АВТОМАТИЗИРОВАННАЯ ОБРАБОТКА ТЕКСТА И ЕЕ ПРИМЕНЕНИЕ
AUTOMATED TEXT PROCESSING AND ITS APPLICATIONS

Вестник Челябинского государственного университета. 2024. № 3 (485). С. 55–65.
ISSN 1994-2796 (print). ISSN 2782-4829 (online)
Bulletin of Chelyabinsk State University. 2024;(3(485):55-65. ISSN 1994-2796 (print). ISSN 2782-4829 (online)

Научная статья

УДК 81’322.4

doi: 10.47475/1994-2796-2024-485-3-55-65

ПРОГРАММНЫЕ ВОЗМОЖНОСТИ ИДЕНТИФИКАЦИИ ТЕКСТОВ: СОПОСТАВЛЕНИЕ НА СХОЖЕСТЬ, УСТАНОВЛЕНИЕ ТОЖДЕСТВА, ПРОВЕРКА НА УНИКАЛЬНОСТЬ

Галина Викторовна Напреенко

Кемеровский государственный университет, Кемерово, Россия, galina_napreenko@mail.ru,
ORCID ID: 0000-0002-4404-0560

Аннотация. Исследование выполнено в рамках идентификационной лингвистики и лингвистической вариантологии. В статье описаны некоторые способы установления степени схожести текстов: алгоритм шинглов, расстояние Левенштейна, системы по выявлению плагиата. Цель работы — описание и апробация программных возможностей сопоставления текстов на схожесть, установления их тождества, проверки на уникальность. В широком смысле данные задачи входят в область идентификации текста. При качественной (ручной) оценке схожести текстов происходит выбор и отбор идентифицирующих параметров специально для исследуемого текста. Использование электронных ресурсов обусловлено стремлением к объективности применяемых методов установления тождества текстов и объективности получаемых результатов. Программные продукты также позволяют установить иную, квантитативную, характеристику — степень схожести текстов друг с другом или степень оригинальности текста. В работе использованы сервисы, в задачи которых входит: 1) сравнение текстов на схожесть; 2) обнаружение заимствования (плагиата). Материалом исследования явился отрывок интервью главы МИД Сергея Лаврова. Вариантами для сравнения с исходным текстом послужили тексты обратного машинного перевода. Обратный машинный перевод как транслятивный продукт — часть искусственного интеллекта и модель процесса понимания и интерпретации естественного языка. Результаты использования предложенных сервисов позволили расположить пять вариантов текстов обратного машинного перевода от наиболее уникального к наиболее тождественному исходному. Исследование показало, что программы в целом дают схожие результаты, которые могут быть применимы для решения исследовательских и прикладных задач, связанных с установлением тождества и различия текстов. Перспектива исследования — выявление лексических параметров, позволяющих классифицировать вторичные тексты обратного машинного перевода как наиболее или наименее тождественные по отношению к первичному варианту.

***Ключевые слова*:** тождество текстов, уникальность текста, идентификация текста, антиплагиат, обратный машинный перевод, транслятивная лингвистика

Для цитирования: Напреенко Г. В. Программные возможности идентификации текстов: сопоставление на схожесть, установление тождества, проверка на уникальность // Вестник Челябинского государственного университета. 2024. № 3 (485). С. 55–65. doi: 10.47475/1994-2796-2024-485-3-55-65

Original article

SOFTWARE CAPABILITIES FOR TEXT IDENTIFICATION: COMPARISON FOR SIMILARITY, IDENTITY ESTABLISHMENT, UNIQUENESS CHECK

Galina V. Napreenko

Kemerovo State University, Kemerovo, Russia, galina_napreenko@mail.ru, ORCID ID: 0000-0002-4404-0560

Abstract. The study was carried out within the framework of identification linguistics and translational linguistics. The article describes some methods for determining the degree of similarity of texts: shingle algorithm, Levenshtein distance, systems for detecting plagiarism. The purpose of the work is to test the software capabilities of comparing texts for similarity, establishing their identity, and checking uniqueness. In a broad sense, these tasks fall within the area of text identification. In a qualitative (manual) assessment of the similarity of texts, identifying parameters are selected and selected specifically for the text under study. The use of electronic resources is determined by the desire for objectivity of the methods used to establish the identity of texts and the objectivity of the results obtained. Software products also make it possible to establish another, quantitative, characteristic — the degree of similarity of texts to each other or the degree of originality of the text. The work used services whose tasks include 1) comparing the similarity of two texts; 2) calculation of the Levenshtein distance; 3) detection of borrowing. The research material was an excerpt from an interview with Foreign Minister Sergei Lavrov. Reverse machine translation texts served as options for comparison with the source text. Reverse machine translation as a translation product is part of artificial intelligence and a model of the process of understanding and interpreting natural language. The results of using the proposed services made it possible to arrange five reverse machine translation options from the most unique text to the text that is most identical to the invariant. The study showed that the programs generally produce similar results, which can be applicable to solving research and applied problems related to establishing the identity and difference of texts. The prospect of the study is to identify lexical parameters that make it possible to classify reverse machine translation texts as the most or least identical with respect to the invariant.

