УДК 004.89
DOI 10.52575/2687-0932-2021-48-1-123-129

Модель учебно-научного текста для разметки корпуса научно-технических текстов

Бутенко Ю.И.
Московский государственный технический университет им. Н.Э. Баумана,
e-mail: iubutenko@bmstu.ru

Аннотация. В статье предложена модель структуры учебно-научного текста для разметки корпуса научно-технических текстов. Обоснована необходимость учитывать композиционную структуру научно-технических текстов при разметке корпуса. Отмечено, что учебно-научные тексты имеют одинаковую для всех текстов этого класса структуру изложения материала, а также содержат ограниченный набор структурных элементов. Охарактеризованы структурные элементы учебно-научного текста. Представлена композиционная структура учебно-научных текстов в нотациях Бекуса-Наура. Предложена модель учебно-научного текста в виде графа, вершинами и ребрами которого являются полноценные структурные элементы учебно-научного текста.

Ключевые слова: структура текста, структурный элемент, модель текста, учебно-научный текст, корпус научно-технических текстов.

Для цитирования: Бутенко Ю.И. 2021. Модель учебно-научного текста для разметки корпуса научно-технических текстов. Экономика. Информатика, 48 (1): 123–129. DOI: 10.52575/2687-0932-2021-48-1-123-129.

Model of educational texts for markup in a corpus of scientific and technical texts

Butenko Iu.I.
Bauman Moscow State Technical University,
e-mail: iubutenko@bmstu.ru

Abstract. The article proposes a model for the structure of scholarly texts for marking up a corpus of scientific and technical texts. The article substantiates the need to take into account the composition structure of scholarly texts when marking up the corpus. The necessity of adding structural markup to the corpus of scientific and technical texts has been shown. It is noted that scholarly texts have the same for all the texts of this class, as well as contain a limited set of structural elements. The structural elements of an scholarly texts are characterized. The approximate content of each element of scholarly texts is described. The composition structure of scholarly text is presented in Bekus-Naur notation. The model of scholarly texts is proposed in the form of graph, the nodes and edges of which are full-fledged structural elements of scholarly texts. It is proved that the representation of scholarly texts in the form of a graph makes it possible to determine the type of a structural element, the degree of nesting, in the process of computer analysis of the text, by presenting the scholarly text as a finite set of its constituent parts.

Key words: text structure, structural element, text model, scholarly text, corpus of scientific and technical texts.

For citation: Butenko Iu.I. 2021. Model of educational texts for markup in a corpus of scientific and technical texts. Economics. Information technologies, 48 (1): 123–129 (in Russian). DOI: 10.52575/2687-0932-2021-48-1-123-129.

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 [Захаров, 2015; Кружков, 2015]. 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 [Нагель, 2008].

Электронный корпус представляет собой коллекции текстов и их разметку, зависящую от типа исследования или задачи, для решения которой они созданы [Соловьева, 2019]. 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 [Лесников, 2019].

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 [Ванюшкин, Гращенко, 2018]. 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 [Бутенко, Семенова, 2019]. Наличие структурной разметки научно-технических текстов позволит не только отбирать для исследования только определенные структурные компоненты научно-технического текста, например, введения или определенные главы или фрагменты текста, но исключать «лишние» для решаемой задачи элементы, например, список литературы, благодарности, предисловие и т. д [Бутенко, 2020].

К источникам текстов, обеспечивающих репрезентативность корпуса научно-технических текстов, следует отнести учебно-научные тексты, представленные учебниками, учебными пособиями, конспектами лекций и пр.

Целью статьи является построение модели учебно-научного текста для структурной разметки корпуса научно-технического текстов.

