Динамика лексического состава корпуса текстов по компьютерной и корпусной лингвистике: вероятностно-статистический подход

Ольга А. Митрофанова

***Аннотация*—Цель исследования заключается в анализе изменений тематики научных статей в корпусе текстов по компьютерной и корпусной лингвистике (ТКиКЛ) с применением современных методов тематического моделирования. Проведенные эксперименты основываются на применении комбинированного нейросетевого алгоритма BERTopic и матричного алгоритма NMF, расширенного за счет частотного анализа употребления слов-тематизаторов в текстах корпуса ТКиКЛ. Значимость работы заключается в формировании лингвистически обоснованной методологии для семантического поиска научной информации и отслеживания тенденций в области компьютерной и корпусной лингвистики, что может быть использовано для совершенствования наукометрических инструментов. Результаты исследования демонстрируют динамику изменений научных интересов в области компьютерной и корпусной лингвистики под влиянием цифровизации общества и технологического прогресса.**

***Ключевые слова*— компьютерная лингвистика, корпусная лингвистика, корпус текстов, динамическое тематическое моделирование, BERTopic, NMF, количественно-статистический анализ**

I. Введение

Материалы научных конференций по компьютерной и корпусной лингвистике представляют собой ценный источник информации о развитии и современном состоянии данных направлений исследований языка, а также об основных характеристиках академических текстов. Чаще всего предметом изучения оказывается научная терминология, которая претерпела существенные изменения за прошедшее двадцатилетие. В зарубежной академической среде активно развиваются инструменты наукометрии. Исследования динамики тем в области компьютерной и корпусной лингвистики сейчас актуальны как никогда, поскольку стремительно меняется ландшафт как академических проектов, так и промышленных прикладных разработок. Более всего чувствительны к изменениям научные конференции с треками по компьютерной и корпусной лингвистике: *Диалог-21* «Компьютерная лингвистика и интеллектуальные технологии» [1], *AINL* «Artificial Intelligence and Natural Language» [2], *AIST* «International Conference on Analysis of Images, Social Networks and Texts» [3], *SPECOM* «International Conference on Speech and Computer» [4], *FRUCT* [5] и т. д. В последние годы доминантой этих конференций становится лингвистика больших языковых моделей (LLM).

Отслеживать тенденции в тематике научных исследований позволяют наукометрические платформы, обеспечивающие многофакторный поиск информации об исследованиях. Для зарубежных изданий это позволяют сделать такие инструменты, как *Dimensions* [6], *OpenAlex* [7], *ResearchRabbit* [8] и ряд других. Для русскоязычного сегмента такие инструменты не столь распространены, поэтому существует потребность в разработке моделей для семантического поиска научной информации не только по ключевым словам и словосочетаниям, но и по рубрикам, связанным со структурой предметной области и исследовательскими задачами. Данная работа частично восполняет существующие пробелы.

II. Исследовательские данные

Целью исследования является описание динамики тем научных статей в корпусе текстов по компьютерной и корпусной лингвистике (ТКиКЛ), сформированном на основе материалов конференции «Корпусная лингвистика» (*Corpora*) с 2002 по 2021 гг. и семинара «Компьютерная лингвистика и вычислительные онтологии» (*CompLing*) с 2011 по 2023 гг. [9; 10; 11; 12]. В настоящее время в состав корпуса входят 643 текста. Общий объем корпуса составляет более 1 млн словоупотреблений. Сегмент корпуса, представляющий материалы конференции *Corpora*, содержит 442 текста, материалы семинара *CompLing* — 201 текст.

Тексты корпуса распределены по годам и представлены в неразмеченном и лемматизированном виде. В корпусе есть следующие типы информации: при сохранении общей структуры статей (авторство, заголовок, набор ключевых выражений, аннотация, текст статьи) проведена автоматическая разметка ключевых выражений с применением библиотеки *RuTermExtract* [13], генерация аннотаций с помощью моделей экстрактивной и абстрактивной суммаризации в библиотеке *sumy* [14] и с помощью модели *ruT5* [15], систематизация и разметка терминологизированных именованных сущностей, мультимодальное

Статья получена 11 ноября 2025 г.

О.А.Митрованова, канд. филол. наук, доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия (e-mail: o.mitrofanova@spbu.ru).

тематическое моделирование корпуса с автоматическим назначением меток тем, а также экспертная разметка рубрик в корпусе. В частности, были выделены следующие рубрики: 1) *Общие вопросы корпусной лингвистики*, 2) *Создание, разработка и применения корпусов*, 3) *Статистические исследования на материале корпусов*, 4) *Корпусы и лексикография*, 5) *Морфология и синтаксис в корпусах*, 6) *Семантика в корпусах*, 7) *Параллельные корпусы и машинный перевод*, 8) *Обучающие корпусы*, 9) *Исторические корпусы*, 10) *Речевые и мультимедийные корпусы*, 11) *Корпусы художественных текстов*. Имеющаяся в корпусе хронологическая разметка позволяет получить данные об изменениях в тематике статей, алгоритмом выбора для анализа данных является динамическое тематическое моделирование (Dynamic Topic Modelling).

