**составление корпуса научных публикаций в сфере оптики**[[1]](#footnote-1)

**Аннотация.** В ходе данной работы были выявлены критерии отбора научных публикаций для формирования сбалансированного и репрезентативного англоязычного текстового корпуса в сфере оптики. Кроме того, были выявлены ограничения используемой методики, возможные направления ее улучшения и дальнейшие шаги по созданию корпуса. Он будет содержать более 3 500 научных работ, опубликованных в период 2010-2020 годов, и позволит анализировать научные статьи, лингвистически характеризовать научные направления и овладевать современной терминологией. На базе этих материалов в дальнейшем можно разрабатывать учебно-методические пособия для курсов ESP (English for Specific Purposes).

**Ключевые слова:** корпусная лингвистика, методика построения текстовых корпусов, англоязычный корпус в области оптики, «оптический корпус», английский язык для специальных целей.

**1. Введение**

Английский язык прочно занял позиции lingua franca в международном научном сообществе [1]. Владея им, пользователь получает доступ к огромной базе актуальной научной информации и получает средство для эффективного взаимодействия с представителями мировых научных школ. В полной мере это касается такой научно-технической области, как оптика. Темп развития современной оптики очень высок. Ежегодно в международных научных базах цитирования появляется порядка 15000 англоязычных научных статей, посвященных оптике, отражающих последние достижения ученых в этой области. Чтобы соответствовать современному уровню, студентам, ученым и специалистам в области оптики необходимо своевременно получать новые знания из надежных источников. В этой связи составление текстового корпуса научных работ в области оптики представляется весьма актуальной задачей, поскольку, используя пополняемые текстовые корпуса и современные инструменты цифровой лингвистики, технические специалисты смогут быстро проводить анализ текстовых данных, выявлять научные тренды и осваивать современную терминологию [2]. Однако, несмотря на важность этой тематики, до сих пор крайне мало исследований, посвященных созданию и изучению корпусов научных текстов по оптике [3, 4], и, насколько известно, существует всего один доступный коммерческий корпус специальных текстов по оптике (на французском языке) [5]. Составление текстового корпуса –непростая задача, потому что неизбежно возникают вопросы, связанные как с методологическими аспектами, так и с практической реализацией. В данной работе мы освещаем некоторые особенности составления специализированного текстового корпуса научных публикаций в сфере оптики.

**2. Материалы и методы**

Поиск и отбор научных публикаций проводился с использованием поискового инструмента библиографической и реферативной базы данных Scopus. Исходные критерии выборки:

o Предметная область: оптика;

o Временной охват: 11 лет – с 2010 по 2020 (этого достаточно для проведения диахронического анализа, в то же время корпус будет содержать наиболее актуальную информацию);

o Язык: английский; авторы текстов из США, Великобритании, Канады, Австралии, Ирландии, Новой Зеландии (этим ограничением в некоторой степени обеспечивается «эталонность» языка и стиля, поскольку с большой долей вероятности среди авторов публикаций есть носители английского языка);

o Принцип доступа к публикациям: открытый (это необходимо, чтобы избежать лишних расходов и дополнительной работы с копирайтингом).

**3. Исследование, результаты и обсуждение**

***3.1.  Отбор публикаций по типу документа***

За основу классификации по типу документа была взята классификация, принятая в базе Scopus; был проведен поиск документов в соответствии с рассматриваемой тематикой. На рисунке 1 показано распределение публикаций с ключевым словом «оптика» по типам документов (количество документов). К примеру, в двух самых больших сегментах количество публикаций с типом «статья» включает 222 276 наименований, в то время как тезисов конференций 140 713. Остальные сегменты намного меньше, чем предыдущие два.

Стоит отметить, что мы сознательно избегали привлечения в качестве источника материала монографий и учебников, поскольку это не в полной мере первоисточники и в них велика инерция представления актуальной информации. Соответственно, рассматривая элементы корпуса в диахронном аспекте, мы бы получили не совсем корректную информацию. Кроме того, количество токенов в книгах значительно больше, чем в средней статье или тезисах конференции, что сделало бы выборку неоднородной с точки зрения объема текстов.



