Весці БДПУ. Серыя 1. 2023. № 2. С. 84–88

УДК 81’33

ЛИНГВИСТИЧЕСКИЙ ПОТЕНЦИАЛ КОРПУСНОГО ИССЛЕДОВАНИЯ АНГЛОЯЗЫЧНОГО ЭКОЛОГИЧЕСКОГО ДИСКУРСА

А. А. Баркович,
доктор филологических наук, заведующий кафедрой информатики и прикладной лингвистики МГЛУ;
О. А. Рипинская,
магистрант МГЛУ

Поступила в редакцию 16.03.23.

UDC 81’33

LINGUISTIC POTENTIAL OF CORPUS RESEARCH OF ENGLISH-LANGUAGE ENVIRONMENTAL DISCOURSE

A. Barkovich,
Doctor of Philology, Head of the Department of Informatics and Applied Linguistics, MSLU;
O. Ripinskaya,
Master Student of MSLU

Received on 16.03.23.

Статья посвящена рассмотрению особенностей лингвистического анализа больших массивов языкового материала с использованием корпусного инструментария. С помощью сплошной выборки был агрегирован представительный массив текстов экологического дискурса *EuroNews Green Corpus*. Понятийно-концептуальная идентичность экологического дискурса обеспечена его востребованностью и постоянным расширением релевантной тематики и, соответственно, лингвистической проблематики. Для изучения данного материала были использованы возможности находящегося в открытом доступе интернет-ресурса *Sketch Engine*. Актуальные корпусные методики были задействованы не только при автоматизации сбора статистических данных, но и для их оценки. Также была выполнена комплексная аналитическая работа по интерпретации, структуризации и моделированию данного дискурса. Использование корпусных методик позволило выполнить метаязыковое описание экологического дискурса с использованием современных программных средств. Задействованный в исследовании потенциал корпусного исследования англоязычного экологического дискурса может быть продуктивно реализован для дальнейшей лингвистической работы.

Ключевые слова: экологический дискурс, корпусный анализ, лингвистический потенциал, *EuroNews Green Corpus*, *Sketch Engine*, статистика, методика.

The article deals with the peculiarities of linguistic analysis of large arrays of linguistic material using corpus tools. A representative array of *EuroNews Green Corpus* environmental discourse texts was aggregated by means of continuous sampling. The conceptual and conceptual identity of the ecological discourse is ensured by its relevance and constant expansion of relevant topics and, accordingly, linguistic issues. To study this material, we used the capabilities of the publicly available Internet resource *Sketch Engine*. Current corpus methods were used not only to automate the collection of statistical data, but also for their evaluation. Also, comprehensive analytical work on the interpretation, structuring, and modeling of this discourse was performed. The use of corpus techniques made it possible to perform a meta-linguistic description of the ecological discourse using modern software tools. The potential of corpus research of English-language environmental discourse involved in the study can be productively implemented for further linguistic work.

Keywords: environmental discourse, corpus analysis, linguistic potential, *EuroNews Green Corpus*, *Sketch Engine*, statistics, methods.

Введение. Экологический дискурс является репрезентативным лингвистическим объектом в силу его ярко выраженной понятийной и концептуальной идентичности и востребованности в современной речевой практике [1–3]. Эта востребованность подтверждается развитием соответствующего контента во многих средствах массовой информации. Подобная практика является широко распространенной и подтверждается наличием множества веб-ресурсов, предоставляющих платформу для распространения новостей и информации об окружающей среде, например, *Sciencedaily.com*, *Phys.org*, *Scientificamerican.com*, *BBC.com*, а также многими другими зарубежными и отечественными порталами, в перечне разделов которых можно встретить заголовки *Ecology* или *Green*. Репрезентативна соответствующая проблематика и в таком известном средстве массовой информации, как *EuroNews* [4]. Данное СМИ демонстрирует максимальный учет предпочтений адресатов своей продукции, что, в частности, подтверждается мультимодальным характером организации референтного дискурса [5]. Многогранность дискурса как научной универсалии, в свою очередь, позволяет объективно оценивать речевую деятельность в контексте разнообразных коммуникационных обстоятельств [6]. В данной связи продуктивны лингвистические исследования, посвященные комплексной характеристике референтной речевой практики.