III. Построение стандартных тематических моделей

Процедуры тематического моделирования направлены на построение семантически интерпретируемой модели корпуса текстов, в которой устанавливаются связи между документами, их словарем и темами (скрытыми факторами). Каждый текст в корпусе с той или иной вероятностью соотносится с одной или несколькими темами, которые могут пересекаться [16].

Из множества алгоритмов тематического моделирования, допускающих расширение до мультимодальных, в том числе и динамических версий (LSA, NMF, pLSA, LDA, LDA2Vec, Top2Vec и т. д.) мы выбрали BERTopic [17; 18] и NMF [19, 20].

Комбинированная модель BERTopic применяет к корпусным данным эмбеддинги трансформера BERT и включает в себя процедуры кластеризации многоступенчатым снижением размерности и ранжированием кандидатов в слова-тематизаторы для формирования тем. Стандартное тематическое моделирование с помощью алгоритма BERTopic включает три шага: 1) векторизация текстов корпуса с помощью предобученной модели BERT, 2) снижение размерности векторов методами UMAP и PCA, кластеризация результирующих эмбеддингов с помощью алгоритма HDBSCAN, 3) из сформированных текстовых кластеров извлекаются *n*-граммы в результате применения видоизмененной метрики c-TF-IDF, ранжирование *n*-грамм методом MMR, сохранение результатов как списка кандидатов в слова-тематизаторы.

В ходе экспериментов с алгоритмом BERTopic, представленным в библиотеке [17], проводится векторизация лемматизированных корпусных данных, фильтрация словаря по частоте лемм и распространенности в текстах корпуса. Параметры обучения тематической модели подбираются эмпирически (минимальный объем темы — 10 слов-тематизаторов, выделение в корпусе *n*-грамм и т. д.). В модель входит 14 тем, соотносимых с группами статей из корпуса ТКиКЛ. Примеры тем: (0) *орган, услуга, правительство, портал, власть,..* (1) *коллокат, граф, кластеризация, онтология,..* (2) *жест, реплика, УРК, разговор, ЭДЕ,..* (8) *стих, слоговой, метрический,..* (9) *житие, цитата, агиографический, СКАТ, рукопись,..* (12) *латышский, румынский, транслитерация,..* Для каждой темы модель выбирает наиболее типичный документ, например, для темы (10) это статья {Рогозина Е. А. «Разметка содержательной структуры житийных текстов в корпусе агиографических текстов СКАТ» // Труды международной конференции «Корпусная лингвистика — 2008). СПб., 2008}. Внутри темы слова-тематизаторы ранжируются по значению меры ассоциации: (9) *житие* (0.129), *цитата* (0.082), *агиографический* (0.071), *СКАТ* (0.068), *рукопись* (0.052), *житийный* (0.049), *словоуказатель* (0.035), *склонение* (0.031), *Алексеев* (0.030), *Евангелие* (0.029) и т. д.

В качестве альтернативы нейросетевому подходу, реализованному в BERTopic, был рассмотрен алгебраический алгоритм NMF, который в предыдущих экспериментах [10] продемонстрировал наиболее интерпретируемые результаты. Для заданной TF-IDF-нормализованной матрицы X, сформированной на основе корпуса текстов, алгоритм NMF формирует матрицы W и H (распределение слов по темам и тем по текстам), произведение которых аппроксимирует исходную матрицу X таким образом, чтобы все значения в матрицах W и H были неотрицательным. Интерпретируемость результатов NMF, обусловлена сокращением отрицательных элементов матриц. В результате NMF формируются темы, позитивно связанные со словами-тематизаторами, которые присутствуют в текстах корпуса.