Keywords: text identity, text uniqueness, identity and difference, text identification, anti-plagiarism, reverse machine translation, translational linguistics

For citation: Napreenko GV. Software capabilities for text identification: comparison for similarity, identity establishment, uniqueness check. *Bulletin of Chelyabinsk State University*. 2024;(3(485):55-65. (In Russ.). doi: 10.47475/1994-2796-2024-485-3-55-65

Введение

Развитие информационных технологий, доступ к большим объемам информации, требования к оригинальности научных работ определяют расширение задач автороведческой идентификации и поиск наиболее результативных методов идентификации текстов. Возможности установления тождества текстов в современное время обширны.

Методы, направленные на идентификацию, в широком смысле — установление тождества и различия между объектами, определяются подходом к обнаружению параметров-идентификаторов и делятся на формальные и содержательные (экспертные); в диссертации [12] эти методы получили названия «этюдные» (индивидуализированные) и универсальные, формально-количественные методы. К «этюдным» методам можно отнести многие идентификационные исследования экспертов в юрислингвистической практике, которые содержат описательный план, основываются на интуитивном выявлении специфических параметров текста. Такие параметры выделяются для каждого исследуемого текста индивидуально.

Использование объективных методов (обращение к электронным ресурсам, осуществляющим техническое обнаружение заимствований, механическую проверку текста на уникальность), интерсемиотический перевод (перевод из одной знаковой системы в другую) объективирует результаты анализа текстов. Кроме качественной оценки тождества и различия текстов, соответствующие программные продукты фиксируют *степень* тождества вариантов. Такой «компьютерный» подход к идентификации текста «заключается в построении классификатора, на входе которого — численные значения различных параметров текста, извлеченных <...> автоматически — слов, частей речи; реже — синтаксических и семантических параметров» [14].

Целью статьи является описание и апробация возможностей некоторых программных продуктов по сопоставлению текстов на схожесть, установлению степени оригинальности текстов и, как следствие, установлению степени тождества между текстами, т. е. их идентификации. В статье описаны методы, лежащие в основе работы ряда компьютерных программ, направленных на решение описанных задач.

Сервисы, используемые в исследовании, — фрагменты искусственного интеллекта, рассматриваемые нами как своего рода моделирование процесса понимания и интерпретации естественного языка, а также поиск инструментария для моделирования. Н. Д. Голев отмечает «надиндивидуальность», «надсубъективность» показаний подобных программ как их преимущество для анализа лингвистических источников и возможности «получения объективных

сведений о естественном языке и речевой деятельности» [8, с. 37].

Исследование включено в парадигму транслятивной лингвистики [9]. В качестве вариантов для сопоставления с первичным текстом в аспекте тождества и различия использованы тексты обратного машинного перевода (далее — ОМП), что является одним из инструментов моделирования естественной речевой переводческой деятельности. А. В. Морозов отмечает, что «межъязыковое пространство становится источником сведений о внутренних семантических процессах, происходящих в изучаемом языке, тем своеобразным зеркалом, в котором преломляется семантико-деривационное развитие лексики родного языка» [11, с. 72]. В процессе двойного машинного перевода посредством различных трансформаций формируется вторичный текст. И. А. Баринова и соавторы отмечают, что программы перевода — своеобразный аналог модели языкового сознания человека [3].

Материалы и методы исследования

Материалом исследования для апробации работы программных продуктов явился отрывок интервью главы МИД Сергея Лаврова[1]. Исходный текст — 136 слов. В качестве объектов для сопоставления с первичным текстом на предмет тождественности были выбраны варианты ОМП. Под ОМП Н. Д. Голев понимает «“трансляционный продукт”, полученный в результате машинного (компьютерного) перевода текста или других единиц (слов, словосочетаний, предложений и т. д.) с языка R на язык N и обратного перевода полученного “продукта” с языка N на язык R» [8, с. 36]. Описанным способом было получено пять вариантов ОМП через следующие языки: арабский, хинди, английский, китайский, испанский. Деривационная цепочка ОМП в данной статье записывается следующим образом (на примере машинного перевода с русского языка на китайский язык и затем обратно с китайского языка на русский язык): ОМП рус.-кит.-рус. Задача исследования — сопоставить исходный текст и его варианты на русском языке в аспекте их схожести (в широком смысле — тождественности) с помощью различных сервисов.

В качестве методов исследования были использованы сервисы, в задачи которых входит: 1) сравнение схожести двух текстов (Ciox; iRewriter 1.0); 2) вычисление расстояния Левенштейна; 3) обнаружение заимствования (TEXT.RU; AdvegoPlagiatus; Антиплагиат.ру).

Методы установления тождества текстов

Алгоритм шинглов. Шингл — это часть текста, состоящая из нескольких слов. Программы сопоставляют шинглы определенного размера в двух текстах и устанавливают, являются ли они тождественными. Размер шингла 3 равен трем словам в шингле, то есть это последовательности из трех соседних слов; шингл 4 — последовательность из четырех соседних слов. Снижение размера шингла приводит к более точной проверке текста. В русском языке много «устойчивых словосочетаний, состоящих из трёх и более слов: терминов, названий документов, различных оборотов речи»[2], в связи с чем минимальный размер шингла, который необходимо устанавливать для проверки текста, равен трем.