*Рис. 1.* **Распределение публикаций, отобранных по ключевому слову «оптика», по типам документов**

С учетом вышесказанного рассматривалось два варианта отбора публикаций по типу документа – пропорциональный отбор от каждого сегмента, либо паритетная выборка из двух самых крупных сегментов. При выборе первого варианта репрезентативность корпуса оказывается выше, но пополнять корпус при этом весьма проблематично. Так как основная часть документов представлена статьями и тезисами конференций и нет значимых отличий по используемой специальной терминологии от жанров обзора, письма и др., было решено привлечь только эти два типа документов. Аналогичное решение принято в корпусе современного американского английского языка (Corpus of Contemporary American English, COCA), в котором тексты поровну разделены между крупными жанрами [6]. Причем этот состав относится как к корпусу в целом, так и к каждому временному периоду (году), который в нем представлен. А корпус исторического американского английского (Corpus of Historical American English, COHA), помимо этого, размечен по десятилетиям в плане жанров (например, книга, статья, обзор) и научных областей [7]. Это означает, что исследователи могут диахронически сравнивать корпусные данные и отслеживать изменения в языке на материале однотипных текстов.

***3.2.  Отбор публикаций по тематической области***

В области оптики исследования часто бывают междисциплинарными, и все очевиднее наблюдается тенденция к усилению этого эффекта, так как мы живем в эпоху активной интеграции научных направлений и развития стыковых предметных областей. Этим объясняется многообразие представленных научных направлений. Первые десять позиций в ранжированном списке количества публикаций в той или иной области знаний (согласно данным базы Scopus), отобранных по ключевому слову «оптика», приведены в таблице 1.

Как видно из таблицы, самые крупные тематические области – физика и астрономия, инженерия и материаловедение – могут сформировать сбалансированный текстовый корпус, так как вклад этих секторов количественно близок. Поэтому мы решили ограничились только этими областями. Таким образом, с учетом всех критериев был составлен алгоритм отбора научных публикаций для составления корпуса текстов сферы «Оптика» (см. Таблицу 2).

*Таблица 1.* **Распределение публикаций в сфере оптики по тематическим областям**

|  |  |  |
| --- | --- | --- |
| **№** | **Тематическая область** | **Количество публикаций** |
| 1 | Физика и астрономия | 239269 |
| 2 | Инженерное дело | 203075 |
| 3 | Материаловедение | 152984 |
| 4 | Информатика | 92131 |
| 5 | Математика | 67907 |
| 6 | Химия | 33458 |
| 7 | Медицина | 22452 |
| 8 | Науки о Земле и планетах | 13917 |
| 9 | Биохимия, генетика и молекулярная биология | 12792 |
| 10 | Химическая инженерия | 10161 |

*Таблица 2.* **Алгоритм отбора научных публикаций для формирования корпуса текстов сферы «Оптика»**

|  |  |
| --- | --- |
| **Алгоритм отбора публикаций** | **Кол-во публ.** |
| Ключевое слово: оптика | 377 741 |
| Доступ: открытый | 79 034 |
| Тип документа: статья, тезисы конференций | 75 664 |
| Тематическая область: физика и астрономия, инженерия, материаловедение | 65 455 |
| Язык: английский | 64 256 |
| Страны: основные англоязычные страны (N = 6) | 30 797 |
| Временной охват: 11 лет, с 2010 по 2020 года | 20 464 |

В итоге, в соответствии с алгоритмом отбора, получилось ограничить число публикаций до 20 464. Распределение публикаций по годам в процентном выражении варьировалось от 8 до 11 %, а в натуральном выражении – от 1549 до 2307. Далее характеристики всех статей были экспортированы в таблицу Excel. Данные содержали информацию: Автор(ы), Название документа, Год, Название источника, том, выпуск, страницы, Количество цитирований, Источник и тип документа, DOI, Издатель, Ключевые слова.

На данном этапе мы выявили ограничение методики: настройки сайта Scopus не позволяют экспортировать данные больше 2000 документов за одно скачивание. Поэтому данные экспортировали отдельно по каждому году и в те периоды, на которые приходилось больше 2000 публикаций, количество метаданных публикаций было лимитировано числом 2000.

***3.3.  Принцип случайности отбора публикаций***

После сбора выходных данных публикаций был рассчитан объем репрезентативной выборки *css* по формуле [8]:

$CSS=\frac{SS}{1+\frac{SS-1}{POP}}$ , (1)

где *ss –* размер выборки в общем случае при заданных точности и погрешности*, pop –* генеральная совокупность.