Эксперименты с алгоритмом NMF проводились на основе библиотеки scikit-learn [21]. Процедура тематического моделирования включает несколько этапов: 1) векторизация текстовых данных с помощью метрики TF-IDF, 2) фильтрация словаря модели с отсевом стоп-слов, низкочастотных и широко распространенных лемм ($min\_df = 2$, $max\_df = 0.8$), 3) подбор оптимального числа тем с учетом когерентности (c_v), 4) построение результирующей модели, включающей ранжированные темы со списками слов-тематизаторов – лемм и биграмм. Наивысшее значение когерентности $c\_v = 14$ достигается при значении $num\_topics = 14$. Полученный список тем ранжируется по важности, оцениваемой через суммирование весов слов-тематизаторов. Например, тема с рангом 1 и с самым высоким суммарным весом слов-тематизаторов 17.97 представлена группой *слово, частота, частотный, биграмма, мера, распределение, коллокация, средний, коллекция, значение, ранговый, объём, статистический, результат* (статистическая обработка корпусных данных), тема с рангом 4 и средним суммарным весом слов-тематизаторов 14,33 включает в себя термины *речь, речевой, устный, ребёнок, коммуникативный, жест, ЭДА, пауза, запись, дискурсивный, фонетический, расшифровка, разговор, звуковой* (корпусы звучащей речи), тема с рангом 6 и средним суммарным весом 12,92 – *форма, морфологический, грамматический, словоформа, слово, словарь, падеж, таблица, существительное, словоизменительный, парадигма, вариант, часть, парадигматический* (компьютерная морфология), тема с рангом 6 и средним суммарным весом 12,92 – *форма, морфологический, грамматический, словоформа, слово, словарь, падеж, таблица, существительное,*

словоизменительный, парадигма, вариант, часть, парадигматический (компьютерная морфология), тема с рангом 9 и низким средним суммарным весом 6.04 – *ижорский, финский, песня, народный, диалектный, диалект, прибалтийский, карельский, вепсский, эпический, топоним, база, говор, Ингерманландия, топонимический* (диалектные корпусы текстов), тема с рангом 10 и низким средним суммарным весом 6.04 – *житие, цитата, агиографический, СКАТ, житийный, рукопись, словоуказатель, написание, разметка, буква, издание, выносной, древнерусский, формат* (диахронические корпусы текстов) и т.д.

Результаты BERTopic и NMF демонстрируют достаточно высокую согласованность несмотря на большую детализацию тем корпуса в BERTopic и меньшее лексическое разнообразие NMF за счет включенных цепочек дериватов.

IV. Динамическое тематическое моделирование

Для построения динамической тематической модели BERTopic в режиме «Topics over Time» определяются временные точки, относительно которых будут оцениваться изменения в наполнении тем. В корпусе ТКиКЛ данными точками являются годы проведения конференций. Результат представлен на рис. 1.

Можно рассмотреть изменение во времени каждой из тем по отдельности. В частности, наблюдается следующая динамика темы (0) *орган, услуга, правительство*,..: в 2002 г. она была сосредоточена на проблемах электронных библиотек (*читальный, зал, навигация*,..), в 2013–2014 гг. доминировала проблематика электронного правительства (*услуга, орган, портал, власть, гражданин, правительство*,..), в 2015 г. фокус тематического наполнения сместился на формальные онтологии, в 2016 г. — на компьютерные тезаурусы (*YARN, WordNet, RussNet, синсет*,..), а к 2022 г. адаптировалась к проблематике чат-ботов и клиент-ориентированной коммуникации (*бот, пациент, настроение*,..). Изменения темы (2) *жест, реплика, УРК*,.. таковы: в 2002 г. в ее составе были термины *фонетика, звукозапись*, в 2004 г. — *диалоговый, тестирование*, в 2013 г. — *коррекция, артикуляторный, самоисправление*, в 2017 г. — *УРК, метакоммуникация, реплика, пересказ* и т. д. Эволюция темы (9) *житие, цитата, агиографический*,.. такова: в 2004 г. тема была связана с исследованием диахронических и диалектных корпусов текстов в целом (*ижорский, песня, рукопись, старолатышский, житие*,..), в 2008 г. акцент сместился в сторону корпуса агиографических текстов и построения словоуказателей (*житие, агиографический, СКАТ, словоуказатель*,..), к 2011 г. — в сторону исследования цитат и морфологической аннотации диахронических корпусов текстов (*житие, цитата, СКАТ, агиографический, склонение*,..).

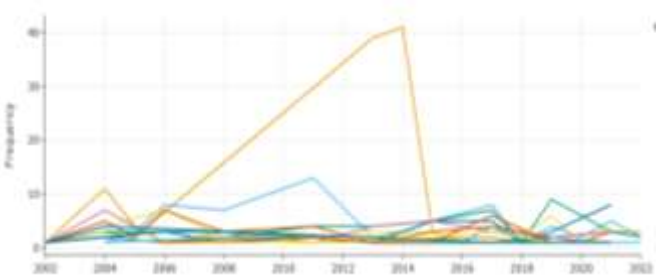

Рис. 1. Динамическое тематическое моделирование корпуса ТКиКЛ (BERTopic)

Динамическое тематическое моделирование на основе NMF обеспечивает выделение нишевых тем с высокой вариативностью слов-тематизаторов в составе тем, что иллюстрируется на рис. 2. Полученные данные были проинтерпретированы на следующем этапе анализа данных.

