МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ISSN 2782-5450

Terra Linguistica

Том 14, № 1, 2023

Инженерно-лингвистические технологии в исследованиях текста

Санкт-Петербургский политехнический
университет Петра Великого
2023

TERRA LINGUISTICA

Сетевое издание публикует научно-исследовательские статьи и рецензии на русском и английском языках в области лингвистических исследований.
С 2002 года входит в Перечень ведущих рецензируемых научных журналов и изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени доктора и кандидата наук.
Сетевое издание зарегистрировано Федеральной службой по надзору в сфере информационных технологий и массовых коммуникаций (Роскомнадзор). Свидетельство о регистрации ЭЛ № ФС77-77377 от 25 декабря 2019 г.
Сведения о публикациях представлены в Реферативном журнале ВИНИТИ РАН, в международной справочной системе «Ulrich`s Periodical Directory», в Российской государственной библиотеке. В базах данных: Российский индекс научного цитирования (РИНЦ), Google Scholar, CNKI, ProQuest, Index Copernicus, КиберЛенинка.
Учредитель и издатель: Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация.

Редакция журнала

д-р филол. наук, профессор В.Е. Чернявская – главный редактор;

Г.А. Пышкина – ответственный секретарь, выпускающий редактор;

А.А. Кононова – компьютерная вёрстка; Д.Ю. Алексеева – перевод на английский язык.

Адрес редакции: Россия, 195251, Санкт-Петербург, ул. Политехническая, д. 29.

Тел. редакции: +7 (812) 552-62-16, e-mail: ntv-human@spbstu.ru

Дата выхода: 31.03.2023

THE MINISTRY OF SCIENCE AND HIGHER EDUCATION OF THE RUSSIAN FEDERATION

ISSN 2782-5450

Terra Linguistica

Vol. 14, No. 1, 2023

Language engineering technologies in text studies

Peter the Great St. Petersburg
Polytechnic University
2023

TERRA LINGUISTICA

The open access journal publishes research papers and reviews on theoretical orientations, and methodological approaches that have a central focus on language in the perspective of theoretical and applied linguistics, linguistic pragmatics, sociolinguistics, linguistic anthropology, discourse analysis, translation studies.

The journal is included in the List of Leading PeerReviewed Scientific Journals and other editions to publish major findings of PhD theses for the research degrees of Doctor of Sciences and Candidate of Sciences.

The journal is indexed by Ulrich's Periodicals Directory, Google Scholar, CNKI, ProQuest, Index Copernicus, VINITI RAS Abstract Journal (Referativnyi Zhurnal), VINITI RAS Scientific and Technical Literature Collection, Russian Science Citation Index (RSCI) database Scientific Electronic Library.

The journal is registered with the Federal Service for Supervision in the Sphere of Telecom, Information Technologies and Mass Communications (ROSKOMNADZOR). Certificate ЭЛ No. ФС77-77377 issued 25.12.2019.

Editorial office

Dr. Sc., Professor V.E. Chernyavskaya – Editor-in-Chief;

G.A. Pyshkina – editorial manager;

A.A. Kononova – computer layout; D.Yu. Alekseeva – English translation.

Address: 195251 Polytekhnicheskaya Str. 29, St. Petersburg, Russia.

+7 (812) 552-62-16, e-mail: ntv-human@spbstu.ru

Release date: 31.03.2023

Содержание

Материалы конференции «Пиотровские Чтения – 2022»

Contents

Conference materials "R. Piotrowski's readings – 2022"

Terra Linguistica. 2023, Том 14, № 1. С. 7–10.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 7–10.

Редакторская заметка
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.14101

ИНЖЕНЕРНЫЕ ЛИНГВИСТИЧЕСКИЕ ТЕХНОЛОГИИ В ИССЛЕДОВАНИИ ТЕКСТА

А.В. Колмогорова ✉

Национальный исследовательский университет «Высшая школа экономики», Санкт-Петербург, Российская Федерация

✉ nastiakol@mail.ru

Аннотация. Публикация посвящена анализу современного состояния инженерной лингвистики, ее основных направлений и исследовательских вызовов. Формулируется определение языковых технологий и их типология по критерию решаемых с их помощью задач. Отмечается, что отечественной школе инженерной лингвистики удается сохранить баланс между технологичностью и лингвистичностью изысканий.

Ключевые слова: языковые технологии, инженерная лингвистика, компьютерная лингвистика, языковые модели.

Для цитирования: Колмогорова А.В. Инженерные лингвистические технологии в исследовании текста // Terra Linguistica. 2023. Т. 14. № 1. С. 7–10. DOI: 10.18721/JHSS.14101

Editorial note
DOI: https://doi.org/10.18721/JHSS.14101

ENGINEERING LINGUISTIC TECHNOLOGIES IN TEXT STUDIES

A.V. Kolmogorova

National Research University Higher School of Economics,
St. Petersburg, Russian Federation

nastiakol@mail.ru

Abstract. The publication is devoted to the analysis of the current state of engineering linguistics, its main directions and research challenges. The definition of language technologies and their typology are formulated according to the criterion of the tasks solved with their help. It is noted that the national school of engineering linguistics manages to maintain a balance between technological and linguistic research.

Keywords: linguistic technologies, engineering linguistics, computational linguistics, linguistics models.