Композиционные особенности учебно-научного текста

Под учебно-научным текстом принято понимать книгу, в которой систематически изложены основы знаний в определенной предметной области на уровне современных достижений науки и техники [Егоров и др., 2008]. К учебно-научным текстам выдвигают такие же требования, как и к научным текстам, а именно: логичность, краткость, ясность, последовательность изложения материала, абстрактность [Тюрина, 2007]. В работе [2005] Тюрина Л.Г. описывает педагогическую модель учебной книги, в которой выделяет три подсистемы: предметную, дидактическую и аксиологическую. Связь между подсистемами представлена следующим образом: сначала излагается вербально или наглядно система знаний (предмет-

ная подсистема) о некоторой предметной области, затем идут материалы, формирующие необходимые навыки и умения – вопросы, задания, упражнения (дидактическая подсистема). При этом элементы двух указанных подсистем строятся с учетом мировоззренческого, воспитательного воздействия на читателя – аксиологическая подсистема.

Стоит отметить, что композиционная структура учебно-научного текста также отражает выше описанную модель и включает в себя текст, как главный компонент, так и внетекстовые, вспомогательные компоненты, к которым относят аппарат организации усвоения (вопросы и задания, памятки или инструктивные материалы, таблицы и шрифтовые выделения, подписи к иллюстративному материалу и упражнения); собственно иллюстративный материал; аппарат ориентировки, включающий предисловие, примечание, приложения, оглавление, указатели [Рыбакова, 2011]. Структура учебно-научного текста показана на рис. 1. Как видно из рисунка, учебно-научные тексты имеют сложную многокомпонентную структуру.

