Научная статья

УДК 004.89
DOI 10.25205/1818-7900-2022-20-3-5-13

Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов

Юлия Ивановна Бутенко

Московский государственный технический университет им. Н. Э. Баумана
Москва, Россия

iubutenko@bmstu.ru

Аннотация

В статье предложена модель текста научно-технической статьи для автоматизации разметки в корпусе научно-технических текстов. Обосновано, что при создании корпуса научно-технических текстов необходимо учитывать структурные особенности текстов научно-технических статей. Показана необходимость добавления структурной разметки в корпус научно-технических текстов. Отмечено, что тексты научно-технических статей имеют одинаковую для всех текстов этого класса структуру изложения материала, а также содержат ограниченный набор структурных элементов. Проанализированы особенности композиционной организации текстов научно-технических статей. Описано примерное содержание каждого из элементов структуры статьи. Представлена композиционная структура текстов научно-технических статей в нотациях Бекуса-Наура. Предложена модель текста научно-технической статьи в виде графа, вершинами и ребрами которого являются полноценные структурные элементы научно-технической статьи. Обосновано, что представление текста научно-технической статьи в виде графа дает возможность в процессе компьютерного анализа текста определить тип структурного элемента, степень вложенности, за счет подачи научно-технической статьи в виде конечного множества ее составных частей. Обосновано, что наличие структурной разметки в корпусе научно-технических текстов значительно расширит его исследовательский потенциал и послужит базой для задач автоматической обработки научно-технических текстов.

Ключевые слова

научная статья, структура текста, структурный элемент, корпус научно-технических текстов, модель текста

Для цитирования

Бутенко Ю. И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 5–13. DOI 10.25205/1818-7900-2022-20-3-5-13

Model of the Text of a Scientific and Technical Article for Markup in the Corpus of Scientific and Technical Texts

Yulia I. Butenko

Bauman Moscow State Technical University
Moscow, Russian Federation

iubutenko@bmstu.ru

Abstract

The paper proposes a model of the text of a scientific and technical article for the automation of markup in the corpus of scientific and technical texts. It is proved that when creating a corpus of scientific and technical texts, it is necessary to take into account the structural features of texts of scientific and technical articles. The necessity of adding structural markup to the corpus of scientific and technical texts has been shown. It is noted that the texts of scientific and technical articles have the same narration structure for all texts in this class, and also contain a limited set of structural elements. The features of compositional organization of the texts of scientific and technical articles are analyzed. The approximate content of each of the elements of article structure is described. Compositional structure of the texts of scientific and

technical articles in Bekus-Naur notation is presented. A model of the text of a scientific and technical article in the form of a graph, the vertices and edges of which are the full-fledged structural elements of a scientific and technical article, is proposed. It is proved that the representation of a text of scientific and technical article in the form of a graph makes it possible to determine the type of structural element and the degree of nesting in the process of computer analysis of the text by presenting the scientific and technical article as a finite set of its constituent parts. It is proved that the presence of structural markup in the corpus of scientific and technical texts significantly expands its research potential and serves as the basis for the tasks of automatic processing of scientific and technical texts.

Keywords

scientific article, text structure, structural element, corpus of scientific and technical texts, text model

For citation

Butenko Yu. I. Model of the Text of a Scientific and Technical Article for Markup in the Corpus of Scientific and Technical Texts. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 5–13. DOI 10.25205/1818-7900-2022-20-3-5-13

Введение

Отличительной особенностью современного мира является накопление огромного фонда различной информации, большая часть которой хранится в электронном виде. Одними из наиболее представительных информационных ресурсов текстов можно назвать электронные корпуса текстов [1–2]. Для описания подъязыка определенной предметной области необходимо использование специальных корпусов узкоспециальных текстов – корпусов научно-технических текстов, так как общие корпуса не подходят для изучения определенных предметных областей в силу их большого объема, разнообразного материала, а также отсутствия специальной терминологии [3].

Тексты, включаемые в корпус, проходят тщательный отбор и группируются в подкорпуса [4]. К источникам текстов, обеспечивающих репрезентативность корпуса научно-технических текстов, прежде всего, следует отнести научно-технические статьи, опубликованные в специализированных журналах.

