Terra Linguistica. 2022, Том 13, № 4. С. 22–40.
Terra Linguistica. 2022, Vol. 13, No. 4. Pp. 22–40.

Научная статья
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.13402

ЭКСПЕРИМЕНТЫ ПО АВТОМАТИЧЕСКОМУ ВЫДЕЛЕНИЮ КЛЮЧЕВЫХ ВЫРАЖЕНИЙ В СТИЛИСТИЧЕСКИ РАЗНОРОДНЫХ КОРПУСАХ РУССКОЯЗЫЧНЫХ ТЕКСТОВ

О.А. Митрофанова[1] ✉, Д.А. Гаврилик[2]

[1] Санкт-Петербургский государственный университет, Санкт-Петербург, Российская Федерация;

[2] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация

✉ o.mitrofanova@spbu.ru

Аннотация. Статья посвящена экспериментальному исследованию методов автоматического выделения ключевых выражений с использованием экспертных оценок. Целью работы является проверка гипотез о распределении ключевых выражений в документе, о дифференциации ключевых выражений с точки зрения используемых алгоритмов их выделения и стилистической принадлежности текстов. Эксперименты по автоматическому выделению ключевых выражений проводятся с помощью девяти алгоритмов различных типов: статистические (Log-Likelihood, TF-IDF, Хи-квадрат), гибридные, или лингвостатистические (RAKE, YAKE, PullEnti, Topia), структурные, или графовые (TextRank), с использованием машинного обучения (KeyBERT). В ходе исследования был подготовлен смешанный корпус объемом около 1 млн с/у, включающий в себя 50 публицистических текстов (новостные сообщения с заголовками), 50 научных текстов (статьи по компьютерной лингвистике с заголовками, аннотациями и задаваемыми вручную наборами ключевых выражений), 50 художественных текстов (главы из прозаических произведений, снабженные авторским описанием содержания). Для проверки гипотезы о пространственно-позиционных и стилистически детерминированных характеристиках ключевых выражений были проведены три серии экспериментов, в результате которых были сопоставлены эталонные ключевые выражения, выделенные экспертами из первого сегмента текстов, и ключевые выражения, извлеченные из второго сегмента автоматическими методами. Количественная оценка совпадений экспертной и автоматической разметки позволила подтвердить гипотезу о различной концентрации ключевых выражений в сравниваемых сегментах текста. Исследование лексико-грамматических и семантических особенностей выделенных ключевых выражений выявило те их признаки, которые определяются стилистическими особенностями текстов. Результаты исследования позволяют усовершенствовать процедуры семантической компрессии, производимые с применением различных методов автоматического выделения ключевых выражений.

Ключевые слова: семантическая компрессия, автоматическое извлечение ключевых выражений, экспертная разметка, корпус текстов, функциональные стили.

Финансирование: НИП СПбГУ № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта», грант РНФ № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики».

Для цитирования: Митрофанова О.А., Гаврилик Д.А. Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Т. 13. № 4. С. 22–40. DOI: 10.18721/JHSS.13402

Research article
DOI: https://doi.org/10.18721/JHSS.13402

EXPERIMENTS ON AUTOMATIC KEYPHRASE EXTRACTION IN STYLISTICALLY HETEROGENEOUS CORPUS OF RUSSIAN TEXTS

O.A. Mitrofanova[1] ✉, D.A. Gavrilic[2]

[1] St. Petersburg State University, St. Petersburg, Russian Federation;

[2] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation

✉ o.mitrofanova@spbu.ru

Abstract. The paper describes the experimental study of automatic keyphrase extraction techniques using expert assessments. The purpose of the study is to confirm the hypotheses on the location of keyphrases within a document and on the differentiation of keyphrases as regards applied algorithms and text styles. Experiments on automatic selection of keyphrases are carried out using nine algorithms of various types, including statistical (Log-Likelihood, TF-IDF, Chi-square), hybrid, also called linguostatistical (RAKE, YAKE, PullEnti, Topia), structural, also called graph-based (TextRank), and machine learning (KeyBERT). In the course of the study a mixed corpus was prepared of about 1 million tokens in size, including 50 social media texts (news reports with headlines), 50 scientific texts (articles on computational linguistics with titles, abstracts and manually specified sets of key expressions), 50 literary texts (chapters from prose works, provided with the author's description of the content). Evaluation procedure implies comparison of keyphrases selected by experts from the first segment of texts and key expressions automatically extracted from the second segment. A quantitative assessment of the matches between expert and automatic markup made it possible to confirm the hypothesis on a different concentration of keyphrases in text segments involved in comparison. The study of lexico-grammatical and semantic features of keyphrases allowed us to reveal features that are determined by text style. The results of the study may improve semantic compression procedures performed using the methods of automatic keyphrase extraction.

Keywords: semantic compression, automatic keyphrase extraction, expert annotation, text corpus, functional styles.

Acknowledgements: Scientific Project of St. Petersburg State University No. 75254082 "Modeling the communicative behavior of residents of a Russian metropolis in the socio-speech and pragmatic aspects with the use of artificial intelligence methods", RSF grant No. 21-78-10148 "Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics."

Citation: O.A. Mitrofanova, D.A. Gavrilic, Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts, Terra Linguistica, 13 (4) (2022) 22–40. DOI: 10.18721/JHSS.13402

Введение

Автоматическое выделение ключевых выражений является базовой процедурой семантической компрессии, способствующей структурированию информации в естественноязыковых текстах. Ключевые выражения также помогают составить быструю оценку содержания документов, которая может быть уточнена и расширена в ходе комплексных процедур индексирования и рубрикации документов, их реферирования, упрощения, перифразирования [1–5].

Ключевые выражения рассматриваются как объект психолингвистического исследования и результат работы алгоритмов обработки текстовых данных.

С точки зрения психолингвистики, ключевые выражения представляют собой особый тип текста, тексты-примитивы [6–8], которые, представляя нестандартный по сравнению с языковой

нормой материал, характеризуются целостным содержанием, но при этом возможным нарушением связности формы. Благодаря цельности ключевые выражения можно считать «смысловыми опорами» в понимании текста [9]. Наборы ключевых слов принято относить ко вторичным текстам-примитивам, по которым можно восстановить содержание исходного текста в результате процессов перифразирования. Эксперименты подтвердили, что ключевые выражения имеют статус самостоятельных языковых единиц, которые подчиняются основным законам построения текста, выступают как компоненты парадигматических синтагматических рядов, организуют ассоциативные поля в исходном тексте, демонстрируют процесс развертывания информации в процессе ее вербализации и проявляют определенную контекстную предсказуемость [10–11]. Учитывая различные регистры функционирования текста (устный и письменный (в том числе и клавиатурно-опосредованный)), следует различать ключевые и опорные выражения [12], под последними понимаются лучше всего воспринимаемые фрагменты текста при его динамическом развертывании. Дискуссионными являются вопросы о процессах порождения ключевых выражений при восприятии информации в текстовом виде или в виде инфографики, при этом эксперименты подтверждают, что при единстве процедур обработки информации носители языка опираются преимущественно на текстовые данные [13].

При работе с текстами ограниченных объемов допустимы ручные методы определения ключевых выражений, однако анализ обширных корпусных данных требует автоматизации. В связи с этим со второй половины 20 века продолжается развитие методов автоматического выделения ключевых выражений, стимулируемое появлением новых информационных ресурсов, а также форм взаимодействия носителей языка и интеллектуальных систем. Типология процедур автоматического выделения ключевых выражений определяется следующими факторами: использование статистических признаков ключевых выражений, их лексико-грамматической организации, ограничения на тип выражений (униграммы, биграммы, триграммы и т.д.), способы их ранжирования в выдаче (учет местоположения в тексте, длина, встречаемость в составе других n-грамм), наличие одного корпуса текстов или пары корпусов – основного и фонового, возможность использования размеченных данных для организации процедур машинного обучения и т.д. [14–18]. Автоматизация выделения ключевых выражений, равно как и ручная их разметка, является предметом дискуссий. Возникающие вопросы связаны с возможным несоответствием лексических единиц в реферативной и основной частях документа: зачастую назначаемые авторами ключевые выражения редко встречаются в тексте или вовсе в нем отсутствуют. В таких случаях неизбежно применение автоматических методов обработки данных. Базовыми количественными характеристиками, по которым можно оценить потенциальную значимость ключевых выражений для читателя, являются их плотность (отношение частоты употребления в тексте по отношению к его общему объему) и пространственно-позиционные признаки (расположение в документе). Принято считать, что наиболее информативны выражения, встречающиеся в заголовке, аннотации, в начальной части текста (первый абзац, первые несколько предложений), а также в конце текста (в заключении) [19–21].

