Terra Linguistica. 2024, Том 15, № 1. С. 20–35.
Terra Linguistica. 2024, Vol. 15, No. 1. Pp. 20–35.

Научная статья
УДК 81'13
DOI: https://doi.org/10.18721/JHSS.15102

КЛЮЧЕВЫЕ ВЫРАЖЕНИЯ В РУССКОЯЗЫЧНЫХ НАУЧНО-ПОПУЛЯРНЫХ ТЕКСТАХ: СРАВНЕНИЕ ВОСПРИЯТИЯ УСТНОЙ И ПИСЬМЕННОЙ РЕЧИ С РЕЗУЛЬТАТАМИ АВТОМАТИЧЕСКОГО АНАЛИЗА

Д.Д. Гусева ✉, **О.А. Митрофанова**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ daria.d.guseva@gmail.com

Аннотация. Процесс передачи информации может осуществляться посредством устной и письменной речи. Механизмы восприятия содержания письменных и устных текстов проявляются на разных уровнях компонентов коммуникации и понимания текста, включая уровень выделения ключевых выражений. Ключевые выражения представляют основополагающую информацию о тексте в компактной форме, способствуя структурированию текстов, их классификации и быстрой оценке содержимого. Цель данного исследования заключается в анализе различий, возникающих при восприятии одного и того же текста, представленного в письменной и устной формах. В рамках исследования были рассмотрены как письменные, так и устные русскоязычные тексты. Исследование включало в себя выделение ключевых выражений как вручную, так и автоматическими методами. Этот подход был выбран с целью выявления алгоритмов, способных приближенно воспроизводить механизмы выбора ключевых выражений, используемых носителями языка. Эксперименты были проведены на материале аудиозаписей и транскриптов выступлений русскоязычных лекторов проекта «Постнаука». Для автоматического выделения ключевых фраз в письменных текстах были применены следующие алгоритмы: статистические (Log-Likelihood, T-test, PMI test, Chi-square), гибридные (RAKE, RuTermExtract, SpaCy), с использованием машинного обучения (KeyBERT) и ChatGPT. Ручная аннотация была получена в ходе перцептивных экспериментов с привлечением русскоязычных участников. Дополнительно было проанализировано распределение ключевых выражений в структуре текстов. Результаты, полученные с применением автоматических алгоритмов выделения ключевых выражений, и результаты перцептивных экспериментов демонстрируют низкий уровень соответствия между выделенными ключевыми выражениями. Были исследованы возможности различных автоматических алгоритмов извлечения ключевых выражений и установлены ограничения при их применении в анализе письменных и устных текстов. Наши наблюдения указывают на то, что для создания эффективных методов выделения ключевых выражений необходимо учитывать типологические характеристики естественных языков, представленных в анализируемых текстах, предметные области текстов, а также наличие необходимых лингвистических и программных ресурсов. Также были получены свидетельства в пользу того, что выбор метода выделения ключевых выражений должен основываться на критериях, связанных не только с устойчивостью и частотностью ключевых выражений, но и с их восприятием.

Ключевые слова: выделение ключевых выражений, восприятие, русский язык, устный текст, письменный текст.

Финансирование: НИП СПбГУ № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта».

Для цитирования: Гусева Д.Д., Митрофанова О.А. Ключевые выражения в русскоязычных научно-популярных текстах: сравнение восприятия устной и письменной речи с результатами автоматического анализа // Terra Linguistica. 2024. Т. 15. № 1. С. 20–35. DOI: 10.18721/JHSS.15102

© Гусева Д.Д., Митрофанова О.А., 2024. Издатель: Санкт-Петербургский политехнический университет Петра Великого

Research article
DOI: https://doi.org/10.18721/JHSS.15102

KEYPHRASES IN RUSSIAN-LANGUAGE POPULAR SCIENCE TEXTS: COMPARISON OF ORAL AND WRITTEN SPEECH PERCEPTION WITH THE RESULTS OF AUTOMATIC ANALYSIS

D.D. Guseva ✉, **O.A. Mitrofanova**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ daria.d.guseva@gmail.com

Abstract. The process of transmitting information can be performed through oral and written speech. The mechanisms of perceiving written and spoken texts manifest themselves at different levels within the components of communication and comprehension of the text, including the level of keyphrases. Keyphrases provide essential information about a text in a compressed form, contributing to the structuring of texts, their classification and rapid assessment of the contents. The aim of this study is to analyze the differences that arise in perceiving the same text presented in written and oral forms. To accomplish this, we have examined both written and oral texts in Russian. The research involved the extraction of keyphrases both manually and automatically. This approach was chosen to determine algorithms that can approximate the mechanisms used by native speakers in selecting keyphrases. Experiments were performed on a dataset containing transcripts and audio recordings of lectures by Russian-speaking participants of the project "Postnauka". The following algorithms were used for automatic keyphrase extraction from written texts: statistical (Log-Likelihood, T-test, PMI test, Chi-square), hybrid linguostatistical (RAKE, RuTermExtract, SpaCy), machine learning-based method (KeyBERT), and ChatGPT. Manual annotation was obtained through perceptual experiments involving Russian-speaking participants. Additionally, keyphrase distribution in the text structure was analyzed. The results obtained during the research on automatic processing and the results of perceptual experiments demonstrate a low level of agreement between extracted keyphrases. The study investigated the capabilities of various automatic extraction algorithms for keyphrases, as well as their limitations when used in the analysis of written and oral texts.Our observations suggest that in order to develop effective techniques for selecting keyphrases, it is essential to consider the typological features of the natural languages represented in the analyzed texts, the subject areas of the texts and the availability of appropriate linguistic and software tools. Additionally, there is evidence that the choice of a method to extract keyphrases should be based not only on criteria related to the frequency and stability of the keyphrases, but also to their perception.

Keywords: keyphrase extraction, perception, Russian language, oral text, written text.

Acknowledgements: Research and Development Project SPbSU No. 75254082 "Modeling the communicative behavior of residents of a Russian metropolis in social, speech and pragmatic aspects using artificial intelligence methods"

Citation: Guseva D.D., Mitrofanova O.A., Keyphrases in Russian-language popular science texts: comparison of oral and written speech perception with the results of automatic analysis, Terra Linguistica, 15 (1) (2024) 20–35. DOI: 10.18721/JHSS.15102

Введение

Речевая коммуникация представляет собой сложный и многоуровневый процесс, обеспечивающий взаимодействие людей. Коммуникация связана с передачей информации, относящейся как к внешнему, так и к внутреннему миру человека [1]. Процесс передачи информации может происходить посредством устной и письменной речи, которые не считаются равноправными [2]. Устная речь первична по отношению к письменной и представляет собой более естественную реализацию языка в процессе общения [3].