Специалисты Advego для работы в своей программе советуют устанавливать следующие настройки для более качественной проверки текста:

«для маленьких текстов (до 1000 символов): размер шингла 3–4;

для средних текстов (1000-2500 символов): размер шингла 5–8;

для больших текстов (свыше 2500 символов): размер шингла 10»[3].

Алгоритм шинглов применяется, например, в исследовании А. А. Платонова, Р. Е. Потапова [16]. В указанной работе ставится задача сопоставления текстовых данных (новостей) на предмет частичного и полного сходства (дублирования). Авторы используют метод шинглов, реализация которого происходит «ручным» способом в следующей последовательности: «канонизация текстов, разбиение текстов на шинглы, нахождение контрольных сумм шинглов, поиск одинаковых последовательностей путем сравнения контрольных сумм» [16, с. 79].

Другим методом, который используется для установления тождества, является измерение ***расстояния Левенштейна***. Это минимальное число преобразований (удаления, вставки или замены), необходимых, чтобы превратить одну

[1] Интервью главы МИД Сергея Лаврова 10 марта 2023 г. URL: https://www.1tv.ru/news/2023-03-10/448829-eksklyuzivnoe_intervyu_glavy_mid_sergeya_lavrova_smotrite_etim_vecherom_na_pervom_kanale (дата обращения: 15.08.2023).

[2] Сидоров С. В. Сайт педагога-исследователя. URL: http://si-sv.com/load/10-1-0-62 (дата обращения: 20.08.2023).

[3] Там же.

последовательность в другую. Расстояние Левенштейна применяется в исследованиях с различной целью. Например, для устранения опечаток в записях баз данных [10]; для исправления ошибок при вводе текста; в приложениях по автоматической обработке текстов и пр.

Примером работы, применяющей данный метод, является статья С. Б. Потемкина [17]. Автор на материале ОМП рассчитал расстояние Левенштейна «между исходным русским текстом и текстом, полученным в результате прямого и обратного переводов» [17. С. 80]. В случае если расстояние между двумя фразами (словами) равно нулю, русский термин признавался эквивалентом иноязычному термину и, как следствие, соответствующим международным терминологическим стандартам; если расстояние не равно нулю, то автор приходит к выводу, что система машинного перевода «не знает» русского термина, а значит, неправильно выбирает его иноязычный эквивалент [17, с. 80–81]. Например:

«Initial (Russian) = вера в справедливый мир

-> Russian to English = just-world hypothesis

-> Back English to Russian = вера в справедливый мир

// Levenshtein distance = 0» [17, с. 81].

С. Б. Потемкин осуществил машинный перевод термина с русского языка на английский язык, а затем обратно с английского на русский. В приведенном примере преобразований между исходной и полученной последовательностями не требуется, расстояние Левенштейна равно нулю, а значит, варианты эквивалентны, то есть тождественны. Применение данного метода позволило автору прийти к выводу о перспективности «использования системы машинного перевода для стандартизации научно-технической терминологии» [17. С. 82]. Описанная работа является примером расчета расстояния Левенштейна исследователем вручную, однако существуют программные продукты, позволяющие измерять расстояние Левенштейна автоматически.

Расстояние Левенштейна наряду с коэффициентом сходства Жаккара используется как способ сравнения текстов из открытой печати на схожесть [1]; как метод сравнения оцифрованных копий деловых документов [2; 18]. В статье «Многоэтапный метод автоматической коррекции искаженных текстов» [5] расстояние Левенштейна выступает в качестве метода выявления искажений в виде ошибочных символов, слов и словосочетаний при автоматическом распознавании речи или оптическом распознавании изображений текстов. Авторы предлагают метод коррекции искаженных текстов: последовательное определение ошибок и исправление искаженных текстов. После установленного признака искаженности отдельных слов «строится список возможных вариантов слов, в который попадают только те словоформы из словаря, которые находятся от исследуемого слова на определенном расстоянии Левенштейна» [5, с. 35].

Использование систем Антиплагиата с целью выявления степени уникальности/оригинальности текстов. В данном случае речь идет об идентификации плагиата и неправомерных заимствований в тексте. Уникальность текста — параметр, отражающий оригинальность текста в процентах. Проблема плагиата актуальна и поднимается во многих современных работах. Программы по обнаружению плагиата широко используются в научно-образовательной среде, в связи с чем существующие исследования систем идентификации плагиата и непосредственно «Антиплагиата» разноаспектно подходят к проблеме оценке их использования: описывают особенности применения программы в науке и образовании [22; 23], осуществляют оценку качества работы программы [7], анализируют применение авторами приемов, влияющих на итоговый процент уникальности [15], формулируют предложения по решению проблем использования программ с целью оптимизации процесса проверки и повышения эффективности ее результатов [19], рассматривают методики борьбы с плагиатом и пр. Важным, на наш взгляд, является следующее положение: «Программные продукты не констатируют факт плагиата, а указывают на наличие в научной или учебной работе текстовых заимствований, не оформленных соответствующим образом. Вывод о наличии в этих заимствованиях плагиата принимается компетентным специалистом — экспертом в предметной области» [19, с. 137].