Электронный корпус представляет собой коллекции текстов и их разметку, зависящую от типа исследования или задачи, для решения которой они созданы [3]. Разметка позволяет сделать корпус гораздо удобнее в использовании и является главной отличительной особенностью корпуса по сравнению с любыми другими коллекциями текстов. Таким образом, создание корпуса научно-технических текстов предполагает наличие лингвистической разметки, которая описывает сугубо лингвистические характеристики языковой выборки корпуса и представляет собой сложный процесс, требующий длительной и кропотливой работы над каждой лексической единицей, представленной в корпусе. Лингвистическая разметка обычно включает в себя разметку морфологическую, синтаксическую, семантическую [5]. Одним из ключевых аспектов проектирования корпусов является также метаразметка текстов – процесс приписывания тексту различных характеристик, описывающих обстоятельства его создания, автора, соотнесенность с определенным жанром и стилем изложения [6]. Основное назначение метаразметки – дать возможность пользователям корпуса настроить внешние параметры поиска текстов: например, осуществлять поиск по текстам, созданным авторами определенного года рождения, страны происхождения, гендерной принадлежности. Метаразметка содержит основную информацию о каждом тексте, включенном в корпус.

Стоит отметить, что научно-технические тексты обладают рядом специфических особенностей, которые требуют использования дополнительных видов разметки. К таким особенностям следует отнести композиционную структуру научно-технических текстов, которая может оказывать существенное влияние на результаты их автоматической обработки. Так, например, поиск информации в текстах стандартов необходимо проводить в два этапа. На первом этапе отбирать стандарты или их разделы, а затем искать в отобранном массиве необходимую информацию. Данный факт объясняется особенностями композиционной структуры текстов

стандартов, с одной стороны, и обобщенно-отвлеченным характером лексики, используемой при изложении требований стандартов, с другой стороны [7]. Наличие структурной разметки научно-технических текстов позволит отбирать для исследования только определенные структурные компоненты научно-технического текста, например, аннотации или введения.

Целью статьи является построение модели текста научно-технической статьи для структурной разметки в корпусе научно-технического текстов.

1. Научно-техническая статья как структурированный текст

Научно-техническая статья – это первичный письменный жанр научного дискурса, задачей которого является постановка и решение одной научной проблемы, имеет средний объем, конвенциальную структуру, системы ссылок и выходные данные [8]. Научно-техническим статьям присущи все стилевые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. Результаты анализа текстов статей представлены в таблице.

Результаты анализа текстов научно-технических статей
Results of the Analysis of Texts of Scientific and Technical Articles

Сфера функционирования и типовая ситуация общения	Наука и техника. Обмен информацией о решении научно-технических проблем и задач
Участники речи	Профессиональное сообщество ученых и инженеров. Специалисты, владеющие обширными знаниями о специализированной предметной области статьи
Функция речи	Профессиональная коммуникация специалистов разных специализированных предметных областей.
Тип содержания и предмет речи (тема)	Содержание – конкретное. Тема – результаты определенного исследования или обзор текущего состояния предмета исследования.
Коммуникативная цель	Обмен ясно, кратко и достоверно изложенной информацией о результатах исследования в некоторой предметной области
Стилеобразующие признаки	Ясность, точность, логичность, объективность и точность. Обезличенность информации. Шаблонность, четкое закрепление места за элементами.
Форма бытия	Письменная, в виде текстов установленной формы в научно-технических или специализированных периодических изданиях определенных предметных отраслей.

К ключевым элементам структуры научнотехнической статьи с точки зрения их функциональных и лексико-грамматических особенностей относят [9–10]:

1. Код УДК.
2. Название.

Отражает содержание статьи, обычно состоит не более одиннадцати слов без учета союзов и предлогов, возможны варианты названий, включающие от двух до пятнадцати слов.

3. Информация об авторах.

Включает имя и фамилию автора, место работы автора и контактные данные. Зачастую количество авторов не ограничено, но некоторые специализированные журналы могут устанавливать ограничения, например, не более пяти авторов. Место работы автора включает название организации с указанием города и страны, реже с полным адресом места работы автора. Контактные данные чаще всего представлены адресом электронной почты. Если авторов несколько, то указывается имя автора ответственного за переписку.

4. Аннотация и ключевые слова.

