оригинальная статья
https://elibrary.ru/vdmfzw

Корпусная интерпретация англоязычного экологического дискурса: лингвистические закономерности и функциональные особенности

Баркович Александр Аркадьевич
Белорусский государственный университет, Беларусь, Минск
eLibrary Author SPIN: 5171-1479
https://orcid.org/0000-0001-8469-8431
albark@tut.by

Аннотация: Исследование посвящено проблематике выявления лингвистической специфики англоязычной речевой практики, ориентированной на экологическую тематику. Экологический дискурс как актуальный в этой связи коммуникационный континуум является вполне самодостаточной и высокоидентичной сферой реализации языка. Цель – провести корпусный анализ англоязычного экологического дискурса, описание выявленных в ходе анализа лингвистических закономерностей и интерпретацию полученных данных в функциональном аспекте. Актуальность работы заключается в рассмотрении не только текстов как статичного лингвистического объекта, но и в системной характеристике функциональных особенностей речевой практики. Англоязычный материал исследования ввиду этого показателен в контексте его аутентичности и типичности для экологической проблематики. Корпусный анализ как базовая методологическая платформа настоящего исследования является общепризнанным и надежным инструментарием изучения *больших данных* языка. Опыт этой работы свидетельствует о результативности корпусно-дискурсивной методики как интердисциплинарной совокупности методов корпусного анализа дискурса. Материалом послужил агрегированный на базе 506 англоязычных текстов экологической проблематики корпус *EuroNews Green Corpus*. Представленный корпус – инновационный тип корпусов, который может быть квалифицирован как инструментально-независимый корпус. На примере указанного корпуса доказала свою эффективность открытая архитектура универсальных корпусных менеджеров, или корпусных оболочек, в частности *Sketch Engine*. При этом использованный корпусный инструментарий показал свою эффективность не только для автоматизации сбора вполне очевидных статистических данных, но и для выполнения достаточно сложной аналитической работы по структуризации, интерпретации и моделированию дискурса. Современный этап развития корпусной лингвистики характеризуется активным поиском качественно новых возможностей для выявления достаточно сложных языковых отношений. Выполненный комплекс исследований позволил разносторонне охарактеризовать экологический дискурс как объект изучения. На базе подробной статистической информации были интерпретированы и оценены категории токенов, ключевых слов и термов. В свою очередь, вышеупомянутые данные были структурированы: такие категории языковой *материи*, как леммы, коллокации и *n*-граммы, были представлены как метаданные. С помощью корпусного инструментария было проведено моделирование метаданных в виде конкордансов, тезаурусов и частотных словарей. Отмечается, что экологическому дискурсу присуща ярко выраженная лексико-семантическая уникальность. В целом в текстах исследованного корпуса идентифицирована специфическая *экологическая* терминология. Однако неожиданным результатом многоаспектного корпусного анализа оказалась, например, крайне низкая частотность ключевой в концептуальном плане лексемы *ecology*. Таким образом, проведенное корпусное исследование англоязычного экологического дискурса позволило создать существенный потенциал для дальнейшей лингвистической работы.

Ключевые слова: корпусное исследование, экологический дискурс, инструментально-независимый корпус, лингвистические закономерности, функциональные особенности, статистика, метаданные

Цитирование: Баркович А. А. Корпусная интерпретация англоязычного экологического дискурса: лингвистические закономерности и функциональные особенности. *Виртуальная коммуникация и социальные сети.* 2024. Т. 3. № 2. С. 111–125. https://doi.org/10.21603/2782-4799-2024-3-2-111-125

Поступила в редакцию 15.04.2024. Принята после рецензирования 04.06.2024. Принята в печать 10.06.2024.

full article

Corpus Interpretation of English Ecological Discourse: Linguistic Regularities and Functional Features

Aleksandr A. Barkovich
Belarusian State University, Belarus, Minsk
eLibrary Author SPIN: 5171-1479
https://orcid.org/0000-0001-8469-8431
albark@tut.by

Abstract: English-language environmental speech has its own linguistic features. Ecological discourse is a relevant, self-sufficient, and highly identical communication continuum. This article introduces a corpus analysis of the English-language environmental discourse with its linguistic patterns and functional interpretation. The texts were considered not as static linguistic objects, but as a system of functional features in speech. The English-language research material was authentic and featured various environmental issues. Corpus analysis is a popular and reliable tool for big linguistic data while the corpus-discursive methodology proved to be an effective interdisciplinary tool. It was applied to the *EuroNews Green Corpus*, aggregated from 506 English-language texts on environmental issues. This innovative type of corpus can be qualified as instrumentally independent. The research proved the effectiveness of universal case managers or case shells of open architecture, e.g., *Sketch Engine*. Not only did it manage to collect the data, but also performed quite a complex structure analysis, interpreting, and discourse modeling. Corpus linguistics is looking for qualitatively new opportunities to identify complex linguistic relations. The case study provided a comprehensive description of environmental discourse as an object of linguistic research. Based on statistic details, the authors analyzed the categories of tokens, keywords, and terms to structure them as lemmas, collocations, and n-grams, which were presented as metadata. The metadata modeling provided concordances, thesauruses, and frequency glossaries. Environmental discourse had a pronounced lexical and semantic uniqueness resulting from specific ecological terminology. However, the multidimensional corpus analysis revealed an extremely low frequency of the key concept of *ecology*. Thus, this corpus study of the English-language ecological discourse demonstrated a good potential for further linguistic research.

Keywords: corpus research, ecological discourse, instrument-independent corpus, linguistic regularities, functional features, statistics, metadata

Citation: Barkovich A. A. Corpus Interpretation of English Ecological Discourse: Linguistic Regularities and Functional Features. *Virtual Communication and Social Networks*, 2024, 3(2): 111–125. (In Russ.) https://doi.org/10.21603/2782-4799-2024-3-2-111-125

Received 15 Apr 2024. Accepted after review 4 Jun 2024. Accepted for publication 10 Jun 2024.

Введение

В целом *дискурс* как научная универсалия позволяет объективно оценивать речевую деятельность в контексте разнообразных коммуникационных обстоятельств [Баркович 2015b; Рассказы о сновидениях 2009; Dijk van 2012]. Многие из них традиционно могли бы быть отнесены к экстралингвистическим, но сегодня лингвистика стала чрезвычайно пластичной, демонстрируя практически полный охват коммуникации, в том числе ее технических и информационных аспектов. Востребованным предметом такой дискурсивно-совместимой лингвистики является *корпусная интерпретация* речи, обеспечивающая исследователей компьютерным инструментарием и статистически репрезентативной информацией о многоаспектном функционировании языка. Конечно, доступ к статистике не исключает насущной необходимости сочетать компьютерные средства с ресурсами человеческой ментальности: «Проще говоря, интуиция лингвиста может обеспечить основу для обоснованных догадок, позволяя ему быстро и эффективно взаимодействовать с корпусом»[1] [McEnery, Hardie 2012: 161].

[1] Здесь и далее по тексту перевод выполнен автором статьи.

И не меньше интуиции востребованы лингвистические знания и опыт. Как показывает практика, современные корпусные исследования требуют вполне определенной компетенции и должны соотноситься с уже верифицированными данными и общелингвистическими реалиями. И такой баланс достижим при целесообразном использовании интердисциплинарного исследовательского инструментария.