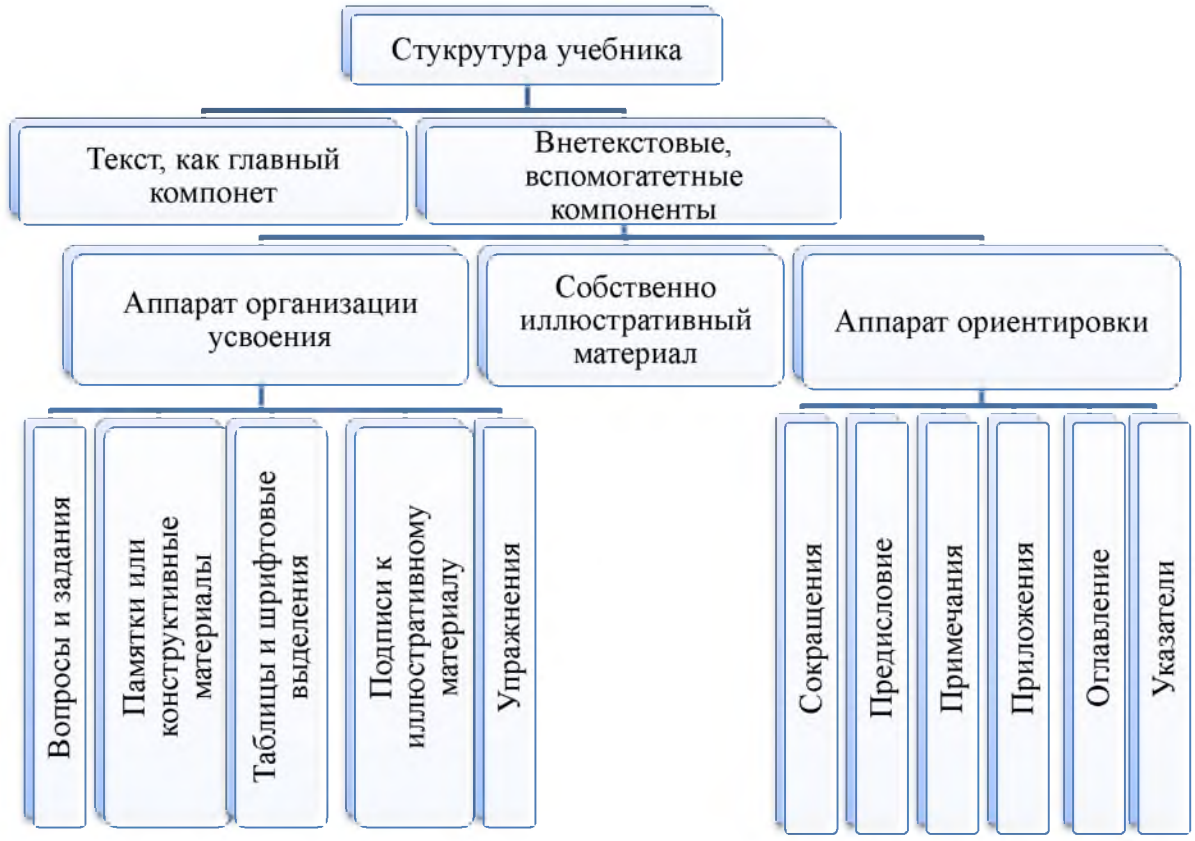

Рис. 1. Структура учебно-научного текста по любой предметной области
Fig. 1. Structure of scholarly texts in a subject field

При реализации структурной разметки корпуса научно-технических текстов необходимо проанализировать каждый компонент учебно-научного текста с целью определения оптимальных вариантов их разметки. Например, необходимость включения текстов упражнений в основную часть корпуса, так как при составлении упражнения могут быть использованы разные предметные области или их комбинации, при этом исследуемая предметная область в силу специфики не будет отражена. Ярким примером могут служить учебники и учебные пособия по научно-техническому переводу, где тематика текстов упражнений зачастую кардинально отличаются от предмета содержания учебно-научного текста.

В результате анализа композиционной структуры учебно-научных текстов выявлено, что они имеют ярко выраженную структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа.

Формальная модель учебно-научных текстов для представления в корпусе научно-технических текстов

В качестве исходных данных выступают результаты композиционного анализа учебно-научных текстов, полученные в предыдущем разделе. Для решения задачи необходимо разработать формальные средства композиционной структуры учебно-научных текстов, использование которых позволит осуществлять структурную разметку в корпусе научно-технических текстов. В результате будет получена модель формального представления учебно-научных текстов, которая даст возможность при разметке корпуса научно-технических текстов учитывать их композиционную структуру.

Композиционная структура учебно-научного текста в первом приближении состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела. При этом структурные элементы научно-учебных текстов можно разделить на обязательные и факультативные, то есть те, которые приводятся в зависимости от необходимости. К обязательным элементам относят название, автор(ы), оглавление, введение, основной текст и ссылки на источники. Несмотря на то, что название и авторы являются элементами метатекстовой разметки, они также несут значимую информацию при автоматической обработке научно-технических текстов и являются полноценными объектами лингвистического исследования.

Оглавление является важным элементом учебно-научного текста, дающим общее представление о структуре и проблематике учебного пособия, отражает взаимосвязи всех компонентов учебника и является средством навигации по научно-учебному тексту. Если у разных разделов учебника разные авторы, то вместо структурного элемента «Оглавление» используют элемент «Содержание» [Лыков, 2008]. Введение в учебно-научном тексте обычно представляет читателю информацию о текущем состоянии проблем и явлений в некоторой предметной области, обзор взглядов и литературных источников, базовую терминологию и др. Основной текст раскрывает содержание, обеспечивает последовательное, полное и аргументированное изложение материала и служит основным источником учебно-научной информации, обязательной для изучения и усвоения. Структурный элемент «Ссылки на источники» содержит основные или рекомендуемые литературные источники для углубленного или самостоятельного изучения определенных тем некоторой предметной области.