Стандартная аннотация содержит информацию об объекте, цели и методах исследования, основных результатах и выводах. Ключевые слова используют для быстрого поиска информации в больших коллекциях документов.

5. Введение.

Во введении акцентируют внимание на новизне и актуальности работы, приводят обзор литературы предметной области исследования. Элемент «Введение» зачастую заканчивается формулировкой цели работы.

6. Основная часть.

Способы организации данного раздела варьируют в зависимости от цели исследования, обычно выделяют разделы «Материалы и методы», «Обсуждение», «Результаты».

7. Заключение.

Отражает факты, сделанные на основе проведенной работы.

8. Слова благодарности.

Благодарят тех, кто оказал помощь в проведении работы.

9. Ссылки на литературу.

Представляет собой список литературы, процитированной в тексте статьи.

В результате анализа композиционной структуры текстов научно-технических статей, выявлено, что они имеют ярко-выраженную структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа, взаимоувязанную систему заголовков разделов и подразделов.

2. Формальная модель текстов научно-технических статей

В качестве исходных данных выступают результаты композиционного анализа текстов стандартов, полученные в выше. Для решения задачи необходимо разработать формальные средства композиционной структуры научно-технических статей, использование которых позволит осуществлять структурную разметку в корпусе научно-технических текстов.

В результате будет получена модель формального представления текстов научно-технических статей, которая даст возможность при разметке корпуса научно-технических текстов учитывать их композиционную структуру.

Прежде всего, необходимо установить общую структуру языка стандартов как совокупности элементов. Структура совокупности выявлена в результате проведения тщательной классификации исследуемых единиц, то есть их иерархического распределения по определенному признаку на основные разделы, которые далее распадаются на подразделы, пункты, подпункты, требования, которые в свою очередь разбиваются на отдельные более мелкие базисные единицы – предложения. Все это вместе образует номенклатуру – полный подробный перечень отдельных элементов изучаемой совокупности [11].

Любую систему можно представить как некоторую совокупность взаимосвязанных элементов. Каждая из таких систем S_j является отделенной системой (научный / официально-деловой стиль) и может быть представлена как некоторая часть (подсистема) более общей системы S (суперсистема – русский язык) $S_j \in S$.

Взаимосвязь между системами S_j и S построена по иерархическому принципу, который предусматривает подчиненность подсистемы S_j суперсистеме S, в плане своего структурного расположения, так и в плане функционально-коммуникативной направленности составных частей. Отсюда вытекает, что любую систему (совокупность) S можно разделить на подсистемы разных рангов $S_1 – S_2 – S_3$ определенный естественный язык – стиль – жанр), проводя процесс членения по определенным признакам до получения составных элементов. При этом каждая операция членения системы порождает отдельные подистемы, что обеспечивает построение некоторого дерева метасистем S, на котором выделены отдельные подсистемы (S_1; S_2; S_3), которые относятся к разным уровням (S_{12}; S_{121}; S_{1211}). Членение системы может быть произведено рядом способов, при этом генерируется разное количество частей (подсистем, базисных элементов) [12].

Композиционная структура научно-технической статьи в первом приближении состоит из трех частей. Первая часть включает реферативный раздел, под которым понимается совокупность основных конвенциональных элементов, используемая для формальной идентификации первичного документа с учетом его природы, элементов и порядка внешних признаков, которые отличают его от других. Компонентами реферативного раздела являются: код УДК, название статьи, информация об авторах, место работы авторов, аннотация (реферат), ключевые слова. Код УДК не является обязательным элементом статьи. Реферативный раздел зачастую имеет переводную версию на английском языке. Вторая часть статьи представляет корпус научно-технической статьи, который обычно состоит из пяти основных структурных элементов: введение, материал и методы, результаты, обсуждение результатов, заключение. Стоит отметить, что названия разделов, кроме «введения» и «заключения» могут отличаться, однако их содержание соответствует по смыслу заявленным. В связи с тем, что разные научные журналы формируют собственные требования к рукописям статей, структурные элементы корпуса научно-технической статьи могут быть разбиты на пункты и подпункы, которые в свою очередь разбиваются на абзацы и затем предложения. Третья часть статьи является вспомогательным аппаратом публикации, который включает примечание и ссылки на источники [13–14].