Экологический дискурс – коммуникационно активная и насыщенная речевая деятельность, ставшая своеобразной визитной карточкой культуры XXI в. Экология – не просто модная тема для обсуждения, а *доминанта*, формирующая менталитет и определяющая образ жизни. Эта динамика подтверждается экспансией напрямую ассоциирующегося с экологической значимостью концепта *зеленый* (*green* – англ. зеленый), который стал трендом во многих сферах человеческой жизни – от экономики до политики. Естественно, что подобная социокультурная востребованность экологической проблематики предопределила внимание к ней лингвистов [Филиппова 2018; Porter 2019; Song, Tang 2020]. Актуальность изучения экологического дискурса как объекта сегодня тесно связана с возрастающей активностью и влиятельностью природоохранной проблематики в современном коммуникационном пространстве.

Методы и материалы

Динамичная актуализация *зеленой* повестки стала атрибутом массового сознания и, соответственно, оказалась в фокусе средств массовой информации еще в XX в. В том числе это явным образом способствовало популяризации экологической тематики в профессиональной журналистике. Сегодня распространением новостей и научно-популярной информации об окружающей среде занимаются практически все средства массовой информации. Показательно, что у целого ряда популярных интернет-ресурсов, таких как *Sciencedaily.com*, *Phys.org*, *Scientificamerican.com*, *BBC.com*, как и у многих других, в рубрикации присутствуют категории *Ecology* или *Green*. Референтная языковая практика характеризуется ярко выраженной идентичностью, формирование которой продолжается, и, что немаловажно, эта практика доступна для рефлексии и обсуждения практически в режиме реального времени.

Не осталась в стороне от этой тенденции и телекомпания *EuroNews*, где *зеленая* тематика – в ряду достаточно компактного и знакового набора основных рубрик: *My Europe*, *Мир*, *Бизнес*, *Спорт*, *Green*, *Next*, *Путешествия* и *Культура*. Глобальная интерязыковая значимость этой проблематики подтверждается тем, что в русскоязычной версии интернет-портала компании рубрика *Green* не переведена и даже не транслитерирована (наряду с *Next* (*будущее*) и *My Europe* (дословно *моя Европа*). В целом *EuroNews* гибко учитывает предпочтения адресатов своей продукции, что, в частности, подтверждается мультимодальным характером организации референтного дискурса. Немаловажным аспектом отражения речевой практики на портале *EuroNews* является ее интерактивный характер, что обеспечивается опцией распространения контента посредством гиперссылок и предложением подписки на рассылку. Подобная ориентация на тенденции развития информационных технологий во многом обусловлена разноплановым и динамичным характером самого объектного дискурса.

Опора на репрезентативный материал в лингвистическом исследовании – один из ключевых критериев объективности его результатов. Обращение к текстам крупного интернет-портала в данной связи вполне закономерно и позволяет практически обеспечить квантитативную и квалитативную безупречность эмпирического материала. Новостной контент в этом случае обладает явными преимуществами перед материалом форумов и блогов, поскольку создается профессиональными журналистами и представляет собой литературную форму языка без ярко выраженных и массовых аномалий. Любые отклонения от нормы в речи оказываются информативными в лингвистическом аспекте, однако не должны заслонять языковой мейнстрим при рассмотрении особенностей языковой практики как таковой.

Репрезентативность нашего исследования обеспечена материалом новостных текстов, размещенных в разделе *EuroNews.green* интернет-портала *EuroNews.com*. Исходя из проведенного анализа динамики контента, на портале публикуется в среднем от одной до четырех статей в день, что свидетельствует об активной работе журналистов и о высокой востребованности дискурса экологической направленности. Здесь новостные сводки по экологической проблематике уже не дифференцированы согласно еще более узкопрофильной тематике, что усложнило бы их тематическую структуризацию. Подбор материала из раздела *EuroNews.green* осуществлялся методом сплошной выборки в хронологических рамках 2022 г.:

первая статья, попавшая в выборку (и первая по факту в 2022 г.), датируется 3 января, последняя статья, включенная в соответствующую выборку, была опубликована 20 ноября. В результате был агрегирован текстовый массив, объединивший в общей сложности 506 текстов разного объема – средний размер каждого составил 720 словоупотреблений. Представленный массив стал основой корпуса *EuroNews Green Corpus*[2]. Предобработка и структуризация собранных текстов уже позволяет квалифицировать их совокупность как корпус.

Итак, объектом настоящего исследования является экологический дискурс. Предметом рассмотрения выступает специфика корпусного исследования англоязычного экологического дискурса в лингвистическом и функциональном аспектах. Исследование направлено на корпусный анализ англоязычного экологического дискурса, описание выявленных в ходе анализа лингвистических закономерностей и интерпретацию полученных данных в функциональном аспекте, что в совокупности и является целью работы. Таким образом, ключевыми компонентами объединенного этой единой целью гносеологического процесса можно назвать анализ, описание и интерпретацию объекта исследования. Конечно, изучение текстов не равноценно изучению дискурса, но именно в текстах воплощается содержательная и экстралингвистическая специфика речевой деятельности. Анализ дискурса невозможен без анализа уже реализованной дискурсивной материи – текста. И лишь на основе обобщения данных и метаданных речевой практики возможен синтез знаний о дискурсе как таковом.

Методология

В целом корпусный формат обеспечивает возможности лингвистического рассмотрения как отдельных языковых единиц или текстов, так и их совокупности, что исключительно важно при исследовании дискурса [Corpora and Discourse… 2004; 2008; Corpora and Discourse Studies… 2015]. Дополнительные возможности, появляющиеся при компьютерной обработке языковых данных – весомый аргумент при верификации результатов. Корпусная методология признана специалистами эффективным инструментарием решения широкого круга лингвистических задач [Плунгян 2008; Discourse patterns… 2004; Flowerdew 2012]. Это обусловлено возможностью благодаря корпусам собирать, пополнять и изучать гигантские объемы языковых данных, для изучения которых еще не так давно требовались огромные материальные ресурсы и временные затраты. Показательна в этой связи позиция А. Н. Баранова, который утверждает о преимуществе использования корпусных методов при обработке статистических данных, зачастую оказывающихся решающим аргументом при оценке той или иной гипотезы [Баранов 2021: 135].

В контексте высокой целесообразности использования корпусов важным аспектом оказывается их оснащенность. Сегодня от корпусов ждут не только классических возможностей собирать очевидную статистику данных, но и возможностей осуществлять их сложную и комплексную обработку [Захаров 2016: 153]. В потоке современных *больших данных*, в том числе языковых, особо ценятся структурированные данные, метаданные и знания. В связи с этим продуктивны лингвистические исследования, посвященные комплексной характеристике референтной речевой практики, анализу ее концептуальных основ, моделированию ее метаструктуры и иным существенным характеристикам объектной проблемной области. Такой подход исключает поверхностность и фрагментарность исследований, актуализируя вопросы совместимости и преемственности разнообразной метаязыковой практики.

Для компьютерной поддержки лингвистических исследований сегодня создаются как узкоспециализированные корпус-менеджеры, так и универсальные программы, которые можно использовать для анализа текстовых ресурсов разного происхождения и состава. В настоящее время подобные программы, так называемые *корпусные оболочки*, как правило, доступны посредством Интернета. Среди апробированных и пользующихся известностью инструментов такого рода можно отметить *AntConc*, *Nooj*, *Tropes*, *WordSmith* и др.

