ИЗДАТЕЛЬСТВО ГРАМОТА

Филологические науки. Вопросы теории и практики
Philology. Theory & Practice

2021. Том 14. Выпуск 12. С. 3986-3990 | 2021. Volume 14. Issue 12. P. 3986-3990

ISSN 1997-2911 (print)

Материалы журнала доступны на сайте (articles and issues available at): philology-journal.ru

RU

Выявление ключевых слов тематического поля «Образование/Education»

Башмакова А. Ю.

Аннотация. Цель исследования - определить состав и особенности ключевой лексики тематического поля «Образование/Education» для русского и английского языков. В статье описываются этапы автоматизированного сбора корпуса новостных статей с сайтов образовательных онлайн-порталов “EDU-Inform” и “Education Today Magazine”. Значительное внимание уделяется лингвистическому анализу выявленной ключевой лексики. Научная новизна исследования заключается в междисциплинарном рассмотрении вопроса изучения ключевых слов и использовании инструментов компьютерного программирования для автоматической обработки текстов на естественном языке. В результате исследования была представлена визуализация тематического поля «Образование/Education» в формате облака слов для русского и английского языков.

EN

Keywords Extraction on the Topic "Образование/Education"

Bashmakova A. Y.

Abstract. The paper aims to identify the keywords features of the thematic field “Образование/Education” in the Russian and English languages. The article describes the stages of the automated parsing of the news articles from the websites of the educational online portals “EDU-Inform” and “Education Today Magazine” for corpus creation. The study also pays significant attention to the linguistic analysis of the extracted keywords. The scientific originality of the research consists in the interdisciplinary consideration of the issue of keywords studying and usage of computer programming instruments for the automated natural language text processing. As a result, the research presents the visualization of the thematic field “Образование/Education” for the Russian and English languages in the form of a word cloud.

Введение

Цифровые технологии уже на протяжении второго десятилетия оказывают значительное влияние на трансформацию подходов и методов гуманитарных исследований. Такие научные направления, как компьютерная лингвистика и цифровая гуманитаристика, позволяют сохранить глобальную парадигму лингвистического и филологического изучения языка, при этом меняя его качество и способствуя повышению верифицируемости результатов исследования. Актуальность настоящей статьи обусловлена повышенным интересом ученых к проблематике исследований направления цифровой гуманитаристики, а также применением инструментов компьютерной лингвистики для изучения особенностей лексического состава языков.

Для достижения поставленной цели необходимо решить следующие задачи: 1) собрать корпус новостных статей по теме «Образование/Education» для русского и английского языков; 2) выявить и описать ключевые слова тематического поля «Образование/Education»; 3) визуализировать полученные результаты на основе частотности употребления ключевой лексики. В ходе исследования использовались следующие методы: теоретико-аналитический метод, метод контент-анализа, метод сопоставительного анализа и метод лингвистического описания и интерпретации.

Материалом исследования стал корпус текстов новостных статей, состоящий из двух частей: *русскоязычная часть корпуса* представлена текстами с сайта образовательного портала *“EDU-Inform”* (http://edu-inform.ru/), *англоязычная часть корпуса* – статьи онлайн-издания *“Education Today Magazine”* (https://www.education-today.co.uk/). Языковой материал данных источников позволяет выявить ключевые лексемы, которые способны наиболее репрезентативно продемонстрировать основные темы, характеризующие тематическое поле «Образование/Education», так как представляют собой разнородные данные, охватывающие все основные актуальные топики, связанные со сферой образования.

Теоретической основой настоящего исследования послужили работы российских и зарубежных ученых в области лексико-семантических исследований (Ахманова, 1957; Филин, 1957; Арнольд, 1966; Глобина, 1995;

Научная статья (original research article) | https://doi.org/10.30853/phil20210598

Лысякова, 2005) и компьютерной и корпусной лингвистики (Kaser, Lemire, 2007; Anandarajan, Hill, Nolan, 2019; Grootendorst, 2020).

Практическая значимость исследования заключается в том, что результаты проведенного в статье анализа могут быть использованы в процессе преподавания таких лингвистических дисциплин, как основы лексикологии, компьютерная лингвистика, лингвострановедение, а также стать материалом для проведения кросскультурных сопоставлений с целью углубления и расширения исследований особенностей процесса образования в эпоху глобализации.