Организация устной речи происходит быстрее, чем письменной. Для передачи смысла и выражения эмоций и отношения к предмету разговора в устной речи могут быть использованы

© Guseva D.D., Mitrofanova O.A., 2024. Published by Peter the Great St. Petersburg Polytechnic University

интонация, паузы, мимика, жесты и другие паралингвистические средства, которые отсутствуют в письменной речи. Такая речь является зафиксированной на материальном носителе, графически оформленной. Однако вместе с тем она позволяет вернуться к написанному и исправить ошибки, что в устной речи сделать сложнее. Таким образом, письменная речь позволяет более тщательно формулировать мысли и выражать их с большей точностью, чем звучащая речь. Кроме того, устная и письменная речь могут использовать разные лексические и синтаксические средства, исходя из того, что удобнее и понятнее для слушателя или читателя. Звучащую речь труднее запоминать без возможности повтора и возврата к определённым её фрагментам, так как при естественных условиях коммуникации она не фиксируется, в отличие от письменной речи. Как следствие, восприятие устной речи требует большего внимания и концентрации [4].

Этими особенностями объясняются и различные механизмы восприятия письменного и устного текста. Существует целый ряд работ, исследующих проблемы, связанные с производством и восприятием как устной, так и письменной речи [5–7]. Воспринимая письменный текст, читатели опираются как на левый (в большей степени), так и на правый контексты, то есть на слова, окружающие рассматриваемое слово. Однако при восприятии устного текста доступен в основном только левый контекст [8]. В процессе устной речи происходит воспроизведение говорящим некоего высказывания, затем воспринимаемого и понимаемого слушающим. Известно, что воспринимаемые слушателем высказывания собеседника могут влиять на последующее производство речи слушающим [9].

Разнообразие восприятия смысла и содержания устных и письменных текстов играет значительную роль в передаче информации, социальном взаимодействии и образовании. Различные стратегии восприятия текста оказывают влияние на социокультурные, психологические и образовательные контексты. Понимание этих механизмов способствует разработке и применению эффективных коммуникативных методов в различных сферах, таких как публичные выступления, создание учебных материалов, выбор методик обучения в преподавательской деятельности, рекламные кампании и другие. Анализ стратегий восприятия текста также способствует исследованиям в области литературы, драматургии и других искусств, где взаимосвязь устного и письменного творчества играет ключевую роль. Кроме того, восприятие текста в разных культурах и языках может оказывать влияние на межкультурное понимание и общение. Особенности восприятия также могут быть проанализированы с учётом когнитивных механизмов, включая функции памяти, внимания и психологические особенности, что представляет интерес для областей когнитивной психологии и лингвистики.

Различия восприятия содержания письменного и устного текста проявляются на разных уровнях компонентов коммуникации и понимания текста, в том числе на уровне выделения ключевых выражений [10]. Цель данного исследования заключается в сопоставлении различий, возникающих при восприятии одного и того же текста, представленного в письменной и устной формах. В рамках статьи рассмотрены как письменные, так и устные тексты, ключевые выражения в которых выделены как вручную, так и автоматическими методами. Такой подход применяется с целью выявления алгоритмов, способных приближенно воспроизводить механизмы выбора ключевых выражений говорящими. Актуальность нашего исследования обусловлена необходимостью учета стратегий выделения ключевых выражений в устных и письменных текстах в задачах, предполагающих автоматизацию этого процесса, в частности, при индексировании текстов в поисковых машинах интернета, при генерации транскриптов и аннотаций и веб-сервисах для семантической обработки текстов.

Ключевые выражения в фундаментальных и прикладных исследованиях

Ключевые выражения рассматриваются в различных научных дисциплинах, включая компьютерную и когнитивную лингвистику, психолингвистику, теорию коммуникации, информатику,

экономику, информационный поиск, филологию и другие [11]. В таком разделе лингвистических исследований, как лингвистика текста, выделение ключевых выражений считается одной из наиболее трудных и дискуссионных задач. Различные области науки проявляют значительные разногласия в подходах к изучению и определению ключевых выражений, что может наблюдаться, например, в «противостоянии» информационного поиска и психолингвистики [12].

Разнообразие подходов к изучению ключевых выражений объясняет многообразие терминов, используемых для обозначения данного феномена, определений и методов их выделения. Помимо термина «ключевое выражение», в работах исследователей из отдельных областей науки встречаются такие понятия, как «ключевое слово», «ключевое словосочетание» [13], «ключевой термин» [14], «опорное слово», «смысловая веха» [15], «слово-концепт», «слово-лейтмотив» [16], «лексическая доминанта», «семантическая доминанта» [17] и другие. В большинстве исследований эти термины рассматриваются как синонимы. В англоязычной терминологии преимущественно используется термины «keyword» [18, 19, etc.] и «keyphrase» [20]. Под термином «ключевое слово» часто понимают как отдельные слова, так и словосочетания. Такой подход обусловлен необходимостью использования именно словосочетаний для более точного и полного отражения содержания текста [21].

В нашем исследовании было принято решение использовать термин «ключевое выражение» как объединяющий текстовые единицы разной структуры (как слова, так и словосочетания), способные в сжатом виде представить основные компоненты семантической структуры текста.

Для обозначения ключевых выражений исследователи предлагают не только различные термины, но и разнообразные определения. В русскоязычных исследованиях, посвящённых вопросам лингвистики, часто опираются на подход, предложенный Л.В. Сахарным и А.С. Штерн [22]. Метод заключается в том, чтобы рассматривать «набор ключевых слов (НКС)» как одну из форм сокращенной версии исходного текста, так называемый «текст-примитив», который передаёт основное содержание, несмотря на возможные нарушения связности формы.

Ключевые выражения не только обеспечивают понимание и восстановление смысла исходного текста, но также представляют собой самостоятельные языковые единицы, которые подчиняются основным принципам функционирования текста. Они несут в себе основополагающую информацию о содержании текста, сообщая её в компактной форме, способствуя структурированию текстов, их классификации и облегчая оценку содержания документов. Иными словами, ключевые слова представляют собой лексические группы, состоящие из одного или более элементов и отражающие содержание документа. Это свидетельствует о непосредственном участии ключевых выражений в восприятии текста.

Список ключевых слов, извлеченный из текста, может использоваться в качестве метаданных, представляющих документ в процессе решения задач информационного поиска, классификации, кластеризации, суммаризации, аннотирования и конспектирования [23, 24]. Например, сейчас актуальна задача индексации информационных сообщений [25, 26]. На новостных онлайн-порталах ключевые выражения способствуют эффективному поиску статей [27]. Однако выделение ключевых выражений вручную — трудоёмкая задача, которая осуществима при анализе текстов ограниченного объёма, но требует автоматизации при обработке корпусных данных. В таких случаях под ключевыми выражениями подразумеваются неслучайно встречающиеся элементы, которые имеют значимость для данной выборки в контексте общего набора данных. Таким образом, автоматическое выделение ключевых выражений в упрощённом виде сводится к выбору слов и словосочетаний, наиболее точно отражающих суть анализируемого документа. Процесс извлечения происходит без участия человека и зависит от применяемой модели.

В настоящее время доступны различные открытые платформы и инструменты для работы с речью и текстом, которые предполагают автоматизацию процесса [28]. Так, сервис KeyPhrases позволяет с помощью одного из выбранных алгоритмов сгенерировать набор ключевых выражений из аннотации статей [29].