Citation: A.V. Kolmogorova, Engineering linguistic technologies in text studies, Terra Linguistica, 14 (1) (2023) 7–10. DOI: 10.18721/JHSS.14101

Инженерные лингвистические технологии в исследовании текста: основные направления и результаты

Цифровой поворот во всех сферах жизни общества стал стимулом для активного роста исследований в области компьютерной лингвистики, которая имеет в отечественной исследовательской практике давние традиции.

60-е годы XX в. были отмечены бурным развитием первых компьютерных технологий и появлением в некотором смысле «идеологии технологичности»: роль искусственного интеллекта стала активно осмысляться обществом. Устоявшиеся правила нормативной науки, будучи перенесены в новый контекст, требовали серьезных изменений. Лингвистика в этом смысле не стала исключением.

В эти годы преимущественно на Западе возникает направление компьютерной (вычислительной) лингвистики, а в СССР Раймондом Генриховичем Пиотровским разрабатывается концепция инженерной лингвистики, в фокусе которой оказываются «проблемы применения вероятностного моделирования для автоматической переработки текста» и разработки необходимых для этого лингвистических технологий [1].

Сегодня мы можем наблюдать, как вычислительные модели с успехом воплощаются в жизнь в виде языковых (лингвистических) технологий, позволяющих менять суть и свойства коммуникативных, информационных, этических, когнитивных и даже мировоззренческих процессов и категорий в нашей повседневной жизни.

В качестве одного из возможных определений лингвистических технологий можно было бы предложить следующее: это инженерные решения, имеющие воспроизводимый алгоритм применения, базирующиеся на вычислительных операциях обработки данных на естественном языке и позволяющие моделировать и воспроизводить широкий диапазон процессов, аналогичных тем, которые запускаются человеческим интеллектом в момент обработки сообщения на естественном языке.

По критерию решаемой задачи сегодняшние языковые технологии формируют несколько значимых кластеров:

1. Технологии автоматической обработки текста, которые делают текстовую ткань лингвистически транспарентной и доступной для дальнейшего машинного анализа: инструменты морфологического и синтаксического анализа и др.

2. Технологии и модели, моделирующие знаниевую компоненту текста: тезаурусы, онтологии, инструменты извлечения сущностей, тем и отношений.

3. Технологии, которые распознают и моделируют оценочную и эмоциональную составляющую текста (осуществляют аффективные вычисления): инструменты сентимент-анализа и эмоционального анализа.

4. Технологии, моделирующие процессы извлечения смысла из текстовых данных с возможностью его дальнейшей обработки, перекодирования и модификации: модели понимания и интерпретации естественного языка, инструменты парафразирования, симплификации и саммаризации, алгоритмы машинного перевода.

5. Технологии, моделирующие коммуникативное поведение человека, включающее не только вербальную составляющую, но и жестовую, мимическую и прагматическую: инструменты, позволяющие разрабатывать модели когнитивных ассистентов, чат-ботов, виртуальных дополненных личностей.

Приведенный список заведомо неполон – новые технологические решения постоянно пополняют имеющийся в распоряжении компьютерных и математических лингвистов арсенал.

Однако магия техники не должна мешать балансу делегирования – четкому пониманию соотношения той работы, которая выполняется технологическими средствами, и операций, которые остаются за человеком [2]. В нашем случае – за профессиональным лингвистом, который понимает, что именно находится «под капотом» у той или иной модели, алгоритма; как это работает; какие ограничения и искажения может давать применяемое технологическое решение; какая комбинация методов покажет наибольшую эффективность и, самое главное, как интерпретировать на языковедческом уровне те результаты, которые мы получаем «на выходе», или те сбои, которые мы наблюдаем, проводя вычислительные эксперименты.

Иными словами, между компьютерным лингвистом и его технологическим инструментом выстраивается своеобразный эвристический диалог о языке [3]: интерпретация поведения вычислительной модели, обрабатывающей языковые данные (например, в ходе машинного или глубинного обучения), результатов применения математических методов к текстовому материалу дает лингвисту непрямые, но значимые ответы на вопросы о внутреннем устройстве языка, как это представлено, например, в публикации [4], функционировании идиолекта [5], языковых средств реализации художественного целого [6].

Подобный сочинительный принцип, соединяющий инженерию и теорию языка, был тонко подмечен еще в работах Р. Г. Пиотровского [7]. Ясно прослеживается он и в исследованиях, результаты которых представлены в тематическом выпуске журнала «*Инженерные лингвистические технологии в исследовании текста*». Содержание номера, совмещающее в себе гносеологическую эталонность и творческий подход [8], позволит всем интересующимся лингвистикой и «доказательной лингвистикой» читателям познакомиться с современными достижениями инженерной лингвистики в России, даст импульс к новым экспериментам и эвристикам на пути познания языка.

СПИСОК ИСТОЧНИКОВ

1. **Беляева Л.Н., Богданов С.И., Горностай Т.** Инженерная лингвистика в контексте современной «Информации 4.0» // Proceedings. Сер. "CEUR-WS Workshop Proceedings" Том 2233. St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Herzen State Pedagogical University of Russia. 2018. С. 48–56.

2. **Вахштайн В.** Техника. СПб.: Издательство Европейского университета в Санкт-Петербурге, 2021. 156 с.

3. **Колмогорова А.В., Калинин А.А., Маликова А.В., Вдовина Л.А.** Методы компьютерной и корпусной лингвистики для решения задач эмоционального анализа интернет-текстов, М.: IPR Media, 2021.