Наше исследование направлено на определение соотношения между ключевыми выражениями, выделяемыми в русскоязычных текстах разных стилей вручную и автоматически. Цель эксперимента заключается в проверке гипотезы о возможности автоматического выявления в основной части текста тех ключевых выражений, которые размечены экспертами вручную в начальной части текста. Объективность результатов эксперимента обеспечивается разнообразием источников, которые включают публицистические, научные и художественные тексты, отличающиеся содержательной структурой, а также выбором группы методов автоматического выделения ключевых выражений. В случае подтверждения рассматриваемой гипотезы можно считать обоснованным использование автоматических методов выделения ключевых выражений в исследованиях семантики текстов и ее восприятия носителями языка. Кроме того, сравнительный анализ методов

автоматического выделения ключевых выражений, проведенный на материале русскоязычных корпусов текстов, позволит дать оценку их эффективности и определить их сферы применения.

Методы автоматического извлечения ключевых выражений

Для автоматического извлечения ключевых выражений в нашем исследовании применены статистические, гибридные (лингвостатистические), структурные (графовые) методы, методы с использованием машинного обучения. В рамках данного исследования рассматривались девять методов, представляющих разные типы, а именно, статистические: Log-Likelihood, TF-IDF, Хи-квадрат; гибридные (лингвостатистические): RAKE, YAKE, PullEnti, Topia; структурные (графовые): TextRank; с использованием машинного обучения (KeyBERT). Лингвистические методы, основанные на лексико-грамматических шаблонах ключевых выражений, использовались как компонент гибридных методов. Рассматриваемый набор методов не является исчерпывающим [22–25].

При отборе методов выделения ключевых выражений мы учитывали возможность их применения в работе с русскоязычными текстами, в также возможность извлечения n-грамм разной структуры (униграмм, биграмм, триграмм и т.д.). Основные методы выделения ключевых выражений учитывают не только их типичность для определенного документа или классов документов, но и их коллокационную природу. В традиционном понимании коллокацией считается устойчивое сочетание двух или более токенов или лемм, проявляющих тенденцию к совместной встречаемости [26–29]. Поэтому ожидается, что ключевые выражения проявляют значимую степень устойчивости [30–31].

Статистические методы

Мера TF-IDF [32–33] (Term Frequency – Inverse Document Frequency) определяет, в какой мере данное выражение характерно для документа внутри корпуса. Большой вес TF-IDF получают слова с высокой частотой в конкретном документе и при этом с низкой частотой в других документах.

Мера ассоциации Log-likelihood (логарифмическая функция правдоподобия) считается классическим показателем силы синтагматической связи между элементами коллокаций. Оперируя наблюдаемыми значениями параметров данных и их вероятностной моделью, можно получить ожидаемые значения параметров, максимально приближенные к реальным.

Критерий Хи-квадрат [34–35] как метод выделения ключевых выражений не требует использования фоновых корпусов текстов или набора сравниваемых документов. Метод основан на построении матрицы совместной встречаемости по тексту и позволяет сократить долю низкочастотных слов, которые получали бы неоправданно высокое значение в силу разреженности матрицы.

В настоящем исследовании были использованы алгоритмические реализации статистических методов в библиотеке NLTK[1]. Существуют и другие статистические методы (T-score, C-value и т.д.), не рассматриваемые в наших экспериментах.

Гибридные (лингвостатистические) методы

Гибридный алгоритм RAKE (Rapid Automatic Keyword Extraction)[2] [36–38], основан на предположении о том, что ключевые элементы могут быть неоднословными, не содержат знаков пунктуации, служебных и десемантизированных слов. Ключевые выражения, выделенные в тексте с учетом словаря разделителей, ранжируются по весу, определяемого как сумма трех метрик (частота, степень (мера совместной встречаемости), отношение частоты к степени).

Алгоритм YAKE (Yet Another Keyword Extractor)[3] [39] во многом сходен с алгоритмом RAKE. Особенностью YAKE является то, что в нем веса ключевых выражений определяются по комбинации из пяти метрик (нормированные частота, местоположение в тексте, число предложений с выражением, число капитализированных употреблений, сходство со стоп-словами).

[1] https://pypi.org/project/nltk/
[2] https://pypi.org/project/rake-nltk/
[3] https://pypi.org/project/yake/

Лингвистический процессор PullEnti[4] разработан для извлечения фактов из корпусов текстов. Процессор является одним из лучших в своем классе [40] благодаря тому, что в нем реализован надежный алгоритм выделения конструкций с учетом морфологического и семантико-синтаксического анализа и с детализацией типов фактической информации. Распознаваемые в PullEnti конструкции являются кандидатами в ключевые выражения.

Алгоритм Topia[5] обеспечивает автоматическое выделение ключевых выражений в тексте на основе процедур токенизации и морфологического анализа корпуса текстов. Важные для текста выражения выделяются в результате применения системы правил и количественного анализа текстов для определения силы связей внутри выражений-кандидатов.

Известны иные гибридные подходы (подход к выделению ключевых выражений в системе SketchEngine и ряд других, в том числе, использующий словарные ресурсы типа WordNet), которые выходят за рамки нашего исследования.

Структурные (графовые) методы

Алгоритм TextRank[6] [41–42] относится к классу неконтролируемых методов ранжирования на графах и является модификацией алгоритма PageRank для ранжирования страниц в поисковой выдаче. Суть подхода TextRank состоит в построении взвешенного графа, в вершинах которого размещаются токены, леммы или фразы, ребра соответствуют связям внутри текста и имеют веса – оценки силы связей и/или метки типов семантических связей. Вершины ранжируются по значению PageRank, полученные кандидаты в ключевые выражения будут иметь высокий ранг. Наряду с TextRank могут применяться и другие графовые подходы, например, DegExt.

Методы с использованием машинного обучения

Контекстуализированная предсказывающая модель распределенных векторов BERT [43] является двунаправленным трансформером, который позволяет преобразовывать предложения и документы в векторы, отражающие их значение. В основе метода KeyBERT [44] лежит процедура определения косинусной близости векторов потенциальных ключевых выражений по отношению к тексту в целом. Кроме KeyBERT существуют иные методы выделения ключевых выражений с машинным обучением, в частности, алгоритм KEA, основанный на вероятностной модели классификации.

Корпусные данные

Исследовательский корпус, задействованный в нашем исследовании, состоит из текстовых документов, отвечающих следующим требованиям:

(1) тексты принадлежат разным функциональным стилям (публицистический, научный, художественный);

(2) тексты снабжены сжатым представлением их содержания (для публицистического подкорпуса – развернутые заголовки статей и лид, для научного подкорпуса – заголовок, аннотация, список ключевых выражений, для художественного подкорпуса – выделенные автором сюжетные составляющие каждой главы).

В каждом подкорпусе содержится 150 текстов (50 новостей, 50 научных статей и 50 глав художественных произведений) общим объемом ~ 1 млн словоупотреблений до предобработки.

Планирование эксперимента

В ходе эксперимента в каждом из 150 текстов был выделен начальный фрагмент (для публицистического подкорпуса – заголовок и первые два-три предложения новости, для научного – название, аннотация, ключевые выражения и первый абзац статьи, для художественного – размеченные автором опорные слова и первый абзац главы). Экспертам была предложена инструкция

[4] https://pypi.org/project/pullenti/
[5] https://pypi.org/project/topia.termextract/
[6] https://pypi.org/project/pytextrank/

по традиционной методике А.С. Штерн [45] с некоторыми правками: "*Прочитайте текст. Подумайте над его содержанием. Выделите 3-7 ключевых выражений, ранжируя их от самого важного к менее важному*". Экспертам были даны инструкции по распознаванию ключевых выражений в текстах разных функциональных стилей:

1) **публицистический стиль:** ключевым выражением для новостных текстов является именная группа, наиболее ярко и полно отражающая суть обозначенной в заголовке текста;

2) **научный стиль:** ключевым выражением для новостных текстов являются термины, именные группы использованных методов и т.д. авторами статьи, а также предмет и объект, обозначенные в статье;

3) **художественный стиль:** ключевым выражением для новостных текстов являются обобщения описываемых автором событий.