EuroNews Green Corpus подтвердил совместимость с целым рядом корпусных менеджеров и – в соответствии с современной тенденцией к универсализации корпусной практики – обрабатывался посредством корпусных менеджеров ряда *открытых* корпусных ресурсов. Использование подобного рода ресурсов обладает таким немаловажным преимуществом, как полная доступность и лингвистическая

[2] Баркович А. А., Рипинская О. А. Иллюстративный корпус текстов экологического дискурса EuroNews Green Corpus. URL: http://surl.li/fmiip (дата обращения: 30.03.2024).

 https://doi.org/10.21603/2782-4799-2024-3-2-111-125

прозрачность самих текстов, что практически исключено в классическом корпусном формате. *Типологическая* особенность корпусов подобного типа – их открытая архитектура, совместимая с неограниченным количеством и качеством корпусных менеджеров или корпусных оболочек. Корпусы данного типа – инновационный тип корпусов. Он идентифицируется как *инструментально-независимый корпус* – корпус, который может обрабатываться корпусными менеджерами различной локализации.

В частности, *Sketch Engine*, который в приоритетном порядке задействовался для этого исследования, представляет собой многопрофильный инструмент для компиляции и обработки языковых данных[3]. Его функциональность позволяет загружать *свои* тексты размером более миллиона словоупотреблений, анализировать их количественный и качественный состав, структуру и выполнять целый ряд сложноаналитических манипуляций. Немаловажно, что *Sketch Engine* совместим со многими языками. Указанный ресурс предоставляет возможность для обработки материала на русском, украинском, других славянских языках, основных европейских языках и ряде экзотических языков. При этом англоязычная аутентичность его интерфейса позволяет уверенно ориентироваться в терминологии и функционале данного корпусного менеджера, а также корректно анализировать англоязычный материал, для которого имеется максимальный набор опций.

Таким образом, в прикладном аспекте для выполнения этого исследования активно использовался широкий арсенал корпусных методов. Однако необходимо отметить и существенное задействование метаязыкового инструментария функционального дискурс-анализа, без которого полученные данные о речевой продукции остались бы без внимания и не были бы должным образом интерпретированы. Функциональный дискурс-анализ характеризуется использованием разнообразного лингвистического инструментария, методологической квинтэссенцией которого является широкий охват речевого континуума и рассмотрение сферы речевой деятельности через призму тенденций ее использования [Баркович 2022]. Опыт данного исследования и тенденции мировой практики свидетельствуют о явной кристаллизации *корпусно-дискурсивной методики* как интердисциплинарной совокупности методов корпусного анализа дискурса. В прикладном лингвистическом исследовании всегда востребованы общенаучные методы анализа, описания и сравнения. В представленном исследовании были задействованы и специализированные методы, такие как контент-анализ, метод непосредственных составляющих, статистический метод и ряд других. Это исследование выполнялось в прикладном интердисциплинарном ключе. Использование комплекса вышеупомянутых методов для решения поставленных задач оказалось полностью целесообразным. Данный подход обеспечил 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 – экологического дискурса.

Результаты

Языковая специфика исследования

Для англоязычного материала *Sketch Engine* предлагает следующие уже традиционные возможности компьютерной обработки: *pos tagging* (англ. *частеречная разметка*), *lemmatization* (англ. *лемматизация*), *word sketches* (англ. *коллокации*), *terms* (англ. *термы*). Под *частеречной разметкой* (*pos tagging*), например, в *Sketch Engine* 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.»[4]. *Аннотирование* (аннотация), тегирование, разметка – это все синонимы, описывающие как процесс, так и результат приписывания единицам корпуса специальных обозначений для их обработки компьютерной программой.

При этом автоматическая *грамматическая* разметка для корпусов текстов обычно дифференцируется на морфологическую и синтаксическую. И это принципиально важно, например, для синтетических языков. Так, в корпусах русского языка такая дифференциация реализуется последовательно (см. например, НКРЯ[5]). Для преимущественно аналитического английского языка морфология в корпусах ассоциируется в первую очередь с частеречной принадлежностью лексических единиц,

[3] Sketch Engine. URL: https://www.sketchengine.eu (accessed 30 Mar 2024).
[4] Ibid.
[5] Национальный корпус русского языка. URL: https://ruscorpora.ru (дата обращения: 03.04.2024).

а грамматика – с более широким набором категорий, описывающих опять же лексические единицы. Эта реальность не полностью совпадает с рамками категориального описания в теоретической лингвистике, где под грамматикой традиционно понимается комплекс морфологических и синтаксических категорий.

Так или иначе инструмент разметки морфологических категорий в *Sketch Engine* работоспособен и полностью совместим с англоязычными текстами. Формализация морфологических отношений является приоритетом практически любого корпусного анализа. Этому есть простое объяснение: морфология выступает наиболее категориально обеспеченной областью языкознания, и отразить соответствующую систему координат в *компьютерном* виде не является невыполнимой задачей. Противоречие здесь есть, но в другом: многие категории создавались в свое время примерно и сейчас весьма приблизительно отражают реальные языковые отношения. Так, в корпусе НКРЯ по необходимости уже 11 падежей, а должно быть и больше. По мнению А. А. Зализняка – 14: «Формально все имена признаются имеющими единую 14-падежную систему» [Зализняк 2002: 54]. А. А. Зализняк выделял четыре родительных падежа, по два дательных, винительных, творительных, предложных, наряду с именительным падежом и счетной формой [Зализняк 2002: 53]. С падежами английского языка все намного проще, но все достаточно сложно, особенно в прикладном аспекте, с классификацией уже самих частей речи. Так, для морфологической (частеречной) разметки текстов анализируемого корпуса соответствующая программа *Sketch Engine* использовала 62 тега. Корректность выполненной работы можно оценить на примере автоматической разметки тестового иллюстративного фрагмента:

"Athens, Barcelona, Paris: The European cities leading on climate adaptation" By Charlotte Elton Updated: 17/11/2022

Many of Europe's cities are leading the way on environmental action, a new report has revealed.

From encouraging public transport use to building more parks, there are plenty of ways that cities can fight climate change…

Для разметки была использована программа *English 3.3 for TreeTagger pipeline v2* – один из наиболее часто используемых инструментов аннотации англоязычных текстов [Olvera-Lobo et al. 2021: 44].

Табл. 1. Образец разметки фрагмента текста *EuroNews Green Corpus* программой *English 3.3 for TreeTagger pipeline v2*
Tab. 1. Sample mark-up of a *EuroNews Green Corpus* text fragment by *English 3.3 for TreeTagger pipeline v2*

"	''	"-x	17/11/2022 **CD** number]-m			encouraging **VVG** encourage-v		
Athens **NP** Athens-n			Many **JJ**	many-j		public **JJ** public-j		
,	,	,-x	of	**IN**	of-i	transport **NN**		transport-n
Barcelona **NP** Barcelona-n			Europe's	**NPZ**	Europe-n	use	**NN**	use-n
,	,	,-x	cities	**NNS**	city-n	to	**TO**	to-x
Paris **NP** Paris-n			are	**VBP**	be-v	building	**VVG**	build-v
:	:	:-x	leading	**VVG**	lead-v	more	**JJR**	more-j
The	**DT**	the-x	the	**DT**	the-x	parks	**NNS**	park-n
European **JJ** European-j			way	**NN**	way-n	,	,	,-x
cities **NNS** city-n			on	**IN**	on-i	there	**EX**	there-x
leading **VVG**		lead-v	environmental **JJ** environmental-j			are	**VBP**	be-v
on	**RP**	on-x	action **NN**		action-n	plenty **NN**		plenty-n
climate **NN** climate-n			,	,	,-x	of	**IN**	of-i
adaptation **NN** adaptation-n			a	**DT**	a-x	a	**DT**	a-x
"	''	"-x	new	**JJ**	new-j	that	**IN**/that	that-i
By	**IN**	by-i	report **NN**		report-n	cities	**NNS**	city-n
Charlotte **NP** Charlotte-n			has	**VHZ**	have-v	can	**MD**	can-x
Elton	**NP**	Elton-n	revealed	**VVN**	reveal-v	fight	**VV**	fight-v
Updated	**VVD** update-v		.	**SENT** .-x		climate **NN**		climate-n
:	:	:-x	From	**IN**	from-i	change **NN**		change-n

 https://doi.org/10.21603/2782-4799-2024-3-2-111-125

И обработанный данной программой *размеченный* текст может быть представлен в виде метаданных (табл. 1).