В нотациях Бекуса-Наура композиционную структуру текстов научно-технических статей можно задать следующим образом:

$$St_i ::= \langle X^1, X^2, X^3 \rangle$$

где X^1– реферативный раздел научно-технической статьи, X^2– корпус научно-технической статьи, X^3– информативный раздел научно-технической статьи.

X^1 – реферативный раздел научно-технической статьи, состоящий из следующих элементов:

$$X^1 ::= \langle x_{11},x_{12},x_{13},x_{14},x_{15},x_{16},x_{17}\rangle|\langle x_{12},x_{13},x_{14},x_{15},x_{16},x_{17}\rangle$$

где x_{11} – код УДК, x_{12} – название статьи, x_{13} – информация об авторах, x_{14} – место работы авторов, x_{15} – контактная информация авторов, x_{16} – аннотация, x_{17} – ключевые слова.

X^2 – корпус научно-технической статьи можно представить в виде набора из следующих элементов:

$$X^2 ::= \langle x_{21},x_{22},x_{23},x_{24},x_{25}\rangle|\langle x_{21},x_{22},x_{23},x_{25}\rangle|\langle x_{21},x_{23},x_{25}\rangle|\langle x_{21},x_{23},x_{24},x_{25}\rangle,$$

где x_{21} – введение, x_{22} – материал и методы, x_{23} – результаты, x_{24} – обсуждение результатов, x_{25} – заключение.

X^3 – информативный раздел научно-технической статьи, для которого справедливо

$$X^3 ::= \langle x_{31},x_{32}\rangle|\langle x_{32}\rangle,$$

где x_{31} – примечания, x_{32} – ссылки на источники.

На рисунке представлена полученная структурная схема элементов текста научно-технической статьи.

На основе проведенного анализа композиционной структуры текстов научно-технических статей модель текста научно-технической статьи *St* целесообразно представить в виде:

$$St = \left\langle E^L, R \right\rangle$$

где E – структурный элемент, R – отношения между структурными элементами, L – уровень структурного элемента. При этом $L = \{l_1,\ldots l_5\}$, где l_1 – раздел, l_2 – пункт, l_3 – подпункт, l_4 – абзац, l_5 – предложение.

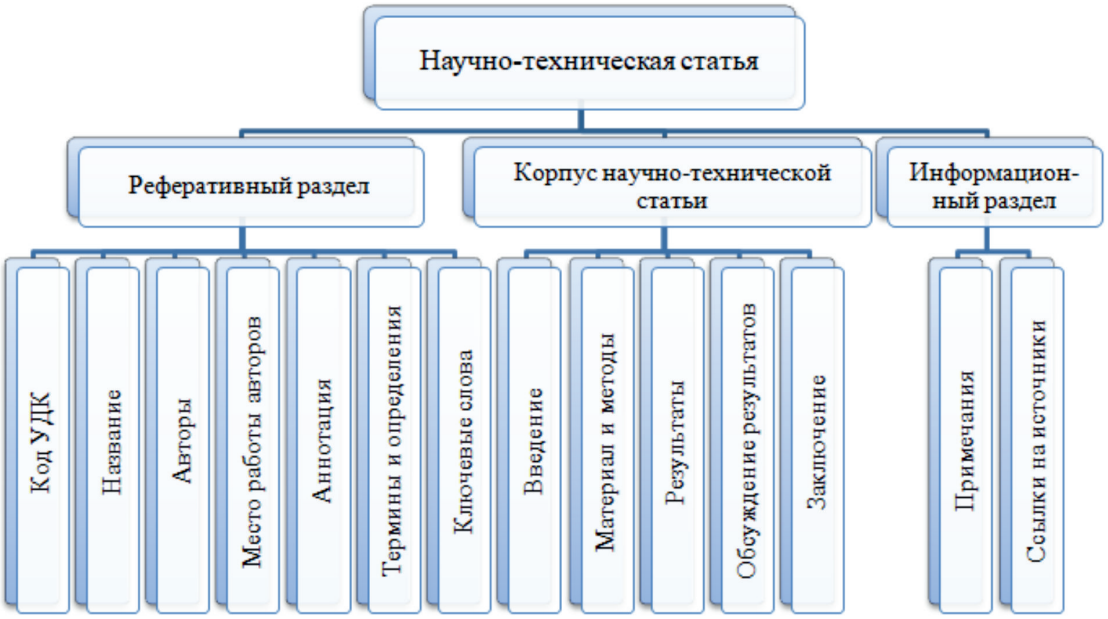

Рис. Структурные элементы текстов научно-технических статей
Fig. Structural elements of texts of scientific and technical articles