В эксперименте по ручной разметке ключевых выражений приняли участие пять экспертов-информантов (далее обозначаются как ЭКС1, ЭКС2, ЭКС3, ЭКС4, ЭКС5). Результаты экспертизы приняты за эталонную разметку. Из оставшихся частей текстовых документов ключевые выражения извлекались автоматическими методами. Такое деление текстов для ручного и автоматического извлечения ключевых выражений было сделано намеренно по причине того, что автоматическое извлечение ключевых выражений из первой части текста – это процедура с ожидаемо положительным результатом, в то время как распознавание их в оставшихся частях текста – это нетривиальная задача, решение которой подтверждает роль тех ключевых выражений, которые выявляются в начале текста. Особенностью проведения эксперимента является то, что тексты перед автоматической обработкой не подвергались лемматизации и фильтрации по стоп-словам, это позволяет сохранить их в том виде, в каком они передавались для оценки испытуемым. В табл. 1 приведены параметры экспериментов.

Таблица 1. Параметры экспериментов
Table 1. Experimental parameters

Параметр	**Работа экспертов**		**Работа алгоритмов**	
Объем текста	*Публицистический*	Развернутый заголовок и лид	*Публицистический*	Максимум 10 предложений
	Научный	Название, аннотация, список ключевых выражений, первый абзац	*Научный*	Основные содержательные разделы и заключение
	Художественный	Авторская аннотация к главе и первый абзац текста	*Художественный*	Основной текст главы
Длина ключевого выражения	Ограничений нет		Ограничения зависят от метода	
Объем списка ключевых выражений	3...7		5...1000	
Способ ранжирования ключевых выражений	По убыванию важности		По убыванию важности	

Результаты экспериментов

Обработка публицистического подкорпуса

Тексты из публицистического подкорпуса характеризуются краткостью, точностью изложения фактов, наличием лида – начального предложения или абзаца, призванного привлечь внимание читателя к содержанию новости. В качестве источников для создания подкорпуса исполь-

зовались новостные порталы «Бумага»[7] и «Медуза»[8], откуда были случайным образом отобраны 50 новостных сообщений. Ниже приведен пример текста публицистического подкорпуса.

“*В Петербурге возобновили плановую вакцинацию детей. Ее приостанавливали из-за коронавируса*”: *В Петербурге сняли запрет на плановую вакцинацию детей, введенный в начале апреля. Постановление главного санитарного врача опубликовано на сайте Роспотребнадзора*.

Вакцинация взрослых пока остановлена. Как пояснили в комитете по здравоохранению, она проводится лишь по эпидемическим показаниям. Например, в поликлинике можно сделать прививку против клещевого энцефалита. Ранее Минздрав приостановил плановую вакцинацию детей и взрослых из-за коронавируса. Пояснялось, что решение не касается прививок новорожденным. Актуальные новости о распространении COVID-19 в городе читайте в рубрике «Бумаги» «Коронавирус в Петербурге».[9]

Экспертам было предложено выделить ключевые выражения из первого абзаца текста (выделен подчеркиванием), автоматическая процедура извлечения ключевых выражений проводилась в отношении второго абзаца текста. В табл. 2 представлены ключевые выражения, полученные в результате оценки экспертов (ЭКС1, ЭКС2, ЭКС3, ЭКС4, ЭКС5), а в табл. 3 – с помощью автоматических методов выделения ключевых выражений.

Таблица 2. Ключевые выражения для текста публицистического подкорпуса, извлеченные экспертами
Table 2. Key expressions for the text of the news subcorpus extracted by experts

ЭКС1	ЭКС2	ЭКС3	ЭКС4	ЭКС5
вакцинация, Роспотребнадзор, коронавирус	*Петербург, коронавирус, вакцинация, детская медицина*	*вакцинация детей, сняли запрет, Петербург, Роспотребнадзор*	*возобновили вакцинацию детей*	*сняли запрет на вакцинацию*

В табл. 3 полужирным шрифтом отмечены точные совпадения автоматически выделенных ключевых выражений (без учета морфологических форм) по сравнению с выражениями, определенными экспертами, а подчеркиванием отмечены их смысловые корреляты или целостные конструкции: экспертная разметка: *вакцинация*, *вакцинация детей* – автоматическая разметка: *плановая вакцинация*, *плановая вакцинация детей*; экспертная разметка: *вакцинация*, *вакцинация детей* – автоматическая разметка: *прививка*, *сделать прививку*, *прививку новороженным*; экспертная разметка: *Петербург*, *коронавирус* – автоматическая разметка: *коронавирус в Петербурге*, и т.д. Ключевые выражения, отмеченные экспертами, воспроизводятся напрямую или косвенно в выдаче алгоритмов Log-Likelihood, TextRank (4 совпадения), PullEnti, YAKE, TF-IDF, KeyBERT (2 совпадения), Topia, RAKE, Хи-Квадрат (по 1 совпадению). Малозначимыми с точки зрения оценки выражений как ключевых являются отдельные прилагательные (*эпидемический*), глагольные формы (*приостановлена*), конструкции с обобщенным содержанием (*актуальные новости, пояснялось решение*) и т.д. Особенностью новостного текста является его соотнесенность с конкретной ситуацией, в которой описываются реальные события и их участники, поэтому приоритет в выборе алгоритмов выделения ключевых выражений будет за теми алгоритмами, которые позволяют распознавать в тексте его фактическое содержание.

Обработка научного подкорпуса

В научный подкорпус вошли тексты 50 статей по корпусной лингвистике, представляющие научный стиль изложения и описывающие экспериментальные исследования. Данным текстам свойственны высокая терминологичность, однозначность, преобладание именных конструкций.

[7] https://paperpaper.ru/
[8] https://meduza.io/
[9] https://vk.com/@paperpaper_ru-preview-1581515435-1847493496?ysclid=la3x9pfzyd839380455

Таблица 3. Ключевые выражения
для текста публицистического подкорпуса, извлеченные автоматически
Table 3. Key expressions for the text of the news subcorpus, extracted automatically

Алгоритм	Ключевые слова и выражения (первые пять результатов)
TF-IDF	***коронавирус** 0.35, **вакцинацию** 0.34, взрослых 0.22, энцефалита 0.18, COVID 0.18 ...*
Log-likelihood	***вакцинацию детей** 32.99, плановую вакцинацию 32.99, пояснили [в] комитете19.32, сделать прививку 19.31, прививок новорожденным 19.31 ...*
Хи-квадрат	*актуальные новости 93.0, решение касается 93.0, пояснялось решение 93.0, прививку клещевого 93.0, прививок новорожденным 93.0 ...*
RAKE	*эпидемическим показаниям 4.0, клещевого энцефалита 4.0, плановую вакцинацию 4.0, актуальные новости 4.0, вакцинация взрослых 3.5 ...*
YAKE	*вакцинация взрослых 0.05, взрослых пока остановлена 0.10, коронавирус в Петербурге 0.12, **вакцинация** 0.12, остановлена 0.12 ...*
PullEnti	*прививок против клещевого энцефалита, плановая вакцинация детей, министерство здравоохранения, вакцинация взрослых, **вакцинация** ...*
Topia	***коронавирус** 2, вакцинация взрослых 1, комитет 1, здравоохранение 1, эпидемические показания 1 ...*
TextRank	***коронавируса**, **коронавирус**, вакцинация взрослых, **вакцинацию**, новорожденным ...*
KeyBERT	*здравоохранению 0.84, **коронавируса** 0.81, эпидемическим 0.81, **коронавирус** 0.80, распространении 0.79...*

При анализе статей были сохранены заголовки, аннотации, ключевые выражения и все текстовые разделы. Ниже приведен пример анализа одного из текстов научного подкорпуса.

И.В. Азарова, Е.Л. Алексеева

ОТ КРИТИЧЕСКОГО ИЗДАНИЯ К СТРУКТУРИРОВАННОМУ КОРПУСУ СЛАВЯНСКИХ ВАРИАНТОВ ЕВАНГЕЛИЯ

Аннотация. В статье рассматривается создание корпуса текстов на базе издания славянского Евангелия, включающего 28 рукописей, представляющих 8 групп славянских списков. Для обеспечения возможности поиска по разным типам текстовых фрагментов предлагается преобразование данных в корпус размеченных типизированных текстов.

Ключевые слова: Корпус, славянское Евангелие, Паратекст.

1. Проект издания славянского евангелия

Работа по проекту началась в 1993 г. при финансовой поддержке Немецкого библейского общества, выделившего средства для критического издания Евангелия от Иоанна, которое вышло в свет в 1998 г. Затем за счет Синодальной библиотеки Московского Патриархата было подготовлено издание Евангелия от Матфея, опубликованное в 2005 г. при поддержке РГНФ]. В настоящее время СПбГУ финансирует работу над Евангелиями от Марка и Луки и подготовку итогового издания в двух томах, содержащего критический текст всех четырех евангелий и результаты научного исследования материала.

План исследования славянского Евангелия был обоснован теоретически А.А. Алексеевым и описан в виде практической процедуры в [].