В статье [21] исследуется методология оценки научных публикаций с помощью системы «Антиплагиат». Авторы, проведя исследование более 200 магистерских диссертаций, 14 кандидатских и трёх докторских диссертаций с помощью программы «Антиплагиат», пришли к выводу об особенностях использования: например, система не приспособлена для автоматической работы, без последующей проверки результатов специалистом.

Однако перспектива использования программ по обнаружению плагиата может заключаться и в их применении в рамках «идентификационной

лингвистики» [13]. Выделенные программой фрагменты текста (плагиата по оценке системы) могут быть подвергнуты оценке в идентификационном аспекте. При этом важным является тот факт, что «результат автоматической проверки текста в системе «Антиплагиат» содержит информацию лишь о наличии совпадений фрагментов проверяемого текста и иных текстов, но при отсутствии экспертной оценки не является показателем оригинальности текста и не позволяет судить об отсутствии или наличии в нем заимствований» [20].

На границе компьютерных, формальных и экспертных методов находятся методы, содержащие интерсемиотический перевод как формально-количественный подход, позволяющий переводить текстовые данные в статистические благодаря частотному анализу (пр., диаграмма сравнения частотных портретов по однограммам трех выбранных писателей в статье [6]). Близостью к методам формального характера является еще и возможность установить степень тождества и различия текстов по заданным исследователем параметрам. Подобный подход представлен в статье Ю. А. Башкатовой [4]. Автор, используя ОМП с целью анализа деривационных деформаций между исходным текстом на русском языке и полученным текстом на русском языке, предлагает уровневый механизм измерения (от лексического уровня к тексту), чем осуществляет перевод содержательных трансформаций в объективные категории. С этой целью автор предлагает применять пятибалльную шкалу: 1 балл — лексико-синонимическая замена слов (пр., невероятно — удивительно) или грамматическая (морфологическая) транспозиция (пр., удивлял — удивил); 2 балла — деривационная трансформация по двум направлениям; 3 балла — к лексическому и грамматическому фактору добавляется синтаксическая трансформация предложения (пр., замена сложного предложения простым); 4 балла — смысловое расхождение между исходным текстом и его вариантом; 5 баллов — деривационная мутация, приводящая к значительному искажению смысла текста. Апробация методов данного типа не входит в задачи настоящей статьи.

Результаты исследования и их обсуждение

Апробация программ по установлению степени тождества между текстами

Результаты сравнения текстов на схожесть с помощью программы, работающей на ***алгоритме шинглов***[1]. Информация в процентах отражает степень схожести исходного текста и его варианта. Данные в таблице расположены от наименее тождественного текста исходному (низкий процент схожести) к наиболее тождественному (высокий процент схожести). Так, вариант ОМП рус. — кит. — рус. оказался наименее схожим с исходным текстом, а вариант ОМП рус. — англ. — рус. — более схожим.

Таблица 1
Table 1

Сравнение текстов ОМП на схожесть Ciox.ru (алгоритм шинглов)
Comparison of reverse machine translation texts for similarity Ciox.ru (shingle algorithm)

	Китайский	Хинди	Арабский	Испанский	Английский
Исходный	35,96 %	41,44 %	43,4 %	46,73 %	54,46 %

Программа iRewriter 1.0[2] также использует алгоритм шинглов, но предоставляет возможность дополнительно выбрать размер шингла от 1 до 10, а также выделяет цветом заимствованные фрагменты текста. Ориентируясь на рекомендации создателей программы о необходимости выбора размера шингла 3–4 для маленьких текстов (до 1000 символов), остановимся на размере шингла 3 (более строгая проверка) с функцией «удалять стоп слова». В таком случае в процессе проверки текстов удаляются некоторые слова и символы, к которым обычно относят числа, указательные слова, междометия, частицы, союзы, предлоги, знаки препинания.

[1] Вычислительные веб-сервисы. CIOX. URL: https://ciox.ru/comparing_the_similarity_of_texts (дата обращения: 20.08.2023).

[2] iRewriter 1.0. URL: http://si-sv.com/load/10-1-0-62 (дата обращения: 20.08.2023).

Таблица 2
Table 2

Сравнение текстов ОМП на схожесть iRewriter 1.0 (алгоритм шинглов, функция «удалять стоп-слова»)
Comparison of reverse machine translation texts for similarity iRewriter 1.0 (shingle algorithm, “remove stop words” function)

	Китайский	Арабский	Хинди	Испанский	Английский
Исходный	3,9 %	11,8 %	14,3 %	15,7 %	23,1 %

Таблица 3
Table 3

Сравнение текстов ОМП на схожесть iRewriter 1.0 (алгоритм шинглов)
Comparison of reverse machine translation texts for similarity iRewriter 1.0 (shingle algorithm)

	Китайский	Арабский	Испанский	Хинди	Английский
Исходный	3, 9 %	16,4 %	22,7 %	22,9 %	31,7 %

В таблице 3 представлены данные по сопоставлению текстов без учета функции «удалять стоп-слова». В таком случае программа не изменяет исходный текст, то есть не исключает слова и символы из текста перед проверкой.