Основная часть. *Объектом* данной работы является лингвистический потенциал корпусного исследования англоязычного экологического дискурса. *Цель* – выявить и охарактеризовать сущность и специфику лингвистического описания англоязычного экологического дискурса посредством корпусного инструментария. *Материалом* исследования выступили англоязычные тексты экологического дискурса. Исследование выполнено на базе прикладной методологии с привлечением корпусной методики и функционального дискурс-анализа. Использование *корпусной методики* как апробированного и эффективного инструментария для решения широкого круга лингвистических задач при изучении англоязычного экологического

дискурса полностью целесообразно. Данный подход позволяет рассчитывать на лингвистическую *репрезентативность*, результативность изучения языковой практики и *актуальность* релевантных выводов как для отдельных текстов экологической тематики, так и для речевой деятельности, касающейся экологической проблематики в целом.

Репрезентативность нашего исследования обеспечена материалом новостных текстов, размещенных на странице *EuroNews.green* интернет-портала *EuroNews.com (https://www.euronews.com/green)*. Данный портал является официальным ресурсом телекомпании *EuroNews* и его наполнение целевым контентом обеспечивается внушительным штатом сотрудников, в том числе профессиональных журналистов. Немаловажным атрибутом речевой практики на портале *EuroNews* является ее интерактивный характер, что обеспечивается наличием опции распространения материалов посредством публикации ссылок на страницах в социальных сетях, а также возможностью подписки на рассылку тематических статей. Непосредственно для подбора эмпирического материала нами был использован контент раздела *News* (англ. 'новости'), включающий в себя новостные сводки экологической проблематики, но не ограниченные строго какой-либо узкопрофильной тематикой (например, темой климата или технологическими «зелеными» изобретениями).

Подбор материала осуществлялся методом сплошной выборки в хронологических рамках 2022 года: первая статья, размещенная в 2022 году, датируется 3 января, последняя включенная в выборку статья была датирована 20 ноября 2022 г. Согласно проведенному анализу динамики контента, на портале публикуется в среднем от одной до четырех статей в день, что позволяет вычислить среднее количество опубликованных за данный период времени статей в день – 1,5. Это свидетельствует как об активной и систематической работе журналистов и авторов новостного портала, так и о высокой востребованности дискурса экологической направленности, что, очевидно, мотивировано обеспокоенностью проблемами окружающей среды в обществе.

В результате проведенных выборки, обработки и форматирования текстов нами был агрегирован текстовый массив, объединяющий в общей сложности 506 текстов (46 статей сопровождаются видеоматериалами), общим объемом 364 303 словоупотребления. Для лингвистической обработки массива текстов целесообразным был признан корпусный формат, обеспечивающий как возможности рассмотрения текстов в виде репрезентативной совокупности, так и возможности компьютерной обработки самих текстов. Корпусный анализ как методика признан специалистами эффективным инструментарием решения прикладных лингвистических задач [7–9]. Действительно, современные корпусы обладают возможностью хранить и исследовать массивы языковых данных, для изучения которых еще несколько десятелетий назад потребовалось бы выполнить огромный объем работы. Показательна в данной связи позиция А. Н. Баранова, который говорит о преимуществе использования методов корпусного анализа ввиду возможности произведения подсчетов и составления статистических данных, подтверждающих или опровергающих гипотезы исследования [10, с. 135–137].

Для компьютерного обеспечения лингвистической функциональности массива текстов в настоящее время имеется достаточно широкий арсенал так называемых «корпусных оболочек». Как правило, сегодня подобные сервисы доступны посредством Интернета. К наиболее популярным из них относятся, в частности, *AntConc*, *Nooj*, *Tropes*, *Sketch Engine* и др. Для решения задач нашего исследования был использован находящийся в открытом доступе ресурс *Sketch Engine* [11].

Sketch Engine представляет собой многопрофильный инструмент для работы с языковым материалом: сервис предоставляет возможность загружать и исследовать массивы текстов объемом более миллиона словоупотреблений, позволяет выявлять языковые особенности тех или иных текстов, анализировать их структуру и систематизировать их характеристики. *Sketch Engine* совместим с целым рядом языков и позволяет обрабатывать параллельные корпусы. Ресурс готов работать с русским, украинским, другими славянскими языками, основными европейскими языками и рядом экзотических языков.