Была проведена коллация двух фрагментов из Евангелия от Иоанна и Евангелия от Марка по 1500 рукописям с базовым текстом издания – Мариинским Евангелием. Данные коллаций были сведены в структуру узлов разночтений, по которым был проведен автоматический кластерный анализ, позволивший выделить 8 неравных по объему групп. В издании представлены 28 рукописей как представители групп списков, что существенно упрощает нахождение чтений, которыми они противопостав-

лены друг другу, позволяет устанавливать генетические отношения между группами, реконструировать архетип, выдвигать текстологические гипотезы...[10]

Эксперты выделили ключевые выражения из заголовка, аннотации, списка ключевых слов и первого абзаца (выделены подчеркиванием), в оставшемся фрагменте статьи, включающем содержательные разделы и заключение, ключевые выражения были выделены автоматически. В табл. 4 представлены ключевые выражения, полученные в результате оценки экспертов (ЭКС1, ЭКС2, ЭКС3, ЭКС4, ЭКС5), в табл. 5 – с помощью автоматических методов.

Таблица 4. Ключевые выражения для текста научного подкорпуса, извлеченные экспертами
Table 4. Key expressions for the text of the scientific subcorpus extracted by experts

ЭКС1	ЭКС2	ЭКС3	ЭКС4	ЭКС5
Евангелие, корпус, поиск, разметка	*корпус текстов, поиск текстовых фрагментов, Евангелие*	*Евангелия от Матфея, Проект издания славянского Евангелия, СПбГУ финансирует работу*	*Славянское Евангелие, создание корпуса, финансирование, поддержка*	*текст четырех Евангелий издается в двух томах*

Таблица 5. Ключевые выражения для текста научного подкорпуса, извлеченные автоматически
Table 5. Key expressions for the text of the scientific subcorpus, extracted automatically

Алгоритм	Ключевые слова и выражения (первые пять результатов)
TF-IDF	***Евангелия*** *0.31, текста 0.22, рукописей 0.18, фрагментов 0.13, славянских 0.10 ...*
Log-likelihood	*структурированный корпус 23.33, узлов разночтений 19.51, фрагментов текста 16.64, каждой [из] рукописей 16.60,* ***славянского Евангелия*** *13.794 ...*
Хи-квадрат	*автоматический кластерный 302.0, проведена коллация 302.0, сведены [в] структуру 302.0, рукописям базовым 302.0, реконструировать архетип 302.0 ...*
RAKE	*подстрочное представление базового Мариинского Евангелия 20.83, проверку итоговых данных издания 13.67, стихам соответствующих глав Евангелия 12.83, структурированный корпус славянских текстов 11.33, автоматический кластерный анализ 9.0 ...*
YAKE	***Евангелия*** *0.03, обоснован теоретически 0.04, план исследования славянского 0.05,* ***славянского Евангелия*** *0.05, текста 0.061 ...*
PullEnti	*текст, издание, рукопись, фрагмент, структурированный корпус ...*
Topia	*структурированный корпус 3, замена 3,* ***славянское Евангелие*** *2, вид 2,* ***Евангелие*** *2 ...*
TextRank	*текста,* ***Евангелием****, славянских текстов, издании, рукописей*
KeyBERT	*Четвероевангелия 0.39, Евангелиями 0.38 подтверждения 0.38 сопоставления;0.36 исследования 0.34...*

В табл. 5 полужирным шрифтом отмечены точные совпадения автоматически выделенных ключевых выражений (без учета морфологических форм) по сравнению с выражениями, определенными экспертами, а подчеркиванием отмечены их смысловые корреляты или целостные конструкции: экспертная разметка: текстовых фрагментов – автоматическая разметка: фрагментов текста; экспертная разметка: корпус, корпус текстов – автоматическая разметка: структурированный корпус, структурированный корпус славянских текстов, текст; экспертная разметка: текст

[10] Азарова И.В., Алексеева Е.Л. От критического издания к структурированному корпусу славянских вариантов Евангелия // Труды международной конференции «Корпусная лингвистика – 2015». СПб., 2015.

четырех Евангелий – автоматическая разметка: Четвероевангелие, и т.д. Ключевые выражения, отмеченные экспертами, в том или ином виде воспроизводятся в выдаче алгоритмов PullEnti, TF-IDF, TextRank (5 совпадений), Log-Likelihood, YAKE, Topia (3 совпадения), KeyBERT (2 совпадения), RAKE (1 совпадение). Для выдачи алгоритма Хи-Квадрат совпадений не обнаружено: все ключевые выражения, выделенные этим способом, описывают внутренние процедуры обработки корпуса, не упомянутые в аннотации. Поскольку научный текст имеет строгую композицию и регламентированный словарь, включающий терминологию предметной области, при выборе алгоритма автоматического выделения ключевых выражений следует отдавать тем методам, которые высокочувствительны к терминам и терминосочетаниям.

Обработка художественного подкорпуса

В художественный подкорпус вошли 50 текстов глав юмористических повестей Джерома К. Джерома «Трое на четырех колесах» (перевод М. Жаринцовой)[11], «Трое в одной лодке, не считая собаки» (перевод М. Салье)[12]; романа-робинзонады Жюля Верна «Таинственный остров» (перевод Н. Немчиновой, А. Худадовой)[13]. Выбор данных источников определен тем, что отдельные главы в них снабжены авторским кратким описанием содержания.

Ниже приводится пример анализа одного из текстов художественного подкорпуса. В табл. 6 представлены ключевые выражения, полученные в результате оценки экспертов (ЭКС1, ЭКС2, ЭКС3, ЭКС4, ЭКС5), а в табл. 7 – с помощью автоматических методов выделения ключевых выражений.

Таблица 6. Ключевые выражения для текста художественного подкорпуса, извлеченные экспертами
Table 6. Key expressions for the text of the fiction subcorpus extracted by experts

ЭКС1	ЭКС2	ЭКС3	ЭКС4	ЭКС5
Рэдинг Этельред набег	*Рэдинг, Эльтеред, Альфред, история городка*	*остановка по дороге, город, история*	*старинный, знаменитый городок*	*Рэдинг – старинный городок, основанный в дни короля Этельреда*

В табл. 7 полужирным шрифтом отмечены точные совпадения автоматически выделенных ключевых выражений (без учета морфологических форм) по сравнению с выражениями, определенными экспертами. В отличие от текстов в публицистическом и научном корпусах смысловые корреляты целостные конструкции, аналогичные экспертной разметке, не были выявлены. Между результатами ручной и автоматической разметки ключевых выражений зарегистрированы единичные совпадения: PullEnti, TF-IDF, TextRank, Topia, YAKE, KeyBERT (1 совпадение). Для выдачи алгоритмов Log-Likelihood, Хи-Квадрат, RAKE совпадения не найдены. Это объясняется композицией художественного текста и выразительными средствами, используемыми при его создании: сюжет произведения оказывается существенно богаче, чем его лаконичное описание, предваряющее повествование.

Обработка результатов экспериментов

Процедура обработки результатов экспериментов заключается в проверке совпадения ключевых выражений, выделенных автоматически из второй части текста, с выражениями, выделенными экспертами из первой части текста.

С одной стороны, несовпадение экспертной и автоматической разметки ключевых выражений подтверждает гипотезу об их пространственно-позиционных свойствах. С другой стороны, частичное пересечение списков ключевых выражений, назначенных экспертами и сформиро-

[11] http://lib.ru/JEROM/chetwero.txt
[12] http://lib.ru/JEROM/troe_w_lodke.txt
[13] http://lib.ru/INOFANT/VERN/ostrow.txt

Таблица 7. Ключевые выражения для текста художественного подкорпуса, извлеченные автоматически
Table 7. Key expressions for the text of the fiction subcorpus, extracted automatically

Алгоритм	Ключевые слова и выражения (первые пять результатов)
TF-IDF	***Рэдинга** 0.10, река 0.08, Стритли 0.07, Горинг 0.07, баркаса 0.04 ...*
Log-likelihood	*шесть шиллингов 23.03, шиллингов неделю 23.03, приблизились нему 14.81, приняла объятия 14.81, принц Оранский 14.81 ...*
Хи-квадрат	*баркасами них 606.0, прибрежных городов 606.0, принялся доказывать 606.0, приняла объятия 606.0, принц Оранский 606.0 ...*
RAKE	*руку коробку грошовых конфет 16.0, какую-нибудь пустяковую чуму 9.0, завсегдатаям картинных выставок 9.0, отпечаток безмятежного спокойствия 9.0, охваченные справедливым негодованием 9.0...*
YAKE	*[в] Лондоне становилось скверно 0.001, [в] Вестминстере объявлялась чума 0.01, [в] Лондоне становилось 0.01, **Рэдинг** всякий 0.02, поздние годы 0.019...*
PullEnti	*Лондон, **город**, парламент, Генрих, Карл*
Topia	***Рэдинга** 5, Горинг 5, время 4, река 4, лодка 3 ...*
TextRank	*реке, **Рэдингом**, Стритли, баркасом, лодки...*
KeyBERT	*Вестминстере, 0.88, **городов** 0.88, равнодушием, 0.87, местечки 0.87, принадлежащим 0.87 ...*

ванных автоматическими методами, позволит определить ожидаемую концентрацию ключевых выражений в основной части текста. Поскольку в исследовании используется девять методов автоматического выделения ключевых выражений, сравнение результатов их работы позволит определить их чувствительность к определению ключевых выражений в нетипичных фрагментах текстов.