В целом, как можно видеть, разметка выполнена посредством совместимого с текстом англоязычного тегсета (набора тегов). Всего в *Sketch Engine* для английского языка доступно 11 вариантов наборов тегов. Кроме собственно морфологического тегирования программа на основе встроенной базы данных выдала отсылки к соответствующим леммам. Так, начальной формой для *Europe's* указана лексема *Europe*, для *cities* – *city*, для *are* – *be* и т. д. Доступность данных такого рода подтверждает эффективность выполнения в *Sketch Engine лемматизации* – ассоциации словоупотреблений корпуса с их так называемой *начальной формой*.

Анализ языкового материала посредством корпусных методов позволяет масштабировать круг исследуемого материала. На примере двух последних предложений (без заголовка и метатекстового сопровождения) вышеупомянутого тестового иллюстративного фрагмента – даже без обращения к общей статистике корпуса – становится очевидным колоссальный перевес в данном дискурсе существительных. Их (13) оказывается более чем в два с половиной раза больше, чем прилагательных (5) и почти в два раза больше, чем глаголов (7). Впрочем, в общих чертах эти пропорции подтверждаются на всем массиве данных: по сведениям *Sketch Engine* общее количество существительных-словоупотреблений в корпусе *EuroNews Green Corpus* – 130787, прилагательных – 32986 (25,22 % в сравнении с существительными) и глаголов – 61271 (46,85 % в сравнении с существительными).

Необходимо отметить, что данные такого рода зависят от характера текстов, программы разметки и других факторов. Например, при оценке учебных текстов в другом исследовании количество существительных почти не отличалось от количества глаголов – 100 / 93 (93 % в сравнении с существительными): «В данном исследовании количество глаголов и существительных оценивалось при помощи онлайн ресурса *Textln-spector*» [Маник, Смирнова 2022: 58]. С одной стороны, обнаружить программу *Textln-spector* или какие-либо упоминания о ней в Интернете не удалось. С другой стороны, известно, что *нарративность* (англ. *narrative / non- narrative feature*) экологического дискурса действительно значительно уступает нарративности исторического дискурса и весьма существенно уступает нарративности академического дискурса [Biber et al. 1998: 162]. Косвенным же показателем нарративности текста как раз и является соотношение существительных и глаголов в тексте. Таким образом, подобная информация в масштабе дискурса вполне лингвистически актуальна. Это также подтверждается нацеленностью на решение нарративных задач ресурсом *Tropes*.

Вместе с тем получение корпусной статистики зачастую обременено и некоторой противоречивостью самих референтных данных. Так, две единицы из отмеченных как существительные *Europe's* и *climate* в тестовом тексте (табл. 1) выполняют явно определительные функции и с позиций функционального синтаксиса вполне могли бы быть отнесены к атрибутивам: *Europe's cities* (*европейские города*) и *climate change* (климатическое изменение). Словоупотребление *climate* должно быть квалифицировано как прилагательное со значимостью *климатический* при любом *ручном* лингвистическом анализе. Но, как можно заметить, проблема компьютерной обработки текста состоит в невозможности корректно идентифицировать подобные случаи. Речь идет не только о банальной омонимии, но и об интерференции категорий. Еще более очевидная противоречивость наблюдается в случае приписывания словоупотреблению *to* тега *TO to in all its uses* (*sorry*) (рус. – *во всех случаях использования (к сожалению)*)[6]. Здесь даже не просматривается стремление разработчиков тегсета разграничить варианты частеречной принадлежности *to*: а это может быть и предлог (как в примере – *use to building*), и частица. Преимущественно аналитический характер английского языка не позволяет программе точно указать принадлежность подобных словоупотреблений. Подобный *шум* в разметке устраним, как правило, лишь при учете семантической структуры всего высказывания. Однако так далеко компьютерные программы пока не видят. Хотя на примере чат-бота *GPT* мы видим в последнее время определенную динамику в понимании подобной проблематики программистами, что не могло не отразиться на эффективности компьютерных программ.

В целом посредством сервиса *Sketch Engine* о корпусе текстов может быть получена лингвистически актуальная информация разного рода. Начавшись с идеи A. Kilgarriff и D. Tugwell о необходимости

[6] Sketch Engine…

учитывать контекст употребления слов и введения термина *word sketch* (*коллокация, практика сочетаемости слова*) [Kilgarriff, Tugwell 2001], сегодня проект *Sketch Engine* значительно продвинулся. При этом корпусный инструментарий позволяет не только аккумулировать доступную непосредственному наблюдению статистику, но и выполнять аналитическую работу. В таком аспекте корпусный анализ предоставляет ценные данные для *структуризации*, *интерпретации* и *моделирования* релевантного языкового материала. Актуальный в данном контексте инструментарий весьма перспективен.

Лингвистическая параметризация корпусного анализа

С помощью корпусных методов в *Sketch Engine* определяется достаточно широкий круг разнообразных языковых данных и метаданных – от фиксации количества и номенклатуры широкого круга элементарных языковых единиц так называемых *орфографических слов*, до квалификации такой сложной и востребованной в семантических исследованиях категории единиц, как *ключевые слова*. К тому же базовой единицей корпусного анализа являются *токены* – 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 далеко не все последовательности символов между пробелами в тексте считаются отдельными токенами. В том числе, как правило, не относятся к токенам сами пробелы (англ. *spase*), хотя, конечно, они выполняют важную роль в разметке.

Подобным образом в базовой разметке зачастую неурегулированным статусом обладает так называемый *регистр*. Учет *регистра* принципиально важен для любого корпусного менеджера, поисковые системы которого работают с учетом регистра. *Sketch Engine* в данном контексте вполне последователен: здесь корпусный менеджер чувствителен к регистру, и, в частности, при разметке текста замена всех заглавных букв на строчные производится автоматически. Согласно *Sketch Engine* в корпусе идентифицировано 21830 таких случаев. Эти данные могут быть использованы, например, для анализа собственных наименований, начинающихся как раз с заглавных букв. Однако на примере указанного выше тестового фрагмента мы видим в нем 12 заглавных букв на 4 предложения и 50 словоупотреблений, что делает необходимой ручную перепроверку представленной *Sketch Engine* противоречивой статистики о замене регистра. Совсем не упрощает ситуацию и то, что заглавные буквы – традиционный атрибут заголовков. По данным корпусного менеджера в корпусе фиксируется 15849 предложений. Что оставляет количество прописных букв для собственных имен (не в начале предложения) и заголовков лишь в сумме 5981, что, конечно, потребует ручного пересчета суммы заглавных букв при необходимости.