В частности, для английского языка ресурс предлагает довольно разнообразный инструментарий обработки: *pos tagging* (англ. 'морфологическая разметка'), *lemmatization* (англ. 'лемматизация'), *word sketches* (англ. 'коллокации'), *terms* (англ. 'термы') [11]. Необходимо отметить, что функциональность практически любого корпусного анализа обеспечивается формализацией грамматических отношений. Пожалуй, ключевой в этом плане является технология *pos tagging* ('частеречная разметка'). Под *pos tagging* в *Sketch Engine* разработчиками подразумевается «… процесс аннотирования каждого токена тегом, несущим информацию о его принадлежности к какой-либо части речи, а также – часто морфологическую и грамматическую информацию, такую как число, род, падеж, время и т. д.» [11].

Итак, посредством инструментария *Sketch Engine* массив тестов был преобразован в полноценный корпус текстов *EuroNews Green Corpus (http://surl.li/fmiip)*. Функциональность разметки была апробирована на тестовом фрагменте, ниже – примерный формат разметки на примере предложения: *From encouraging public transport use to building more parks, there are plenty of ways that cities can fight climate change* (таблица 1).

Для разметки была использована программа *English 3.3 for TreeTagger pip eline v2* – один из наиболее часто используемых инструментов разметки для англоязычных текстов [12, с. 44]. Кроме кодирования самих токенов здесь сразу указывается их частеречная принадлежность и «начальная форма слова». Для морфологической разметки текстов нашего корпуса *EuroNews Green Corpus* сервис *Sketch Engine* использовал 62 тега. В целом, как можно видеть, разметка выполнена посредством совместимого с текстом англоязычного тегсета (набора тегов).

Таблица 1* – Образец разметки фрагмента текста *EuroNews Green Corpus* программой *English 3.3 for TreeTagger pipeline v2

From → **IN** → from-i	more → **JJR** → more-j	ways → **NNS** → way-n
encouraging → **VVG** → encourage-v	parks → **NNS** → park-n	that → **IN / that** → that-i
public → **JJ** → public-j	, → , → ,-x	cities → **NNS** → city-n
transport → **NN** → transport-n	there → **EX** → there-x	can → **MD** → can-x
use → **NN** → use-n	are → **VBP** → be-v	fight → **VV** → fight-v
to → **TO** → to-x	plenty → **NN** → plenty-n	climate → **NN** → climate-n
building → **VVG** → build-v	of → **IN** → of-i	change → **NN** → change-n

Однако корпусный инструментарий позволяет не только аккумулировать доступную наблюдению статистику, но и выполнять аналитическую работу. В частности, корпусная методика является эффективной и полностью совместимой с *функциональным дискурс-анализом* [13]. В данном аспекте корпусный анализ предоставляет ценные данные для интерпретации, структуризации и моделирования релевантного языкового материала. Возможности выявления как традиционных («очевидных») корпусных данных дискурса, так и основанных на них метаданных и моделей могут быть систематизированы в функциональном ключе.

Итак, базовой единицей корпусного анализа являются, как известно, ***токены***. Их в нашем корпусе сервис *Sketch Engine* насчитал 426 467. Токены представляют собой минимальные значимые единицы текстов корпуса. *Sketch Engine* способен определять два типа токенов: «слова» (англ. *words*) и «неслова» (англ. *non-words*), к последним относятся разного рода символы, знаки препинания, цифры и др. Но, так или иначе, токенов в любом тексте больше, чем слов. Разделение текста на токены производится специальной программой, токенизатором, адаптированной под каждый язык, поддерживаемый корпусным менеджером. «Слов» согласно терминологии *Sketch Engine*, или ***«словоупотреблений»*** в значимости прикладной лингвистики – в нашем корпусе 364 303. Однако, это – именно количество словоупотреблений в тексте, токенов-слов, а не уникальных единиц языка. ***Слов*** как уникальных *лексических* единиц языка в нашем корпусе сервис насчитал 26 638.

Существенный лингвистический потенциал изучения *лексических* единиц в текстах корпуса может быть задействован посредством технологии выявления и ранжирования ***ключевых слов***. При этом данные могут быть распределены с учетом частотности той или иной «ключевой» единицы в исследуемом корпусе *(Frequency (focus))*, его частотности в референтном корпусе (корпусе-референсе) *(Frequency (reference))*, относительной частотности ключевого слова в данном корпусе *(Relative frequency (focus))* и его относительной частотности в референтном корпусном массиве *(Relative frequency (reference))*.