Основная часть

Лексический состав языка и его семантические характеристики являются одним из самых важных и наиболее обширных направлений лингвистических исследований. Именно поэтому к настоящему моменту сформировались различные классификации для группировки лексики языка на основе системных семантических отношений, существующих между словами. Так, например, учеными выделяются: *семантические классы, синонимические* и *антонимические ряды, лексико-семантические группы* и *поля*, *тематические поля* и *группы* и другие (Глобина, 1995). Первоначально теория семантического поля разрабатывалась немецкими филологами: в *ономасиологическом* подходе Йостом Триром, который делал акцент на парадигматические отношения лексем, и в *семасиологическом* подходе Вальтером Порцигом, который считал важным выявлять синтагматические отношения лексики (цит. по: Лысякова, 2005, с. 5). Позднее при описании лексических единиц их стали также объединять в тематические группы, которые были описаны в работах О. С. Ахмановой (1957), И. В. Арнольд (1966) и других ученых. Так, Ф. П. Филин противопоставлял тематические группы (ТГ) лексико-семантическим (ЛСГ). Основная причина состояла в том, что лексемы, входящие в ЛСГ, объединены по принципу семантической однородности значений, а в ТГ такие отношения основываются «не на лексико-семантических связях, а на классификации самих предметов и явлений» (Филин, 1967, с. 526). В связи с этим в тематические группы допустимо объединять слова с экстралингвистической семантикой, которые обладают только общностью предметной сферы и не обязательно схожим лексическим значением. Отличие тематического поля от группы в большинстве случаев проявляется в большем размере первого. Для целей настоящего исследования мы будем использовать объединение лексики в **тематическое поле (группу),** которое понимается как совокупность слов разных частей речи, семантически объединенных одной общей темой или связанных с определенным понятием в широком смысле. Выбор такого формата обусловлен тем, что тематические поля и группы обладают наибольшей системообразующей силой, так как лексемы, входящие в них, могут находиться не только в отношениях синонимии и антонимии, но и в связях ассоциативного характера.

Работа по сбору корпуса, автоматической обработке текстов, выявлению и анализу ключевых слов, описанная в настоящей статье, включала в себя три основных этапа:

1 этап. *Автоматический сбор корпуса текстов новостных статей образовательных порталов "EDU-Inform" и "Education Today Magazine"*. Использование текстовых данных, доступных в сети Интернет, в исследованиях естественного языка становятся все более популярным ввиду многочисленных преимуществ такого ресурса: открытость, бесплатный доступ, удобный формат хранения информации. Например, если данные содержатся на веб-странице, то они могут быть автоматически собраны с помощью *веб-скрейпинга (web-scraping)* или, другими словами, *парсинга* сайтов (Anandarajan, Hill, Nolan, 2019, с. 35). Данный метод был использован в настоящем исследовании.

2 этап. *Предобработка текстов и выявление ключевых слов*. Предварительная обработка текста необходима для приведения языковых данных к единому формату и уменьшения их размерности. Процесс предобработки включает в себя: *токенизацию*, *стандартизацию*, *удаление стоп-слов* и *лемматизацию*. Для автоматизированной обработки текстов новостных статей на русском и английском языках и дальнейшего выявления ключевых слов были использованы предобученные модели **BERT** *(Bidirectional Encoder Representations from Transformers)* по методу голландского исследователя Маартена Гроотендорста (Grootendorst, 2020).

3 этап. *Анализ результатов и визуализация*. Данный этап включал в себя определение объема выборки ключевых слов на основе их частотности, изучение лингвистических особенностей и сопоставление полученных данных для русского и английского языков. Результаты частотного анализа ключевой лексики были оформлены в формат облака слов с использованием библиотеки wordcloud для языка программирования Python (Kaser, Lemire, 2007).

В настоящем исследовании для изучения русского языка используются материалы новостных статей образовательного портала "EDU-Inform" (http://edu-inform.ru/). Данный ресурс является информационно-справочным порталом, который содержит сведения об образовательных заведениях России всех уровней, их структуре, направлениях деятельности и перечне предоставляемых услуг, включая дошкольное, школьное, высшее, профессиональное и дополнительное образование. В том числе на сайте портала представлены дополнительные сервисы: доска объявлений, пресс-релизы и дни открытых дверей вузов, актуальные новости образования в регионах РФ и в России в целом. Таким образом данный ресурс наиболее оптимально удовлетворял целям и задачам настоящего исследования.