4. **Пиотровский Р. Г.** Инженерная лингвистика и теория языка. Л.: Наука, 1979. 111 с.

5. **Митрофанова О.А., Гаврилик Д.А.** Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Т. 13. № 4. С. 22–40. DOI: 10.18721/JHSS.13402

6. **Андреев В.С.** Экспоненциальное распределение частей речи в стихотворном тексте: опыт стилеметрического анализа // Общество. Коммуникация. Образование. 2021. Т. 12. № 4. С. 94–104. DOI: 10.18721/JHSS.12407

7. **Се Линьи, Загайнов А.И.** Моделирование характеристик персонажей и их взаимосвязей в сюжете художественного произведения методами численного фрактального анализа // Terra Linguistica. 2022. Т. 13. № 3. С. 36–47. DOI: 10.18721/JHSS.13304

8. **Беляева Л.Н., Чернявская В.Е.** Доказательная лингвистика: метод в когнитивной парадигме // Вопросы когнитивной лингвистики. 2016. № 3 (48). С. 77–84. DOI: 10.20916/1812-3228-2016-3-77-84

REFERENCES

[1] **L. Beliaeva, S. Bogdanov, T. Gornostay,** Engineering in the Framework of Modern "Information 4.0", Proceedings. Cep. "CEUR-WS Workshop Proceedings" Vol. 2233. St. Petersburg Institute for Informatics and Automation of the Russian Academy of Sciences, Herzen State Pedagogical University of Russia. 2018. Pp. 48–56.

[2] **V. Vakhshtayn,** Technics, Saint-Petersburg, Editions of European University in Saint-Petersburg, 2021.

[3] **A.V. Kolmogorova, A.A. Kalinin, A.V. Malikova, L.A. Vdovina,** Methods of computational and corpus linguistics in tasks of emotional analysis of the internet-texts, M.: IPR Media, 2021.

[4] **R.G. Piotrovskiy,** Inzhenernaya lingvistika i teoriya yazyka [Engineering linguistics and theory of language]. L.: Nauka, 1979. 111 p.

[5] **O.A. Mitrofanova, D.A. Gavrilic,** Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts, Terra Linguistica, 13 (4) (2022) 22–40. DOI: 10.18721/JHSS.13402

[6] **V.S. Andreev,** Exponential distribution of parts of speech in verse text: experience in stylometric analysis, Society. Communication. Education, 12 (4) (2021) 94–104. DOI: 10.18721/JHSS.12407

[7] **Xie Linyi, A.I. Zagaynov,** Modeling of character characteristics and their relationships in a novel plot by methods of numerical fractal analysis, Terra Linguistica, 13 (3) (2022) 36–47. DOI: 10.18721/JHSS.13304

[8] **L.N. Belyayeva, V.Ye. Chernyavskaya,** Evidence-based linguistics: Methods in cognitive paradigm, Issues of Cognitive Linguistics. 3 (48) (2016) 77–84. DOI: 10.20916/1812-3228-2016-3-77-84

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Колмогорова Анастасия Владимировна
Anastasia V. Kolmogorova
E-mail: nastiakol@mail.ru
ORCID: https://orcid.org/0000-0002-6425-2050

Поступила: 14.02.2023; Одобрена: 17.03.2023; Принята: 17.03.2023.
Submitted: 14.02.2023; Approved: 17.03.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 11–20.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 11–20.

Научная статья
УДК 811'32
DOI: https://doi.org/10.18721/JHSS.14102

ЭВОЛЮЦИЯ ОБРАЗНОЙ СИСТЕМЫ ВЛАДИМИРА НАБОКОВА: КВАНТИТАТИВНЫЙ АНАЛИЗ

В.С. Андреев ✉

Смоленский государственный университет,
г. Смоленск, Российская Федерация

✉ vadim.andreev@ymail.com

Аннотация. В работе рассматривается эволюция важного аспекта индивидуального стиля Владимира Набокова – образной системы. Набоков более известен как прозаик, однако он начинал писать именно как поэт и продолжал писать стихи на протяжении всей жизни. Используемый традиционный в когнитивной лингвистике подход к пониманию образа как трансфера концептуальных свойств позволяет вести количественный анализ используемых автором концептов и сопоставить их частотность в текстах созданных в раннем и зрелом периодах творчества. В качестве статистического метода для сопоставления стиля раннего и зрелого поэта применяется многомерный дискриминантный анализ, который позволил сопоставить группы текстов одновременно по большому количеству переменных (частотности различных концептов-источников). Полученные результаты свидетельствуют, что имеют место значительные изменения в концептосфере автора, а, следовательно, в его мировоззрении. Полученная признаковая модель, включающая восемь маркирующих изменения стиля концептов, позволяет со стопроцентной точностью атрибутировать произведение к соответствующему периоду. Содержательный анализ изменений частотности концептов показывает сложный полифонический характер изменений, в котором сочетаются переход от сложных концепций к более простым, с одной стороны, и усложнение понимания живого существа, с другой.

Ключевые слова: стилехронометрия, индивидуальный стиль, образная система, дискриминантный анализ, Набоков.

Финансирование: Грант СПбГУ «Литературные тексты и их язык vs количественные, корпусные и компьютерные методы: взаимное тестирование (Набоков и сопоставительный материал)» проект № 92565342.