Еще одной категорией «слов» в корпусе являются ***леммы*** – начальные формы слов, на базе которых группируются их словоизменительные варианты. Таких единиц в корпусе выявлено 17 079. Такого рода информация напрямую может использоваться при характеристике *морфологических* особенностей единиц корпуса текстов.

Автоматическое определение ***термов*** (типичных словосочетаний) в сервисе *Sketch Engine* возможно на базе специализированных программ *English Terms 3.1* (выбранный нами вариант для обработки корпуса), *English Terms 3.0, English (TreeTagger – PennTB) for term extraction 2.3) (Term Grammar)*. Для корпусного анализа нашего корпуса *(EuroNews Green Corpus)* целесообразным был признан *English Terms 3.1.* Анализ и определение термов как *синтаксических* единиц в исследуемом корпусе производится путем обращения корпусного менеджера к корпусу-референсу, например, загруженным в базу данных *Sketch Engine* корпусам *English Web (enTenTen20), British National Corpus (BNC)* и *Brown Corpus* – и последующего сравнения лексического состава корпуса *EuroNews Green Corpus* с составом корпуса-референса. Работа с термами требует привязки корпуса-референса (доступны в библиотеке сервиса, но могут быть автоматически выбраны, если пользователь не определился с выбором). Данная технология подразумевает структуризацию метаданных в виде таблиц (таблица 2) – аналогично вышеупомянутой технологии выявления и анализа ключевых слов.

Таблица 2* – Образец сравнительного анализа термов корпуса *EuroNews Green Corpus

Item	Frequency (focus)	Frequency (reference)	Relative frequency (focus)	Relative frequency (reference)
fossil fuel	645	290405	1512,42651	6,734
climate change	646	1232153	1514,77136	28,57153
renewable energy	179	403357	419,72766	9,35316
greenhouse gas	175	288761	410,34827	6,69587
climate crisis	171	37513	400,9689	0,86986

Используя данный инструмент, в корпусе был определен ряд термов, или устойчивых в речи синтаксически несвободных словосочетаний. В текстах статей экологической тематики, входящих в состав исследуемого корпуса *EuroNews Green Corpus* таких единиц было идентифицировано 90.

Еще одна лингвистическая функция синтаксической направленности – ***Word Sketch***, подразумевающая поиск *коллокаций* – используемых валентностей слова. Выявление ***n-грамм***, типичных последовательностей знаков (в данном контексте – сложных знаков, слов) – еще одна возможность для лингвистической характеристики массива текстов посредством *Sketch Engine*.

Кроме описанных выше данных лексического, морфологического и синтаксического свойства посредством *Sketch Engine* можно получить и *металингвистические* данные в виде конкордансов, тезаурусов и частотных словарей. Составление ***конкордансов*** – по запросу пользователя показывает все случаи контекстуального употребления той или иной леммы (рисунок 1).

Заключение. Итак, проведение лингвистического анализа больших массивов языковых данных с помощью современной корпусной методики оказывается весьма информативным. Корпусный инструментарий с использованием актуальных программ может быть задействован не только для автоматизации сбора вполне очевидных статистических данных, но и для выполнения достаточно сложной аналитической работы по интерпретации, структуризации и моделированию дискурса. Методом сплошной выборки нами был агрегирован репрезентативный массив текстов по экологической тематике. Этот массив, являя собой достаточно экстенсивный и неупорядоченный эмпирический материал, представлял сложный для изучения объект. Однако использование корпусных методик позволило в рамках компактного исследования выполнить существенный объем лингвистического анализа, в частности провести интерпретацию, структурирование и моделирование дискурса. После фиксации и формализации языковых данных в компьютерно-опосредованном формате посредством сервиса *Sketch Engine* был