Ручное выделение ключевых выражений из текстов может осуществляться авторами текста (например, при подаче научной статьи на конференцию), экспертами по данной тематике или с привлечением широкого круга испытуемых [30]. Методику выделения ключевых выражений вручную можно представить в виде следующей схемы: после изучения документа эксперт (профессиональный или независимый участник перцептивного эксперимента) выбирает несколько слов или словосочетаний, которые, по его субъективному мнению, максимально точно отражают основную суть прочитанного или услышанного. При этом необходимо учитывать возможность возникновения случаев, когда ключевые выражения, выделенные авторами или экспертами, редко или совсем не встречаются в тексте [31].

В звучащем тексте ключевые выражения обычно обладают определенной просодической специфичностью, что позволяет сформулировать исследовательскую задачу сравнения стратегий выделения ключевых слов в письменном и устном тексте. Это может помочь лучше понять, как говорящие воспринимают и анализируют информацию, а также как ключевые выражения влияют на понимание текста.

Различия в подходах к определению ключевых выражений подтверждает предположение о том, что изучение данного явления требует использования различных методологических подходов. В частности, представляет интерес исследование того, как распространённые методы выделения ключевых выражений, ориентированные на анализ письменных текстов, применимы при анализе восприятия устного текста.

Экспериментальные данные и процедуры

Исследование проведено на материале аудиозаписей выступлений русскоязычных лекторов, опубликованных в открытом доступе в рамках проекта «Постнаука» [32]. Аудиозаписи, размещённые на образовательной платформе, сопровождены текстовым транскриптом. Для исследования были отобраны аудиозаписи двух дикторов мужского и женского пола – профессиональных лекторов, работающих с разновозрастной аудиторией и являющихся экспертами в области лингвистики [33, 34].

Выбор материала обусловлен тем, что лекции принадлежат к научно-популярному стилю и адресованы широкой аудитории, не обладающей специализированными знаниями. Это позволяет привлекать к участию в опросах информантов независимо от их профессиональной сферы деятельности. Кроме того, была проведена оценка читабельности (удобочитаемости) текстов с применением ряда метрик, адаптированных для русского языка [35]. Эти метрики основаны на различных характеристиках текста, чаще всего средней длины слова, предложения и доли общих слов. Оба текста имеют приблизительно одинаковый уровень сложности, который, согласно различным метрикам, соответствует уровню ученика старших классов школы, студента или выпускника университета (табл. 1).

Автоматизированное извлечение ключевых выражений из текста представляет собой многоступенчатый процесс, предполагающий многообразие стратегий его реализации. Общая схема практически идентична для всех используемых методов и включает в себя этапы формирования начального списка кандидатов в ключевые выражения и их последующей фильтрации для получения итогового набора ключевых выражений [36]. Обычной практикой перед выделением ключевых выражений является удаление из текста стоп-слов — элементов, которые не несут смысловой нагрузки (союзы, артикли, предлоги, местоимения, частицы, вводные слова, междометия и т.д.). Различия в методиках определяются способами обработки текста на каждом этапе, а также требованиями к лингвистической аннотации текстов.

В данной работе для автоматического извлечения ключевых выражений из письменного текста были применены следующие методы: (1) статистические (Хи-квадрат, Log-Likelihood, PMI-test, T-test), (2) гибридные (RAKE, RuTermExtract, SpaCy), (3) с использованием машинного обучения

Таблица 1. Оценка читабельности текстов
Table 1. Evaluation of the text readability

Метрика (оценочная шкала)	Значение для текста №1	Значение для текста №2	Зависимость сложности от значения	Уровень сложности
Тест Флеша-Кинкейда (0–20)	9.04	10.82	Чем выше, тем сложнее	Довольно сложно
Индекс удобочитаемости Флеша (0–100)	38.97	32.01	Чем выше, тем легче	Довольно сложно
Индекс Колман-Лиау (0–∞)	9.59	10.74	Чем выше, тем сложнее	Довольно сложно
Индекс SMOG (0–∞)	16.69	18.48	Чем выше, тем сложнее	Довольно легко
Автоматический индекс удобочитаемости (1–14)	9.59	10.74	Чем выше, тем сложнее	Довольно легко
Индекс удобочитаемости LIX (0–100)	65.28	67.64	Чем выше, тем сложнее	Довольно сложно

(KeyBERT), а также (4) ChatGPT, языковая модель на базе искусственного интеллекта, разработанная OpenAI.

Кратко обсудим особенности алгоритмов из разных групп.

Статистические алгоритмы для выделения ключевых выражений базируются на анализе относительных частот морфологических, лексических и синтаксических единиц, а также их комбинаций, что позволяет сократить вычислительную сложность процедур выделения ключевых выражений [37]. Такие методы могут быть достаточно эффективны при анализе текстов научных статей [38]. Тем не менее, при учёте только частотных характеристик ключевых выражений статистические алгоритмы могут упустить наиболее значимые слова и сочетания в тексте, что приводит к снижению точности извлечения ключевых выражений.

Гибридные методы извлечения ключевых выражений комбинируют в себе статистические и лингвистические подходы, объединяя их преимущества. Такие алгоритмы, учитывая как статистические характеристики текста, так и его лингвистические особенности (семантика и синтаксис), могут обеспечить более точные результаты по сравнению со статистическими методами. Однако гибридные методы обычно более сложны в реализации и требуют больше расчётов и ресурсов.

Алгоритмы, основанные на методах машинного обучения, могут также рассматриваться как подгруппа гибридных методов. Например, при использовании машинного обучения с учителем (как в случае алгоритма KEA) требуется заранее подготовленная база данных с размеченными ключевыми выражениями для формирования обучающей выборки и создания классификатора. В процессе обучения модели ключевые выражения помечаются как положительные примеры, а все остальные — как отрицательные. Каждому слову сопоставляются вектора значений определённых параметров (например, меры TF-IDF, длины слова, его позиции в тексте), и вычисляется вероятность отнесения выражений к набору ключевых и задается порог вхождения в этот набор. После завершения обучения процесс выделения ключевых выражений осуществляется с помощью вычисления степени релевантности слов по векторам параметров и их вероятности являться ключевыми в соответствии с обученной моделью. Методы машинного обучения без учителя, а также методы глубинного обучения (например, KeyBERT [39] и другие модели на основе BERT [40, 41]) опираются на выделение статистически значимых признаков в текстах, а также

учитывают совместную встречаемость слов при формировании моделей распределенных векторных вложений [42], в этом смысле они также сближаются с гибридными методами автоматического выделения ключевых выражений.

При проведении экспериментов на основе сформированного нами набора данных перед применением всех автоматических методов была выполнена предварительная обработка текста с целью адаптировать его в формат, подходящий для последующего анализа: произведена токенизация (идентификация словоформ в тексте) и удаление стоп-слов. Необходимо отметить, что стоп-слова могут являться дискурсивными маркерами, специфичными для конкретных типов текстов и коммуникативных целей. Так, в данном исследовании в список стоп-слов были дополнительно внесены числительные. Также была предпринята попытка лемматизировать тексты, однако при работе с данным материалом лемматизация ухудшила результаты. Из ключевых выражений, выделенных с применением автоматических методов, рассматривались только первые 10 результатов. Пример работы алгоритмов представлен в таблице (табл. 2).