С другой стороны, если общее количество словоупотреблений в корпусе – 364303 (согласно корпус-менеджеру), то сведения о количестве *предложений* (15849) нуждаются в не менее тщательной верификации. Простая арифметика показывает, что среднее число слов в предложении в таком случае составит приблизительно 23 (22,98), что аномально много даже для аналитического английского языка. В преимущественно синтетическом русском языке – для сравнения – средняя длина предложения по данным С. Шарова – 10,38 слова[7]. Вместе с тем в англоязычном тестовом фрагменте, указанном выше, на 2 предложения – 36 словоупотреблений: в среднем 18 слов на предложение, что существенно меньше, чем 23. При этом в заголовке слов оказалось 10, а в метатекстовом описании – 4. По данным исследования англоязычных учебных текстов средняя длина предложения составила 14,67 слов [Маник, Смирнова 2022: 58]. И по обобщенным данным среднестатистическая длина английского предложения как такового составляет 15–20 слов [Cutts 2020: 16]. При ручной перепроверке трех первых текстов корпуса *EuroNews Green Corpus* (112 предложений) медиана составила 17,62 слова на предложение (без заголовков и метатекстовых данных). Эти сведения подтверждают тезис о небезупречности сложных алгоритмов обработки текста и актуальности развития категорий разметки, например, при обработке слов, содержащих заглавные буквы. Перечень сложноаналитического функционала от *Sketch Engine* на сегодня следующий:

1. *Word Sketch* – коллокации и комбинации слов.
2. *Sketch difference* – сравнение сочетаемости слов.
3. *Thesaurus* – выявление синонимов, антонимов и семантически близких слов (*similar words*).
4. *Concordance* – примеры использования лексических единиц в контексте.

[7] Lingvisto. URL: http://lingvisto.org/artikoloj/ru_stat.html (дата обращения: 30.03.2024).

 https://doi.org/10.21603/2782-4799-2024-3-2-111-125

5. *Parallel Concordance* – параллельные контексты.
6. *Wordlist* – частотные перечни и лингвистические базы данных.
7. *N-grams* – типичные последовательности слов.
8. *Keywords and Terms* – выявление ключевых слов и типичных словосочетаний.
9. *Bilingual Term Extraction* – извлечение двуязычной терминологии.
10. *Trends* – выявление неологизмов и диахронический анализ словоупотребления.
11. *OneClick Dictionary* – универсальный словарь.
12. *Text Type Analysis* – анализ типа текста[8].

Учитывая традиции метаописания и специфику нашего исследования, доступная посредством вышеперечисленного инструментария информация может быть классифицирована на *лексически*, *грамматически* (*морфологически* и *синтаксически*) и *металингвистически* релевантную. Особую ценность такого рода данные представляют в связи с тем, что в процессе лингвистического анализа оказывается возможной не только ее квантитативная (статистическая) фиксация, но и квалификативная (качественная) *интерпретация*. Рамки настоящего исследования и жанр статьи не предполагают детального рассмотрения всего заявленного в *Sketch Engine* инструментария, однако основные параметры соответствующего корпусного анализа и релевантные дискурсивной проблематике аспекты вполне могут быть систематизированы с учетом традиций уровневого рассмотрения языковых данных.

Лексический аспект корпусного анализа

Разделение текста на базовые лексические единицы в корпусах выполняется специальной программой (токенизатором), адаптированной под каждый язык, поддерживаемый корпусным менеджером. Слов (в терминологии *Sketch Engine*) или *словоупотреблений* (в значимости традиционной прикладной лингвистики) в анализируемом корпусе – 364303. Однако в этом случае речь идет количестве использованных в тексте словоупотреблений, а не уникальных лексических единиц языка. Естественно, *слов* как уникальных единиц языка в корпусе идентифицировано лишь 26638. Собственно, *токенов* в любом тексте должно быть больше, чем словоупотреблений, и на порядок больше, чем слов. Всего токенов в корпусе *EuroNews Green Corpus* корпус-менеджер *Sketch Engine* насчитал 426467.

Еще одной категорией лексических ресурсов в корпусе являются *леммы* – начальные формы слов и группировки их словоизменительных вариантов. Для идентификации леммы требуется специальный алгоритм, и это на порядок сложнее идентификации токенов. Далеко не каждый корпусный менеджер решает подобные задачи. Тем не менее подобных сложноаналитических единиц в корпусе выявлено 17079. То, что лемм оказалось почти на 10000 меньше (64 %), чем слов – вполне реалистично: например, с учетом отнесения к разным *словам* однокоренных слов с разной частеречной принадлежностью. Как уже отмечалось выше, заметным лингвистическим потенциалом обладает такая опция корпусного анализа, как выявление *ключевых слов*. Рассмотрим фрагмент рейтинга ключевых слов корпуса *EuroNews Green Corpu*, выявленных корпус-менеджером *Sketch Engine*: 1 – *cop27*, 2 – *canva*, 3 – *euronews*, 4 – *txt*, 5 – *heatwave*, 6 – *cop26*, 7 – *greenwashing*, 8 – *guterres*, 9 – *sharm*, 10 – *afp*, 11 – *dejong*, 12 – *pfas*, 13 – *fossil*, 14 – *reuters*, 15 – *deforestation*, 16 – *maeve*, 17 – *microplastic*, 18 – *chernobyl*, 19 – *ap*, 20 – *el-sheikh*.

Известно, что компьютерные программы пока не справляются с задачами углубленного семантического анализа текста. Эта закономерность нарастает пропорционально уровню сложности *семантических* задач. В подтверждение данному тезису корпус-менеджер *Sketch Engine* решил семантическую задачу на лексическом уровне тоже небезупречно. В выше представленном рейтинге ключевых слов (20 единиц) *Sketch Engine*, даже на первый взгляд, собрал достаточно много *шума*. При этом нужно отдать должное, т. к. определенный набор ключевых слов программа отобразила корректно: это, например, *heatwave* (аномальная жара), *greenwashing* (зеленый пиар), *fossil* (ископаемое), *deforestation* (вырубка леса), *microplastic* (микропластик), *chernobyl*[9] (Чернобыль). Все они, безусловно, являются семантическими маркерами экологического дискурса. В частности, эти единицы – после минимальной постобработки предоставленных сервисом *Sketch Engine* метаданных – вполне могут быть напрямую использованы для анализа концептосферы экологического дискурса.

Что касается остальных 14 позиций рейтинга, то они нуждаются в дополнительном толковании, дополнительном объяснении или рассмотрении с учетом контекста. Так, *cop26* и *cop27*

[8] Sketch Engine…
[9] Здесь и далее сохранена орфография оригинала.

(названия конференций), *pfas* (название опасного вещества) – акронимы, тесно связанные с экологической проблематикой. Однако уже *Euronews*, *reuters*, *ap* и *afp* – просто названия информационных агентств; *guterres* – фамилия политика, *sharm* и *el-sheikh* – части названия населенного пункта *Sharm el-Sheikh*, который *Sketch Engine* не идентифицировал как единое целое. Как ни странно, но частотность токена *sharm* зафиксирована в 68 случаях, а *el-sheikh* – лишь в 35. И это при том, что лексемы *sharm* в английском языке нет. Что касается единиц *canva*; *txt*; *maeve* и *dejong* – они полностью зависимы от метаязыкового контекста. Такие единицы, как *canva* или *txt*, никакой ценности для характеристики экологической составляющей дискурса не имеют. Это некие метатекстовые идентификаторы. Подобным образом *maeve* и *dejong* могут представлять, например, фамилию журналиста или фотографа.