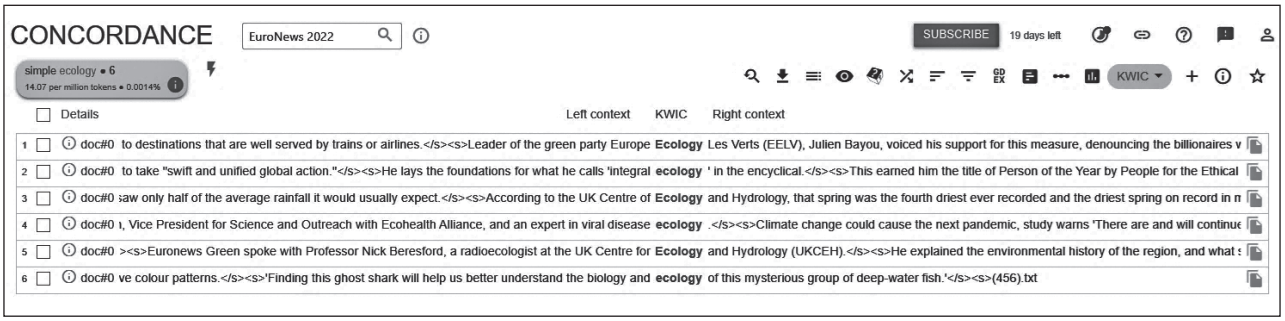

Рисунок 1 – Фрагмент конкорданса контекстов использования лексемы ecology (в корпусе EuroNews Green Corpus)

Составление ***тезаурусов*** – по запросу пользователя подразумевает создание словарей синонимов, антонимов или семантически «близких слов» *(similar words)*, которые могут быть представлены в виде списка-рейтинга или в формате понятийного облака. Эта аналитическая деятельность также реализуема посредством корпусного анализа. Составление ***частотного словаря*** – еще одна металингвистическая практика, возможная с помощью инструментария *Sketch Engine*. В результате данного анализа был составлен каталог всех встречающихся в корпусе лексических единиц (26 638) с указанием их частотных характеристик. Данная технология позволяет отразить статистику использования тех или иных словоупотреблений, в том числе, рассмотреть их относительную частотность – как в аспекте использования в корпусе лемм, так и в аспекте представленности в корпусе частей речи.

выполнен комплекс исследований, позволивший разносторонне охарактеризовать экологический дискурс как объект изучения. Так, в частности, была получена статистическая информация о количестве единиц в корпусе, в том числе, о количестве словоупотреблений и предложений. Такого рода данные были интерпретированы, в частности, рассмотрены как токены, ключевые слова, и термы. В свою очередь, вышеупомянутые данные были структурированы: в том числе, такие категории языковых метаданных, как леммы, коллокации и *n*-граммы были систематизированы и подготовлены для дальнейшего анализа в виде таблиц. С помощью корпусного инструментария была проведено моделирование данных в виде конкордансов, тезаурусов и частотных словарей. Задействованный в исследовании потенциал корпусного исследования англоязычного экологического дискурса может быть продуктивно реализован для дальнейшей лингвистической работы.

Литература

1. *Alexander, R., Stibbe, A.* From the analysis of ecological discourse to the ecological analysis of discourse / R. Alexander, A. Stibbe // Language Sciences, 2014. – p. 104–110.
2. *Knott, J.* A roadmap for exploring the thematic content of ecology journals / J. Knott, E. LaRue, S. Ward, E. McCallen, K. Ordonez, F. Wagner, I. Jo, J. Elliott, and S. Fei // Ecosphere, 2019. 10(8). – P. 1–10.

References

1. *Alexander, R., Stibbe, A.* From the analysis of ecological discourse to the ecological analysis of discourse / R. Alexander, A. Stibbe // Language Sciences, 2014. – p. 104–110.
2. *Knott, J.* A roadmap for exploring the thematic content of ecology journals / J. Knott, E. LaRue, S. Ward, E. McCallen, K. Ordonez, F. Wagner, I. Jo, J. Elliott, and S. Fei // Ecosphere, 2019. 10(8). – P. 1–10.