Таблица 2. Пример выделения ключевых выражений (униграмм) автоматическими методами
Table 2. Example of automatic keyphrase extraction (unigrams)

Автоматический метод выделения ключевых выражений	Ключевые выражения (униграммы)
RuTermExtract	язык 8, детали 8, ребёнок 7, слово 6, человек 5, структура 5, универсальная грамматика 4, половина 4, люди 4, языковые выражения 3
SpaCy	слова 1; детали 1; языка 0,75; структуру 0,75; гены 0,625; грамматика 0,5; языке 0,5; ребенок 0,5; люди 0,5; образом 0,5
KeyBERT	американский 0,4446; грамматический 0,425; головоломки 0,4045; гумилевских 0,3947; английском 0,3913; английского 0,3809; мнемотехника 0,3782; великий 0,3778; европейской 0,3692; говорили 0,3681

Ручная разметка ключевых выражений была получена в результате серии перцептивных экспериментов, в которых участвовали испытуемые. Для создания и проведения исследования использовалась платформа Google Forms. В эксперименте с использованием письменного текста аудиторы должны были после прочтения выделить из текста 10 ключевых выражений, наиболее полно передающих основное содержание документа, и проранжировать их от наиболее значимого к менее важному. При работе с устным текстом испытуемые выполняли аналогичное задание после прослушивания аудиозаписи. Опросы были построены таким образом, что аудиторы имели возможность несколько раз прочитать текст или прослушать аудиозапись.

Количество ключевых выражений в наборе может изменяться в значительных пределах, однако оптимальным, особенно при работе с устным текстом, является ограниченный список. В различных источниках упоминается, что оптимальный набор обычно включает от 5 до 15 или от 8 до 10 выражений. Таким образом, выделение именно 10 ключевых выражений позволяет сохранить достаточный уровень их значимости и релевантности в рамках рассматриваемого текста. Согласно инструкции, участники экспериментов должны выбирать в качестве ключевых выражений униграммы, биграммы или триграммы (ключевые выражения, состоящие из одного, двух или трёх элементов соответственно).

Процедура анализа результатов исследования включала сопоставление ключевых выражений, выделенных (1) различными автоматическими методами на материале письменного текста, (2) автоматически и вручную в письменном тексте участниками перцептивного эксперимента, (3) автоматическими методами в письменном тексте и с участием испытуемых в устном тексте. В табл. 3 приведены параметры проведённых экспериментов.

Таблица 3. Параметры экспериментов
Table 3. Experimental parameters

Параметр	Эксперименты с автоматическим выделением ключевых выражений	Эксперименты с выделением ключевых выражений участниками перцептивных экспериментов (вручную)
Функциональный стиль текста	Научно-популярный стиль	Научно-популярный стиль
Функциональный регистр текста	Письменный	Письменный и устный
Длина ключевых выражений	Ограничения зависят от метода	Униграммы, биграммы, триграммы
Количество ключевых выражений	1...10	1...10
Метод ранжирования ключевых выражений	В порядке убывания значимости	В порядке убывания значимости

Кроме того, было проанализировано, как ключевые выражения распределены в структуре текстов. Существенный аспект, который следует учитывать при рассмотрении ключевых выражений, заключается в следующем: распределение ключевых выражений в тексте неравномерно. Считается, что выражения, присутствующие в заголовке, аннотации, вводной и заключительной частях текста, обладают особой информативностью. Это объясняется тем, что указанные части текста обычно содержат ключевую информацию, которая быстрому восприятию основной идеи и содержания текста [43].

Результаты экспериментов

1. Распределение ключевых выражений в текстах

В перцептивном эксперименте на материале письменных текстов приняло участие 66 испытуемых. В эксперименте на материале первого устного текста приняло участие 34 слушателя, второго текста – 30.

Первый текст, использованный в экспериментах, состоит из 8 абзацев, второй текст включает в себя 10 абзацев.

В первом письменном тексте отмечается сосредоточение выделенных аудиторами ключевых слов в начальной (первый абзац) и заключительной части (последний абзац). При этом к набору ключевых выражений, выделенных из этих частей текста, относятся практически все наиболее частотные среди ответов аудиторов выражения (т. е. выбранные более чем 10 участниками эксперимента). Таким образом, ключевое выражение «универсальная грамматика», выделенное 53 участниками, встречается в первом предложении первого абзаца текста. Ключевые выражения «грамматический взрыв» и «когнитивный взрыв», появляющиеся в последнем абзаце текста, были выбраны 51 и 56 читателями соответственно. Напротив, ключевые выражения, выделенные в основной части текста, в основном были выделены только отдельными аудиторами, то есть в общем наборе ответов встречались всего один раз. Так, в четвёртом абзаце есть выражение «память», которое было выделено только одним читателем, а в пятом абзаце — «экспериментатор», также выделенное одним аудитором.

Аналогичные наблюдения были сделаны при анализе второго письменного текста. Например, в первом абзаце присутствует выражение «язык», которое выбрали 32 участника эксперимента, а выражение «лексическая сочетаемость», находящееся в последнем абзаце, было выбрано 25 читателями.

Кроме того, в обоих экспериментах некоторые аудиторы выделили слова и словосочетания, которых не было в исходных текстах. К таким выражениям относятся «лингвистический эксперимент», «научиться», «порождение речи», «лексика», «оценка языка» и другие.

Анализ первого устного текста показывает, что выделенные ключевые выражения распределены в нём более равномерно. Помимо начала и конца текста, выражения, выделенные более чем

одним аудитором, были встречены во втором, третьем и предпоследнем абзацах, хотя наиболее частотные среди выделенных выражения всё равно находятся в начале и заключении текста. Выражения «универсальная грамматика» и «грамматические гены» (первый абзац) были выделены 26 и 24 слушателями, «грамматический взрыв» и «когнитивный взрыв» (последний абзац) – 21 и 20 слушателями соответственно. Однако выражение «изучение иностранного языка», выбранное 8 участниками, находится в третьем абзаце. Также в этом абзаце есть слово «структура», которое было выделено 8 слушателями.

Анализ второго устного текста демонстрирует схожие результаты. Наиболее частотное сочетание «лексическая функция», выделенное 15 слушателями, впервые встречается в тексте в третьем абзаце. Ключевое выражение «взаимная информация» (8 слушателей) находится в восьмом абзаце, а в шестом появляется выражение «модель мешка слов», выделенное 7 слушателями.

Результаты свидетельствуют о том, что в основном в набор ключевых выражений попадают выражения, не самые частотные для языка, но встречаемость которых в изучаемом тексте выше их лингвистической вероятности. Это подтверждают эксперименты с научными текстами по лингвистике, в которых испытуемые-нелингвисты, плохо понимавшие содержание, выбирали в качестве ключевых выражений лингвистические термины, наиболее часто встретившиеся в данных текстах.

2. Сопоставление результатов выделения ключевых выражений различными автоматическими методами и вручную

В рамках статистических методах алгоритмы группируются по близости результатов выделения ключевых выражений. Из рассмотренных это пары методов Хи-квадрат – PMI-test и Log-Likelihood – T-test. При работе с первым текстом данные методы дали одинаковые или почти одинаковые группы ключевых выражений – как биграмм, так и триграмм.

Униграммы были выделены с помощью гибридных методов и метода KeyBERT. Совпадений между наборами, полученными с помощью разных групп методов, не было обнаружено, однако 6 из 10 ключевых выражений, выделенных методами RuTermExtract и SpaCy, совпали (совпадающие ключевые выражения выделены жирным шрифтом, см. табл. 4).