Для цитирования: Андреев В.С. Эволюция образной системы Владимира Набокова: квантитативный анализ // Terra Linguistica. 2023. Т. 14. № 1. С. 11–20. DOI: 10.18721/JHSS.14102

Research article
DOI: https://doi.org/10.18721/JHSS.14102

EVOLUTION OF VLADIMIR NABOKOV'S IMAGE SYSTEM: QUANTITATIVE ANALYSIS

V.S. Andreev ✉

Smolensk State University,
Smolesk, Russian Federation

✉ vadim.andreev@ymail.com

Abstract. The article deals with the evolution of such important aspect of Vladimir Nabokov's individual style as image system. Navokov is much better known as prose writer. However, he began his creative career as a poet and continued to write verse all his life. The utilized traditional approach to the definition of image as a transfer of conceptual characteristics makes it possible to carry out quantitative analysis of the concepts used by the author and to compare their quantity in early and mature periods of creative activity. Multivariate discriminant analysis is used as a statistical method to differentiate between the periods simultaneously on the basis of a larger number of variables (frequencies of concepts in the function of source domain). The obtained results demonstrate that there are significant changes in the individual style of the poet, and, consequently, in his worldview. The obtained discriminant model, which includes eight characteristics (concepts marking style alteration), makes it possible to automatically attribute the text to the right period in 100% of cases. Qualitative analysis of changes in the frequencies of concepts reveals a complex polyphonic character of style alteration, which includes both the transition from complicated to simpler concepts and a change to more complex understanding of living beings.

Keywords: stylochronometry, individual style, image system, discriminant analysis, Nabokov.

Acknowledgements: Grant of St. Petersburg State University "Literary texts and their language vs quantitative, corpus and computer methods: mutual testing (Nabokov and comparative material)" project No. 92565342.

Citation: V.S. Andreev, Evolution of Vladimir Nabokov's image system: quantitative analysis, Terra Linguistica, 14 (1) (2023) 11–20. DOI: 10.18721/JHSS.14102

Введение и постановка проблемы

В настоящее время в филологии, как и в других гуманитарных науках, усиливается внимание к количественным методам анализа. Широкое использование мер, индексов и более сложных методик стало необходимым из-за накопленного огромного массива наблюдений: объем уже имеющихся результатов таков, что невозможно выделить скрытые тенденции и закономерности без привлечения статистических методов и информационных технологий. Примером успеха квантитативных методов может стать бурное развитие стилеметрии, изначально ориентированной на точные (количественные) методы анализа [1–4]. В рамках стилеметрии уже разработаны эффективные методики, позволяющие решать практические задачи по определению авторства, гендера автора, количества авторов текста и др. [5–11]. Показателем успешности применения точных методов анализа стали не только практические успехи, но также то, что в рамках стилеметрии началось выделение новых направлений. Так, можно с уверенностью говорить о появлении стилехронометрии как отдельного направления исследований [12, 13], ставящего такие задачи как периодизация творчества авторов, датировка произведений и изучение эволюции стиля. Последняя из перечисленных задач решается в предлагаемом исследовании.

В данной работе ставится цель оценить изменения, произошедшие в стиле всемирно известного писателя Владимира Набокова на материале его стихотворных текстов. Набоков более известен как прозаик, однако он начинал писать именно как поэт, продолжал писать стихи и считал

себя поэтом (на что сам неоднократно указывал) на протяжении всей жизни. Поэзия, таким образом, была существенной и неотъемлимой частью его творчества. Более сложная система взаимозависимостей в стихотворном тексте, где элементы связаны не только синтаксически, но и в вертикальном плане в силу соотносимости строк в целом и соответствующих ритмических позиций, в частности [14], уменьшает вероятность случайного выбора автором языковых и речевых элементов. Все это делает анализ индивидуального стиля Набокова на материале стихов более сложной, но необходимой и интересной задачей.

Если в стилеметрии основное внимание уделяется формальным признакам, то мы рассматриваем эволюцию стиля с точки зрения изменения образной системы. Опыт исследований показывает, что анализ образов позволяет не только констатировать изменения, но и сделать содержательные выводы о видении мира автором и изменениях в его мировоззрении [15, 16].

Материал

В качестве материала исследования нами привлекаются все стихи Набокова длиной не менее 30 строк из сборников Горний путь (Сборник-1), принадлежащий раннему периоду, и Стихотворения (1929–1951) (Сборник 2), предствляющий зрелый период [17]. Из второго сборника взяты произведения, написанные позже 1939. Включение текстов достаточного объема позволяет обеспечить сопоставимость материала и избежать случайных флуктуаций при использовании поэтом образов.

Нами привлекались следующие произведения (в скобках даны краткие обозначения, используемые далее)

Детство (Т1);
«Звени, мой верный стих, витай, воспоминанье...» (Т2);
Крым (Т3);
Лес (Т4);
Сон на акрополе (Т5);
M. W. (Т6);
Два корабля (Т7);
Лестница (Т8);
«Феина дочь утонула в росинке...» (Т9);
Поэты (Т10);
Слава (Т11);
Парижская поэма (Т12).

Тексты с первого по девятый принадлежат первому периоду творчества, с десятого по одиннадцатый – зрелому периоду.