В трех сериях экспериментов с публицистическим, научным и художественным корпусами текстов были получены данные о количественном соотношении ключевых выражений, выделенных экспертами в первой части текстов и автоматически определенных во второй части текстов. Данные оценки были обобщены для каждого из 150 текстов в трех подкорпусах и для каждого из девяти методов автоматического выделения ключевых выражений. В итоге были рассчитаны средние значения доли пересечений между ключевыми выражениями, отобранными в ходе эксперимента с испытуемыми и в ходе применения автоматических методов.

Наилучшие показатели у лингвистического процессора PullEnti (усредненная доля совпадений в публицистических текстах 1,8 слова на текст, в научных текстах – 2,32 слова на текст, в художественных текстах – 1,82 слова на текст). Самые низкие оценки продемонстрировал алгоритм KeyBERT (усредненная доля совпадений в публицистических текстах 0,26 слова на текст, в научных текстах – 0,06 слова на текст, в художественных текстах – 0,08 слова на текст). Остальные методы занимают промежуточное положение.

Интерпретация полученных результатов состоит в том, что метод выделения ключевых выражений, реализованный в процессоре PullEnti, совмещает большое число лингвистических параметров текстов и их статистические характеристики, тогда как алгоритм KeyBERT, основанный на модели распределенных векторов типа Трансформер и использующий многоэтапную процедуру сокращения размерности, обладает сверхобобщающей способностью, вследствие чего может уступать по качеству методам, использующим более простые вычислительные приемы и системы правил. Лингвистические факторы, подтверждающие данные наблюдения, это тенденция метода PullEnti к выделению именных униграмм или коллокаций, выражающих фактическую инфор-

Рис. 1. Диаграмма количественных оценок совпадений ключевых выражений в публицистическом подкорпусе

Fig. 1. Diagram of quantitative estimates of the coincidence of key expressions in the news subcorpus

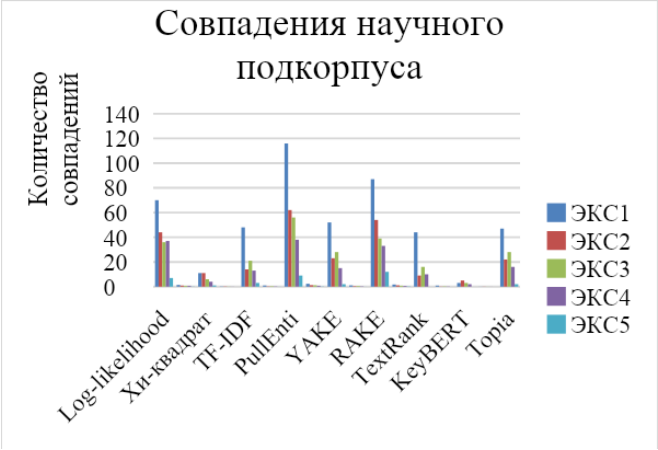

Рис. 2. Диаграмма количественных оценок совпадений ключевых выражений в научном подкорпусе

Fig. 2. Diagram of quantitative estimates of the coincidence of key expressions in the scientific subcorpus

Рис. 3. Диаграмма количественных оценок совпадений ключевых выражений в художественном подкорпусе

Fig. 3. Diagram of quantitative estimates of the coincidence of key expressions in the fiction subcorpus

мацию (одна из задач процессора – выделение именованных сущностей), и смешанная выдача алгоритма KeyBERT, включающая абстрактную лексику, глагольные формы, и кроме этого, автоматически устанавливающая ограничения на число предсказываемых ключевых выражений.

На рис. 1–3 представлены диаграммы с количественными оценками совпадений экспертной и автоматической разметки ключевых выражений.

Заключение

В ходе экспериментов по сравнению ручной и автоматической разметки ключевых выражений в русскоязычных корпусах текстов была подтверждена гипотеза о возможности автоматического распознавания в основной части текста тех ключевых выражений, которые были выделены экспертами вручную в начальной части текста. Это обосновывает возможность автоматизации выделения ключевых выражений в исследованиях, направленных на моделирование семантической организации корпусов текстов, где ручная разметка неприменима. В то же время, наличие расхождений между экспертными и автоматически сгенерированными наборами ключевых выражений объясняется стилистическими особенностями текстов (публицистическом, научным и худоджественных), а также различиями в работе алгоритмов, использованных в экспериментах (Log-Likelihood, TF-IDF, Хи-квадрат, RAKE, YAKE, PullEnti, Topia, структурныеTextRank, KeyBERT). В рамках экспериментов были впервые получены сопоставительные данные о результатах применения данного набора алгоритмов на материале русскоязычных текстов.

Полученные результаты исследования представляют высокую ценность для исследований, проводимых в области семантической компрессии текстов, в особенности, в сфере автоматического моделирования тематики корпусов текстов с использованием мультимодальных тематических моделей [45–48].

СПИСОК ИСТОЧНИКОВ

1. **Барахнин В.Б., Ткачев Д.А.** Кластеризация текстовых документов на основе составных ключевых термов // Вестник НГУ. Серия: Информационные технологии. 2010. Т. 8 (2). С. 5–14. URL: https://cyberleninka.ru/article/n/klasterizatsiya-tekstovyh-dokumentov-na-osnove-sostavnyh-klyuchevyh-termov?ysclid=laftym3w15838783872

2. **Гуляев О.В., Лукашевич Н.В.** Автоматическая классификация текстов на основе заголовка рубрики // Новые информационные технологии в автоматизированных системах. 2013. Т. 16. С. 238–244. URL: https://cyberleninka.ru/article/n/avtomaticheskaya-klassifikatsiya-tekstov-na-osnove-zagolovka-rubriki?ysclid=lafu0o2df9558148953

3. **Москвитина Т.Н.** Методы выделения ключевых слов при реферировании научного текста // Вестник Томского государственного университета. 2018. Т. 8 (197). С. 45–50. URL: https://vestnik.tspu.edu.ru/files/vestnik/PDF/articles/moskvitina_t._n._45_50_8_197_2018.pdf?ysclid=lafuynavpj156594902

4. **Попова С.В., Данилова В.В.** Представление документов в задаче кластеризации аннотаций научных текстов // Научно-технический вестник информационных технологий, механики и оптики. 2014. Т. 1. № 89. С. 99–107. URL: https://cyberleninka.ru/article/n/predstavlenie-dokumentov-v-zadache-klasterizatsii-annotatsiy-nauchnyh-tekstov?ysclid=lafuznumz3740161495

5. **Шмулевич М.М., Пивоваров В.С., Киселев М.В.** Метод кластеризации текстов, учитывающий совместную встречаемость ключевых терминов, и его применение к анализу тематической структуры новостного потока, а также ее динамики. 2005. URL: http://elar.urfu.ru/handle/10995/1421

6. **Мурзин Л.Н., Штерн А.С.** Текст и его восприятие. Свердловск, 1991.

7. **Сахарный Л.В.** Тексты-примитивы и закономерности их порождения // Человеческий фактор в языке: язык и порождение речи. М., 1991. URL: https://elibrary.ru/item.asp?id=27284933&pff=1&ysclid=lafv4co0ns297089297

8. **Сахарный Л.В., Сибирский С.А., Штерн А.С.** Набор ключевых слов как текст // Психолого-педагогические и лингвистические проблемы исследования текста. Пермь, 1984. URL: https://elibrary.ru/item.asp?id=30045397&ysclid=lafv5bewh8847279895

9. **Филиппов К.А.** Лингвистика текста. СПб., 2003.

10. **Петрова Т.Е.** Контекстная предсказуемость ключевых слов текста. СПб., 2006.

11. **Ягунова Е.В.** Вариативность стратегий восприятия звучащего текста (экспериментальное исследование на материале русскоязычных текстов разных функциональных стилей). Пермь, 2008.