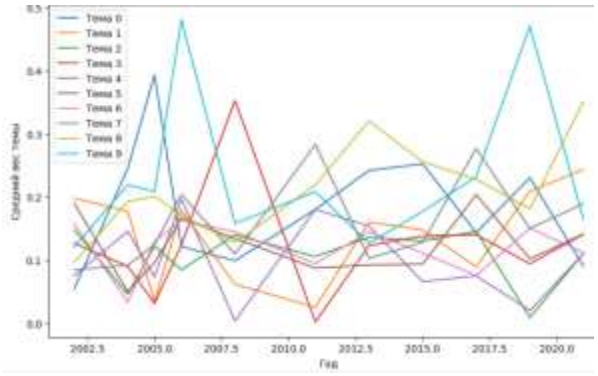

Рис. 2. Динамическое тематическое моделирование корпуса ТКиКЛ (NMF)

V. Количественно-статистический анализ динамики тем

В результате построения динамической тематической модели NMF был сформирован объединенный список слов-тематизаторов, представленный во всех хронологических периодах корпуса ТКиКЛ. В данный набор вошло 655 лемм, из них 212 присутствует как минимум в двух тематических наборах. Тепловая карта, отражающая схему повторов слов-тематизаторов в тематических моделях для разных временных периодов представлена на рис. 3. Больше всего пересечений между наборами слов-тематизаторов для 2008 и 2011 годов (43 леммы), 2008 и 2011 годов (37 лемм), 2011 и 2013 годов (37 лемм), 2011 и 2021 годов (37 лемм). Можно высказать предположение, что тематика статей 2011 года в корпусе ТКиКЛ охватывает наиболее широкий круг тем компьютерной и корпусной лингвистики. Наименьшее число пересечений наборов слов-тематизаторов зарегистрировано между 2005 и 2019 (16 лемм) и 2002 и 2019 годами (18 лемм), возможной причиной является специфичность тематики статей в исходных данных. Более наглядно данная информация представлена в графе связей, представленном на рис. 4 (узлы графа помечены ссылками на хронологические периоды, дуги графа имеют метки числа совпадений слов-тематизаторов в паре тематических моделей разных хронологических

периодов).

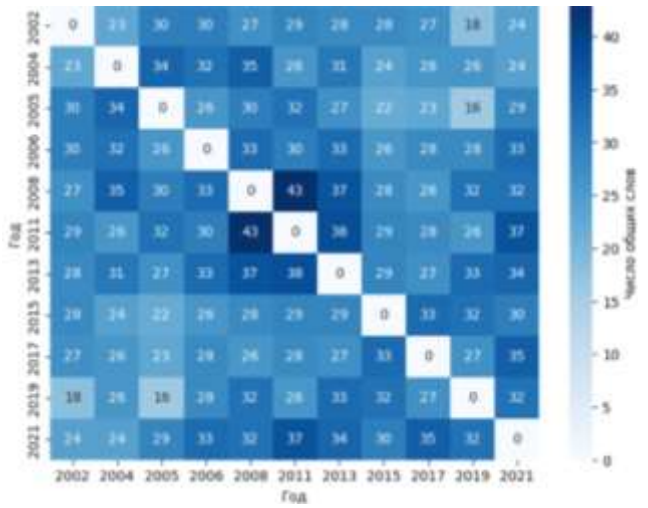

Рис. 3. Тепловая карта пересечений наборов слов-тематизаторов в тематических моделях для разных временных периодов в корпусе ТКиКЛ

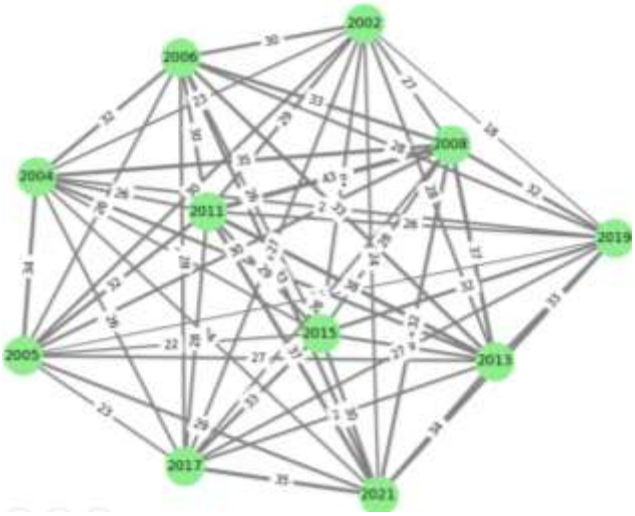

Рис. 4. Граф связей между наборами слов-тематизаторов в тематических моделях для разных временных периодов в корпусе ТКиКЛ