Методы исследования

Отдельно следует остановиться на методике выделения и количественного анализа образов. При всем многообразии подходов к определению образа можно выделить три основных направления. В рамках первого из них образ это ассоциации, возникающие в сознании человека, который воспринимает произведение. Такой подход является плодотворным в психологии, но в филологии не позволяет собрать объективные данные в силу сильных различий в восприятии (даже у подготовленных исследователей), вызванных различным жизненным опытом. Второй подход основывается на понимании образа какого-либо объекта как совокупности характеристик этого объекта, содержащихся в тексте. Этот подход активно и плодотворно используется в литературоведении (образ женщины в стихах Некрасова, Блока и др.). Третий подход, применяемый нами, ставит задачу вскрыть структуру концептосферы автора и описать скрытые закономерности его видения мира. Образ в этом случае понимается как фрагмент текста, который реализует пере-

нос свойств с одного концепта на другой. Такое определение опирается на классическую схему переноса концептуальных свойств с концепта-источника на концепт-цель в когнитивной лингвистике. Этот подход позволяет провести подсчет концептов в различных функциях, определить авторские предпочтения и концепты, которые автор избегает. Кроме того, в огромном многообразии образов можно выявить устойчивые сочетания концептов – модели (например, Орган – Существо, Растения – Ткань). Различия в частотности и сочетаемости концептов показывают эволюцию стиля автора и изменения в его мировоззрении.

Онтология, применяемая для описания образной системы, не совпадает с традиционными семантическими классификациями. Она получена в результате индуктивного исследования большого объема художественных текстов [18], затем апробирована нами на значительном объеме стихотворных текстов на русском и английском языке и скорректирована [15, 16]. Полный список выделяемых концептов приводится в табл. 1. Ниже остановимся на ряде методологических особенностей анализа.

Таблица 1. Концепты, выделяемые при анализе
Table 1. Concepts highlighted in the analysis

Концепт	**Примеры лексических репрезентантов**
Вещество	Глина, железо, пепел, песок, стекло
Вода	Вода, волна, озеро, океан, поток, река, роса, слеза
Время	Время, век, весна, вечер, день, миг, осень, прошлое, час
Драгоценность	Золото, диадема, жемчуг, корона, серебро, янтарь
Еда	Еда, вино, овощи, фрукты, хлеб
Звук	Звук, голос, мелодия, музыка, песня, симфония
Информация	Слово, легенда, рассказ, сказка, фраза
Контейнер	Вместилище, коробка, кубок, чашка
Мир	Мир, вселенная
Музыкальный инструмент	Барабан, колокол, лира, лютня, рожок, струна, флейта
Огонь	Огонь, пламя, искра, пожар
О́рган	Волосы, глаз, губы, живот, крыло, нога, рука, сердце, язык
Орудие	Дротик, иголка, лук, меч, молоток, мотыга, ножницы, пушка, щит
Предмет	Знамя, флаг, лестница, парус, паутина, распятие, статуя, цепь
Природа	Природа
Пространство – небесное, воздушное – земное – строения	 Небо, облако, туман Земля, территория, поле, горы, равнина, пустыня Дом, арка, здание, крыша, окно, палатка, храм
Психическая сфера (ПС) – чувства, состояния – ментальная сфера	 Гнев, любовь, надежда, ненависть, печаль, радость, сон, страх, чувство Идея, мысль, память, размышления
Растение	Растение, ветви, дерево, дуб, листок, сосна, ствол, трава, цветок
Свет	Звезда, лампа, луна, луч, свет, сияние, солнце, свеча, темнота
Социальный феномен (СФ)	Власть, война, мир, равенство, свобода
Стихия	Ветер, гроза, мороз, снегопад, ураган, шторм
Существо	Человек, брат, воин, мать, читатель, певец, птица, рыба, жук, ангел, ласточка
Ткань	Ткань, вуаль, вышивка, занавес, нитка, платье, сутана, шарф
Транспорт	Автобус, корабль, поезд, паровоз, телега
Экзистенция	Жизнь, рождение, смерть, судьба

Ряд лексических единиц может реализовывать различные концепты. Так, слово *море* может быть репрезентантом концептов Вода или Пространство, в зависимости от того, выступает ли на первый план значение воды или безбрежного простора.

Среди рассматриваемых концептов встречаются единицы высокого уровня обобщения – мегаконцепты (Существо, Растение, Психическая сфера, Социальный феномен). Их выделение обусловлено тем, что в сознании авторов эти обобщенные категории очевидным образом существуют и авторы оперируют ими как целостными сущностями. Об этом говорит целый ряд фактов. Рассмотрим проблему на примере мегаконцепта Существо. Анализ лексических репрезентантов Существа и их сочетаемости говорит о том, что авторы склонны видеть различные живые существа как обладающие в высокой степени сходными качествами. Широко распространена метафора, реализуемая предикатом, в которой точное наименование существа отсутствует вовсе. Вместо этого имеется указание на свойства и функции живого существа вообще (например, умение перемещаться, выступать деятелем, смертность – кланяться, прыгать, реветь, умирать). Частотными являются и случаи, когда вместо предиката на Существо указывает эпитет (также без прямого указания на конкретное существо): живой, испуганный и др. Таким образом, различия между различными видами живых существ в художественном мире нейтрализуются.

Как указывалось выше, образное видение мира не совпадает с энциклопедическим. Примером может служить концепт Орган.

Во многих случаях авторы строят образы на словах, несомненно, являющихся органами (*сердце*, *глаз*). Однако в соответствии с нашим подходом к органам относятся многие части тела, не являющиеся органами в медицинском смысле: *рука*, *плечо*, *лицо*, *волосы* и др. Опыт исследований показывает, что указанная лексика используется для метонимического отражения человека так же, как собственно названия органов. Более того, аналогичную функцию выполняют и такие слова как *взор*, *взгляд*, *улыбка* и др.