Использование двух программ, основанных на работе алгоритмов шингла, показало, что вариант ОМП рус. — кит. — рус. является менее тождественным исходному, а вариант ОМП рус. — англ. — рус. — более тождественным исходному тексту. Расхождения в результатах (порядок расположения ОМП в таблице), представленных в таблицах 2 и 3, незначительны и относятся к вариантам ОМП рус. — хинди. — рус. и ОМП рус. — исп. — рус.

В таблице 4 представлены результаты проверки программы, работающей по другому принципу — ***измерению расстояния Левенштейна***[1]. ОМП в таблице расположены от наименее тождественного первичному тексту к наиболее тождественному, то есть от текста с наибольшим числом расстояния Левенштейна к наименьшему.

Таблица 4
Table 4

Сравнение текстов ОМП на схожесть с помощью измерения расстояния Левенштейна
Comparing of reverse machine translation texts for similarity using Levenshtein distance measurements

	Китайский	Хинди	Арабский	Испанский	Английский
Исходный	545	359	358	314	253

Данные, отраженные в таблицах 1 и 4, а также с небольшим расхождением в таблицах 2 и 3, показывают одинаковые результаты: наиболее схожим вариантом с исходным текстом, а значит, вариантом, содержащим наибольшее количество тождественных параметров, является ОМП, полученный посредством обратного перевода рус. —англ. — рус.; вариантом, имеющим меньшее количество тождественных параметров с исходным текстом, является ОМП, полученный посредством обратного перевода рус. — кит. — рус.

Работа программ по ***проверке текстов на уникальность*** (обнаружению плагиата) отличается от проверки, основанной на методе шинглов и измерении расстояния Левенштейна. В описанных выше программах тексты были сопоставлены друг с другом, варианты — с исходным текстом. В программах по выявлению плагиата текст сопоставляется с источниками из открытого доступа в Интернете, а значит, параметры тождества и различия выявляются не между заданными текстами, а между одним заданным текстом и множеством, представленным в Интернете.

Проверка исходного текста и его вариантов в программе TEXT.RU[2] позволила получить результаты, отраженные в таблице 5. Расположение

[1] Онлайн-калькулятор для расчета расстояния Левенштейна. URL: https://hostciti.net/calc/it/distance-lowenstein.html (дата обращения: 05.09.2023).

[2] Сервис проверки текста на уникальность и биржа контента. URL: https://text.ru/antiplagiat/unauthorized (дата обращения: 08.09.2023).

вариантов ОМП в таблице следующее: от более уникального (имеющего больше различий с исходным текстом, чем тождества) к менее уникальному (имеющему большее количество тождественных параметров с исходным текстом, чем различий).

Таблица 5
Table 5

Проверка текстов ОМП на уникальность в программе TEXT.RU
Checking of reverse machine translation texts for uniqueness in the TEXT.RU program

Китайский	Испанский	Арабский	Хинди	Английский
67,57 %	44,16 %	29,83 %	27,82 %	13,72 %

Программа для проверки уникальности текста AdvegoPlagiatus[1] анализирует текст по нескольким алгоритмам: алгоритм шинглов, алгоритм лексических совпадений, алгоритм псевдоуникализации (наличие символов, повышающих уникальность текста). Сервис выдает два результата: первый — процент уникальности текста, второй — процент рерайта (таблица 6). ОМП рус. — кит. — рус. является наиболее уникальным, ОМП рус. — англ. — рус. — наименее уникальным.

Таблица 6
Table 6

Проверка текстов ОМП на уникальность в программе AdvegoPlagiatus
Checking of reverse machine translation texts for uniqueness in the AdvegoPlagiatus program

Китайский	Арабский	Испанский	Хинди	Английский
97 % / 94 %	76 % / 50 %	70 % / 33 %	66 % / 33 %	58 % / 14 %

В таблице 7 представлены результаты проверки текстов ОМП в системе «Антиплагиат»[2]. Данные выражены в виде процента оригинальности текста. Система «Антиплагиат» выдает также следующие результаты: процент совпадения, самоцитирования и цитирования. ОМП расположены в таблице от более оригинального (текст имеет меньшее количество параметров, тождественных параметрам в других текстах) к наименее оригинальному (текст имеет большее количество параметров, тождественных параметрам в других текстах).

Таблица 7
Table 7

Проверка текстов ОМП на уникальность в программе «Антиплагиат.ру»
Checking of reverse machine translation texts for uniqueness in “The Antiplagiat.ru” program

Арабский	Китайский	Хинди	Английский	Испанский
66, 74 %	66,52 %	44,99 %	44,99 %	24,6 %

Результаты апробации программных продуктов по сопоставлению текстов ОМП различаются в зависимости от целей использования этих сервисов: проверка текстов на схожесть или проверка текстов на уникальность.