На момент сбора корпуса раздел сайта во вкладке «Новости образования» насчитывал **112 страниц** и **1119 новостных статей**, охватывающих период с июня 2011 года по август 2021 года. Русскоязычная часть корпуса

насчитывает: **34960 предложений**, **779793 словоупотребления** (tokens), из которых **49199 уникальных слов**, в том числе содержатся метаданные к каждой статье: *заголовок статьи, дата публикации, ссылка на web-страницу, текст статьи, предобработанный текст статьи* и *список ключевых слов*. Общий список ключевых слов составляет **11190 единиц**, по 10 слов для каждой статьи в подкорпусе. На основе частотного анализа и группировки списка ключевых слов была выявлена 2281 уникальная лексема, которые наиболее характерно отражают тематику публикаций в сфере образования на русском языке.

Важно также отметить, что из-за особенностей публицистического стиля русского языка тексты новостных статей, используемых в настоящем исследовании, были насыщены канцеляризмами и именами собственными, которые из-за высокой частотности употребления попадали в список ключевых слов. В связи с этим некоторые из таких лексем были дополнительно добавлены к стандартному списку стоп-слов библиотеки *pymorphy2* с целью диверсификации результатов, хотя частично общеупотребительная лексика всё же осталась и рассматривается в качестве погрешности. Результаты подсчета частотности объединенного списка ключевых слов тематического поля «Образование» для русского языка были оформлены в облако слов (Рис. 1).

По получившемуся облаку слов можно заметить, что образовательный дискурс русского языка подвержен сильному влиянию особенностей официально-делового и публицистического стилей, а также специфике репрезентации информации в прессе государственных структур. Так, например, можно выделить преобладание сложных прилагательных по сравнению с другими частями речи, обилие канцеляризмов и нейтральной общеупотребительной лексики: *«пресс-служба», «необходимость», «использование», «продолжительность», «большинство», «обязательный», «опубликовать», «предоставлять», «организовать», «соответствовать», «воспользоваться», «обсуждение», «завершиться», «прокомментировать»*.

Ключевую лексику на примере наиболее частотных слов можно разделить тематически на несколько подгрупп. Одна из них описывает основную действительность процесса образования, его типы, особенности и участников: *«образование», «просвещение», «профобразование», «специальность», «преподаватель», «руководитель», «образовательный», «студенческий», «общеобразовательный», «педагогический», «преподавательский», «учебно-методический», «университетский», «экзаменационный», «вступительный», «научно-исследовательский», «обществознание», «контрольно-измерительный», «профессиональный», «технологический»*. В связи с тем, что в контексте российской образовательной традиции государство играет важную роль в регулировании процесса обучения, формируется следующая подгруппа ключевых слов: *«премьер-министр», «муниципальный», «правоохранительный», «законодательство», «демографический», «подведомственный»*. Кроме этого, важно отметить, что ситуация, сложившаяся в современном мире в связи с глобальной эпидемиологической обстановкой, оказала своё влияние в том числе и на процесс образования, что также отразилось в изменении новостных топиков, в которых частотными стали следующие слова, отражающие новый формат обучения и проведения различных образовательных мероприятий: *«безопасность», «распространение», «компьютерный», «психологический», «международный», «взаимодействие», «самостоятельный»*.

Следующим шагом стал сбор подкорпуса текстовых данных для английского языка. Для определения ключевых лексем тематического поля “Education” в английском языке в настоящем исследовании используются материалы новостных статей с сайта онлайн-издания журнала “Education Today Magazine” (https://www.education-today.co.uk/). Данный британский журнал публикует статьи на различные темы, охватывающие весь сектор образования: текущие новости, особенности учебных программ, правительственные стратегии, обзоры выставок, конкурсы и прочие. Таким образом данный ресурс оптимально удовлетворял целям и задачам настоящего исследования и обладал схожей направленностью с русскоязычным источником.

На момент сбора корпуса новостной раздел сайта во вкладке “Latest News” насчитывал **54 страницы** и **534 новостных статьи**, которые охватывали период с июня 2014 года по август 2021 года. Статистическая информация англоязычной части корпуса: **8071 предложение**, **285320 словоупотреблений** (токенов), из которых **18853 уникальных слова**, в том числе содержатся метаданные к каждой статье: *заголовок, тематическая категория новости, дата публикации, ссылка на web-страницу, оригинальный текст статьи, предобработанный текст* и *список ключевых слов*. Общий список ключевых слов составляет **5340 единиц**, по 10 слов для каждой статьи в корпусе. На основе частотного анализа и группировки списка ключевых слов было выявлено **1755 уникальных лексем**, которые наиболее характерно отражают тематику публикаций в сфере образования на английском языке.

