ТЕОРЕТИЧЕСКАЯ, ПРИКЛАДНАЯ И СРАВНИТЕЛЬНО-СОПОСТАВИТЕЛЬНАЯ ЛИНГВИСТИКА / THEORETICAL, APPLIED AND COMPARATIVE LINGUISTICS

DOI: https://doi.org/10.18454/RULB.2023.48.17

БИБЛИОМЕТРИЧЕСКИЙ АНАЛИЗ ТРЕНДОВЫХ ТЕМ В АНГЛОЯЗЫЧНОЙ НАУЧНОЙ ЛИТЕРАТУРЕ ПО ЛИНГВИСТИКЕ

Научная статья

Шарнин М.М.[1, *]

[1] ORCID : 0000-0003-0450-5156;

[1] Федеральный исследовательский центр «Информатика и управление» Российской академии наук, Москва, Российская Федерация

* Корреспондирующий автор (mc[at]keywen.com)

Аннотация

В работе представлен библиометрический анализ трендовых тем в англоязычной научной литературе по лингвистике. Анализ выполнен на основе коллекций научных статей DBLP и PubMed на начало 2023 года, которые содержат суммарно более 39 миллионов статей. В результате анализа выявлен значительный рост (более чем в 3 раза) количества англоязычных работ по лингвистике за 10 лет с 2012 по 2022 год. Также выявлены трендовые ключевые слова с прогнозируемым долгосрочным ростом трендов. Группы трендовых слов, часто встречающиеся вместе в заголовках статей, образуют трендовые темы. Рассмотрены три трендовые темы в лингвистике:

1) естественный язык, интеграция и принятие решений;
2) предубеждения и сложность;
3) образование и компетенции.

Приведены ключевые слова из каждой трендовой темы и рассмотрены соответствующие научные работы.

Ключевые слова: библиометрический анализ, долгосрочные прогнозы растущих трендов, трендовые ключевые слова, трендовые темы в лингвистике, англоязычная научная литература.

A BIBLIOMETRIC ANALYSIS OF TRENDING TOPICS IN THE ENGLISH LANGUAGE LINGUISTICS RESEARCH LITERATURE

Research article

Sharnin M.M.[1, *]

[1] ORCID : 0000-0003-0450-5156;

[1] Federal Research Center "Computer Science and Control" of the Russian Academy of Sciences, Moscow, Russian Federation

* Corresponding author (mc[at]keywen.com)

Abstract

The work presents a bibliometric analysis of trending topics in the English-language scientific literature in linguistics. The analysis is based on DBLP and PubMed collections of scientific articles as of the beginning of 2023, which contain more than 39 million articles in total. The analysis shows a significant increase (more than 3-fold) in the number of English-language papers on linguistics over the 10 years from 2012 to 2022. Trending keywords with predicted long-term trend growth were also identified. Groups of trending words that frequently occur together in article titles form trending topics. Three trending topics in linguistics are examined:

1) natural language, integration and decision-making;
2) biases and complexity;
3) education and competences.

Keywords from each trending topic are provided, and relevant research works are examined.

Keywords: bibliometric analysis, long-term forecasts of growing trends, trending keywords, trending topics in linguistics, English-language scientific literature.

Введение

В данной статье представлен библиометрический анализ трендовых тем в англоязычной научной литературе по лингвистике. Анализ выполнен на основе коллекций научных статей DBLP и PubMed на начало 2023 года, которые содержат суммарно более 39 миллионов статей и свободно представлены в Интернете. Библиометрический анализ использует прогноз трендов ключевых слов, выполненный по методике, описанной в работах [1], [2] с применением пакета машинного обучения CatBoost [3]. В результате библиометрического анализа выявлены трендовые ключевые слова, которые имеют долгосрочный прогноз роста их трендов, и построен рейтинг таких слов. Результаты прогноза были визуализированы с помощью алгоритма t-SNE, что позволило определить трендовые темы, содержащие кластеры трендовых слов.

Статья имеет следующую структуру. В разделе ниже описаны результаты библиометрического анализа, далее приводится обзор научной литературы по выявленным трендовым темам в лингвистике. В конце дается заключение.

Библиометрический анализ

Библиографический анализ был проведен на базе коллекций PubMed и DBLP-v13. Коллекция DBLP-v13 содержит 5,354,309 статей в области компьютерных наук, а коллекция PubMed по состоянию на начало 2023 года содержала более 34 миллионов статей по медицине, биологии и связанным наукам. Из этих коллекций было выделено 13640 заголовков научных статей, содержащих слова linguistic (лингвистический), linguistics (лингвистика) или linguistically (лингвистически). Эти заголовки статей были найдены по запросу Linguistic*, где звездочка (*) означает любую последовательность букв. Данные 13640 статей мы называем в дальнейшем Выборочной коллекцией. Выборочная коллекция безусловно связана с лингвистикой, т.к. все её статьи в заголовках содержат упоминание лингвистики. В заголовок обычно выносятся слова, имеющие существенное отношение к теме статьи.