В результате частотного анализа объединенного списка слов-тематизаторов, представленных во всех хронологических периодах корпуса ТКиКЛ, были получены данные об изменении частот употребления данных терминов в текстах корпуса. На основании обработки ранжированного частотного списка слов-тематизаторов по годам была произведена оценка суммарных значений Δ для рангов. На рис. 5-10 отрицательные суммарные значения Δ для рангов указывают на рост частотности употребления слов-тематизаторов в текстах корпуса (зеленые столбцы), тогда как положительные соответствуют снижению частоты встречаемости (красные столбцы). Рассмотрим динамику употребления слов-тематизаторов в отдельных частично пересекающихся терминологических группах. В группе общих терминов компьютерной и корпусной лингвистики (*алгоритм, интерфейс, информация, информационный, интернет, контекст, контекстный, корпус, корпусный, корпусной, модель, подкорпус, поиск, поисковый, текст, текстовый*), см. Рис. 5, регистрируется конкуренция вариантов написания терминов корпусный и корпусной с тенденцией снижения частотности первого и повышения частотности второго (*корпусный* – 26 контекстов, *корпусной* – 253 контекста с конкуренцией форм именительного и родительного падежей, в том числе *корпусной анализ, корпусной материал, корпусной подход*), падение частоты употребления терминов *корпус, подкорпус, интернет, интерфейс* и повышением частоты терминов, связанных с интеллектуальными технологиями (*поисковый, алгоритм, информационный, модель* и т.д.).

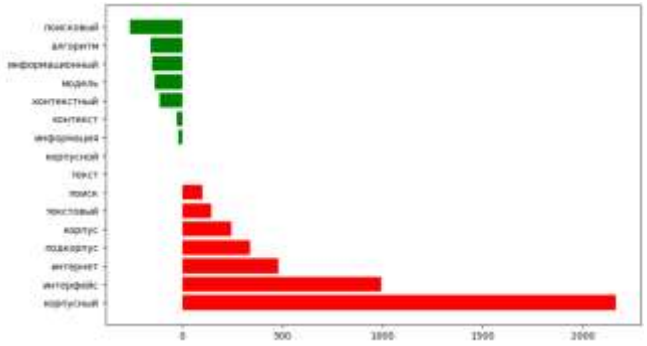

Рис. 5. Оценка суммарных значений Δ для рангов частот в группе общих терминов компьютерной и корпусной лингвистики

Группа терминов анализа лингвистических данных (*аннотация, аннотирование, дизамбигуация, корпус, лингвистический, морфологический, неоднозначность, подкорпус, парсер, прагматический, процессор, риторический, разметка, семантический, синтаксический, тег, теггер, тегсет, текст*), см. Рис. 6, содержит лексику, которая с течением времени стала реже употребляться в статьях корпуса ТКиКЛ (в частности, *процессор*, *дизамбигуация*), наряду с терминами, которые упоминаются чаще в связи с ростом интереса к соответствующей проблематике (*прагматический, риторический* – например, *разметка риторических структур*). Примечательно преобладание употребления термина *аннотирование* (*процесс разметки*) по сравнению со словом *аннотация* (в значении *аннотирование* и в значении *резюме*), употребление термина *парсер* (как более общего по значению) по сравнению с термином *теггер*, и т.д.

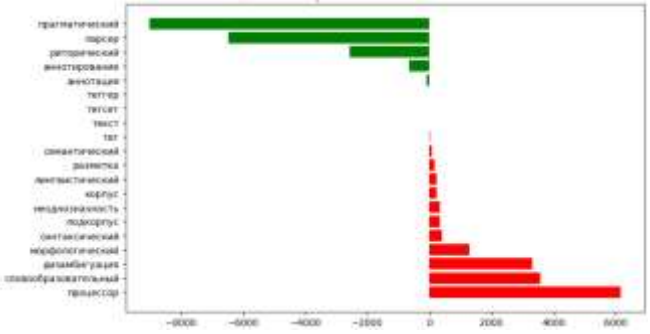

Рис. 6. Оценка суммарных значений Δ для рангов частот в группе терминов анализа лингвистических данных

Весьма показательно соотношение частотности терминов, характеризующих типы методов анализа

лингвистических данных (*алгебраический, ассоциативный, вероятностный, гибридный, интеллектуальный, когнитивный, машинный, нейросетевой, правиловый, психолингвистический, социолингвистический, статистический, формальный*), см. рис. 7. В отношении динамики частот употребления в статьях корпуса ТКиКЛ термины интеллектуальный, алгебраический, вероятностный, нейросетевой, с одной стороны, и психолингвистический, формальный, статистический, машинный, с другой стороны, проявляют разные тенденции – первая подгруппа с тенденцией к росту частоты, а вторая – с тенденцией к снижению. В паре коррелирующих по смыслу терминов *вероятностный* и *статистический* в последние годы предпочтение отдается первому.

