10 совпадений, например: «1 2 3 4 5»;

– названия организаций, учреждений и т. п. – 20 совпадений, например: «министерство Российской Федерации», «Федеральная служба по надзору»;

– названия документов, нормативных актов и т. п. – 200 совпадений, например: «указ президента»; «постановление правительства»; «федеральный закон»;

– описание полномочий – 55 совпадений, например: «осуществление комплекса мероприятий по нейтрализации»; «внедрение государственной информационной системы»; «совершенствование нормативного правового регулирования».

Таким образом, считаем возможным использовать предлагаемую модель при проведении исследований текстов больших объемов на предмет их совпадения в рамках комплексной компьютерно-технической и лингвистической экспертизы. Предложенный методический подход позволит оптимизировать работу эксперта-лингвиста, сократить сроки производства экспертизы подобных объектов, а привлечение эксперта компьютерно-технической лаборатории позволит добиться точных и максимально объективных выводов. Предлагаемый методический подход уже сейчас показывает положительные результаты в совместной экспертной практике структурных подразделений ФБУ РФЦСЭ при Минюсте России (ЛСЛЭ и ЛСКТЭ).

Необходимо также отметить, что компьютерные технологии в лингвистической экспертизе могут использоваться не только для решения задач по обработке больших массивов данных и исследованию этих массивов, но и задач, связанных со сравнением текстов в целом. Авторы планируют развивать и дополнять описанный методический подход, в частности в настоящий момент разрабатываются подходы по оптимизации комплексных компьютерно-технических и лингвистических экспертиз, а именно:

1) разрабатывается механизм автоматизированного установления изменений словоформ в сравниваемых абзацах (механизм автоматизированной лемматизации)[7];

2) разрабатывается механизм автоматизированного определения синонимических замен с использованием электронных словарей синонимов (в том числе путем индексации имеющихся печатных изданий);

3) изучаются возможности комбинирования предложенных выше механизмов для достижения наиболее оптимальных и объективных результатов исследования.

Развитие предложенного подхода представляется авторам перспективным не только в практическом плане (для производства экспертиз текстов больших объемов), но и в целом для судебной экспертологии как науки.

[7] Для этого планируется использовать грамматические словари русского языка (проиндексировать указанные в словарях словоформы, для того чтобы иметь возможность, например, привести все слова в сравниваемых абзацах к единой форме: все существительные – в именительном падеже и единственном числе, все глаголы – в инфинитиве и др.). Такое решение позволит рассчитывать расстояния Левенштейна (с соответствующими коэффициентами) для сравниваемых абзацев в приведенных формах – с учетом и без учета порядка слов.

СПИСОК ЛИТЕРАТУРЫ

1. Баранов А.Н. Лингвистическая экспертиза текста. Теоретические основания и практика: учебное пособие. М.: Флинта: Наука, 2007. 592 с.
2. Эджубов Л.Г., Карпухина Е.С., Усов А.И., Хатунцев Н.А. Производство судебной компьютерно-технической экспертизы. I. Общая часть. II. Диагностические и идентификационные исследования аппаратных средств. Методическое пособие / Под ред. А.И. Усова. М.: РФЦСЭ, 2009. 117 с.
3. Эджубов Л.Г., Карпухина Е.С., Усов А.И., Хатунцев Н.А., Нехорошев А.Б., Шухнин М.Н., Яковлев А.Н., Юрин И.Ю. Производство судебной компьютерно-технической экспертизы. (IV. Актуальные комплексные экспертные задачи). Методическое пособие / Под ред. А.И. Усова. М.: РФЦСЭ, 2011. 295 с.