Полученные количественные данные по частотности образов в текстах (количество выявленных концептов отдельно для Целей и Источников делилось на число строк для получения сопоставимых данных) обрабатывались с помощью многомерного дискриминантного анализа. Эта методика позволяет сопоставить группы объектов (в нашем случае две группы – тексты первого периода и тексты зрелого периода творчества) одновременно по большому количеству признаков (признаками являются данные о количестве концептов, входящих в состав образов). Метод устойчив к отклонениям от нормального распределения, поэтому получил значительное распространение в естественных науках, а затем стал успешно применяться в лингвистике [19, 20 и др.].

Результаты

Использование многомерного дискриминантного анализа позволило на базе шестнадцати концептов, представленных в функции источника в образах, построить признаковую модель, которая дифференцирует стиль поэта различных периодов.

В дискриминантную модель вошли такие концепты как Вода, Информация, Предмет, Пространство, ПС, и Растение. Таким образом, эти пять признаков различают стиль раннего и зрелого Набокова-поэта. Однако прежде чем более подробно рассматривать различия, следует оценить эффективность модели – насколько успешно модель дифференцирует стиль двух периодов. Для этого проведем *post hoc* классификацию: при этом принадлежность текста тому или иному периоду определяется без учета реального времени его написания исключительно на основе частотности вошедших в модель концептов. Результаты показаны в табл. 2, где в строках даны тексты так, как они в действительности распределены по сборникам, а в столбцах – так, как тексты распределились на основании выделенной признаковой модели.

Как мы видим, наблюдается полное совпадение полученной группировки с имеющей место в действительности. Полученный результат неожиданно высок. Можно сделать вывод, что роль

элементов в концептосфере Набокова претерпела в течение жизни кардинальные изменения. Рассмотрим вклад каждого концепта в дифференциацию стиля раннего и зрелого поэта.

Таблица 2. Результаты post hoc классификации
Table 2. Post hoc classification results

Сборник	Процент правильной классификации	Сборник 1	Сборник 2
Сборник 1	100	9	0
Сборник 2	100	0	3
Всего	100	9	3

Наиболее сильно дифференцирует стиль раннего и зрелого поэта использование концептов Вода (коэффициент –5,20), Информация (–4,14) и ПС (3,28). За ними следуют Предмет (3,03), Растение (2,59) и Существо (–2,42), и, наконец, Пространство (1,41) и Свет (1,19). Отрицательные коэффициенты указывают на то, что высокая частотность данного концепта-цели свойственна первому периоду творчества, положительный коэффициент – что зрелому периоду. Следует отметить, что значение коэффициента по модулю не свидетельствует о высокой либо низкой частотности – абсолютная величина коэффициента указывает на масштаб изменений в использовании данного концепта автором. Таким образом, для молодого Набокова следующие образы гораздо более типичны, чем для зрелого:

Свет – Вода
Раскрыты окна в сад. На кресло, на паркет
широкой полосой янтарный льется свет (Т2)

Время – Вода
День мирно протекал. Я вспоминаю вновь (Т1)

Пространство – Информация
на темных крылышках... Текла
от тени к тени золотистой,
подобна музыке волнистой,
неизъяснимая яйла! (Т3)

Орган – Информация
и встала бархатная тайна
в твоих языческих глазах. (Т6)

Транспорт – Существо
У мирной пристани, блестя на солнце юга,
с дремотной влагой в лад снастями шевеля,
задумчивы, стояли друг близ друга
два стройных корабля. (Т7)

ПС – Существо
дрожь нисходящую отчаянья и ровный
шаг равнодушия, шаг немощи скупой,
мечтательности шаг, взволнованный, слепой,

всегда теряющий две или три ступени,
и поступь важную самодовольной лени (Т8)

Для зрелого периода творчества типичными становятся следующие образы:

Свет – ПС
... укоризны вечерней зари (Т10)

Предмет – ПС
головные уборы, как мысли вовне (Т11)

Вода – ПС
А мосты — это счастье навеки,
счастье черной воды. Посмотри (Т12)

Существо – Предмет
И распутать себя осторожно,
как подарок, как чудо, и стать (Т12)

Информация – Растение
эти триста листов беллетристики праздной
разлетятся — но у настоящей листвы
...
а бедные книги твои,
без земли, без тропы, без канав, без порога,
опадут в пустоте, где ты вырастил ветвь (Т11)

Существо – Свет
что я страны менял, как фальшивые деньги,
торопясь и боясь оглянуться назад,
как раздваивающееся привиденье,
как свеча меж зеркал, уплывая в закат. (Т11)

Экзистенция – Пространство
Сейчас переходим с порога мирского
в ту область... как хочешь ее назови:
пустыня ли, смерть, отрешенье от слова,
иль, может быть, проще: молчанье любви. (Т10)

Заключение

Таким образом, налицо существенные изменения в списке типичных концептов-источников, т.е. тех единиц концептосферы, которые рассматриваются автором как интуитивно понятные и в силу этого используемые в качестве доноров концептуальных свойств. Для природных источников свойств Набоков переходит от Воды к Пространству и Свету: от текучести и изменчивости к простору и ясности. Вместо Существа поэт обращается к психической сфере и Растению – здесь одновременно смена фауны флорой и обобщенного видения живого существа (в качестве существ выступают не только люди) его состоянием (эмоциональные и ментальные феномены, естественно, уже исключительно человеческие). От абстрактной Информации совершается по-

ворот к вещественному Предмету. Изменения характеризуются разнонаправленностью, имеют характер сложной полифонии. Одновременно реализуются тенденции к упрощению (смена Информации Предметом) и усложнения (смена Существа как единого легко наблюдаемого целого его психическим состоянием).