Таблица 4. Ключевые выражения, выделенные с использованием RuTermExtract и SpaCy (абсолютная частота)
Table 4. Keyphrase extraction using RuTermExtract and SpaCy (absolute frequency)

RuTermExtract	SpaCy
язык 8, **детали 8**, **ребёнок 7**, **слово 6**, человек 5, **структура 5**, универсальная грамматика 4, половина 4, **люди 4**, языковые выражения 3	**слова 1**, **детали 1**, **языка 0,75**; **структуру 0,75**; гены 0,625; грамматика 0,5; **языке 0,5**; **ребенок 0,5**; **люди 0,5**; образом 0,5

Не было выявлено совпадений в наборах ключевых выражений, выделенных статическими методами, RAKE и KeyBERT. Таким образом, сходные группы ключевых выражений дали кластеры статистических методов, а также гибридные методы RuTermExtract и SpaCy.

При работе со вторым текстом наблюдалась схожая картина. Отличия заключаются в том, что при выделении биграмм Log-Likelihood – T-test дали одно совпадение с Rake, а при выделении униграмм KeyBERT дал одно совпадение с результатами, полученными с помощью RuTermExtract, и одно – с полученными в результате применения метода SpaCy.

Были проведены эксперименты с применением технологии ChatGPT. Цель заключалась в получении от генеративной модели искусственного интеллекта контрольного набора ключевых выражений, который мог бы служить в качестве синтетического эталона для оценки результатов, полученных после применения автоматических методов и в ходе перцептивных экспериментов. Ключевые выражения, выделенные с помощью ChatGPT, приведены в табл. 5.

Таблица 5. Ключевые выражения, выделенные с использованием ChatGPT
Table 5. Keyphrase extraction using ChatGPT

Текст №1	Текст №2
Ноам Хомский, универсальная грамматика, языки, грамматические гены, дети, родной язык, грамматика, гены, классическая европейская хореографическая традиция, индийская хореографическая традиция, классическая музыка, русская поэзия, структура, мнемотехника, эксперимент, Саймон Кирби, итерированное обучение	грамматика, словосочетания, предложения, значение, лексическая функция, сочетаемость, стандартные значения, magn, перевод, английский, русский, языки

В результате экспериментов получены данные, свидетельствующие о том, что наборы ключевых выражений, полученные с применением различных автоматических методов, демонстрируют мало совпадений между собой. Чтобы выяснить, какие из них дают результаты, наиболее приближенные к реальному человеческому восприятию, были проведены перцептивные эксперименты на материале письменного и устного текста.

При сопоставлении наборов ключевых выражений, выделенных с применением автоматических методов и участниками экспериментов, не учитывались ответы аудиторов, указавших в качестве родного языка не русский язык, а также ответы участников младше 18 лет.

При анализе результатов перцептивного эксперимента на материале письменных текстов были рассмотрены ответы 41 испытуемого. Читатели должны были выделить 10 ключевых выражений – соответственно, всего было получено 410 ответов. По первому тексту было получено 60 уникальных ключевых выражений, которые были отранжированы. В сумме на них приходится 250 ответов, что составляет 61% от общего числа. В случае с первым текстом есть совпадения между результатами выделения ключевых выражений вручную и с помощью алгоритмов RuTermExtract, Rake, KeyBERT, SpaCy и ChatGPT (табл. 6). Последний метод при этом показывает лучший результат.

Таблица 6. Сопоставление результатов выделения ключевых выражений вручную и автоматическими методами в письменном тексте №1
Table 6. Comparison of the results of manual and automatic keyphrase extraction from the written text №1

Ключевые выражения, выделенные информантами	Частотность	Ранг	Встречаемость	Автоматические методы, выделившие данные ключевые выражения
Универсальная грамматика	0,088	59	36	RuTermExtract, Rake, ChatGPT
Грамматический взрыв	0,088	59	36	KeyBERT
Когнитивный взрыв	0,088	59	36	
Грамматические гены	0,083	57	34	ChatGPT
Структурные закономерности	0,063	56	26	
Ноам Хомский	0,046	54,5	19	KeyBERT, ChatGPT
Устройство овладения языком	0,046	54,5	19	Rake
Ищи структуру в хаосе	0,039	53	16	
Структура	0,034	51,5	14	RuTermExtract, SpaCy, ChatGPT
Языковые высказывания	0,034	51,5	14	

По второму тексту было получено 78 уникальных выражений. Первые 10 выражений составили 217 ответов (53%). Совпадения дают те же самые алгоритмы. Как при работе с первым текстом,

так и при работе со вторым текстом четыре выражения из набора ключевых выражений не имеют совпадений в выдаче автоматических алгоритмов.

При обработке результатов эксперимента на материале первого устного текста учитывались ответы 20 испытуемых, второго текста – 21 испытуемого. По первому тексту было получено 88 уникальных ключевых выражений. Первые 10 выражений составили 86 ответов от общего числа (43%). По второму тексту получено 108 уникальных ключевых выражений. Первые 10 составили 67 ответов (32%).

В случае с первым текстом есть совпадения между результатами выделения ключевых выражений вручную и с помощью алгоритмов RuTermExtract, Rake, KeyBERT, SpaCy и ChatGPT (табл. 7).

Таблица 7. Сопоставление результатов выделения ключевых выражений вручную и автоматическими методами в устном тексте №1
Table 7. Comparison of the results of manual and automatic keyphrase extraction from the oral text №1

Ключевые выражения, выделенные информантами	Частотность	Ранг	Встречаемость	Автоматические методы, выделившие данные ключевые выражения
Универсальная грамматика	0,074	87,5	14	RuTermExtract, Rake, ChatGPT
Грамматический взрыв	0,074	87,5	14	KeyBERT
Когнитивный взрыв	0,068	86	13	
Грамматические гены	0,063	85	12	ChatGPT
Структура в хаосе	0,037	84	7	
Ищи структуру в хаосе	0,032	83	6	
Гены	0,026	80,5	5	SpaCy
Язык	0,026	80,5	5	RuTermExtract, SpaCy, ChatGPT
Изучение инопланетного языка	0,026	80,5	5	
Структура	0,026	80,5	5	RuTermExtract, SpaCy, ChatGPT

При работе с первым текстом четыре выражения из набора не имеют совпадений в выдаче автоматических алгоритмов. При работе со вторым текстом совпадения даёт меньшее число алгоритмов – RuTermExtract, KeyBERT и ChatGPT (табл. 8). При этом было выявлено шесть выражений, не имеющих совпадений.

Наименее удовлетворительные результаты показали статистические алгоритмы, для них совпадения с ответами аудиторов не были зарегистрированы. Это указывает на существующие ограничения в применимости статистических методов. Такие методы обычно выдают сочетания, которые характеризуются устойчивостью для данного текста. Методы достаточно универсальны, но область их применения ограничена языками с бедной морфологией, где частотность словоформ одной лексемы велика. К таким языкам относят английский, датский, современный китайский и другие, однако русский язык обладает богатой морфологией [44].