По результатам такой постобработки включение в список ключевых слов экологического дискурса *heatwave*, *greenwashing*, *fossil*, *deforestation*, *microplastic*, *Chernobyl*, *cop26*, *cop27* и *pfas* (9 единиц, 45 %) должно быть признано актуальным. А вот рассмотрение функциональности в экологическом дискурсе единиц *Euronews*, *reuters*, *ap*, *afp*, *guterres*, *sharm*, *el-sheikh*, *canva*, *txt*, *maeve* и *dejong* (11 единиц, 55 %) не является актуальным. Последний набор единиц – очевидный *шум*, непригодный для анализа, например, экологической концептосферы.

В целом интерпретация языкового материала в лексическом аспекте оказалась лингвистически релевантной. Выявленные ошибки и противоречия – следствие не столько несовершенства конкретных корпусных методов и средств обработки материала, сколько несовершенства *компьютерных* представлений о языке.

Грамматический аспект корпусного анализа

Кроме вышеописанного лексического функционала сервиса *Sketch Engine* в нем есть возможности и для более сложной обработки текстов, получения данных грамматического характера. Здесь также возможна идентификация и систематизация данных о *термах*, *коллокациях* (функция *Word Sketch*) и *n- граммах*.

Идентификация *термов* (типичных словосочетаний) корпусным менеджером *Sketch Engine* возможна посредством таких специализированных программ, как *English Terms 3.1*, *English Terms 3.0*, *English* (*TreeTagger – PennTB*) *for term extraction 2.3*) (*Term Grammar*). Для корпусного анализа *EuroNews Green Corpus* наиболее лингвистически актуальной оказалась программа *English Terms 3.1*. При этом алгоритм выявления и анализа термов включает обращение корпусного менеджера к корпусу-референсу (например, загруженным в базу данных *Sketch Engine* корпусам *English Web* (*enTenTen*20), *British National Corpus* (*BNC*) и *Brown Corpus* – и соответствующее сопоставление *EuroNews Green Corpus* с корпусом-референсом. Корпус-референс может быть выбран в библиотеке *Sketch Engine* или назначается автоматически, если пользователь не определился с выбором.

Список термов в текстах статей экологической тематики, входящих в состав исследуемого корпуса *EuroNews Green Corpus*, составил 90 единиц. Алгоритм, указанный выше, позволил получить и существенную для лингвистической оценки дополнительную информацию, в частности о специфике употребления термов этого корпуса в сопоставлении с неким тематически неспециализированным речевым континуумом. Эта информация доступна благодаря показателям частотности терма в анализируемом корпусе (*Frequency* (*focus*)), его частотности в референтном корпусе (корпусе-референсе) (*Frequency* (*reference*)), относительной частотности терма в данном корпусе (*Relative frequency* (*focus*)) и его относительной частотности в референтном корпусном массиве (*Relative frequency* (*reference*)) (табл. 2).

Проведенный анализ позволил получить исчерпывающую информацию о присутствии в текстах экологической направленности достаточно большой группы высокоидентичных для экологического дискурса термов. Такими являются *fossil fuel*, *climate change*, *renewable energy*, *greenhouse gas* и др. И если, например, терм *climate change* является характерным и для тематически нейтрального дискурса – с показателем 28,57153 при индексе 1514,77136 в специализированном *экологическом* корпусе, то *climate crisis* имеет относительную референтную частотность уже лишь 0,86986 – при индексе 400,9689 в специализированном *экологическом* корпусе. И терм *climate change* в данном контексте является исключением, подтверждающим общую высокую идентичность остальных выявленных в корпусе *EuroNews Green Corpus* термов и их относительно редкую частотность в тематически немаркированной речи: *fossil fuel* – 6,734; *renewable energy* – 9,35316; *greenhouse gas* – 6,69587; *climate crisis* – 0,86986 и т.д. Характерно, что другие термы с компонентом *crisis* демонстрируют сопоставимую с *climate crisis* высокую уникальность в экологическом дискурсе: *energy crisis* – 0,44264; *living crisis* – 0,00921;

 https://doi.org/10.21603/2782-4799-2024-3-2-111-125

Табл. 2. Фрагмент выдачи списка термов корпуса *EuroNews Green Corpus* программой *English Terms 3.1*
Tab. 2. Sample term list output from *EuroNews Green Corpus* by *English Terms 3.1*

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

cost of living crisis – 0,00751. Аналогичная картина наблюдается и при анализе речевой распространенности сверхактуального для экологического дискурса компонента *climate*: *climate conference* – 0,17298; *climate action* – 0,83631; *un climate* – 0,29864; *climate activist* – 0,20447; *climate talk* – 0,22214; *climate finance* – 0,19193; *climate justice* – 0,3504; *climate target* – 0,11348; *climate policy* – 0,62228; *climate goal* – 0,14374; *impact of climate* – 1,17808. Таким образом, в экологическом дискурсе лексема *climate* оказывается весьма востребованной как компонент термов.

Еще одна фирменная для *Sketch Engine* функция синтаксической направленности – *Word Sketch*. Такая функциональность подразумевает поиск коллокаций – используемых синтаксических валентностей слова. И здесь мы видим удивительную аномалию – крайне низкую сочетаемостную реализацию концепта *ecology*. Таких коллокаций специализирующийся на поиске коллокаций лингвистический инструмент *Word Sketch* в экологическом корпусе *EuroNews Green Corpus* находит всего 6. Это три атрибутивно-номинативных коллокации с подчинительной связью компонентов – *integral ecology*; *viral disease ecology*; *disease ecology* (модель *modifiers of* "*ecology*"); номинативная коллокация с подчинительной связью компонентов – *Ecology Les Vertes* (модель *nouns modified by* "*ecology*") и две номинативных коллокации с сочинительной связью компонентов – *Ecology and Hydrology*; *ecology and biology* (модель "*ecology*" *and / or*...). Для экологического дискурса это выглядит как минимум неожиданно. Вероятно, единственное научно обоснованное объяснение подобной аномалии кроется в перераспределении достаточно представительной и разнообразной экологической семантики в пользу других концептов, присущих обозначенной дискурсивной тематике.

Можно предположить конкуренцию для *ecology* единицы *environment* – в силу определенной синонимичности базовой семантики *ecology* – *environment*, в том числе терминов *ecological discourse* (экологический дискурс) и *environmental discourse* (дискурс окружающей среды) [Alexander, Stibbe 2014]. И по сведениям словаря *Collins* к ключевым терминалам понятийного фрейма *ecology* действительно можно уверенно отнести *environment* (*среда*) – наряду с такими понятиями, как *habitat* (*местообитание*), *context* (*обстановка*), *surroundings* (*окрестности*), *scene* (*явление*), *situation* (*ситуация*) и *conditions* (*обстоятельства*)[10]. Однако все они не представлены высокочастотными лексемами. Хотя лучший результат из вышеперечисленных единиц принадлежит как раз концепту *environment* – 310 релевантных словоупотреблений в анализируемом корпусе.

Выявление *n*-грамм, типичных последовательностей знаков – еще одно возможное направление синтаксической характеристики дискурса посредством корпус-менеджера *Sketch Engine*. В данном контексте *знаками* выступают сложные знаки или слова. Не будем подробно останавливаться на анализе биграмм (двухэлементных *n*-грамм): в аналитическом английском языке в их составе типичным образом оказывается множество единиц, принадлежащих к служебным частям речи. Это снижает ценность данных о специфике словоупотребления тех или иных концептуально существенных для характеристики дискурса лексем. Такого рода данные мало чем будут отличаться от простых рейтингов частотности. Служебные части речи присутствуют и в абсолютном большинстве триграмм (трехэлементных *n*-грамм), однако триграммы уже позволяют более содержательно оценить и сочетаемостную специфику единиц, представляющих классы самостоятельных частей речи. Суммарно программой найдено

[10] Collins Online Dictionary. URL: https://www.collinsdictionary.com (accessed 30 Mar 2024).