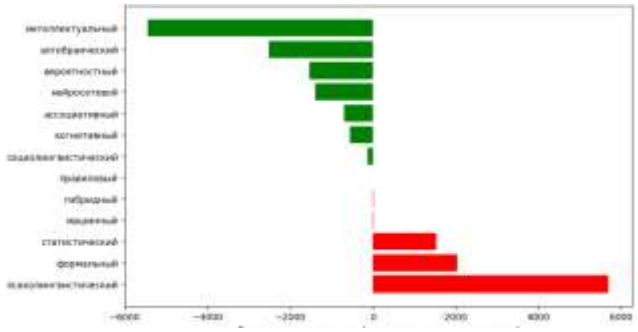

Рис. 7. Оценка суммарных значений Δ для рангов частот в группе терминов, характеризующих типы методов анализа лингвистических данных

В группе терминов, характеризующих морфологический и синтаксический анализ лингвистических данных (*аннотирование, вид, глагол, глагольный, группа, грамматика, грамматический, зависимость, дерево, именной, морфологический, морфология, падеж, время, существительное, синтаксис, синтаксический, составляющий, структура*) положительную динамику проявляют только термины *зависимость, именной, падеж* и *синтаксис*, см. Рис. 8.

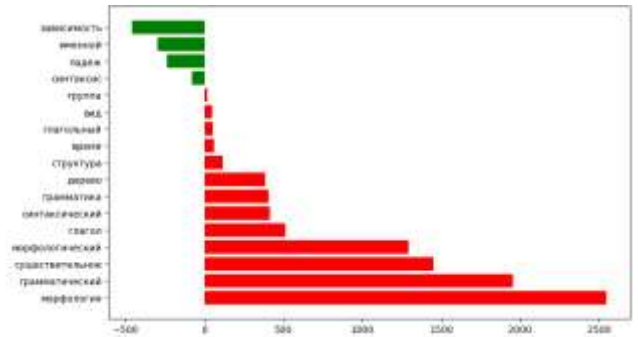

Рис. 8. Оценка суммарных значений Δ для рангов частот в группе терминов, характеризующих морфологический и синтаксический анализ лингвистических данных

Для группы терминов, характеризующих семантический анализ лингвистических данных (*актант, аргумент, валентность, значение, конструкция, лексема, пропозиция, пропозициональный, семантика, семантический, сценарий, ситуация, смысл, слово, тема, тематический, фрейм*) повышение частотности со временем проявляют леммы *валентность, аргумент, конструкция, актант, фрейм, смысл, семантика, тема*. Обратная тенденция наблюдается в отношении остальных терминов в группе, в том числе и для прилагательного *тематический*.

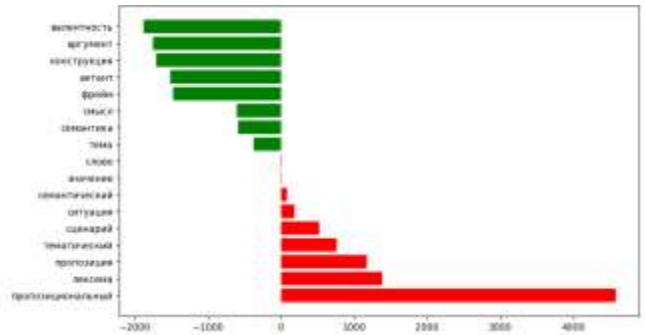

Рис. 9. Оценка суммарных значений Δ для рангов частот в группе терминов, характеризующих семантический анализ лингвистических данных

В группе терминов – названий языков (*белорусский, болгарский, бурятский, венгерский, вепсский, вьетнамский, английский, ижорский, итальянский, карельский, китайский, ливвиковский, людиковский, монгольский, немецкий, польский, румынский, русский, словацкий, украинский, финский, французский, церковнославянский, чешский, якутский, японский, банту, древнерусский*), см. Рис. 10, повышение частотности и снижение рангов зарегистрировано для терминов *церковнославянский, белорусский, итальянский, китайский, украинский, ижорский, древнерусский* , что объясняется ростом интереса к корпусам текстов на указанных языках.

Рис. 10. Оценка суммарных значений Δ для рангов частот в группе терминов – названий языков

Полученные данные указывают на динамические тенденции в тематике текстов корпуса ТКиКЛ.

VI. Заключение

В ходе экспериментов были применены расширенные возможности алгоритмов BERTopic и NMF в отношении настройки обучения тематических моделей с учетом когерентности тем, ранжирования тем по значимости с учетом весов слов-тематизаторов, хронологических меток документов, а также режим визуализации как всего набора тем, так и групп тем и отдельных тем по выбору пользователей.