12. **Ягунова Е.В.** Набор опорных слов как вид свертки текста (в сопоставлении с набором ключевых слов // Язык и речевая деятельность. 2008. Т. 8. С. 225–235. URL: https://www.dialog-21.ru/media/1811/91.pdf

13. **Петрова Т.Е., Риехакайнен Е.И., Кузнецова А.С., Мараев А.В., Шаталов М.А.** Выделение ключевых слов в вербальных и невербальных паттернах // Социо- и психолингвистические исследования. Вып. 5, 2017. С. 149–156. URL: https://cyberleninka.ru/article/n/vydelenie-klyuchevyh-slov-v-verbalnyh-i-neverbalnyh-patternah?ysclid=lafv8ev06b194106481

14. **Абрамов Е.Г.** Подбор ключевых слов для научной статьи // Научная периодика: проблемы и решения. 2011. № 2. С. 35–40. URL: https://cyberleninka.ru/article/n/podbor-klyuchevyh-slov-dlya-nauchnoy-stati?ysclid=lafv9byvdz763665197

15. **Ванюшкин А.С., Гращенко Л.А.** Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. Т. 19. С. 85–93. URL: https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov?ysclid=lafva1k1bi186577723

16. **Дубинина Е.Ю.** 2020 Выделение ключевых слов текста научной статьи в процессе создания автоматического реферата // Вестник ВГУ. Серия: Филология. Журналистика. 2020. № 1. С. 26–28. URL: http://www.vestnik.vsu.ru/pdf/phylolog/2020/01/2020-01-06.pdf

17. **Тихонова Е.В., Косычева М.А.** Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology, 2022. Vol. 3 (4). Pp. 7–15. URL: https://cyberleninka.ru/article/n/effektivnye-klyuchevye-slova-strategii-formulirovaniya?ysclid=lafvbrf2bd330329275

18. **Шереметьева С.О., Осминин П.Г.** Методы и модели автоматического извлечения ключевых слов // Вестник Южно-Уральского государственного университета. № 12 (1). 2015. С. 76–81. URL: https://cyberleninka.ru/article/n/metody-i-modeli-avtomaticheskogo-izvlecheniya-klyuchevyh-slov?ysclid=lafvcmbxhe291903213

19. **Камшилова О.Н.** Анализ выбора и состава списков ключевых слов (по материалам научных публикаций) // Прикладная лингвистика в науке и образовании. СПб., 2012. С. 136–142. URL: https://elibrary.ru/item.asp?id=21530271&ysclid=lafvdyd3ia359073087

20. **Камшилова О.Н.** Малые формы научного текста: ключевые слова и аннотация (информационный аспект) // Известия Российского государственного педагогического университета им. А.И. Герцена. № 156. 2013. С. 106–117. URL: https://cyberleninka.ru/article/n/malye-formy-nauchnogo-teksta-klyuchevye-slova-i-annotatsiya-informatsionnyy-aspekt

21. **Kamshilova O., Beliaeva L., Geikhman L.** Author's Choice for Keyword List: Research Aspect // PRLEAL-2019. R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL–2019). CEUR Workshop Proceedings. Saint Petersburg, Russia, November 27, 2019. 2020. Pp. 47–59. URL: http://ceur-ws.org/Vol-2552/Paper5.pdf

22. **Браславский П., Соколов Е.** Сравнение пяти методов извлечения терминов произвольной длины // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог–2008». М., 2008. URL: https://www.dialog-21.ru/media/1732/11.pdf

23. **Виноградова Н.В., Иванов В.К.** Современные методы автоматизированного извлечения ключевых слов из текста // Информационные ресурсы России. 2016(4). С. 13–18. URL: https://elibrary.ru/item.asp?id=27036419&ysclid=lafvhiu0wr488919615

24. **Лукашевич Н.В., Логачев Ю.М.** Комбинирование признаков для автоматического извлечения терминов // Вычислительные методы и программирование. 2010. Т. 11 (4). С. 108–116. URL: https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=vmp&paperid=345&option_lang=rus&ysclid=lafvih6e1i417544486

25. **Papagiannopoulou E., Tsoumakas G.** A Review of Keyphrase Extraction // https://arxiv.org/pdf/19-05.05044.pdf

26. **Захаров В.П., Хохлова М.В.** Выделение терминологических словосочетаний из специальных текстов на основе различных мер ассоциации // XVII Всероссийская объединенная конференция «Интернет и современное общество» (IMS-2014). СПб., 2014. URL: https://ojs.itmo.ru/index.php/IMS/article/view/268

27. **Залесская В.В.** Программа выявления в тексте двучленных статистически значимых осмысленных коллокаций (на материале русского языка) // XVII Всероссийская объединенная конференция «Интернет и современное общество» (IMS-2014). СПб., 2014. URL: https://ojs.itmo.ru/index.php/IMS/article/download/267/263

28. **Brezina V., McEnery T., Wattam S.** Collocations in context: A new perspective on collocation networks // International Journal of Corpus Linguistics. Vol. 20 (2). August 2015. DOI: 10.1075/ijcl.20.2.01bre

29. **Evert S.** Corpora and collocations // Corpus Linguistics. An International Handbook. Article 58. Mouton de Gruyter, 2008. Pp. 1212–1248. URL: https://stephanie-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf

30. **Андреева Д., Митрофанова О.А.** Эксперименты по кластеризации русскоязычных новостных текстов на основе списков лексических конструкций // Структурная и прикладная лингвистика. Вып. 13. СПб., 2019. С. 141–157. URL: https://elibrary.ru/item.asp?id=44751350&ysclid=lafvm63kog77269386

31. **Букия Г.Т.** Автоматическая кластеризация новостных сообщений с опорой на ключевые слова и биграммные конструкции // Структурная и прикладная лингвистика. Вып. 12. СПб., 2018. URL: https://elibrary.ru/item.asp?id=41085947&ysclid=lafvnf50p0798429448

32. **Luhn H.P.** A Statistical Approach to Mechanized Encoding and Searching of Literary Information // IBM Journal of Research and Development. Vol. 4. Issue 1. 1957. Pp. 309–317. URL: http://openlib.org/home/krichel/courses/lis618/readings/luhn57.pdf

33. **Luhn H.P.** The Automatic Creation of Literature Abstracts // IBM Journal of Research and Development.Vol. 2. Issue 2. 1958. Pp. 159–165. URL: https://courses.ischool.berkeley.edu/i256/f06/papers/luhn58.pdf

34. **Красавина В.Д., Мирзагитова А.Р.** Оптимизация поиска в системе LeadScanner с помощью автоматического выделения ключевых слов и словосочетаний // Труды международной конференции «Корпусная лингвистика–2015». СПб., 2015. URL: https://events.spbu.ru/eventsContent/files/corpling/corpora2015/Krasavina,%20Mirzagitova.pdf?ysclid=lafvqg594z137492449

35. **Tomokiyo T., Hurst M.** A language model approach to keyphrase extraction // ACL 2003 Workshop on Multiword expressions. 2003. Vol. 18. P. 33–40. URL: https://aclanthology.org/W03-1805.pdf

36. **Москвина А.Д., Ерофеева А.Р., Митрофанова О.А., Харабет Я.К.** Автоматическое выделение ключевых слов и словосочетаний из русскоязычного корпуса текстов с помощью алгоритма RAKE // Труды Международной конференции «Корпусная лингвистика–2017» (Санкт-Петербург, 27–30 июня 2017 г.). Изд-во СПбГУ, 2017. С. 268–275. URL: https://elibrary.ru/item.asp?id=32425675&ysclid=lafvs71ipx62334401

37. **Moskvina A., Sokolova E., Mitrofanova O.** KeyPhrase extraction from the Russian corpus on Linguistics by means of KEA and RAKE algorithm // Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS. С. 369–372. URL: https://elibrary.ru/item.asp?id=41112843&ysclid=lafvsyrid2180400661

38. **Rose S.J., Cowley W.E., Crow V.L., Cramer N.O.** Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. 2009. URL: http://www.google.co.ve/patents/US8131735

39. **Campos R., Mangaravite V., Pasquali A., Jatowt A., Jorge A., Nunes C., Jatowt A.** YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences Journal. Vol. 509. 2020. Pp. 257–289. DOI: 10.1016/j.ins.2019.09.013

40. **Abrosimov K.I., Mosyagina A.G.** Sodner for Russian nested named entity recognition // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference «Dialogue». Issue 21. Moscow, 2022. URL: https://www.dialog-21.ru/media/5744/abrosimovkiplusmosyaginaag037.pdf