Заключение

Полученные в ходе исследования результаты свидетельствуют о незначительных совпадениях между результатами автоматической обработки текстов и перцептивных экспериментов. Были исследованы возможности различных автоматических алгоритмов извлечения ключевых выражений и выявлены ограничения на их использование в работе с письменными и устными текстами. Наши наблюдения указывают на то, что для создания эффективных методов выделения ключевых выражений необходимо учитывать типологические характеристики естественных языков, представленных в анализируемых текстах, предметные области текстов, а также наличие в распоряжении

Таблица 8. Сопоставление результатов выделения ключевых выражений вручную и автоматическими методами в устном тексте №2
Table 8. Comparison of the results of manual and automatic keyphrase extraction from the oral text №2

Ключевые выражения, выделенные информантами	Частотность	Ранг	Встречаемость	Автоматические методы, выделившие данные ключевые выражения
Лексическая функция	0,06	108	12	ChatGPT
Сочетаемость	0,05	107	10	ChatGPT
Коллокация	0,04	105,5	8	
Словосочетание	0,04	105,5	8	RuTermExtract, KeyBERT, ChatGPT
Грамматика	0,035	104	7	ChatGPT
Взаимная информация	0,025	102,5	5	
Модель мешка слов	0,025	102,5	5	
Корпусная лингвистика	0,02	99,5	4	
Слова	0,02	99,5	4	
Коллокации	0,02	99,5	4	

исследователей необходимых лингвистических и программных ресурсов. Также были получены свидетельства в пользу того, что при выборе метода выделения ключевых выражений нужно руководствоваться соображениями, связанными не только с устойчивостью и частотностью ключевых выражений, но и с их восприятием. Кроме того, при проведении перцептивных экспериментов были отмечены различия в восприятии текстов информантами, относящимися к разным возрастным группам. В связи с этим также планируется провести сравнительный анализ результатов по группам, расширив состав участников опросов за счёт школьников старше двенадцати лет и иностранцев, изучающих русский язык.

СПИСОК ИСТОЧНИКОВ

1. **Кодзасов С.В., Кривнова О.Ф.** Общая фонетика. М., 2001. 590 с.
2. **Леонтьев А.А.** Некоторые вопросы лингвистической теории письма // Вопросы общего языкознания. М., 1964.
3. **Васильева В.В., Коньков В.И.** Устная речь: практикум // С.-Петерб. гос. ун-т, Ин-т «Высш. шк. журн. и мас. коммуникаций». 2015. 100 с.
4. **Трошева Т.Б.** Устная речь // Стилистический энциклопедический словарь русского языка / под ред. М.Н. Кожиной. Москва. 2003. 567 с.
5. **Мурзин Л.Н., Штерн А.С.** Текст и его восприятие. Свердловск. 1991.
6. **Ван Дейк Т.А., Кинч В.** Стратегии понимания связного текста // Новое в зарубежной лингвистике. Вып. XXIII: Когнитивные аспекты языка. М.: Прогресс. 1988. С. 153–211.
7. **Леонтьев А.А.** Основы психолингвистики. М.: Academia. 2005. 287 с.
8. **Касевич В.Б.** Семантика. Синтаксис. Морфология. М., 1988.
9. **Pardo J.S., Jordan K., Mallari R., Scanlon C., Lewandowski E.** Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures // Journal of Memory & Language. 2013. 69. Pp. 183–195.
10. **Ягунова Е.В.** Роль ключевых слов при восприятии звучащего и письменного текста (на материале русского языка) // Человек пишущий и читающий: проблемы и наблюдения. Материалы международной конференции 14-16 марта 2002, Санкт-Петербург. СПб: Издательство СПбГУ. 2004. С. 197–204.

11. **Москвитина Т.Н.** Ключевые слова и их функции в научном тексте // Вестник ЧГПУ. 2009. № 11. С. 270–283.

12. **Ягунова Е.В.** Эксперимент и вычисления в анализе ключевых слов художественного текста // Философия языка. Лингвистика. Лингводидактика №1. Пермь. 2010. С. 83–89.

13. **Абрамов Е.Г.** Подбор ключевых слов для научной статьи // Научная периодика: проблемы и решения. 2011. № 2. С. 35–40.

14. **Гринева М., Гринев М.** Анализ текстовых документов для извлечения тематически сгруппированных ключевых терминов // Труды Института системного программирования РАН. 2009. Т. 16. С. 155–165.

15. **Папуша И.С.** Сложное синтаксическое целое: ключевые слова или гермы // Вестник Ассоциации ВУЗов туризма и сервиса, 2008, № 3. С. 48–54.

16. **Светозарова Н.Д., Штерн А.С.** Ключевые и фонетически выделенные слова текста // Экспериментальная фонетика. М., 1989. С. 157–170.

17. **Шехтман Н.А.** Понимание речевого произведения и гипертекст. Оренбург: Изд-во ОГПУ. 2005. 168 с.

18. **Dostal M.** Automatic Keyphrase Extraction Based on NLP and Statistical Methods // Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects. Pisek, Czech Republic. 2011. Pp. 140–145.

19. **Zhang C., Wang H., Liu Y. et al.** Automatic keyword extraction from documents using conditional random fields // Journal of Computational Information Systems. 2008. Vol. 4, No. 3. Pp. 1169–1180.

20. **Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G.** KEA: Practical Automatic Keyphrase Extraction // Proceedings of the 4th ACM conference on Digital libraries. 1999. URL: http://www.cs.waikato.ac.nz/~eibe/pubs/chap_Witten-et-al_Windows.pdf

21. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учебное пособие. М.: МИЭМ. 2011. 272 с.

22. **Сахарный Л.В., Штерн А.С.** Набор ключевых слов как тип текста // Лексические аспекты в системе профессионально-ориентированного обучения иноязычной речевой деятельности. Пермь: Перм. политехн. ун-т, 1988. С. 34–51.

23. **Шереметьева С.О., Осминин П.Г.** Методы и модели автоматического извлечения ключевых слов // Вестник ЮУрГУ. Серия «Лингвистика». 2015. Т. 12. № 1. С. 76–81.

24. **Song M., Feng Y., Jing L.** A Survey on Recent Advances in Keyphrase Extraction from Pre-trained Language Models // Findings of the Association for Computational Linguistics: EACL 2023. 2023 Pp. 2153–2164. DOI: 10.18653/v1/2023.findings-eacl.161

25. **Piskorski J., Stefanovitch N., Jacquet G., Podavini A.** Exploring Linguistically-Lightweight Keyword Extraction Techniques for Indexing News Articles in a Multilingual Set-up // Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021. Pp. 35–44. URL: https://aclanthology.org/2021.hackashop-1.6 (дата обращения: 08.02.2024).

26. **Verma Y., Jangra A., Saha S., Jatowt A., Roy D.** 2022. MAKED: Multi-lingual Automatic Keyword Extraction Dataset // Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. Pp. 6170–6179. URL: https://aclanthology.org/2022.lrec-1.664 (дата обращения: 08.02.2024).

27. **Koloski B., Pollak S., Škrlj B., Martinc M.** Extending Neural Keyword Extraction with TF-IDF tagset matching // Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021. Pp. 22–29. URL: https://aclanthology.org/2021.hackashop-1.4 (дата обращения: 08.02.2024).