3. *Porter, J. H.* Evaluating a thesaurus for discovery of ecological data / J. H. Porter. // Ecological Informatics Volume 51, May 2019. – P. 151–156.
4. Euronews. – Режим доступа: URL: https://www.euronews.com. – Дата доступа: 15.12.2022.
5. *Баркович, А. А.* Мультимодальность и медиатекст: специфика «Euronews» / А. А. Баркович, А. И. Еронская // Молодые ученые в инновационном поиске : сб. науч. ст. IX Международной научной конференции, 27–28 мая 2020 г., Минск. – Минск : МГЛУ, 2021. – С. 41–46.
6. *Кибрик, А. А.* Анализ дискурса в когнитивной перспективе: дисс. ... д-ра филол. наук : 10.02.19 / А. А. Кибрик. – М., 2003. – 90 с.
7. *Баркович, А. А.* Методологический аспект изучения компьютерно-опосредованного дискурса / А. А. Баркович // Вестник Нижегородского государственного лингвистического университета им. Н. А. Добролюбова. – 2015. – Вып. 30. – С. 38–48.
8. *Райскина, В. А.* Современные методы корпусной лингвистики при анализе текста (на примере корпуса BFM) / В. А. Райскина, О. А. Дубнякова // Актуальные вопросы современной науки: сборник научных трудов. – Новосибирск : Издательство ЦРНС, 2015. – Вып. 40. – С. 146–155.
9. *Захаров, В. П.* Корпусная лингвистика : учеб. для студентов направления «Лингвистика» / В. П. Захаров, С. Ю. Богданова. – 2-е изд., перераб. и дополн. – СПб. : СПбГУ. Филологический факультет, 2013. – 148 с.
10. *Баранов, А. Н.* Корпусная лингвистика / Баранов А. Н. // Введение в прикладную лингвистику. – М., 2001. – С. 112–137.
11. Sketch Engine. – Режим доступа: URL: https://www.euronews.com. – Дата доступа: 15.12.2022.
12. *Olvera-Lobo, M. D.* Innovative Perspectives on Corporate Communication in the Global World / M. D. Olvera-Lobo, J. Gutiérrez-Artacho, I. Rivera-Trigueros, M. Díaz-Millón. – Hershey, PA: IGI Global, 2021. – 319 p.
13. *Баркович, А. А.* Функциональный дискурс-анализ: модели формализации и структуризации / А. А. Баркович // Труды БГТУ. Сер. 4, Принт- и медиатехнологии. – 2022. – № 2 (261). – С. 29–35.

3. *Porter, J. H.* Evaluating a thesaurus for discovery of ecological data / J. H. Porter. // Ecological Informatics Volume 51, May 2019. – P. 151–156.
4. Euronews. – Rezhim dostupa: URL: https://www.euronews.com. – Data dostupa: 15.12.2022.
5. *Barkovich, A. A.* Mul'timodal'nost' i mediatekst: specifika «Euronews» / A. A. Barkovich, A. I. Eronskaya // Molodye uchenye v innovacionnom poiske : sb. nauch. st. IX Mezhdunarodnoj nauchnoj konferencii, 27–28 maya 2020 g., Minsk. – Minsk : MGLU, 2021. – S. 41–46.
6. *Kibrik, A. A.* Analiz diskursa v kognitivnoj perspektive: diss. ... d-ra filol. nauk : 10.02.19 / A. A. Kibrik. – M., 2003. – 90 s.
7. *Barkovich, A. A.* Metodologicheskij aspekt izucheniya komp'yuterno-oposredovannogo diskursa / A. A. Barkovich // Vestnik Nizhegorodskogo gosudarstvennogo lingvisticheskogo universiteta im. N. A. Dobrolyubova. – 2015. – Vyp. 30. – S. 38–48.
8. *Rajskina, V. A.* Sovremennye metody korpusnoj lingvistiki pri analize teksta (na primere korpusa BFM) / V. A. Rajskina, O. A. Dubnyakova // Aktual'nye voprosy sovremennoj nauki: sbornik nauchnyh trudov. – Novosibirsk : Izdatel'stvo CRNS, 2015. – Vyp. 40. – S. 146–155.
9. *Zaharov, V. P.* Korpusnaya lingvistika : ucheb. dlya studentov napravleniya «Lingvistika» / V. P. Zaharov, S. Yu. Bogdanova. – 2-e izd., pererab. i dopoln. – SPb. : SPbGU. Filologicheskij fakul'tet, 2013. – 148 s.
10. *Baranov, A. N.* Korpusnaya lingvistika / Baranov A. N. // Vvedenie v prikladnuyu lingvistiku. – M., 2001. – S. 112–137.
11. Sketch Engine. – Rezhim dostupa: URL: https://www.euronews.com. – Data dostupa: 15.12.2022.
12. *Olvera-Lobo, M. D.* Innovative Perspectives on Corporate Communication in the Global World / M. D. Olvera-Lobo, J. Gutiérrez-Artacho, I. Rivera-Trigueros, M. Díaz-Millón. – Hershey, PA: IGI Global, 2021. – 319 p.
13. *Barkovich, A. A.* Funkcional'nyj diskurs-analiz: modeli formalizacii i strukturizacii / A. A. Barkovich // Trudy BGTU. Ser. 4, Print- i mediatekhnologii. – 2022. – № 2 (261). – S. 29–35.