В процессе анализа были рассчитаны графики роста количества статей за последние годы (см. рисунок 1), а также характерные/релевантные ключевые слова в Выборочной коллекции и трендовые ключевые слова среди характерных.

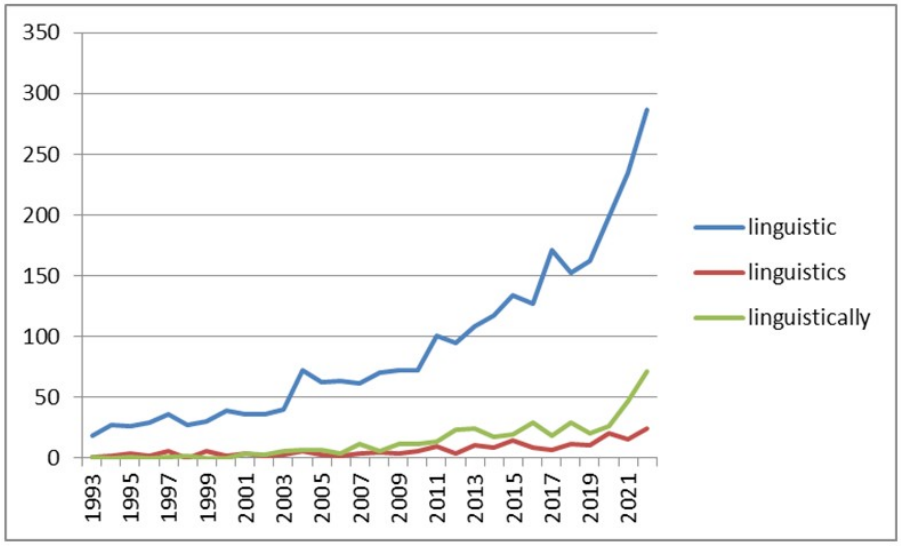

Рисунок 1 - Графики количества научных статей со словами linguistic (лингвистический), linguistics (лингвистика) и linguistically (лингвистически) в коллекциях DBLP и PubMed за различные годы

DOI: https://doi.org/10.18454/RULB.2023.48.17.1

Из рисунка 1 видно, что наблюдается значительный рост количества ежегодно публикуемых научных статей со словами linguistic (лингвистический), linguistics (лингвистика) и linguistically (лингвистически) в коллекциях DBLP и PubMed за последние годы. Так, за последние 10 лет с 2012 по 2022 количество таких статей выросло в 3.13 раза, за последние 20 лет – в 9.31 раз, а за последние 30 лет – в 17.36 раз.

Характерными/релевантными мы считаем слова, которых относительно много в Выборочной коллекции по сравнению с коллекцией PubMed. Список/рейтинг наиболее характерных ключевых слов в Выборочной коллекции: linguistic (лингвистический), linguistics (лингвистика), linguistically (лингвистически), cross-linguistic (межлингвистический), culturally linguistically (культурно-лингвистический), linguistically diverse (лингвистически разнообразный), fuzzy linguistic (нечеткая лингвистика), computational linguistics (компьютерная лингвистика), linguistic analysis (лингвистический анализ), linguistic validation (лингвистическая проверка), linguistic information (лингвистическая информация), culturally (культурно), linguistic features (лингвистические особенности), linguistic knowledge (лингвистические знания), non-linguistic (нелингвистический), hesitant (нерешительный), linguistic term (лингвистический термин), linguistic approach (лингвистический подход), linguistic preference (лингвистические предпочтения), cultural linguistic (культурно-лингвистический), probabilistic linguistic (вероятностный лингвистический), linguistic processing (лингвистическая обработка), uncertain linguistic (неопределенный лингвистический), hesitant fuzzy (колеблющийся нечеткий), linguistic resources (лингвистические ресурсы), term sets (наборы терминов), linguistic structure (лингвистическая структура), group decision (групповое решение), linguistic data (лингвистические данные), cognitive linguistic (когнитивный лингвистические), linguistic variables (лингвистические переменные), cross-linguistic study (кросс-лингвистическое исследование), linguistic fuzzy (лингвистический нечеткий анализ), linguistic summaries (лингвистические резюме), linguistic cultural (лингвокультура), decision making (принятие решений), multiple attribute (множественный атрибут), preference relations (отношения предпочтений), linguistic diversity (языковое разнообразие), linguistic environment (языковая среда), linguistic complexity (лингвистическая сложность), group decision-making (групповое принятие решений) и т.д.