Такое представление научно-технической статьи порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста научно-технической статьи в целом. Таким образом, модель композиционной структуры текста научно-технической статьи – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки текста научно-технической статьи при создании корпуса научно-технических текстов значительно расширит исследовательский потенциал корпуса, что в свою очередь позволит при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности.

Заключение

В настоящее время для описания подъязыка определенной предметной области необходимо использование специальных корпусов узкоспециальных текстов – корпусов научно-технических текстов. К источникам текстов для корпуса научно-технических текстов отнесены научно-технические статьи. Показано, что электронный корпус представляет собой коллекции

текстов и их разметку: морфологическую, синтаксическую, семантическую, метаразметку. Выявлено, научно-технические тексты обладают рядом специфических особенностей в композиционной структуре.

Научно-техническая статья – это первичный письменный жанр научного дискурса, задачей которого является постановка и решение одной научной проблемы, имеет средний объем, конвенциальную структуру, системы ссылок и выходные данные. Научно-техническим статьям присущи все стилевые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. К ключевым элементам структуры научно-технической статьи с точки зрения их функциональных и лексико-грамматических особенностей относят: название, информацию об авторах, аннотацию и ключевые слова, введение, основную часть, заключение, слова благодарности и ссылки на литературу.

Композиционная структура научно-технической статьи состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела. Компонентами реферативного раздела являются: код УДК, название статьи, информация об авторах, место работы авторов, аннотация (реферат), ключевые слова. Код УДК не является обязательным элементом статьи. Корпус научно-технической статьи состоит из «Введения», «Материалов и методов», «Результатов», «Обсуждения результатов», «Заключения». Информационный раздел включает примечание и ссылки на источники. Композиционная структура текстов научно-технических статей задана в нотациях Бекуса-Наура. Построена модель текста научно-технической статьи для структурной разметки в корпусе научно-технического текстов, которая порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста научно-технической статьи в целом.

Список литературы

1. **Захаров В. П.** Корпуса русского языка // Труды института русского языка имени В. В. Виноградова. 2015. Т.6. С. 20–65.
2. **Кружков М. Г.** Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. 2015. Т. 25, № 2. С. 140–159.
3. **Нагель О. В.** Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. № 4. С. 53–59.
4. **Соловьева А. Е.** Англоязычные тексты военной авиации как основа лингвистического корпуса // Балтийский гуманитарный журнал. 2019. № 3 (28). С. 369–372.
5. **Лесников В. С.** Виды разметок текстовых корпусов русского языка // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2019. №9. С. 27–30.
6. **Ванюшкин А. С.** О разметке корпусов текстов ключевыми словами // Новые информационные технологии в автоматизированных системах. 2018. № 21. С. 207–211.
7. **Бутенко Ю. И.** Влияние лингвистических особенностей текстов стандартов на информационный поиск // Филологические науки. Научные доклады высшей школы. 2019. № 6. С. 29–35. DOI: 10.20339/PhS.6-19.029
8. **Попова Т. Г.** Структура испанской научно-технической статьи как первичного жанра научного дискурса // Вестник Российского университета дружбы народов. Серия: Русский и иностранные языки, и методика их преподавания. 2004. № 1. С. 108–115.
9. **Романов Д. А.** Кратко о структуре экспериментальной научной статьи на английском языке // Вестник Казанского технологического университета. 2014. Т. 17, № 6. С. 325–327.
10. **Раицкая Л. К.** Структура научной статьи по политологии и международным отношениям в контексте качества научной информации // Полис. Политические исследования. 2019. № 1. С. 167–181.