Необходимо отметить, что англоязычная часть корпуса в сравнении с русскоязычной обладает рядом отличий, такими как временной период публикаций, общее количество статей и, как следствие, статистические показатели корпуса. В обоих случаях, при сборе русскоязычной и англоязычной частей корпуса, был использован весь имеющийся текстовый материал опубликованных статей, поэтому для сохранения однородности данных, стиля и формата публикаций было принято решение не добавлять и не использовать дополнительно другие англоязычные источники с целью создания эквивалентного объема статистических показателей. В связи с тем, что данная особенность не оказывает негативного влияния на работу модели машинного обучения по выявлению ключевых слов, в настоящем исследовании описанные различия рассматриваются как индивидуальные характеристики источников текстового материала разных языков. Результаты подсчета частотности объединенного списка ключевых слов тематического поля “Education” для английского языка были оформлены в облако слов (Рис. 2). Доля нерелевантной лексики в списке ключевых слов англоязычной части корпуса не составляла большого количества, в связи с этим использовался классический список стоп-слов библиотеки NTLK, а оставшаяся общеупотребительная лексика рассматривается в качестве погрешности.

Рисунок 1. *Облако ключевых слов по теме «Образование»*

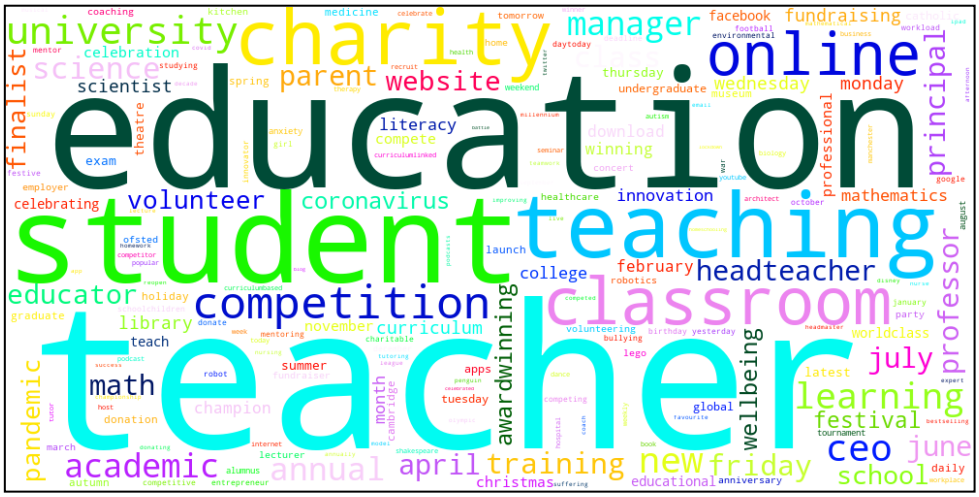

Рисунок 2. *Облако ключевых слов по теме “Education”*

При рассмотрении наиболее частотных лексем в списке ключевых слов, их можно разделить на несколько основных тематических категорий, отражающих культурные особенности образовательного процесса в Великобритании. Первая подгруппа обозначает участников и людей, вовлеченных в процесс образования и научные исследования: *«teacher / учитель», «student / студент», «professor / профессор», «educator / педагог», «principal / директор школы», «parent / родитель», «scientist / ученый»*. Следующей подгруппой выделяются лексемы, описывающие процесс обучения, дисциплины, места и названия образовательных учреждений: *«education / образование», «science / наука», «literacy / грамотность», «teaching / преподавание», «learning / обучение», «class / урок», «classroom / класс», «school / школа», «university / университет», «library / библиотека», «academic / учебный», «professional / профессиональный»*. Также важно отметить, что благотворительность является неотъемлемым от повседневной жизни процессом в Великобритании, в котором участвуют как взрослые, так и дети. Здесь можно говорить о различных ежегодных конкурсах, ярмарках, концертах и других мероприятиях, чем объясняется частотность следующих лексем: *«charity / благотворительность», «fundraising / сбор средств», «festival / фестиваль», «volunteer / волонтер»*. Кроме этого, в последние годы важным становится направление образовательного менеджмента и развития управленческих и лидерских качеств обучающихся и сотрудников учебных учреждений, это можно проследить по лексемам: *«CEO / исполнительный директор», «manager / менеджер», «competition / конкурс», «finalist / финалист», «champion / чемпион», «winning / победа»*. Отдельной группой в ключевой лексике английского языка выделяются слова, характерные для публикаций в контексте пандемии коронавируса, которые отражают новые цифровые реалии образовательного процесса: *«coronavirus / коронавирус», «pandemic / пандемия», «wellbeing / благополучие», «online / онлайн»*.