3485 триграмм и выявлено 34007 случаев их использования в тексте. Самой высокой частотностью обладает триграмма *per cent of*, зафиксированная 443 раза. Среди найденных программой *n*-грамм не все являются маркерами *экологичности*, но очевидно экологически-специфические *oil and gas*, *the climate crisis*, *the climate change* и ряд других также демонстрируют высокую востребованность в контекстах: 201, 129 и 106 случаев использования. И это вполне сопоставимо с такими актуальнейшими для всеобъемлющей языковой практики *n*-граммами, как *per cent of* (частотность 443), *one of the* (частотность 157), *the end of* (частотность 99) и рядом других.

Таким образом, формальные на первый взгляд статистические показатели – с учетом надлежащей постобработки данных – предоставляют ценную информацию, будучи обобщенными как метаданные и будучи верифицированными далее в качестве знаний.

Металингвистический аспект корпусного анализа

Кроме описанных выше данных лексического и грамматического характера посредством *Sketch Engine* можно получить и металингвистические обобщения в виде конкордансов, тезаурусов и частотных словарей. Так, выдача конкорданса – построения по запросу пользователя списка всех случаев контекстуального употребления той или иной лексемы – является уже широко распространенной и пользующейся спросом опцией работы развитого корпусного менеджера. Важным достоинством указанной функциональности можно назвать возможность оценить специфику употребления слова в широком металексическом контексте [Баркович 2015а]. Следующий конкорданс был построен по запросу *ecology* в анализируемом корпусе.

На данном примере мы снова имеем подтверждение ограниченной функциональности лексемы *ecology*. Согласно вышеописанным данным по показателю *Word Sketch* (поиск *коллокаций*) – ее сочетаемость оказалась крайне низкой: 6 коллокаций. Аналогична и ее представительность в широком контексте конкорданса – 100 знаков с пробелами справа и слева. При этом подобная перспектива обеспечивает возможность учесть место той или иной лексемы в окружении в среднем около 15 слов до (слева) и после (справа) нее. Это дает возможность расширить коннотацию того или иного слова до максимально широких рамок – фактически учитывать контекст не только одного объектного предложения, но и соседних. Например, в данном конкордансе можно видеть, что описанная выше коллокация *Ecology Les Vertes* должна быть совсем не трех-, а минимум четырехкомпонентной – *Europe Ecology Les Vertes*; а желательно и восьмиэлементной – *the green party Europe Ecology Les Vertes* (*EELV*). Подобным образом *Ecology and Hydrology* в идеале должна рассматриваться как семикомпонентная коллокация – *the UK Centre of Ecology and Hydrology* (*UKCEH*). И в этом случае, по крайней мере, не будет стоять вопрос об обоснованности заглавного регистра в словоупотреблениях *Ecology*, *Les*, *Vertes* и *Ecology*, *Hydrology*. Такого рода информация будет полезной и при идентификации на корпусном материале, например, именованных сущностей.

Функция составления *тезауруса* не менее лингвистически актуальна и обладает значимым потенциалом для лексикографической работы современного формата. Создание словарей синонимов, антонимов и семантически близких слов (англ. *similar words*) посредством *Sketch Engine* может быть воплощено как в формате списка-рейтинга, так и в формате *понятийного облака*. Особый металингвистический интерес представляет настойчиво формируемая рядом ученых категория *близких слов* [Lafayette 2015; Lin 1998]. Идентификация *близких слов* не только является важным шагом в изучении понятий, но и позволяет «конструировать тезаурус, используя размеченный корпус» [Lin 1998: 768]. Не менее инновационна и сама опция создания понятийных облаков – принципиально новый уровень металингвистического описания, позволяющий апеллировать к образности человеческой ментальности.

Характеристика частотности единиц корпуса – традиционный параметр корпусного анализа. В свое время корпусная лингвистика с подобных задач и начиналась. На сегодняшний день процедура создания *частотных словарей* посредством корпусных менеджеров хорошо отработана. При этом *Sketch Engine* демонстрирует в рамках данной функциональности целый спектр дополнительных возможностей. Так, при помощи корпусного менеджера можно получить статистику по количеству словоупотреблений (в том числе отразить относительную частотность) как для лемм, так и для лексем, принадлежащих каким-либо частям речи, а также чисел. Например, в топе рейтинга существительных находятся в основном ожидаемые лексемы *climate* (2188 (5130,52593 – относительная частотность)), *energy* (1645 (3857,27383)), *gas* (1192 (2795,058)), *fuel* (879 (2061,12079)), *emission* (766 (1796,15304)), *oil* (656 (1538,21984)), *water* (531 (1245,11392)),

 https://doi.org/10.21603/2782-4799-2024-3-2-111-125

food (495 (1160,69942)), *carbon* (443 (1038,76736)), *air* (381 (893,38683)), *pollution* (380 (891,04198)), *environment* (310 (726,90267)) и т. д. Лексемы *climate* и *energy* занимают первые два места, лексема *environment* – 48 (по убыванию) в достаточно плотном рейтинге. И данная статистика позволяет уверенно идентифицировать тот или иной текст по его референтности экологической проблематике, степени выраженности экологической семантики и другим содержательным параметрам, вплоть до определения тональности текста. Опция сентимент-анализа, кстати, была бы вполне логичным развитием лингвистической функциональности *Sketch Engine*, но пока не реализована.

Заключение

Экологический дискурс является вполне самодостаточной и высокоидентичной сферой функционирования языка. Англоязычные тексты экологической тематики рассматривались на материале современной публицистики (2022 г.). Отмечается, что указанному сегменту речевой практики присуща ярко выраженная лексико-семантическая уникальность. Аналитический строй английского языка обусловливает и существенную грамматическую (особенно синтаксическую) специфику изучения соответствующих текстов. Металингвистический потенциал описания экологического дискурса значительно расширяет горизонт лингвистических параметров описания первичных языковых данных. Вместе с тем, являясь надстроечным, вторичным воплощением корпусного анализа, металингвистический инструментарий в целом подтверждает данные уровневого анализа дискурса. Агрегированный на базе 506 англоязычных текстов экологической проблематики корпус *EuroNews Green Corpus* в процессе анализа посредством разнообразного инструментария *Sketch Engine* полностью подтвердил свою репрезентативность и функциональность. В частности, в нем идентифицируется специфическая экологическая терминология. К неожиданным результатам следует отнести крайне низкую частотность ключевой в концептуальном плане лексемы *ecology*. На примере представленного корпуса доказала свою эффективность открытая архитектура универсальных корпусных оболочек. Этот корпус принадлежит к инновационному типу корпусов, который может быть квалифицирован как инструментально-независимый корпус.

В целом проведение лингвистического анализа больших массивов языковых данных с помощью современных корпусных методов обладает огромным потенциалом. Опыт настоящего исследования показал продуктивность корпусно-дискурсивной методики как интердисциплинарной совокупности методов корпусного анализа дискурса. Корпусный инструментарий эффективен не только для автоматизации сбора вполне очевидных статистических данных, но и для выполнения достаточно сложной аналитической работы по структуризации, интерпретации и моделированию дискурса. Выполненный комплекс исследований позволил разносторонне охарактеризовать экологический дискурс как объект изучения. На базе подробной статистической информации были интерпретированы и оценены категории токенов, ключевых слов и термов. В свою очередь, вышеупомянутые сведения были структурированы: такие категории языковой материи, как леммы, коллокации и *n*-граммы, были представлены как метаданные. С помощью корпусного инструментария было проведено моделирование кластеров данных в виде конкордансов, тезаурусов и частотных словарей. Таким образом, проведенное корпусное исследование англоязычного экологического дискурса позволило создать существенный потенциал для дальнейшей лингвистической работы.