Долгосрочные прогнозы трендов для характерных ключевых слов были рассчитаны, используя различные показатели для групп статей, содержащих эти ключевые слова. Наиболее важным индикатором тренда ключевого слова является количество цитирований ключевого слова (КЦКС). Для расчета этого показателя мы сначала находим все статьи с этим ключевым словом в определенном году, а затем подсчитываем все цитирующие ссылки на эти статьи. Для каждого слова мы рассчитали продолжительность его трендового роста, равную количеству лет непрерывного роста его средней цитируемости. Такая продолжительность роста тренда была целью для алгоритма машинного

обучения CatBoost [3]. Регрессионная модель CatBoost была обучена для 20 параметров тренда ключевого слова в текущем и предыдущих годах, включая: текущий КЦКС, общее количество статей с ключевым словом, количество лет роста предыдущего тренда, общий рост цитируемости для предыдущего тренда, количество лет от начала тренда, количество цитирующих связей между статьями со словом, время от трендовой ситуации до текущего года. Более подробно данная методика описана в работах [1], [2].

С помощью алгоритма машинного обучения CatBoost были рассчитаны долгосрочные прогнозы растущих трендов для характерных слов и построен следующий рейтинг трендовых ключевых слов в Выборочной коллекции: deficit (дефицит), competence (компетентность), preschool (дошкольный возраст), bias (предвзятость), integrating (интеграция), aggregation (агрегирование), patient-reported (по сообщениям пациентов), decision making (принятие решений), retrieval (извлечение), ***, annotation (аннотация), predicts (прогнозирует), arabic (арабский язык), autism (аутизм), complexity (сложность), genetic diversity (генетическое разнообразие), online (онлайн), fuzzy (нечеткий), questionnaire (анкета), foreign language (иностранный язык), context (контекст), dyslexia (дислексия), listening (аудирование), natural language (естественный язык), verbal (вербальный), *****, autism spectrum (спектр аутизма), culturally (культурно), emotional (эмоциональный), overactive (гиперактивный), brainstem (ствол мозга), individual differences (индивидуальные различия), linguistic (лингвистический), dialogue (диалог), speech therapy (логопед), parsing (синтаксический анализ), language learning (изучение языка), lateralization (латерализация), working memory (рабочая память), linguistics (лингвистика), linguistically (лингвистически), group (групповой), fuzzy linguistic (нечеткой лингвистики), computational linguistics (компьютерной лингвистики), cross-linguistic (кросслингвистический) и т.д. В начале данного рейтинга расположены ключевые слова с наиболее долгосрочным прогнозом роста трендов. Рейтинг разделен на три группы с помощью меток *** и *****, которые стоят в конце первой и второй группы. Первая группа содержит 9 терминов, вторая – 15, а третья – 19. Деление на группы условно и определяется наглядностью визуализации. Термины из первой группы отображаются на рисунке 2 красным цветом, из второй – синим, из третьей – черным.

Далее рассчитали совместную встречаемость характерных ключевых слов/терминов в заголовках статей в Выборочной коллекции, а также их меру близости – чем чаще термины встречаются вместе, тем меньше семантическое расстояние. По этим данным была построена семантическая карта с помощью алгоритма t-SNE (см. рисунок 2).