28. **Тихонова Е.В., Косычева М.А.** Эффективные ключевые слова: стратегии формулирования // Health, Food & Biotechnology. 2021. № 4 (3). С. 7–15. DOI: 10.36107/hfb.2021.i4.s122

29. **Морозов Д.А., Глазкова А.В., Тютюльников М.А., Иомдин Б.Л.** Генерация ключевых слов для аннотаций русскоязычных научных статей // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. 2023. №1. URL: https://cyberleninka.ru/article/n/generatsiya-klyuchevyh-slov-dlya-annotatsiy-russkoyazychnyh-nauchnyh-statey (дата обращения: 11.03.2024).

30. **Ванюшкин А.С., Гращенко Л.А.** О разметке корпусов текстов ключевыми словами // Новые информационные технологии в автоматизированных системах. 2018. № 21. С. 207–211.

31. **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

32. Постнаука. URL: https://postnauka.ru/ (дата обращения: 08.02.2024).
33. Грамматические гены. URL: https://postnauka.org/video/61500 (дата обращения: 08.02.2024).
34. Лексическая сочетаемость. URL: https://postnauka.org/video/57524 (дата обращения: 08.02.2024).
35. ruTS, a library for statistics extraction from texts in Russian. 2023. URL: https://github.com/SergeyShk/ruTS (дата обращения: 08.02.2024).
36. **Song M., Jing L., Xiao L.** Importance Estimation from Multiple Perspectives for Keyphrase Extraction // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 2726–2736. DOI: 10.18653/v1/2021.emnlp-main.215
37. **Ushio A., Liberatore F., Camacho-Collados J.** Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 8089–8103. DOI: 10.18653/v1/2021.emnlp-main.638
38. **Papagiannopoulou E., Tsoumakas G.** A review of keyphrase extraction // Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery. 2020. 10 (2). DOI: 10.1002/WIDM.1339
39. **Grootendorst M.** KeyBERT: Minimal Keyword Extraction with BERT. 2020. URL: http://doi.org/10.5281/zenodo.4461265 (дата обращения: 11.03.2024).
40. **Song M., Feng Y., Jing L.** Utilizing BERT Intermediate Layers for Unsupervised Keyphrase Extraction // Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022). 2022 Pp. 277–281. URL: https://aclanthology.org/2022.icnlsp-1.32 (дата обращения: 11.03.2024).
41. **Kulkarni M., Mahata D., Arora R., Bhowmik R.** Learning Rich Representation of Keyphrases from Text // Findings of the Association for Computational Linguistics: NAACL 2022. 2022. Pp. 891–906. DOI: 10.18653/v1/2022.findings-naacl.67
42. **Liang X., Wu S., Li M., Li Z.** Unsupervised keyphrase extraction by jointly modeling local and global context // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 155–164. DOI: 10.18653/v1/2021.emnlp-main.14
43. **Щерба Л.В.** Языковая система и речевая деятельность. Л., 1974.
44. **Tsarfaty R., Seddah D.é, Kübler S., Nivre J.** Parsing Morphologically Rich Languages: Introduction to the Special Issue // Computational Linguistics. 2013. 39 (1). Pp. 15–22.

REFERENCES

[1] **Kodzasov S.V., Krivnova O.F.** Obshchaya fonetika [General phonetics]. M., 2001.
[2] **Leontyev A.A.** Nekotoryye voprosy lingvisticheskoy teorii pisma [Some questions of the linguistic theory of writing], Voprosy obshchego yazykoznaniya. M., 1964.
[3] **Vasilyeva V.V., Konkov V.I.** Ustnaya rech: praktikum [Oral speech: practicum], St. Petersburg State University, In-t "Higher School of Journalism. and mass communications", 2015.
[4] **Trosheva T.B.** Ustnaya rech [Oral speech], Stylistic encyclopedic dictionary of the Russian language / edited by M.N. Kozhina, Moskva, 2003.
[5] **Murzin L.N., Shtern A.S.** Tekst i yego vospriyatiye [The text and its perception]. Sverdlovsk, 1991.
[6] **Van Deyk T.A., Kinch V.** Strategii ponimaniya svyaznogo teksta [Strategies for understanding a coherent text], Novoye v zarubezhnoy lingvistike. Vyp. XXIII: Kognitivnyye aspekty yazyka [New in foreign linguistics. Issue XXIII: Cognitive aspects of language]. M.: Progress. 1988. Pp. 153–211.
[7] **Leontyev A.A.** Osnovy psikholingvistiki [Fundamentals of psycholinguistics], M.: Academia, 2005.
[8] **Kasevich V.B.** Semantika. Sintaksis. Morfologiya [Semantics. The syntax. Morphology], M., 1988.
[9] **Pardo J.S., Jordan K., Mallari R., Scanlon C., Lewandowski E.** Phonetic convergence in shadowed speech: The relation between acoustic and perceptual measures, Journal of Memory & Language, 69 (2013) 183–195.
[10] **Yagunova Ye.V.** Rol klyuchevykh slov pri vospriyatii zvuchashchego i pismennogo teksta (na materiale russkogo yazyka) [The role of keywords in the perception of a sounding and written text (based on the material of the Russian language)], A person writing and reading: problems and observations. Proceedings of the international conference on March 14–16, 2002. St. Petersburg. St. Petersburg: St. Petersburg State University Publishing House. 2004. Pp. 197–204.

[11] **Moskvitina T.N.** Klyuchevyye slova i ikh funktsii v nauchnom tekste [Keywords and their functions in a scientific text], Vestnik ChGPU, 11 (2009) 270–283.

[12] **Yagunova Ye.V.** Eksperiment i vychisleniya v analize klyuchevykh slov khudozhestvennogo teksta [Experiment and calculations in the analysis of keywords of a literary text], Filosofiya yazyka. Lingvistika. Lingvodidaktika №1. Perm. 2010. Pp. 83–89.

[13] **Abramov Ye.G.** Podbor klyuchevykh slov dlya nauchnoy stati [Selection of keywords for a scientific article], Nauchnaya periodika: problemy i resheniya, 2 (2011) 35–40.

[14] **Grineva M., Grinev M.** Analiz tekstovykh dokumentov dlya izvlecheniya tematicheski sgruppirovannykh klyuchevykh terminov [Analysis of text documents for the extraction of thematically grouped key terms], Trudy Instituta sistemnogo programmirovaniya RAN [Proceedings of the Institute of System Programming of the Russian Academy of Sciences], 16 (2009) 155–165.

[15] **Papusha I.S.** Slozhnoye sintaksicheskoye tseloye: klyuchevyye slova ili germy [A complex syntactic whole: keywords or hermes], Vestnik Assotsiatsii VUZov turizma i servisa [Bulletin of the Association of Universities of Tourism and Service], 3 (2008) 48–54.

[16] **Svetozarova N.D., Shtern A.S.** Klyuchevyye i foneticheski vydelennyye slova teksta [Key and phonetically highlighted words of the text], Eksperimentalnaya fonetika [Experimental phonetics], M., 1989. Pp. 157–170.

[17] **Shekhtman N.A.** Ponimaniye rechevogo proizvedeniya i gipertekst [Understanding speech works and hypertext]. Orenburg: OGPU Publishing House, 2005.