Конфликт интересов: Автор заявил об отсутствии потенциальных конфликтов интересов в отношении исследования, авторства и / или публикации данной статьи.

Conflict of interests: The author stated that there are no potential conflicts of interest regarding the research, authorship and / or publication of this article.

Литература / References

Баранов А. Н. Введение в прикладную лингвистику. М.: URSS, 2021. 368 с. [Baranov A. N. *Introduction to Applied Linguistics*. Moscow: URSS, 2021, 368. (In Russ.)]

Баркович А. А. Функциональный дискурс-анализ: модели формализации и структуризации. *Труды БГТУ. Серия 4: Принт- и медиатехнологии*. 2022. № 2. С. 29–35. [Barkovich A. A. Functional discourse analysis: Models of formalization and structurization. *Trudy BGTU. Seriia 4: Print- i mediatekhnologii*, 2022, (2): 29–35. (In Russ.)] https://doi.org/10.52065/2520-6729-2022-261-2-29-35

Баркович А. А. Компьютерно-опосредованная коммуникация: потенциал металексической значимости. *Ученые записки Петрозаводского государственного университета*. 2015a. № 7. С. 38–43. [Barkovich A. A. Computer-mediated communication: Potential for metalexical significance. *Proceedings of Petrozavodsk State University*, 2015a, (7): 38–43. (In Russ.)] https://elibrary.ru/uxkarj

Баркович А. А. Функциональность диады «коммуникационный – коммуникативный»: дискурсивный аспект. *Вестник Томского государственного университета. Филология*. 2015b. № 5. С. 37–52. [Barkovich A. A. Functionality of the "communicational–communicative" dyad: Discursive aspect. *Tomsk State University Journal of Philology*, 2015b, (5): 37–52. (In Russ.)] https://doi.org/10.17223/19986645/37/3

Зализняк А. А. «Русское именное словоизменение» с приложением избранных работ по современному русскому языку и общему языкознанию. М.: ЯСК, 2002. 752 с. [Zaliznyak A. A. *"Russian nominal inflection" with the application of selected works on the modern Russian language and general linguistics*. Moscow: YaSK, 2002, 752. (In Russ.)]

Захаров В. П. Пролегомены к корпусной лингвистике. *Вопросы психолингвистики*. 2016. № 28. С. 150–161. [Zakharov V. P. Prolegomena to corpus linguistics. *Questions of psycholinguistics*, 2016, (28): 150–161. (In Russ.)] https://elibrary.ru/wddxwl

Маник С. А., Смирнова В. Л. Современные подходы к анализу сложности учебного текста на материале учебников английского языка. *Вестник Ивановского государственного университета. Серия: Гуманитарные науки*. 2022. № 2. С. 53–63. [Manik S. A., Smirnova V. L. Contemporary approaches to analysis of complexity of educational text based on data of English textbooks. *Vestnik Ivanovskogo gosudarstvennogo universiteta. Serija: Gumanitarnye nauki*, 2022, (2): 53–63. (In Russ.)] https://doi.org/10.46726/H.2022.2.7

Плунгян В. А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики. *Русский язык в научном освещении*. 2008. № 2. С. 7–20. [Plungian V. A. Corpus as a tool and as an ideology: lessons of modern corpus linguistics. *Russian Language and Linguistic Theory*, 2008, (2): 7–20. (In Russ.)] https://elibrary.ru/mtbalv

Рассказы о сновидениях: Корпусное исследование устного русского дискурса, ред. А. А. Кибрик, В. И. Подлесская. М.: ЯСК, 2009. 735 с. [*Dream stories: Corpus study of oral Russian discourse*, eds. Kibrik A. A., Podlesskaya V. I. Moscow: YaSK, 2009, 735. (In Russ.)] https://elibrary.ru/rbonut

Филиппова Т. А. Понятие и основные характеристики экологического дискурса (на материале англоязычных СМИ). *Известия Волгоградского государственного педагогического университета*. 2018. № 2. С. 97–101. [Filippova T. A. Concept and main characteristics of ecological discourse (by the material of English-speaking media). *Izvestia of the Volgograd State Pedagogical University*, 2018, (2): 97–101. (In Russ.)] https://elibrary.ru/yqxpnk

Alexander R., Stibbe A. From the analysis of ecological discourse to the ecological analysis of discourse. *Language Sciences*, 2014, 41: 104–110. http://dx.doi.org/10.1016/j.langsci.2013.08.011

Biber D., Conrad S., Reppen R. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press, 1998, 300.

*Corpora and Discourse. Proceedings of CamConf 2002. Linguistic Insights 9. Studies in Language and Communicatio*n, eds. Partington A., Morley J., Haarman L. Bern: Peter Lang, 2004, 420. http://dx.doi.org/10.1093/llc/fqi061

Corpora and discourse studies: Integrating discourse and corpora, eds. Baker P., McEnery A. Basingstoke: Palgrave Macmillan, 2015, 310.

Corpora and discourse: The challenges of different settings, eds. Ädel A., Reppen R. Amsterdam: John Benjamins, 2008, 295.

Cutts M. *Oxford guide to plain English*. Oxford: Oxford University Press, 2020, 368.

Dijk van T. A. Discourse and Knowledge. *The Routledge Handbook of Discourse Analysis*, eds. Gee J. P., Handford M. London: Routledge, 2012, 587–603.

Discourse patterns in spoken and written corpora, eds. Aijmer K., Stenström A.-B. Amsterdam: John Benjamins, 2004, 279.

Flowerdew L. Corpus-based discourse analysis. *Routledge handbook of discourse analysis*, eds. Gee J. P., Handford M. London: Routledge, 2012, 174–187.

Kilgarriff A., Tugwell D. Word sketch: Extraction and display of significant collocations for lexicography. *Collocation: Computational Extraction, Analysis and Exploitation*: Proc. 39th ACL & 10th EACL workshop. Toulouse, 2001, 32–38.

Lafayette M. De. *Thesaurus and lexicon of similar words and synonyms in 21 dead and ancient languages and dialects*. NY: Times Square Press, 2015, 536.

Lin D. Automatic retrieval and clustering of similar words. *17th International Conference on Computational Linguistics*: Proc. 17th Inter. Conf., Montréal, 10–14 Aug 1998. NJ: Association for Computational Linguistics, 1998, 768–774. http://dx.doi.org/10.3115/980432.980696

McEnery T., Hardie A. *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press, 2012, 294.

Olvera-Lobo M. D., Gutiérrez-Artacho J., Rivera-Trigueros I., Díaz-Millón M. *Innovative perspectives on corporate communication in the global world*. Hershey, PA: IGI Global, 2021, 319.

Porter J. H. Evaluating a thesaurus for discovery of ecological data. *Ecological Informatics*, 2019, 51: 151–156. http://dx.doi.org/10.1016/j.ecoinf.2019.03.002

Song J., Tang M. Ecological discourse analysis from the perspective of systemic functional linguistics. *5th International Conference on Education Science and Development (ICESD 2020)*: Proc. 5th Intern. Conf., Bangkok, 6–7 Jan 2020. Lancaster, Pennsylvania: DEStech Publications, Inc., 2020, 558–563. http://dx.doi.org/10.12783/dtssehs/icesd2020/34131