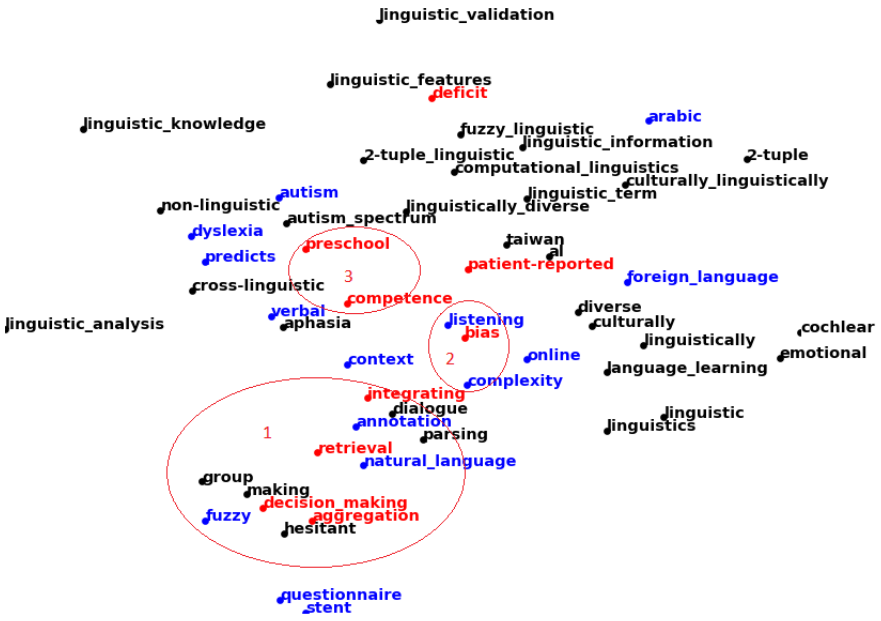

Рисунок 2 - Семантическая карта терминов из предметной области «Лингвистика»:
красным цветом выделены наиболее перспективные/трендовые термины, имеющие долгосрочные тренды; *синим* – средние тренды; *черным* – минимальные тренды
DOI: https://doi.org/10.18454/RULB.2023.48.17.2

Примечание: чем чаще термины встречаются вместе в заголовках статей, тем ближе они расположены на рисунке; группы рядом расположенных трендовых терминов образуют трендовые темы

На рисунке 2 красным и синим цветами представлены трендовые термины. Красным цветом выделены термины с наиболее долгосрочным прогнозом роста трендов. Также видны кластеры этих терминов, например, (integrating, decision making, retrieval, aggregation), (bias, complexity) и (preschool, competence). Этим кластерам соответствуют трендовые темы, т.к. одинаковые/близкие прогнозы рядом встречающихся слов подтверждают друг-друга. Трендовые темы определяют в Выборочной коллекции группы статей, в заголовках которых присутствуют трендовые термины.

Такие группы статей будут со временем расти в плане количества входящих статей и количества ссылок цитирования. Таким образом, трендовая тема в лингвистике – это группа трендовых терминов с близкими по долгосрочности трендами, которые часто встречаются вместе в заголовках статей, связанных с лингвистикой.

Итак, на рисунке 2 видны три трендовые темы в лингвистике. Этим трендовым темам можно дать следующие названия:

1) естественный язык, интеграция и принятие решений;
2) предубеждения и сложность;
3) образование и компетенции.

В следующем разделе опишем более подробно эти трендовые темы.

Обзор трендовых тем

Рассмотрим три трендовых темы в лингвистике, выявленные в процессе библиометрического анализа.

К первой теме «естественный язык, интеграция и принятие решений» относятся статьи из области компьютерной лингвистики и обработки естественного языка. Эта тема имеет следующие ключевые слова, выделенные красным цветом на рисунке 2: интеграция (integrating), принятие решений (decision making), поиск (retrieval), агрегирование (aggregation).

Рассмотрим некоторые статьи из этой темы. В статье [4] рассматривается интеграция нескольких типов неполных отношений лингвистических предпочтений в процессе принятия решений несколькими людьми. Работа [5] посвящена методам лингвистической агрегации при поиске по блогам. В статье [6] рассматривается метод на основе кластеризации для принятия решений в больших группах с неуверенной нечеткой лингвистической информацией. Метод разделяет экспертов на несколько кластеров. Использование кластеризации одновременно обеспечивает сплоченность кластеров и постепенное повышение уровня коллективного консенсуса при принятии решений.

Ко второй трендовой теме «предубеждения и сложность» относятся статьи, в которых рассматриваются вопросы, связанные с построением лингвистических моделей, а также с лингвистическими предубеждениями и лингвистической сложностью. Эта тема имеет следующие ключевые слова, выделенные красным и синим цветами: предвзятость (bias), сложность (complexity). Термин «complexity» изображен на рисунке 2 синим цветом, но он также является трендовым и к тому же находится ближе к началу второй группы в рейтинге трендовых ключевых слов.

Рассмотрим некоторые статьи из этой трендовой темы. В статье [7] обсуждаются лингвистические модели для анализа и обнаружения предвзятой речи. Работа [8] посвящена методам обнаружения предвзятости посредством анализа настроений и простых лингвистических функций. В статье [9] рассматривается влияние алгоритмической предвзятости на лингвистическую сложность машинного перевода. Авторы предполагают, что «алгоритмическая погрешность», т.е. обострение часто наблюдаемых закономерностей в сочетании с утратой менее частых не только усугубляет социальные предубеждения, присутствующие в текущих наборах обучающих текстовых данных, но также приводит к искусственно обеднённому языку машинного перевода. Авторы оценивали лингвистическое богатство (на лексическом и морфологическом уровне) переводов, созданных с помощью различных парадигм машинного перевода и показали, что происходит потеря лексического и морфологического богатства в переводах, произведенных всеми исследованными парадигмами машинного перевода для двух языковых пар (EN ↔ FR и EN ↔ ES).