[18] **Dostal M.** Automatic Keyphrase Extraction Based on NLP and Statistical Methods, Proceedings of the Dateso 2011: Annual International Workshop on Databases, Texts, Specifications and Objects. Pisek, Czech Republic. 2011. Pp. 140–145.

[19] **Zhang C., Wang H., Liu Y. et al.** Automatic keyword extraction from documents using conditional random fields, Journal of Computational Information Systems, 4 (3) (2008) 1169–1180.

[20] **Witten I.H., Paynter G.W., Frank E., Gutwin C., Nevill-Manning C.G.** KEA: Practical Automatic Keyphrase Extraction, Proceedings of the 4th ACM conference on Digital libraries. 1999. Available at: http://www.cs.waikato.ac.nz/~eibe/pubs/chap_Witten-et-al_Windows.pdf

[21] Avtomaticheskaya obrabotka tekstov na yestestvennom yazyke i kompyuternaya lingvistika: uchebnoye posobiye [Automatic text processing in natural language and computational linguistics: a textbook], M.: MIEM. 2011.

[22] **Sakharnyy L.V., Shtern A.S.** Nabor klyuchevykh slov kak tip teksta [A set of keywords as a type of text], Lexical aspects in the system of professionally oriented teaching of foreign language speech activity. Perm: Perm Polytechnic University. un-t, 1988. Pp. 34–51.

[23] **Sheremetyeva S.O., Osminin P.G.** Metody i modeli avtomaticheskogo izvlecheniya klyuchevykh slov [Methods and models of automatic keyword extraction], Vestnik YuUrGU. Seriya «Lingvistika» [Bulletin of SUSU. The series "Linguistics"], 12 (1) (2015) 76–81.

[24] **Song M., Feng Y., Jing L.** A Survey on Recent Advances in Keyphrase Extraction from Pretrained Language Models, Findings of the Association for Computational Linguistics: EACL 2023. 2023, Pp. 2153–2164. DOI: 10.18653/v1/2023.findings-eacl.161

[25] **Piskorski J., Stefanovitch N., Jacquet G., Podavini A.** Exploring Linguistically-Lightweight Keyword Extraction Techniques for Indexing News Articles in a Multilingual Set-up, Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021. Pp. 35–44. Available at: https://aclanthology.org/2021.hackashop-1.6 (accessed 08.02.2024).

[26] **Verma Y., Jangra A., Saha S., Jatowt A., Roy D.** 2022. MAKED: Multi-lingual Automatic Keyword Extraction Dataset, Proceedings of the Thirteenth Language Resources and Evaluation Conference. 2022. Pp. 6170–6179. Available at: https://aclanthology.org/2022.lrec-1.664 (accessed 08.02.2024).

[27] **Koloski B., Pollak S., Škrlj B., Martinc M.** Extending Neural Keyword Extraction with TF-IDF tagset matching, Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation. 2021. Pp. 22–29. Available at: https://aclanthology.org/2021.hackashop-1.4 (accessed 08.02.2024).

[28] **Tikhonova E.V., Kosycheva M.A.** Effective Keywords: Strategies for their Formulation, Health, Food & Biotechnology, 4 (3) (2021) 7–15. DOI: 10.36107/hfb.2021.i4.s122

[29] **Morozov D.A., Glazkova A.V., Tyutyulnikov M.A., Iomdin B.L.** Generatsiya klyuchevykh slov dlya annotatsiy russkoyazychnykh nauchnykh statey [Generasia klyuchevix Slov dlya annotasiy russkoyazichnix nauchnix statey], Vestnik NGU. Seriya: Lingvistika i mezhkulturnaya kommunikatsiya [Bulletin of the NSU. Series: Linguistics and Intercultural Communication], 2023. №1. Available at: https://cyberlen-

inka.ru/article/n/generatsiya-klyuchevyh-slov-dlya-annotatsiy-russkoyazychnyh-nauchnyh-statey (accessed 11.03.2024).

[30] **Vanyushkin A.S., Grashchenko L.A.** O razmetke korpusov tekstov klyuchevymi slovami [On marking text corpora with keywords], Novyye informatsionnyye tekhnologii v avtomatizirovannykh sistemakh [New information technologies in automated systems], 21 (2018) 207–211.

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

[32] Postnauka [Post-science], URL: https://postnauka.ru/ (accessed 08.02.2024).

[33] Grammaticheskiye geny [Grammatical genes], Available at: https://postnauka.org/video/61500 (accessed 08.02.2024).

[34] Leksicheskaya sochetayemost [Lexical compatibility], Available at: https://postnauka.org/video/57524 (accessed 08.02.2024).

[35] ruTS, a library for statistics extraction from texts in Russian 2023. Available at: https://github.com/SergeyShk/ruTS (accessed 08.02.2024).

[36] **Song M., Jing L., Xiao L.** Importance Estimation from Multiple Perspectives for Keyphrase Extraction, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 2726–2736. DOI: 10.18653/v1/2021.emnlp-main.215

[37] **Ushio A., Liberatore F., Camacho-Collados J.** Back to the Basics: A Quantitative Analysis of Statistical and Graph-Based Term Weighting Schemes for Keyword Extraction. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 8089–8103. DOI: 10.18653/v1/2021.emnlp-main.638

[38] **Papagiannopoulou E., Tsoumakas G.** A review of keyphrase extraction, Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 10 (2) (2020). DOI: 10.1002/WIDM.1339

[39] **Grootendorst M.** KeyBERT: Minimal Keyword Extraction with BERT. 2020, Available at: http://doi.org/10.5281/zenodo.4461265 (accessed11.03.2024).

[40] **Song M., Feng Y., Jing L.** Utilizing BERT Intermediate Layers for Unsupervised Keyphrase Extraction, Proceedings of the 5th International Conference on Natural Language and Speech Processing (ICNLSP 2022). 2022 Pp. 277–281. Available at: https://aclanthology.org/2022.icnlsp-1.32 (accessed 11.03.2024).

[41] **Kulkarni M., Mahata D., Arora R., Bhowmik R.** Learning Rich Representation of Keyphrases from Text, Findings of the Association for Computational Linguistics: NAACL 2022. 2022. Pp. 891–906. DOI: 10.18653/v1/2022.findings-naacl.67

[42] **Liang X., Wu S., Li M., Li Z.** Unsupervised keyphrase extraction by jointly modeling local and global context, Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. Pp. 155–164. DOI: 10.18653/v1/2021.emnlp-main.14

[43] **Shcherba L.V.** Yazykovaya sistema i rechevaya deyatelnost [Language system and speech activity]. L., 1974.

[44] **Tsarfaty R., Seddah D.é, Kübler S., Nivre J.** Parsing Morphologically Rich Languages: Introduction to the Special Issue, Computational Linguistics, 39 (1) (2013) 15–22.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Гусева Дарья Дмитриевна
Daria D. Guseva
E-mail: daria.d.guseva@gmail.com
https://orcid.org/0009-0005-1753-9904

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

Поступила: 03.02.2024; Одобрена: 21.03.2024; Принята: 25.03.2024.
Submitted: 03.02.2024; Approved: 21.03.2024; Accepted: 25.03.2024.