В целом можно заключить, что использование квантитативного анализа позволило выявить тенденции в традиционно сложном для анализа аспекте индивидуально-авторского стиля на стихотворном материале. Установлены концепты, дифференцирующие различные периоды творчества Владимира Набокова, получена количественная оценка их вклада в эволюцию стиля, сформирована признаковая модель, позволяющая с высокой степенью точности определять период создания произведения.

Перспективами исследования являются как экстенсивное расширение признаковой модели с целью охватить больше граней авторского стиля, так и формирование агрегированных признаков (комбинаций характеристик) для снижения порогового значения объема текста. Кроме того, полученные данные могут быть сопоставлены с результатами аналогичного анализа образов в стиле других поэтов.

СПИСОК ИСТОЧНИКОВ

1. **Мартыненко Г.Я.** Основы стилеметрии. Л.: Изд-во Ленинградского ун-та, 1988. 176 с.

2. **Argamon S., Whitelaw C., Chase P., Hota S., Garg N., Levitan S.** Stylistic text classification using functional lexical features // Journal of the American Society for Information Science and Technology. 2007. Vol. 58 (6). P. 802–821.

3. **Rudman J.** Cherry picking in nontraditional authorship attribution studies // Chance. 2003. Vol. 16, No. 2. P. 26–32.

4. **Андреев В.С.** Квантитативный анализ стиля: распределение частотности морфологических классов слов в текстах Г. Лонгфелло // Известия Смоленского государственного университета. 2019. № 4 (48). С. 222–231.

5. **McMenamin G.R.** Forensic stylistics: Advances in forensic stylistics / by Gerald R. Boca Raton. London, New York, Washington D.C.: CRC Press LLC, 2002. 334 p.

6. **Andreev S.** Verbal vs. adjectival style in long poems by A.S. Pushkin // Glottometrics. 2016. T. 33. C. 25–31.

7. **Andreev S., Místecký M.** Activity in Czech and Russian Nineteenth-century sonnets: A Contrastive Study // Glottotheory. International Journal of Theoretical Linguistics. 2018. Vol. 9. Issue 1. P. 89–104.

8. **Milička J., Kubát M.** Vocabulary Richness Measure in Genres // Journal of Quantitative Linguistics. Vol. 20. No. 4. 2013. P. 339–349.

9. **Bruster D., Smith G.** A new chronology for Shakespeare's plays // Digital Scholarship in the Humanities. 2016. No. 2. Vol. 31. P. 301–320.

10. **Hoover D.L.** The microanalysis of style variation // Digital Scholarship in the Humanities. 2017. Vol. 32. Supplement 2. P. ii17–ii30.

11. **Yu B.** Language and gender in congressional speech // Literary and Linguistic Computing. 2014. No. 1 (29). P. 118–132.

12. **Holmes D.I.** Stylometry and the civil war: The case of the Pickett letters // Chance. 2003. Vol. 16, No. 2, P. 18–26.

13. **Stamou C.** Stylochronometry: Stylistic development, sequence of composition, and relative dating // Literary & Linguistic Computing. 2008. Vol. 23, No. 1. P. 181–199.

14. **Gasparov M.L.** Exact methods of grammar analysis in verse // M.L. Gasparov. Selected Works. 2012. V.4. S. 23–35.

15. **Андреев В.С.** Малый текст: опыт квантитативного анализа // Известия Смоленского государственного университета. 2020. № 4 (52). С. 117–126.

16. **Андреев В.С.** «Светлый» Лонгфелло: концепт свет в меняющемся стиле // Известия Смоленского государственного университета. 2019. № 3 (47). С. 201–210.

17. **Набоков В.В.** Стихотворения. СПб.: Академический проект, 2002. 475 с.

18. **Павлович Н.В.** Язык образов: парадигмы образов в русском поэтическом языке. М.: РАН ИРЯ, 1995. 491 с.

19. **Karlgren J., Cutting D.** Recognizing text genres with simple metrics using discriminant analysis // Proceedings of the 15th Conference on Computational Linguistics. Kyoto, Japan. 1994. P. 1071–1075.

20. **Mikros G.K.** Content words in authorship attribution: An evaluation of stylometric features in a literary corpus // Studies in Quantitative Linguistics 5: Issues in Quantitative Linguistics / Ed. Reinhard Köhler. RAM-Verlag, 2009. P. 61–75.

REFERENCES

[1] **G.J. Martynenko,** Osnovy stilemetrii [Basis of stylometry]. L.: Izd-vo Leningradskogo un-ta, 1988. 176 p.

[2] **S. Argamon, C. Whitelaw, P. Chase, S. Hota, N. Garg, S. Levitan,** Stylistic text classification using functional lexical features, Journal of the American Society for Information Science and Technology. 58 (6) (2007) 802–821.

[3] **J. Rudman,** Cherry picking in nontraditional authorship attribution studies, Chance. 16 (2) (2003) 26–32.