Третья трендовая тема «образование и компетенции» содержит статьи, в которых, в частности, рассматриваются вопросы, связанные с анализом лингвистических компетенций учащихся и оценки профессионального развития компетентности у преподавателей. Эта тема имеет следующие ключевые слова: дошкольный возраст (preschool), преподаватели (teachers), языковые компетенции (linguistic competence).

Рассмотрим некоторые статьи, относящиеся к этой трендовой теме. В статье [10] исследуется взаимодействие зрительной информации, вербальной информации и языковой компетенции у ребенка дошкольного возраста. Работа [11] посвящена методам принятия решений по множественным признакам с неуверенной нечеткой неопределенной лингвистической информацией и их применению для оценки профессионального развития компетентности у преподавателей английского языка колледжей.

Заключение

В последние десятилетия наблюдается значительный рост количества англоязычных работ по лингвистике. Так, за 10 лет с 2012 по 2022 количество таких работ в многомиллионных коллекциях PubMed и DBLP выросло более чем в три раза. Стремительно развиваются такие направления лингвистики, как компьютерная лингвистика, нечеткая лингвистика, построение лингвистических моделей, анализ лингвистических предубеждений, методы обучения иностранным языкам и другие. В результате прогнозного библиометрического анализа выявлены следующие три трендовых направления в англоязычной научной литературе по лингвистике:

1) естественный язык, интеграция и принятие решений;
2) предубеждения и сложность;
3) образование и компетенции.

Конфликт интересов

Не указан.

Рецензия

Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Conflict of Interest

None declared.

Review

All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы на английском языке / References in English

1. Charnine M. Research Trending Topic Prediction as Cognitive Enhancement / M. Charnine, A. Klokov, L. Kochiev et al. — Caen, 2021. — p. 217-220. — DOI: 10.1109/CW52790.2021.00044.

2. Charnine M. Visualization of Research Trending Topic Prediction: Intelligent Method for Data Analysis / M. Charnine, A. Tishchenko, L. Kochiev // CEUR Workshop Proceedings. — 2021. — Vol. 3027. — p. 1028–1037.

3. Prokhorenkova L. CatBoost: Unbiased Boosting with Categorical Features / L. Prokhorenkova, G. Gusev, A. Vorobev et al. // Advances in Neural Information Processing Systems. — 2018. — Vol. 31.

4. Xu Z. Integrating Multiple Types of Incomplete Linguistic Preference Relations in Multi-person Decision Making / Z. Xu. — Xi'an: Springer Berlin Heidelberg, 2006. — p. 300-309.

5. Keikha M. Linguistic Aggregation Methods in Blog Retrieval / M. Keikha, F. Crestani // Information Processing & Management. — 2012. — Vol. 48. — №. 3. — p. 467-475.

6. Zhong X. Clustering-based Method for Large Group Decision Making with Hesitant Fuzzy Linguistic Information: Integrating Correlation and Consensus / X. Zhong, X. Xu // Applied Soft Computing. — 2020. — Vol. 87. — p. 105973.

7. Recasens M. Linguistic Models for Analyzing and Detecting Biased Language / M. Recasens, C. Danescu-Niculescu-Mizil, D. Jurafsky // Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. — 2013. — Vol. 1. — p. 1650-1659.

8. Anthonio T. Team Kermit-the-frog at SemEval-2019 Task 4: Bias Detection Through Sentiment Analysis and Simple Linguistic Features / T. Anthonio, L. Kloppenburg // Proceedings of the 13th International Workshop on Semantic Evaluation. — 2019. — p. 1016-1020.

9. Vanmassenhove E. Machine Translationese: Effects of Algorithmic Bias on Linguistic Complexity in Machine Translation / E. Vanmassenhove, M. Gwilliam // arXiv preprint. — 2021.

10. Perry F.L. Interaction of Visual Information, Verbal Information, and Linguistic Competence in the Preschool-aged Child / F.L. Perry, A. Shwedel // Journal of Psycholinguistic Research. — 1979. — Vol. 8. — p. 559-566.

11. Zheng X.M. Methods for Multiple Attribute Decision Making with Hesitant Fuzzy Uncertain Linguistic Information and Their Application for Evaluating the College English Teachers' Professional Development Competence / X.M. Zheng // Journal of Intelligent & Fuzzy Systems. — 2015. — Vol. 28. — №. 3. — p. 1243-1250.