REFERENCES

1. Baranov A.N. *Linguistic examination of texts. Theoretical foundations and practices: a manual*. Moscow: Flinta: Nauka, 2007. 592 p. (In Russ.)
2. Edzhubov L.G., Karpukhina E.S., Usov A.I., Khatuntsev N.A. *The computer forensics process. I. The general part. II. Hardware examinations for diagnostic and identification purposes. Methodology guide*. Moscow: RFCFS, 2009. 117 p. (In Russ.)
3. Edzhubov L.G., Karpukhina E.S., Usov A.I., Khatuntsev N.A., Nekhoroshev A.B., Shukhnin M.N., Yakovlev A.N., Yurin I.Yu. *The computer forensics process. (IV. Currently relevant complex forensic objectives). Methodology guide.* Moscow: RFCFS, 2011. 295 p. (In Russ.)

4. Эджубов Л.Г., Карпухина Е.С., Усов А.И., Хатунцев Н.А., Комраков И.Л., Костин П.В., Шаров В.И., Нехорошев А.Б., Шухнин М.Н., Яковлев А.Н., Юрин И.Ю. Производство судебной компьютерно-технической экспертизы. (V. Актуальные задачи исследования компьютерной информации). Методическое пособие / Под ред. А.И. Усова. М.: РФЦСЭ, 2011. 270 с.
5. Розенталь Д.Э., Теленкова М.А. Словарь-справочник лингвистических терминов. М.: Астрель, АСТ, 2001. 624 с.
6. Кобозева И.М. Лингвистическая семантика. М.: Эдиториал УРСС, 2000. 352 с.

4. Edzhubov L.G., Karpukhina E.S., Usov A.I., Khatuntsev N.A., Komrakov I.L., Kostin P.V., Sharov V.I., Nekhoroshev A.B., Shukhnin M.N., Yakovlev A.N., Yurin I.Yu. *The computer forensics process (V. Current issues in digital evidence examination). Methodology guide.* Moscow: RFCFS, 2011. 270 p. (In Russ.)
5. Rozental' D.E., Telenkova M.A. *A reference book of linguistic terms.* Moscow: Astrel', AST, 2001. 624 p. (In Russ.)
6. Kobozeva I.M. *Linguistic semantics*. Moscow: Editorial URSS, 2000. 352 p. (In Russ.)

ИНФОРМАЦИЯ ОБ АВТОРАХ

Бойцов Алексей Александрович – заместитель заведующего лабораторией судебной лингвистической экспертизы ФБУ РФЦСЭ при Минюсте России, магистрант кафедры теории и истории государства и права Юридического института Московского городского педагогического университета; e-mail: alexey.boytsov@yandex.ru.

Кулешова Александра Андреевна – государственный судебный эксперт лаборатории судебной лингвистической экспертизы ФБУ РФЦСЭ при Минюсте России, магистрант кафедры теории и истории права Национального исследовательского университета «Высшая школа экономики»; e-mail: kuleshov0061@gmail.com.

Лизоркин Алексей Михайлович – ведущий государственный судебный эксперт лаборатории судебной компьютерно-технической экспертизы ФБУ РФЦСЭ при Минюсте России; e-mail: Liz4@yandex.ru.

ABOUT THE AUTHORS

Boitsov Aleksei Aleksandrovich – Deputy Head of the Laboratory of Forensic Linguistics, the Russian Federal Centre of Forensic Science of the Russian Ministry of Justice, Master's Student of the Department of Theory and History of State and Law of the Law Institute of the Moscow City Pedagogical University; e-mail: alexey.boytsov@yandex.ru.

Kuleshova Aleksandra Andreevna – State Forensic Examiner of the Laboratory of Forensic Linguistics, the Russian Federal Centre of Forensic Science of the Russian Ministry of Justice, Master's Student of the Department of Theory and History of Law of the National Research University Higher School of Economics (HSE); e-mail: kuleshov0061@gmail.com.

Lizorkin Aleksei Mikhailovich – Lead State Forensic Examiner of the Laboratory of Computer Forensics, the Russian Federal Centre of Forensic Science of the Russian Ministry of Justice; e-mail: Liz4@yandex.ru.

Статья поступила 05.06.2018
Received 05.06.2018