Согласно данным по результатам проверки в программах, сравнивающих тексты по степени схожести (таблицы 1–4), наиболее схожим с первичным текстом является ОМП рус. — англ. — рус., так как имеет большее количество параметров, которые оцениваются программами как тождественные, наименее схожим — ОМП рус. — кит. — рус., так как процент уникального текста в нем превалирует над тождественными отрезками.

Результаты проверки текстов на уникальность (таблицы 5–7) совпадают только в части ОМП рус. — кит. — рус. и ОМП рус. —англ. — рус. с данными в таблицах 1–4. Отдельно необходимо выделить значительные отличия в результатах анализа ОМП системой «Антиплагиат»

[1]Advego Антиплагиат — проверка уникальности текста. URL: https://advego.com/antiplagiat/ (дата обращения: 08.09.2023).

[2]Антиплагиат. Система обнаружения текстовых заимствований. URL: https://antiplagiat.ru/ (15.09.2023).

по сравнению с результатами других систем, что обусловлено назначением сервиса, базой проверки (открытые, закрытые источники) и пр. и требует отдельного обзора и апробации.

Заключение

Преимущество использования программы сравнения текстов на схожесть (установления степени схожести / тождества текстов) как один из способов моделирования естественного языка обнаруживается в том, что они сопоставляют два конкретных текста: первичный текст и его вариант. Результатом такого сопоставления является процент схожести текстов, то есть наличие тождественных признаков в обоих текстах, выраженных в процентах.

Выявление уникальности текста (как параметра, отражающего его оригинальность) работает иначе: это сопоставление исследуемого варианта с множеством текстов, представленных в свободном доступе в интернете. Результатом является процент оригинальности текста, то есть процент наличия в нем параметров, тождественных с отрезками в других текстах.

Проверка одного и того же текста разными сервисами может давать различные результаты, однако, как показало исследование, разница не всегда значительна. Данные, отраженные в таблицах, показали, что наиболее схожим вариантом с исходным текстом является ОМП рус. —англ. — рус.; наиболее оригинальным, уникальным — ОМП рус. — кит. — рус.

Описанные программные возможности позволяют устанавливать объективные параметры: ***степень*** тождества текстов и ***степень*** их уникальности. В рамках идентификационной лингвистики такие программы могут быть использованы в качестве первого этапа оценки текстов на предмет их тождества или различия с последующим качественным (экспертным) анализом. Выявлению параметров, позволяющих определить степень тождества текстов ОМП как транслятивного варианта и исходного текста в лексическом аспекте, будут посвящены последующие исследования.

Список источников

1. Аёшин И. Т., Федоров В. А., Городов А. А., Абдулаев Н. А. Методики сравнительного анализа текстовых документов на близость // Научно-технический вестник Поволжья. 2022. № 6. С. 46–50.

2. Андреева Е. И., Манжиков Т. В., Славин О. А. Сравнение оцифрованных страниц деловых документов на основе распознавания // Сенсорные системы. 2018. Т. 32. № 1. С. 35–41.

3. Баринова И. А., Нестерова Н. М., Овчинникова И. Г. «Языковое сознание»: к вопросу об определении и интерпретации термина // Вестник Пермского национального исследовательского политехнического университета. Проблемы языкознания и педагогики. 2010. № 4. С. 10–21.

4. Башкатова Ю. А. Обратный машинный перевод как способ измерения смыслового тождества / различия вариантов текста // Современная парадигма анализа языка и межкультурной коммуникации и ее аппликативный потенциал в обучении родному и иностранному языкам : материалы нац. научн. конф. (Барнаул, 18–19 сентября 2019 г.). Барнаул : Алтайск. гос. пед. ун-т, 2020. С. 18–23.

5. Вахлаков Д. В., Мельников С. Ю., Пересыпкин В. А. Многоэтапный метод автоматической коррекции искаженных текстов // Известия Южного федерального университета. Технические науки. 2020. № 7 (217). С. 35–45.

6. Волынкин П. А., Гянджиев Э. Э. Идентификация авторства текста при помощи частотных портретов // European Scientific Conference : сб. статей X Междунар. науч.-практ. конф. Пенза, 2018. Ч. 1. С. 150–155.

7. Гельман В. Я. Проблемы формально-механистического подхода к выявлению плагиата в научных работах // Экономика науки. 2020. Т. 6, № 3. С. 180–185.

8. Голев Н. Д. Источниковый потенциал обратного машинного перевода // КРСУ жарчысы. 2018. Т. 18. № 1. С. 36–45.

9. Голев Н. Д. Транслятивная лингвистика (аспектуализированный обзор исходных положений). Часть 1. Гносеология перевода // Вестник Кемеровского государственного университета. 2022. Т. 24. № 6 (94). С. 717–734.

10. Карахтанов Д. С. Программная реализация алгоритма Левенштейна для устранения опечаток в записях баз данных // Молодой ученый. 2010. Т. 1, № 8 (19). С. 158–162.