11. **Sidnyaev N. I.** Mathematical apparatus for engineering-linguistic models // AIP Conference Proceedings. 2019. Vol. 2195. No. 1. P. 020033. DOI: 10.1063/1.5140133
12. **Бутенко Ю. И.** Модель текста стандарта при информационном поиске в коллекции документов нормативной базы // Вестник компьютерных и информационных технологий. 2020. – Т. 17, № 11. С. 23–32. DOI: 10.14489/vkit. 2020.11
13. **Попова Н. Г.** Введение к научной статье на английском языке: структура и композиция // Высшее образование в России. 2015. № 6. С. 52–58.
14. **Иванов В. П.** Как написать научную статью (структура материала и организация работы) // Вестник Полоцкого государственного университета. Серия В. Промышленность. Прикладные науки. 2016. № 3. С. 195.

References

1. **Zakharov V. P.** Russian corpora. *Proceedings of Vinogradov Institute of the Russian Language*, 2015. Vol. 6, pp. 20–65. (in Russ.)
2. **Nagel O. V.** Corpus linguistics and its use in computerized language learning. *Language and Culture*, 2008. No. 4, pp. 53–59. (in Russ.)
3. **Kruzhkov M. G.** Information resources of contrastive linguistic research: electronic corpus of texts. *Systems and means of informatics*, 2015. Vol. 25, no. 2, pp. 140–159. (in Russ.)
4. **Lesnikov V. S.** Types of markup of text corpus of the Russian language. *Scientific and Technical Information. Series 2. Information processes and systems*, 2019. No. 9, pp. 27–30. (in Russ.)
5. **Butenko Iu. I.** Model of the text of the standard in the information search in the collection of documents of the normative base. *Bulletin of Computer and Information Technologies*, 2020. Vol. 17, no. 11, pp. 23–32. DOI: 10.14489/vkit. 2020.11 (in Russ.)
6. **Butenko Iu. I., Semenova E. L.** Influence of linguistic features of standards texts on information search. *Philological Sciences. Scientific reports of higher school*, 2019. No. 6, pp. 29–35. DOI: 10.20339/PhS.6-19.029 (in Russ.)
7. **Sidnyaev N. I., Butenko J. I., Garazha V. V.** Mathematical apparatus for engineering-linguistic models. AIP Conference Proceedings, 2019. Vol. 2195, no. 1, p. 020033. DOI: 10.1063/1.5140133
8. **Romanov D. A.** Briefly about the structure of the experimental scientific article in English. *Bulletin of Kazan Technological University*, 2014. Vol. 17, no. 6, pp. 325–327. (in Russ.)
9. **Raitskaya L. K.** Structure of a scientific article on political science and international relations in the context of the quality of scientific information. Polis. *Politicheskie issledovaniye*, 2019. No. 1, pp. 167–181. (in Russ.)
10. **Popova T. G.** Structure of the Spanish scientific and technical article as a primary genre of scientific discourse. *Bulletin of the Peoples' Friendship University of Russia. Series: Russian and foreign languages and the methodology of their teaching*, 2004. No. 1, pp. 108–115. (in Russ.)
11. **Popova N. G.** Introduction to the scientific article in English: structure and composition. *Higher Education in Russia*, 2015. No. 6, pp. 52–58. (in Russ.)
12. **Ivanov V. P.** How to write a scientific article (material structure and work organization). *Bulletin of Polotsk State University. Series B. Industry. Applied sciences*, 2016. No. 3, p. 195. (in Russ.)
13. **Vanyushkin A. S., Grashchenko L. A.** On the markup of corpus texts with keywords. *New Information Technologies in Automated Systems*, 2018. No. 21, pp. 207–211. (in Russ.)
14. **Solov'eva A. E.** English-language texts of military aviation as the basis of linguistic corpus. *Baltic humanitarian journal*, 2019. No. 3(28), pp. 369–372. (in Russ.)

Информация об авторе

Бутенко Юлия Ивановна, кандидат технических наук, доцент кафедры «Романо-германские языки», Московский государственный технический университет им. Н. Э. Баумана (Москва, Россия)

Information about the author

Butenko I. Yulia, Candidate of Technical Sciences, Associate Professor of the Department of Romano-Germanic Languages, Bauman Moscow State Technical University (Moscow, Russian Federation)

Статья поступила в редакцию 27.07.2022;
одобрена после рецензирования 07.11.2022; принята к публикации 07.11.2022
The article was submitted 27.07.2022;
approved after reviewing 07.11.2022; accepted for publication 07.11.2022