41. **Усталов Д.А.** Извлечение терминов из русскоязычных текстов при помощи графовых моделей // CSEDays: Теория графов и приложения. Екатеринбург. 2012. URL: https://scholar.google.com/citations?view_op=view_citation&hl=ru&user=wPD4g7AAAAAJ&citation_for_view=wPD4g7AAAAAJ:3fE2CSJIrl8C

42. **Mihalcea R., Tarau P.** TextRank: Bringing Order into Text // Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing/ Barselona, 2004. Pp. 404–411. URL: https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

43. **Devlin J., Chang M.-W., Lee K., Toutanova K.** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // arxiv.org. 2018. URL: http://arXiv:1810.04805v2

44. **Grootendorst M.** KeyBERT: Minimal keyword extraction with BERT. 2020. URL://https://doi.org/10.5281/zenodo.4461265

45. **Митрофанова О.А.** Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Труды международной конференции «Корпусная лингвистика–2019». СПб., 2019. С. 387–394. URL: https://elibrary.ru/item.asp?id=39449562&ysclid=lafvzcnxb6152046749

46. **Седова А.Г., Митрофанова О.А.** Тематическое моделирование русскоязычных текстов с опорой на леммы и лексические конструкции // Компьютерная лингвистика и вычислительные онтологии: Труды XX Международной Объединенной научной конференции «Интернет и современное общество» (Санкт-Петербург, 21–24 июня 2017 г.). СПб.: Изд-во ИТМО, 2017. С. 132–143. URL: https://openbooks.itmo.ru/ru/file/6518/6518.pdf?ysclid=lafvzxrclb789988181

47. **Mitrofanova O., Kriukova A., Shulginov V., Shulginov V.** E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts – 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Springer Nature, 2021. Pp. 102–114. URL: https://link.springer.com/chapter/10.1007/978-3-030-71214-3_9

48. **Mitrofanova O., Sampetova V., Mamaev I., Moskvina A., Sukharev K.** Topic Modelling of the Russian Corpus of Pikabu Posts: Author–Topic Distribution and Topic Labelling // Proceedings of the International Conference «Internet and Modern Society», IMS 2020. CEUR Workshop Proceedings. 2021. Pp. 101–116. URL: http://ceur-ws.org/Vol-2813/rpaper08.pdf

REFERENCES

[1] **V.B. Barakhnin, D.A. Tkachev,** Clustering of text documents based on compound key terms, Vestnik NSU. Series: Information technologies. 8 (2) (2010) 5–14. Available at: https://cyberleninka.ru/article/n/klasterizatsiya-tekstovyh-dokumentov-na-osnove-sostavnyh-klyuchevyh-termov?ysclid=laftym3w15838783872

[2] **O.V. Gulyaev, N.V. Lukashevich,** Automatic classification of texts based on rubric heading, New information technologies in automated systems. 16 (2013) 238–244. Available at: https://cyberleninka.ru/article/n/avtomaticheskaya-klassifikatsiya-tekstov-na-osnove-zagolovka-rubriki?ysclid=lafu0o2df9558148953

[3] **T.N. Moskvitina,** Methods of highlighting keywords when summarizing a scientific text. Bulletin of the Tomsk State University. 8 (197) (2018) 45–50. Available at: https://vestnik.tspu.edu.ru/files/vestnik/PDF/articles/moskvitina_t._n._45_50_8_197_2018.pdf?ysclid=lafuynavpj156594902

[4] **S.V. Popova, V.V. Danilova,** Representation of documents in the task of clustering annotations of scientific texts, Scientific and technical bulletin of information technologies, mechanics and optics. 1 (89) (2014) 99–107. Available at: https://cyberleninka.ru/article/n/predstavlenie-dokumentov-v-zadache-klasterizatsii-annotatsiy-nauchnyh-tekstov?ysclid=lafuznumz3740161495

[5] **M.M. Shmulevich, V.S. Pivovarov, M.V. Kiselev,** Text clustering method based on the joint occurrence of key terms, and its application to the analysis of the thematic structure of the news flow, as well as its dynamics, 2005. Available at: http://elar.urfu.ru/handle/10995/1421

[6] **L.N. Murzin, A.S. Shtern,** Text and its perception. Sverdlovsk, 1991.

[7] **L.V. Sakharny,** Primitive texts and patterns of their generation, Human factor in language: language and speech generation. M., 1991. Available at: https://elibrary.ru/item.asp?id=27284933&pff=1&ysclid=lafv4co0ns297089297

[8] **L.V. Sakharny, S.A. Sibirsky, A.S. Shtern,** A set of keywords as a text, Psychological-pedagogical and linguistic problems of text research. Perm, 1984. Available at: https://elibrary.ru/item.asp?id=30045397&ysclid=lafv5bewh8847279895

[9] **K.A. Filippov,** Linguistics of the text. SPb., 2003.

[10] **T.E. Petrova,** Contextual predictability of text keywords. SPb., 2006.

[11] **E.V. Yagunova,** Variability of Strategies for Perceiving Oral Text (Experimental Study on the Material of Russian Texts of Different Functional Styles). Perm, 2008.

[12] **E.V. Yagunova,** A set of key words as a type of text convolution (in comparison with a set of keywords, Language and speech activity. 8 (2008) 225–235. Available at: https://www.dialog-21.ru/media/1811/91.pdf

[13] **T.E. Petrova, E.I. Riyekhakaĭnen, A.S. Kuznetsova, A.V. Marayev, M.A. Shatalov,** Identification of keywords in vebral and non-verbal patterns, Socio- and psycholinguistic studies. 5 (2017) 149–156. Available at: https://cyberleninka.ru/article/n/vydelenie-klyuchevyh-slov-v-verbalnyh-i-neverbalnyh-paternah?ysclid=lafv8ev06b194106481

[14] **Ye.G. Abramov,** Selection of keywords for a scientific article, Scientific periodicals: problems and solutions. 2 (2011) 35–40. Available at: https://cyberleninka.ru/article/n/podbor-klyuchevyh-slov-dlya-nauchnoy-stati?ysclid=lafv9byvdz763665197

[15] **A.S. Vanyushkin, L.A. Grashchenko,** Methods and algorithms for extracting keywords, New information technologies in automated systems. 19 (2016) 85–93. Available at: https://cyberleninka.ru/article/n/metody-i-algoritmy-izvlecheniya-klyuchevyh-slov?ysclid=lafva1k1bi186577723

[16] **E.Yu. Dubinina,** Highlighting keywords in the text of a scientific article in the process of creating an automatic abstract, Bulletin of the VSU. Series: Philology. Journalism. 1 (2021) 26–28. Available at: http://www.vestnik.vsu.ru/pdf/phylolog/2020/01/2020-01-06.pdf

[17] **E.V. Tikhonova, M.A. Kosycheva,** Effective keywords: formulation strategies, Health, Food & Biotechnology, 3 (4) (2022) 7–15. Available at: https://cyberleninka.ru/article/n/effektivnye-klyuchevye-slova-strategii-formulirovaniya?ysclid=lafvbrf2bd330329275

[18] **S.O. Sheremetyeva, P.G. Osminin,** Methods and models of automatic extraction of keywords, Bulletin of the South Ural State University. 12 (1) (2015) 76–81. Available at: https://cyberleninka.ru/article/n/metody-i-modeli-avtomaticheskogo-izvlecheniya-klyuchevyh-slov?ysclid=lafvcmbxhe291903213

[19] **O.N. Kamshilova,** Analysis of the choice and composition of keyword lists (based on scientific publications), Applied Linguistics in Science and Education. SPb., 2012. Pp. 136–142. Available at: https://elibrary.ru/item.asp?id=21530271&ysclid=lafvdyd3ia359073087

[20] **O.N. Kamshilova,** Small forms of scientific text: keywords and abstract (informational aspect), Proceedings of the Russian State Pedagogical University named after AI Herzen. 156 (2013) 106–117. Available at: https://cyberleninka.ru/article/n/malye-formy-nauchnogo-teksta-klyuchevye-slova-i-annotatsiya-informatsionnyy-aspekt

[21] **O. Kamshilova, L. Beliaeva, L. Geikhman,** Author's Choice for Keyword List: Research Aspect, PRLEAL-2019. R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Proceedings of the III International Conference on Language Engineering and Applied Linguistics (PRLEAL–2019). CEUR Workshop Proceedings. Saint Petersburg, Russia, November 27, 2019. 2020. Pp. 47–59. Available at: http://ceur-ws.org/Vol-2552/Paper5.pdf