К факультативным структурным элементам учебно-научных текстов относят «Предисловие», «Вопросы», «Задания и упражнения», «Примечания» и «Приложения». Предисловие представляет собой текст, предваряющий изложение основного материала, содержит цель и особенности издания, отражает структуру и краткую характеристику всех разделов. В зависимости от типа литературы и вида издания выделяют ряд разновидностей «Предисловия»: «От автора», «От редактора», «Вместо предисловия» и др. Структурные элементы «Вопросы» и «Задания и упражнения» относят к аппарату организации усвоения материала, призванные стимулировать познавательную деятельность в процессе усвоения материала. Структурный элемент «Примечания» являются краткими дополнениями, пояснениями и уточнениями к основному учебно-научному тексту, бывают внутритекстовые, подстрочные и затекстовые. Авторы используют этот структурный элемент с целью дополнения основного учебно-научного текста [Лупачев, Павлюк, 2011]. В «Приложения» включают материал, служащий дополнением основного текста, куда входят официальные и справочные материалы – таблицы, схемы, словари, чертежи, списки, вклейки, иллюстрации, карты, рисунки.

В нотациях Бекуса-Наура композиционную структуру учебно-научных текстов можно задать следующим образом:

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

где X^1 – реферативный раздел учебно-научного текста, X^2 – корпус учебно-научного текста, X^3 – информативный раздел научно-учебного текста.

X^1 – реферативный раздел учебно-научного текста, состоящий из следующих элемен-

тов:

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

где x_{11} – название, x_{12} – автор(ы), x_{13} – оглавление , x_{14} – введение, x_{15} – предисловие.

X^2 – корпус учебно-научного текста, который можно представить в виде набора из следующих элементов:

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

где x_{21} – основной текст, x_{22} – вопросы, x_{23} – задания и упражнения.

X^3 – информативный раздел учебно-научного текста, для которого справедливо

$$X^3 ::= \langle x_{31}, x_{32}, x_{33} \rangle | \langle x_{31}, x_{33} \rangle | \langle x_{32}, x_{33} \rangle | \langle x_{33} \rangle,$$

где x_{31} – примечания, x_{32} – приложения, x_{33} – ссылки на источники.

На рис. 2 представлена полученная структурная схема элементов учебно-научного текста.