Исследование, проведенное на материале корпуса ТКиКЛ, позволило сформулировать следующие выводы:

- динамическое моделирование, проведенное

средствами BERTopic в режиме «Topics over Time» позволяет отследить изменение фокуса внимания исследователей, работающих над узкими темами, например, в области диахронической корпусной лингвистики, а также зарегистрировать резкие скачки в тематике статей, связанные с социально-политическими процессами и технологическим прогрессом, прежде всего, с цифровизацией общественной жизни;

- динамическое тематическое моделирование на основе алгоритма NMF дает разнородные нишевые темы, допускающие интерпретацию в ходе статистического анализа списка слов-тематизаторов, присутствующих во всех хронологических срезах. Тенденции развития тем статей корпуса ТКиКЛ были изучены в ходе оценки суммарных значений Δ для рангов частот в группах терминов компьютерной и корпусной лингвистики и терминов – названий языков.

Благодарности

Автор выражает глубокую признательность доктору филологических наук, профессору кафедры иностранных языков и лингводидактики Иркутского государственного университета Светлане Юрьевне Богдановой за ценное обсуждение проблемы использования и изменений терминологии компьютерной и корпусной лингвистики в русском языке.

Библиография

[1] Диалог-21 «Компьютерная лингвистика и интеллектуальные технологии». URL: https://dialogue-conf.org (дата обращения: 11.11.2025).
[2] AINL «Artificial Intelligence and Natural Language». URL: https://ainlconf.ru/ (дата обращения: 08.07.2025).
[3] AIST «International Conference on Analysis of Images, Social Networks and Texts». URL: https://aistconf.org/ (дата обращения: 11.11.2025).
[4] SPECOM «International Conference on Speech and Computer». URL: https://specom.nw.ru/ (дата обращения: 11.11.2025).
[5] FRUCT. URL: https://fruct.org/ (дата обращения: 11.11.2025).
[6] Dimensions. URL: https://www.dimensions.ai/ (дата обращения: 11.11.2025).
[7] OpenAlex. URL: https://openalex.org/ (дата обращения: 11.11.2025).
[8] ResearchRabbit. URL: https://www.researchrabbit.ai/ (дата обращения: 11.11.2025).
[9] Митрофанова О. А., Адамова М. А., Букреева Л. А., Зернова А. К., Литвинова А. А., Павликова В. С., Сологуб П. С. Корпус текстов по корпусной лингвистике: состав и этапы формирования // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 13-29. DOI: 10.17586/2541-9781-2024-8-13-29.
[10] Митрофанова О. А., Голубев Р. В., Гусяцкая П. А., Макеев К. В., Плюснина Е. А., Сухан Д. Д., Трошина А. В., Уткина А. А. Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 30-44. DOI: 10.17586/2541-9781-2024-8-30-44.
[11] Сухан Д.Д., Плюснина Е.А. Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 45-60. DOI: 10.17586/2541-9781-2024-8-45-60.
[12] Митрофанова О.А., Адамова М. А., Букреева Л. А., Голубев Р. В., Гусяцкая П. А., Зернова А. К., Литвинова А. А., Макеев К. В., Павликова В. С., Плюснина Е. П., Сологуб П. С., Сухан Д. Д., Трошина А. В., Уткина А. А. Интеллектуальный анализ данных в корпусе текстов по корпусной и компьютерной лингвистике // International Journal of Open Information Technologies. 2024. Т. 12, № 12. С. 11-26.
[13] RuTermExtract. URL: https://pypi.org/project/rutermextract/ (дата обращения: 11.11.2025).
[14] sumy. URL: https://github.com/miso-belica/sumy (дата обращения: 11.11.2025).
[15] ruT5. URL: https://huggingface.co/ai-forever/ruT5-base (дата обращения: 11.11.2025).
[16] Воронцов К. В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf — 2025 (дата обращения: 11.11.2025).
[17] BERTopic. URL: https://github.com/MaartenGr/BERTopic (дата обращения: 11.11.2025).
[18] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint. URL: https://arxiv.org/abs/2203.05794/ — 2022 (дата обращения: 11.11.2025).
[19] Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina M. Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction // Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020. Vol. 12469. P. 134–152.
[20] Kuang D., Choo J., Park H. Nonnegative matrix factorization for interactive topic modeling and document clustering // Partitional clustering algorithms. 2015. P. 215–243.
[21] Scikit-Learn. URL: https://scikit-learn.org/ (дата обращения: 11.11.2025).

О.А.Митрованова, канд. филол. наук, доцент, Санкт-Петербургский государственный университет, Санкт-Петербург, Россия (e-mail: o.mitrofanova@spbu.ru).