[22] **P. Braslavskiy, E. Sokolov,** Comparison of five methods for extracting terms of arbitrary length, Computer Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue-2008". M., 2008. Available at: https://www.dialog-21.ru/media/1732/11.pdf

[23] **N.V. Vinogradova, V.K. Ivanov,** Modern methods of automated extraction of keywords from the text, Information resources of Russia. (4) (2016) 13–18. Available at: https://elibrary.ru/item.asp?id=27036419&ysclid=lafvhiu0wr488919615

[24] **N.V. Lukashevich, Yu.M. Logachev,** Combining features for automatic term extraction // Computational methods and programming. 11 (4) (2010) 108–116. Available at: https://www.mathnet.ru/php/archive.phtml?wshow=paper&jrnid=vmp&paperid=345&option_lang=rus&ysclid=lafvih6e1i417544486

[25] **E. Papagiannopoulou, G. Tsoumakas,** A Review of Keyphrase Extraction, Available at: https://arxiv.org/pdf/1905.05044.pdf

[26] **V.P. Zakharov, M.V. Khokhlova,** Selection of terminological phrases from special texts based on various association measures, XVII All-Russian United Conference "Internet and Modern Society" (IMS-2014). St. Petersburg, 2014. Available at: https://ojs.itmo.ru/index.php/IMS/article/view/268

[27] **V.V. Zalesskaya,** The program for identifying two-term statistically significant meaningful collocations in the text (based on the Russian language), XVII All-Russian United Conference "Internet and Modern Society" (IMS-2014). St. Petersburg, 2014. Available at: https://ojs.itmo.ru/index.php/IMS/article/download/267/263

[28] **V. Brezina, T. McEnery, S. Wattam,** Collocations in context: A new perspective on collocation networks // International Journal of Corpus Linguistics. 20 (2) 2015. DOI: 10.1075/ijcl.20.2.01bre

[29] **S. Evert,** Corpora and collocations, Corpus Linguistics. An International Handbook. Article 58. Mouton de Gruyter, 2008. Pp. 1212–1248. Available at: https://stephanie-evert.de/PUB/Evert2007HSK_extended_manuscript.pdf

[30] **D. Andreyeva, O.A. Mitrofanova,** Experiments on clustering Russian-language news texts based on lists of lexical constructions, Structural and Applied Linguistics. 13 (2019) 141–157. Available at: https://elibrary.ru/item.asp?id=44751350&ysclid=lafvm63kog77269386

[31] **G.T. Bukia,** Automatic clustering of news messages based on keywords and bigram constructions, Structural and Applied Linguistics. 12 (2018). Available at: https://elibrary.ru/item.asp?id=41085947&ysclid=lafvnf50p0798429448

[32] **H.P. Luhn,** A Statistical Approach to Mechanized Encoding and Searching of Literary Information, IBM Journal of Research and Development. 4 (1) (1957) 309–317. Available at: http://openlib.org/home/krichel/courses/lis618/readings/luhn57.pdf

[33] **H.P. Luhn,** The Automatic Creation of Literature Abstracts, IBM Journal of Research and Development. 2 (2) (1958) 159–165. Available at: https://courses.ischool.berkeley.edu/i256/f06/papers/luhn58.pdf

[34] **V.D. Krasavina, A.R. Mirzagitova,** Search optimization in the LeadScanner system using automatic selection of keywords and phrases, Proceedings of the international conference "Corpus Linguistics-2015". SPb., 2015. Available at: https://events.spbu.ru/eventsContent/files/corpling/corpora2015/Krasavina,%20Mirzagitova.pdf?ysclid=lafvqg594z137492449

[35] **T. Tomokiyo, M. Hurst,** A language model approach to keyphrase extraction, ACL 2003 Workshop on Multiword expressions. 18 (2003) 33–40. Available at: https://aclanthology.org/W03-1805.pdf

[36] **A.D. Moskvina, A.R. Yerofeyeva, O.A. Mitrofanova, Ya.K. Kharabet,** Automatic selection of keywords and phrases from the Russian-language corpus of texts using the RAKE algorithm, Proceedings of the International Conference "Corpus Linguistics-2017" (St. Petersburg, June 27–30, 2017). Publishing house of St. Petersburg State University, 2017. Pp. 268–275. Available at: https://elibrary.ru/item.asp?id=32425675&ysclid=lafvs71ipx62334401

[37] **A. Moskvina, E. Sokolova, O. Mitrofanova,** KeyPhrase extraction from the Russian corpus on Linguistics by means of KEA and RAKE algorithm, Data Analytics and Management in Data Intensive Domains: XX International Conference DAMDID/RCDL'2018 (October 9–12, 2018, Moscow, Russia): Conference Proceedings / ed. by L. Kalinichenko, Y. Manolopoulos, S. Stupnikov, N. Skvortsov, V. Sukhomlin. FRC CSC RAS. Pp. 369–372. Available at: https://elibrary.ru/item.asp?id=41112843&ysclid=lafvsyrid2180400661

[38] **S.J. Rose, W.E. Cowley, V.L. Crow, N.O. Cramer,** Rapid Automatic Keyword Extraction for Information Retrieval and Analysis. 2009. Available at: http://www.google.co.ve/patents/US8131735

[39] **R. Campos, V. Mangaravite, A. Pasquali, A. Jatowt, A. Jorge, C. Nunes, A. Jatowt,** YAKE! Keyword Extraction from Single Documents using Multiple Local Features, Information Sciences Journal. 509 (2020) 257–289. DOI: 10.1016/j.ins.2019.09.013

[40] **K.I. Abrosimov, A.G. Mosyagina,** Sodner for Russian nested named entity recognition, Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Issue 21. Moscow, 2022. Available at: https://www.dialog-21.ru/media/5744/abrosimovkiplusmosyaginaag037.pdf

[41] **D.A. Ustalov,** Extracting terms from Russian texts using graph models, CSEDays: Graph Theory and Applications. Yekaterinburg. 2012. Available at: https://scholar.google.com/citations?view_op=view_citation&hl=ru&user=wPD4g7AAAAAJ&citation_for_view=wPD4g7AAAAAJ:3fE2CSJIrl8C

[42] **R. Mihalcea, P. Tarau,** TextRank: Bringing Order into Text, Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing/ Barselona, 2004. Pp. 404–411. Available at: https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf

[43] **J. Devlin, M.-W. Chang, K. Lee, K. Toutanova,** BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arxiv.org. 2018. Available at: http://arXiv:1810.04805v2

[44] **M. Grootendorst,** KeyBERT: Minimal keyword extraction with BERT. 2020. Available at: // https://doi.org/10.5281/zenodo.4461265

[45] **O.A. Mitrofanova,** Study of the structural organization of a work of art using thematic modeling: experience with the text of the novel "The Master and Margarita" by M.A. Bulgakov, Proceedings of the international conference "Corpus Linguistics-2019". SPb., 2019. Pp. 387–394. Available at: https://elibrary.ru/item.asp?id=39449562&ysclid=lafvzcnxb6152046749

[46] **A.G. Sedova, O.A. Mitrofanova,** Topic modeling of Russian-language texts based on lemmas and lexical constructions, Computational Linguistics and Computational Ontologies: Proceedings of the XX

International Joint Scientific Conference “Internet and Modern Society” (St. Petersburg, June 21–24, 2017). StP: ITMO, 2017. Pp. 132–143. Available at: https://openbooks.itmo.ru/ru/file/6518/6518.pdf?ysclid=lafvzxrclb789988181

[47] **O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov,** E-hypertext Media Topic Model with Automatic Label Assignment, Recent Trends in Analysis of Images, Social Networks and Texts – 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Springer Nature, 2021. Pp. 102–114. Available at: https://link.springer.com/chapter/10.1007/978-3-030-71214-3_9

[48] **O. Mitrofanova, V. Sampetova, I. Mamaev, A. Moskvina, K. Sukharev,** Topic Modelling of the Russian Corpus of Pikabu Posts: Author–Topic Distribution and Topic Labelling, Proceedings of the International Conference “Internet and Modern Society”, IMS 2020. CEUR Workshop Proceedings. 2021. Pp. 101–116. Available at: http://ceur-ws.org/Vol-2813/rpaper08.pdf

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Митрофанова Ольга Александровна
Olga A. Mitrofanova
E-mail: o.mitrofanova@spbu.ru
ORCID: https://orcid.org/0000-0002-3008-5514

Гаврилик Дарья Александровна
Daria A. Gavrilic
E-mail: 2702gavrilic@mail.ru

Поступила: 15.11.2022; Одобрена: 19.12.2022; Принята: 26.12.2022.
Submitted: 15.11.2022; Approved: 19.12.2022; Accepted: 26.12.2022.