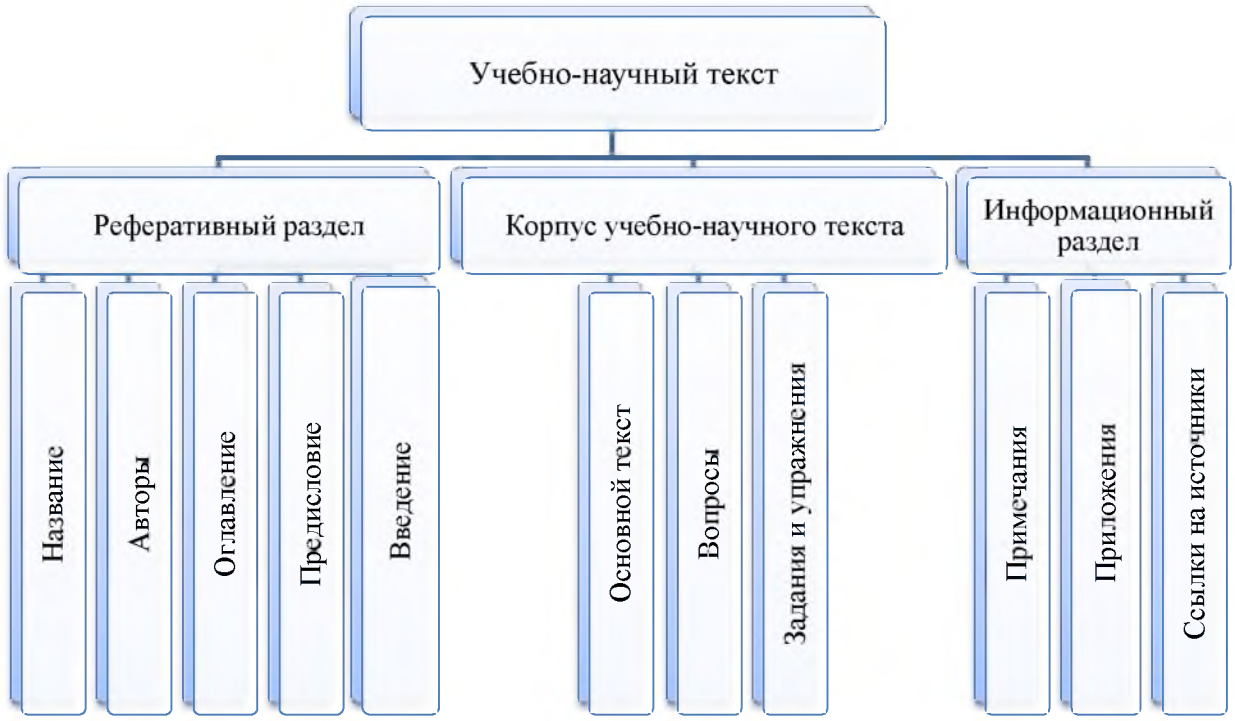

Рис. 2. Структурные элементы учебно-научных текстов
Fig. 2. Structure of scholarly texts in a subject field

На основе проведенного анализа композиционной структуры учебно-научных текстов модель учебно-научного текста *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.

Представление текста в виде упорядоченного набора структурных элементов дает возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего учебно-научного текста. Таким образом, модель композиционной структуры учебно-научного текста – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки учебно-научных текстов при созда-

нии 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 учебно-научные тексты. 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. Учебно-научным текстам присущи все стилевые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. Композиционная структура учебно-научного текста состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела. К ключевым элементам структуры учебно-научных текстов с точки зрения их функциональных и лексико-грамматических особенностей относят: название, информацию об авторах, введение, основной текст и список источников. Композиционная структура учебно-научных текстов задана в нотациях Бекуса-Наура. Построена модель учебно-научного текста для структурной разметки в корпусе научно-технических текстов, которая порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста в целом.

Список литературы

1. Бутенко Ю.И. 2020. Модель текста стандарта при информационном поиске в коллекции документов нормативной базы. Вестник компьютерных и информационных технологий, 17 (11): 23–32. DOI: 10.14489/vkit. 2020.11.pp.023-032.

2. Бутенко Ю.И., Семенова Е.Л. 2019. Влияние лингвистических особенностей текстов стандартов на информационный поиск. Филологические науки. Научные доклады высшей школы, 6: 29-35. DOI: 10.20339/PhS.6-19.029.

3. Ванюшкин А.С., Гращенко Л.А. 2018. О разметке корпусов текстов ключевыми словами. Новые информационные технологии в автоматизированных системах, 21: 207–211.

4. Егоров В.В., Скибицкий Э.Г., Храпченков В.Г. 2008. Педагогика высшей школы. Новосибирск. САФБД: 260.

5. Захаров В. П. 2015. Корпуса русского языка. Труды института русского языка имени В.В. Виноградова, 6: 20–65.

6. Кружков М.Г. 2015. Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов. Системы и средства информатики, 25 (2): 140–159.

7. Лесников В.С. 2019. Виды разметок текстовых корпусов русского языка. Научно-техническая информация. Серия 2. Информационные процессы и системы, 9: 27–30.

8. Лупачев В.Г., Павлюк С.К. 2011. Методические основы и принципы разработки учебной литературы: методическое пособие для слушателей курсов повышения квалификации и переподготовки кадров; под ред. В.А. Сидорова. Минск. БНТУ: 63.

9. Лыков М.Н. 2008. Оглавление как структурный элемент вузовского учебника (на примере учебника по истории отечества для высшей школы). Альманах современной науки и образования, 10-1 (17): 102–105.

10. Нагель О.В. 2008. Корпусная лингвистика и ее использование в компьютеризированном языковом обучении. Язык и культура, 4: 53–59.

11.Рыбакова Г.Р. 2011. О категории «учебный текст» в научной литературе. Научное обозрение. Серия 2: Гуманитарные науки, 6: 64–73.

12.Соловьева А.Е. 2019. Англоязычные тексты военной авиации как основа лингвистического корпуса. Балтийский гуманитарный журнал, 3 (28): 369–372.