11. Морозов А. В. Обратный лексикографический перевод как метод исследования деривационного потенциала русского слова в межъязыковом пространстве // Вестник Воронежского государственного университета. Серия: Лингвистика и межкультурная коммуникация. 2004. № 1. С. 71–74.

12. Напреенко Г. В. Лексико-квантитативное моделирование языковой личности в идентификационном аспекте (на материале русскоязычных интернет-дневников) : автореф. дис. … канд. филол. наук. Кемерово, 2015. 26 с.

13. Напреенко Г. В. Феномен идентификации и идентификационная лингвистика // Мир науки, культуры, образования. 2022. № 5 (96). С. 365–368.

14. Огорелков И. В. Авторoведческая идентификация текстов политического дискурса: эволюция методов // Лингвополитическая персонология: дискурсивный поворот : материалы Междунар. науч. конф. Екатеринбург, 2019. С. 159–161.

15. Павлов А. А. О применении авторами научных текстов технических приемов, искажающих результаты проверки уникальности текстов. Обзор проблемы, опыт выявления и анализ подобных текстов // Научная периодика: проблемы и решения. 2020. Т. 9. № 3–4. URL: https://nppir.ru/01NP320.html (дата обращения: 08.05.2023).

16. Платонов А. А., Потапов Р. Е. Обнаружение дубликатов статей в системе автоматического сбора информации из открытых источников об экологической обстановке // Известия Волгоградского государственного технического университета. 2015. № 6 (163). С. 79–82.

17. Потемкин С. Б. Машинный перевод как средство стандартизации терминологии // Вестник Московского государственного областного университета. Серия: Лингвистика. 2017. № 5. С. 77–84.

18. Славин О. А., Андреева Е. И., Арлазаров В. В. Поиск фальсификаций в копиях деловых документов // Математические методы в технике и технологиях — ММТТ. 2020. Т. 6. С. 96–100.

19. Тимофеев В. В. Анализ современных тенденций использования системы «Антиплагиат» при проверке учебных и научных работ // Вестник Калининградского филиала Санкт-Петербургского университета МВД России. 2023. № 2 (72). С. 136–140.

20. Усачева Е. А. К вопросу о допустимости использования системы «Антиплагиат» для определения авторства и оценки оригинальности произведения // Образование и право. 2019. № 4. С. 204–210.

21. Харченко С. Г., Докукин П. А., Кучер Д. Е. К вопросу о методологии оценки научных публикаций // Самарская Лука: проблемы региональной и глобальной экологии. 2022. Т. 31. № 4. С. 61–68.

22. Хованская Т. В., Сандирова М. Н. Использование системы «Антиплагиат» в высшей школе // Проблемы современного образования. 2019. № 3. С. 51–58.

23. Чиркин Е. С. Использование систем антиплагиата в образовании // Вестник российских университетов. Математика. 2013. № 6 (2). URL: https://cyberleninka.ru/article/n/ispolzovanie-sistem-antiplagiata-v-obrazovanii (дата обращения: 08.09.2023)

References

1. Aeshin IT, Fedorov VA, Gorodov AA, Abdulaev NA. Methods for comparative analysis of text documents for proximity. *Nauchno-tekhnicheskiy vestnik Povolzh'ya = Scientific and technical bulletin of the Volga region*. 2022;(6):46–50. (In Russ.).

2. Andreeva E, Manzhikov T, Slavin O. Comparing digitized pages of business documents using an indicator. *Sensornye sistemy = Sensory systems*. 2018; 32, (1): 35–41. (In Russ.).

3. Barinova I. A., Nesterova N. M., Ovchinnikova I. G. "Linguistic consciousness": on the issue of definition and interpretation of the term. *Bulletin of the Perm National Research Polytechnic University. Problemy yazykoznaniya i pedagogiki = Bulletin of the Perm National Research Polytechnic University. Problems of linguistics and pedagogy*. 2010;(4):10-21. (In Russ.).

4. Bashkatova YuA. Reverse machine translation as a way of measuring semantic identity / differences of text variants. In: Sovremennaya paradigma analiza yazyka i mezhkul'turnoy kommunikatsii i ee applikativnyy potentsial v obuchenii rodnomu i inostrannomu yazykam = Modern paradigm of language analysis and intercultural communication and its applicative potential in teaching native and foreign languages. Barnaul; 2020. Pp. 18-23. (In Russ.).

5. Vakhlakov DV, Mel'nikov SYu, Peresypkin VA. Multi-stage method for automatically correcting distorted texts. *Izvestiya Southern Federal University. Tekhnicheskie nauki = News of the Southern Federal University. Technical science*. 2020;(7(217):35-45. (In Russ.).

6. Volynkin PA, Gyandzhiev EE. Identification of the author of a text using frequency portraits. In: European Scientific Conference. Penza; 2018. Part 1. Pp. 150-155. (In Russ.).

7. Gel'man VYa. Problems of a formal-mechanistic approach to identifying plagiarism in scientific works. *Ekonomika nauki = Economics of Science*. 2020;6(3):180-185. (In Russ.).