Заключение

Подводя итог, можно сделать вывод, что методы компьютерной лингвистики, такие как контент-анализ, веб-скрейпинг и машинное обучение, кардинально меняют принципы изучения языковых текстовых данных

и могут выступить дополнением к традиционному лингвистическому анализу, обеспечивая верифицируемую статистическую информацию. Материалы автоматически собранного корпуса новостных статей на русском и английском языках по теме «Образование/Education» демонстрируют культурные сходства и различия в тематике образовательных текстов на двух языках. Так, проведённый контент-анализ ключевых слов, выявленных методом машинного обучения с использованием моделей BERT, подтверждает, что для русского языка характерна государственная повестка новостей, отражающаяся в высоком уровне канцеляризмов и преобладании сложных прилагательных в списке ключевых слов. Англоязычная часть корпуса, в свою очередь, затрагивает более широкий тематический спектр новостных статей. Интересной особенностью для английского языка является преобладание слов, связанных с темой благотворительности. В общем для обоих языков наблюдается взаимное изменение тематики публикаций в условии глобальной пандемии. Таким образом можно составить зеркальный список из десять ключевых лексем, которые наиболее репрезентативно отражают общие тренды в языке новостных статей в сфере образования в русском и английском языках: *«образование / education», «университет / university», «школа / school», «студент / student», «преподаватель / teacher», «учебный / academic», «наука / science», «экзамен / examination», «профессиональный / professional», «руководитель / manager»*. В завершении исследования результаты контент-анализа собранного корпуса статей были оформлены в облако слов на основе частотности, что стало наглядной визуализацией тематического поля «Образование/Education» в русском и английском языках.

Перспективы дальнейшего исследования проблемы мы видим в более подробном сравнительно-историческом и сопоставительном анализе зеркального списка ключевых слов с изучением их этимологии и выявлением культурных особенностей семантики лексем. В дальнейшем полученные данные также могут стать материалом для создания модели компьютерного учебного мультимодального словаря историко-культурологического типа для изучающих иностранный язык.

Источники | References

1. Арнольд И. В. Семантическая структура слова в современном английском языке и методика ее исследования. Л.: Просвещение, 1966.
2. Ахманова О. С. Очерки по общей и русской лексикологии. М.: Государственное учебно-педагогическое Издательство Министерства Просвещения РСФСР, 1957.
3. Глобина Л. В. Лексико-семантическое поле партитивной лексики в современном русском языке: автореф. дисс. … к. филол. н. Воронеж, 1995.
4. Лысякова М. В. Лексико-семантические парадигмы: лингвистический статус, критерии разграничения // Russian Journal of Linguistics. 2005. № 7.
5. Филин Ф. П. О лексико-семантических группах слов // Езиковедскиі изъследования в чест на академик Стефан Младенов. София: Бълг. акад. на науките, 1967.
6. Anandarajan M., Hill C., Nolan T. Practical Text Analytics. Maximizing the Value of Text Data. Advances in Analytics and Data Science. Springer Nature Switzerland, Cham, 2019.
7. Grootendorst M. Keybert: Minimal keyword extraction with bert. 2020. URL: https://github.com/MaartenGr/KeyBERT
8. Kaser O., Lemire D. Tag-cloud drawing: Algorithms for cloud visualization // Proceedings of the World Wide Web Workshop on Tagging and Metadata for Social Information Organization. Coleman, 2007.

Информация об авторах | Author information

RU **Башмакова Анастасия Юрьевна**[1]
[1] Тюменский государственный университет

EN **Bashmakova Anastasiia Yurievna**[1]
[1] University of Tyumen

[1] *a.y.bashmakova@utmn.ru*

Информация о статье | About this article

Дата поступления рукописи (received): 26.10.2021; опубликовано (published): 28.12.2021.

Ключевые слова (keywords): компьютерная лингвистика; извлечение ключевых слов; образование; облако слов; computational linguistics; keyword extraction; education; BERT; word cloud.