Lexical Dynamics in the Text Corpus on Computational and Corpus Linguistics: Probabilistic and Statistical Approach

Olga A. Mitrofanova

***Anstract*—The aim of the study is to analyze changes in the topics of scientific articles within the text corpus on Computational and Corpus Linguistics using modern topic modeling algorithms. The experiments are based on the application of a combined neural network algorithm BERTopic and the matrix-based algorithm NM, enhanced by a frequency analysis of the use of topical words in the corpus. The significance of the work lies in the development of linguistically grounded methodology for semantic search of scientific information and tracking trends in the field of computational and corpus linguistics, which can be used to improve scientometric tools. The results of the study demonstrate the dynamics of changes in scientific interests in the field of computational and corpus linguistics under the influence of the society digitalization and technological progress.**

***Keywords*—computational linguistics, corpus linguistics, text corpus, dynamic topic modeling, BERTopic, NMF, quantitative and statistical analysis**

REFERENCES

[1] Dialogue-21, Computational Linguistics and Intellectual Technologies. Available: https://dialogue-conf.org.
[2] AINL Artificial Intelligence and Natural Language. Available: https://ainlconf.ru/.
[3] AIST «International Conference on Analysis of Images, Social Networks and Texts». Available: https://aistconf.org/.
[4] SPECOM «International Conference on Speech and Computer». Available: https://specom.nw.ru.
[5] FRUCT. Available: https://fruct.org/.
[6] Dimensions. Available: https://www.dimensions.ai/.
[7] OpenAlex. Available: https://openalex.org/.
[8] ResearchRabbit. Available: https://www.researchrabbit.ai/.
[9] O. A. Mitrofanova, M. A. Adamova, L. A. Bukreeva, A. K. Zemova, A. A. Litvinova, V. S. Pavlikova and P. S. Sologub, "Text Corpus on Corpus Linguistics: Composition and Stages of Formation," In *Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024*, St. Petersburg, June 24–26, 2024), St. Petersburg, ITMO University, pp. 13-29, 2024, doi: 10.17586/2541-9781-2024-8-13-29. (In Russian)
[10] O. A. Mitrofanova, R. V. Golubev, P. A. Gusyatskaya, K. V. Makeev, E. A. Pliusnina, D. D. Sukhan, A. V. Troshina and A. A. Utkina, "Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment," In *Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024*, St. Petersburg, June 24–26, 2024), St. Petersburg, ITMO University, pp. 30-44, 2024, doi: 10.17586/2541-9781-2024-830-44. (In Russian)
[11] D. D. Sukhan and E. A. Pliusnina, "Meta Tagging and Visualization for the Corpora Linguistics Texts Corpora," In *Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024*, St. Petersburg, June 24–26, 2024), St. Petersburg, ITMO University, pp. 45-60, 2024, doi: 10.17586/2541-9781-2024-8-45-60. (In Russian)
[12] O. A. Mitrofanova, M. A. Adamova, L. A. Bukreeva, R. V. Golubev, P. A. Gusyatskaya, A. K. Zemova, K. V. Makeev, A. A. Litvinova, V. S. Pavlikova, E. P. Plyusnina, P. S. Sologub, D. D. Sukhan, A. V. Troshina and A. A. Utkina, "Data Mining in the Text Corpus on Corpus and Computational Linguistics," *International Journal of Open Information Technologies*, Vol. 12, No. 12, pp. 11-26, 2024. (In Russian)
[13] RuTermExtract. Available: https://pypi.org/project/rutermextract/.
[14] sumy. Available: https://github.com/miso-belica/sumy.
[15] ruT5. Available: https://huggingface.co/ai-forever/ruT5-base.
[16] K. V. Vorontsov, "Probabilistic topic modeling: ARTM regularization theory and BigARTM open source library," 2025. Available: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf.
[17] BERTopic. Available: https://github.com/MaartenGr/BERTopic.
[18] M. Grootendorst, "BERTopic: Neural topic modeling with a class-based TF-IDF procedure," *arXiv preprint*, 2022. Available: https://arxiv.org/abs/2203.05794.
[19] T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova and M. Kirina, "Topic Modelling with NMF vs Expert Topic Annotation: The Case Study of Russian Fiction," In *Advances in Computational Intelligence: 19th Mexican International Conference on Artificial Intelligence, MICAI 2020*, Vol. 12469. pp. 134–152, 2020.
[20] D. Kuang, J. Choo and H. Park, "Nonnegative matrix factorization for interactive topic modeling and document clustering," *Partitional clustering algorithms*, pp. 215–243, 2015.
[21] Scikit-Learn. Available: https://scikit-learn.org/.

O. A. Mitrovanova, Candidate of Philology, Associate Professor, Saint Petersburg State University, Saint Petersburg, Russia (e-mail: o.mitrofanova@spbu.ru).