13.Тюрина Л.Г. 2007. Особенности текста учебной книги. Известия высших учебных заведений. Проблемы полиграфии и издательского дела, 3: 70–73.

14.Тюрина Л.Г. 2005. Состав и структура учебной книги как педагогической системы. Известия высших учебных заведений. Проблемы полиграфии и издательского дела, 4: 78–88.

References

1. Butenko Ju.I. 2020. Model of the text of the standard in the information search in the collection of documents of the normative base. Vestnik komp'juternyh i informacionnyh tehnologij. 17 (11): 23–32. DOI: 10.14489/vkit. 2020.11.pp.023-032 (in Russian)

2. Butenko Ju.I., Semenova E.I. 2019. Influence of linguistic features of standards texts on information retrieval. Filologicheskie nauki. Nauchnye doklady vysshej shkoly, 6: 29–35. DOI: 10.20339/PhS.6-19.029 (in Russian)

3. Vanyushkin A.S., Grashchenko L.A. 2018. On keyword markup of corpora of texts. Novye informacionnye tehnologii v avtomatizirovannyh sistemah, 21: 207–211. (in Russian)

4. Egorov V.V., Skibitsky E.G., Khrapchenkov V.G. 2008. Pedagogy of higher school. Novosibirsk. SAFBD: 260. (in Russian)

5. Zakharov V.P. 2015. Corpus of the Russian language. Trudy instituta russkogo jazyka imeni V.V. Vinogradova, 6: 20–65. (in Russian)

6. Kruzhkov M.G. 2015. Information resources of contrastive linguistic research: electronic corpus of texts. Systems and Means of Informatics, 25 (2): 140–159. (in Russian)

7. Lesnikov V.S. 2019. Types of markings of text corpuses of the Russian language. Nauchno-tehnicheskaja informacija. Serija 2. Informacionnye processy i sistemy, 9: 27–30. (in Russian)

8. Lupachev V.G., Pavlyuk S.K. 2011. Methodological Principles and Principles of Developing Educational Literature: Methodological Handbook for Students' Professional Development and Retraining Courses, ed. by V.A. Sidorov. Minsk. BNTU: 63. (in Russian)

9. Lykov M.N. 2008. Table of contents as a structural element of a university textbook (by the example of a textbook on the history of the Fatherland for high school). Al'manah sovremennoj nauki i obrazovanija, 10-1 (17): 102–105. (in Russian)

10.Nagel O.V. 2008. Corpus linguistics and its use in computerized language teaching. Language and Culture, 4: 53-59. (in Russian)

11.Rybakova G.R. 2011. On the category of "learning text" in scientific literature. Scientific Review. Series 2: Humanities, 6: 64–73. (in Russian)

12.Solovyova A.E. 2019. English-language military aviation texts as the basis of a linguistic corpus. Baltijskij gumanitarnyj zhurnal, 3 (28): 369–372. (in Russian)

13.Tyurina L.G. 2007. Peculiarities of the text of a textbook. Izvestia vyssheye izuchennykh uchebnykh obrazovatel'nykh. Problemy poligrafii i izdatel'skogo dela, 3: 70–73. (in Russian)

14.Tyurina L.G. 2005. Composition and structure of an educational book as a pedagogical system. Izvestia vysshee uchebnykh obrazovaniye. Problemy poligrafii i izdatel'skogo dela, 4: 78–88. (in Russian)

ИНФОРМАЦИЯ ОБ АВТОРЕ

Бутенко Юлия Ивановна, кандидат технических наук, доцент кафедры теоретической информатика и компьютерных технологий Московского государственного технического университета им. Н.Э. Баумана, г. Москва, Россия

INFORMATION ABOUT THE AUTHOR

Iuliia I. Butenko, Candidate of Technical Sciences, Associate professor of the Department Theoretical Informatics and Computer Technologies, Bauman Moscow State Technical University, Moscow, Russia