8. Golev ND. Source potential of reverse machine translation. *KRSU жарчысы = Herald of KRSU.* 2018;18(1):36-45. (In Russ.).

9. Golev ND. Translational linguistics (an aspectualized review of the starting points). Part 1. Epistemology of translation. *Vestnik Kemerovskogo gosudarstvennogo universiteta = Bulletin of Kemerovo State University.* 2022;24(6(94):717-734. (In Russ.).

10. Karakhtanov DS. Software implementation of the Levenshtein algorithm for eliminating typos in database records. *Molodoy uchenyy = Young scientist*. 2010;1(8(19):158-162. (In Russ.).

11. Morozov AV. Reverse lexicographic translation as a method for studying the derivational potential of the Russian word in the interlingual space. *Vestnik Voronezhskogo gosudarstvennogo universiteta. Seriya: Lingvistika i mezhkul'turnaya kommunikatsiya = Bulletin of Voronezh State University. Series: Linguistics and intercultural communication*. 2004;(1):71-74. (In Russ.).

12. Napreenko GV. Leksiko-kvantitativnoe modelirovanie yazykovoy lichnosti v identifikatsionnom aspekte (na materiale russkoyazychnykh internet-dnevnikov) = Lexico-quantitative modeling of linguistic personality in the identification aspect (based on Russian-language Internet diaries). Abstract of thesis. Kemerovo; 2015. 26 p. (In Russ.).

13. Napreenko GV. The phenomenon of identification and identification linguistics. *Mir nauki, kul'tury, obrazovaniya = World of science, culture, education.* 2022;(5(96):365-368. (In Russ.).

14. Ogorelkov IV. Author's identification of political discourse texts: evolution of methods. In: Lingvopoliticheskaya personologiya: diskursivnyy povorot = Linguistic and political personology: a discursive turn. Ekaterinburg; 2019. Pp. 159-161. (In Russ.).

15. Pavlov AA. On the use by authors of scientific texts of technical techniques that distort the results of checking the uniqueness of texts. Review of the problem, experience in identifying and analyzing similar texts. *Nauchnaya periodika: problemy i resheniya = Scientific periodicals: problems and solutions*. 2020;9(3-4). Available from: https://nppir.ru/01NP320.html (accessed: 08.05.2023) (In Russ.).

16. Platonov AA, Potapov RE. Detection of duplicate articles in a system for automatically collecting information from open sources about the environmental situation. *Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta = News of Volgograd State Technical Universit*. 2015;(6(163):79-82. (In Russ.).

17. Potemkin SB. Machine translation as a means of terminology standardization. *Vestnik Moskovskogo gosudarstvennogo oblastnogo universiteta. Seriya: Lingvistika = Bulletin of Moscow State Regional University. Series: Linguistics*. 2017;(5):77-84. (In Russ.).

18. Slavin OA, Andreeva EI, Arlazarov VV. Finding falsifications in copies of business documents. *Matematicheskie metody v tekhnike i tekhnologiyakh — MMTT = Mathematical methods in engineering and technology — MMET.* 2020;6:96-100. (In Russ.).

19. Timofeev VV. Analysis of modern trends in the use of the Anti-Plagiarism system when checking educational and scientific works. *Vestnik Kaliningradskogo filiala Sankt-Peterburgskogo universiteta MVD Rossii = Bulletin of the Kaliningrad branch of the St. Petersburg University of the Ministry of Internal Affairs of Russia*. 2023;(2(72):136-140. (In Russ.).

20. Usacheva EA. On the issue of the admissibility of using the Anti-Plagiarism system to determine authorship and assess the originality of a work. *Obrazovanie i parvo = Education and law*. 2019;(4):204-210. (In Russ.).

21. Kharchenko SG, Dokukin PA, Kucher DE. On the issue of methodology for assessing scientific publications. *Samarskaya Luka: problemy regional'noy i global'noy ekologii = Samarskaya Luka: problems of regional and global ecology*. 2022;31(4):61-68. (In Russ.).

22. Khovanskaya TV, Sandirova MN. Using the Anti-Plagiarism System in Higher Education. *Problemy sovremennogo obrazovaniya = Problems of modern education*. 2019;(3):51-58. (In Russ.).

23. Chirkin ES. Using anti-plagiarism systems in education. *Vestnik rossiyskikh universitetov. Matematika = Bulletin of Russian Universities. Mathematics*. 2013;(6(2). Available from: https://cyberleninka.ru/article/n/ispolzovanie-sistem-antiplagiata-v-obrazovanii (accessed: 08.09.2023). (In Russ.).

Информация об авторе

Г. В. Напреенко — кандидат филологических наук, доцент кафедры стилистики и риторики.

Information about the author

Galina V. Napreenko — Cand. Sci. (Philology), Associate Professor, Department of Stylistics and Rhetoric.

Статья поступила в редакцию 20.11.2023; одобрена после рецензирования 27.11.2023; принята к публикации 22.03.2024.

The article was submitted 20.11.2023; approved after reviewing 27.11.2023; accepted for publication 22.03.2024.

Автор заявляет об отсутствии конфликта интересов.

The author declares no conflicts of interests.