При точности (доверительной вероятности) 95% и погрешности (доверительном интервале) 5% *ss* составляет 384,16 [8]. Таким образом, для верхнего предела генеральной совокупности (2000 публикаций) объем репрезентативной выборки равен 322. Несмотря на то, что объем репрезентативной выборки в некоторые годы меньше 322, необходимо было применить это число публикаций как верхний предел, так как сбалансированность корпуса обеспечивается равным вкладом всех временных периодов. Для соблюдения принципа репрезентативности при отборе публикаций был использован генератор случайных чисел.

***3.4.  Ограничения методики и возможности ее корректировки***

Оказалось, что не все публикации, отмеченные в базе Scopus как публикации открытого доступа, на самом деле доступны для скачивания. В связи с этим необходимо обрабатывать гораздо больший массив выходных данных публикаций, чем предполагалось ранее. Поэтому в дальнейшем необходимо будет решить вопрос автоматизации загрузки статей. М. Дэвис [6] использовал VB.NET (интерфейс и язык программирования), с помощью которого создал определенный сценарий загрузки текстов. Подобная автоматизация процесса построения корпуса была бы оптимальна и в нашем случае. Однако существует проблема легитимности парсинга веб-страниц, поскольку его запрещено применять для подавляющего большинства издательских сайтов (см., например, лицензию Elsevier Text and Data Mining (TDM) [9]).

**4. Выводы**

В рамках данной работы составлен алгоритм отбора научных публикаций для формирования сбалансированного и репрезентативного англоязычного текстового корпуса в сфере оптики. Создана база данных, содержащая информацию о более чем 3 500 научных работ, опубликованных в период с 2010 по 2020 гг. В соответствии с этими данными проведен отбор текстового материала за 2020 год. Выявлены ограничения предлагаемой методики и способы ее улучшения. Дальнейшие шаги по созданию корпуса будут включать решение вопроса автоматизации загрузки документов, сбор всего необходимого материала, а также разработку архитектуры корпуса.

**5. Литература**

1. *O'Neil D.* (2018), English as the lingua franca of international publishing, World Englishes, 37 (2), pp. 146–165.

2. *Wu L.-F.* (2014), Motivating college students' learning English for specific purposes courses through corpus building, English Language Teaching, 7 (6), pp. 120–127.

3. *Louvigne S., Shi J., Sharmin S.* (2014), A Corpus-based Analysis of the Scientific RA Genre and RA Introduction, International Conference on Advanced Mechatronic Systems, № 6911636, pp. 123–127.

4. *Louvigne S., Jie S.* (2016), Data-Driven Analysis of the Development of Linguistic Features in Research Articles on Optics, International Conference on Advanced Mechatronic Systems, № 7813502, pp. 516–520.

5. *Mapelli V.* VERBA Polytechnic and Plurilingual Terminological Database – B – MP Optics, ISLRN: 642-374-718-061-9 // ELRA Catalogue of Language Resources, available at: http://catalog.elra.info/en-us/repository/browse/ELRA-T0126/.

6. *Davies M.* (2009), The 385+ million word Corpus of Contemporary American English (1990-2008+): Design, architecture, and linguistic insights, International Journal of Corpus Linguistics, 14 (2), pp. 159–190.

7. *Davies M.* (2012), Expanding horizons in historical linguistics with the 400-million word Corpus of Historical American English, Corpora, 7 (2), pp. 121–157.

8. Sample size calculation / available at: https://socioline.ru/rv.php.

9. Policies / Text and data mining / Elsevier TDM License available at: https://www.elsevier.com/about/policies/text-and-data-mining/elsevier-tdm-license.

**Designing a corpus of scientific publications in optics**

E.A. Korsakova

**Abstract.** In this study, we identify the criteria for selecting scientific publications to design a balanced and representative English-language textual corpus in the field of optics. In addition, we discuss the limitations of the method used, possible directions for its improvement and further steps toward building the corpus. It will contain papers published from 2010 to 2020, and it will allow analyzing scientific articles, revealing scientific trends and mastering state-of-the-art terminology. Basing on these data, one can develop teaching/learning materials for ESP courses (English for Specific Purposes).

**Keywords:** corpus linguistics, methodology for building textual corpora, English-language corpus in optics, "optical corpus", ESP.

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Корсакова Елена Анатольевна**

Уральский федеральный университет им. первого Президента России Б. Н. Ельцина (Россия).

**Korsakova Elena Anatolievna**

Ural Federal University named after the first President of Russia B. N. Yeltsin (Russia).

***E-mail: korsakovaea@mail.ru***

1. Работа выполнена при финансовой поддержке гранта Президента РФ для магистрантов № 330-М. [↑](#footnote-ref-1)