[4] **V.S. Andreev,** Kvantitativnyj analiz stilja: raspredelenie chastotnosti morfologicheskih klassov slov v tekstah G. Longfello [Quantitative Analysis of Style: Distribution of Part of Speech Frequencies in Texts by H. Longfellow], Izvestija Smolenskogo gosudarstvennogo universiteta. 4 (48) (2019) 222–231. (In Russian)

[5] **G.R. McMenamin,** Forensic stylistics: Advances in forensic stylistics / by Gerald R. Boca Raton. London, New York, Washington D.C.: CRC Press LLC, 2002. 334 p.

[6] **S. Andreev,** Verbal vs. adjectival style in long poems by A.S. Pushkin, Glottometrics. 33 (2016) 25–31.

[7] **S. Andreev, M. Místecký,** Activity in Czech and Russian Nineteenth-century sonnets: A Contrastive Study, Glottotheory. International Journal of Theoretical Linguistics. 2018. Vol. 9 (1) (2018) 89–104.

[8] **J. Milička, M. Kubát,** Vocabulary Richness Measure in Genres, Journal of Quantitative Linguistics. 4 (20) (2013) 339–349.

[9] **D. Bruster, G. Smith,** A new chronology for Shakespeare's plays, Digital Scholarship in the Humanities. 2 (31) (2016) 301–320.

[10] **D.L. Hoover,** The microanalysis of style variation, Digital Scholarship in the Humanities. 32 (2) (2017) ii17–ii30.

[11] **B. Yu,** Language and gender in congressional speech, Literary and Linguistic Computing. 1 (29) (2014) 118–132.

[12] **D.I. Holmes,** Stylometry and the civil war: The case of the Pickett letters, Chance. 16 (2) (2003) 18–26.

[13] **C. Stamou,** Stylochronometry: Stylistic development, sequence of composition, and relative dating, Literary & Linguistic Computing. 23 (1) (2008) 181–199.

[14] **M.L. Gasparov,** Exact methods of grammar analysis in verse, M.L. Gasparov. Selected Works. 4 (2014) 23–35.

[15] **V.S. Andreev,** Malyj tekst: opyt kvantitativnogo analiza [Small text: the experience of quantitative analysis], Izvestiya Smolenskogo gosudarstvennogo universiteta. 2020. No. 4 (52) (2020) 117–126. (in Russian)

[16] **V.S. Andreev,** "Light" Longfello: concept Light in the changing style [«Svetlyj» Longfello: koncept svet v menyayushchemsya stile], Izvestiya Smolenskogo gosudarstvennogo universiteta. 3 (47) (2019) 201–210. (in Russian)

[17] **V.V. Nabokov,** Stihotvoreniya [Poems], Saint-Petersburg: Academichsskij proekt, 2002. 475 p.

[18] **N.V. Pavlovich,** Jazyk obrazov: paradigmy obrazov v russkom pojeticheskom jazyke [Language of images: paradigms of images in Russian poetic language]. M.: RAN IRJa, 1995. 491 p. (in Russian)

[19] **J. Karlgren, D. Cutting,** Recognizing text genres with simple metrics using discriminant analysis, Proceedings of the 15th Conference on Computational Linguistics. Kyoto, Japan. 1994. Pp. 1071–1075.

[20] **G.K. Mikros,** Content words in authorship attribution: An evaluation of stylometric features in a literary corpus, Studies in Quantitative Linguistics 5: Issues in Quantitative Linguistics / Ed. Reinhard Köhler. RAM-Verlag, 2009. Pp. 61–75.

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Андреев Вадим Сергеевич
Vadim S. Andreev
E-mail: vadim.andreev@ymail.com
ORCID: https://orcid.org/0000-0001-7580-4386

Поступила: 09.02.2023; Одобрена: 03.03.2023; Принята: 17.03.2023.
Submitted: 09.02.2023; Approved: 03.03.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 21–29.
Terra Linguistica. 2023, Том 14, № 1. С. 21–29.

Research article
UDC 81'33
DOI: https://doi.org/10.18721/JHSS.14103

MAPPING WORD FREQUENCIES IN FICTION ON SOCIOPOLITICAL CONTEXT: THE CASE OF EARLY 20TH CENTURY RUSSIAN SHORT STORIES

A.O. Grebennikov ✉ , **N.M. Marusenko** , **T.G. Skrebtsova**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ a.grebennikov@spbu.ru

Abstract. The paper deals with the language of Russian short stories written in the period from 1900–1930. It is based on the Russian Short Stories Corpus, an ongoing research project aimed to collect, digitally process, and present the Russian literature of the early 20th century in an electronic form. The Corpus contains the stories written by thousands of Russian authors, both well-known and almost forgotten ones. From the corpus, a sample was taken to serve as a testbed for linguists, lexicographers and literary scholars, enabling them to check their intuitions concerning the language and style of the epoch. The sample has been divided into three subsamples along the lines set by the dramatic turns of Russian history. The first subsample contains the stories produced from the onset of the 20th century up to WWI (1900–1913), the second one refers to the tumultuous period of wars and revolutions (1914–1922), and the third accounts for the stories written in the Soviet Union (1923–1930). The Corpus has proved instrumental in detecting manifold changes in language use, including grammar, vocabulary, syntactic patterns, collocations, and stylistics. In the present paper, frequency-sorted word lists are used to bring out relevant changes in Russian vocabulary, linking them to the sociopolitical context. The results obtained will provide valuable data for the lexicographers compiling Russian dictionaries of the above-mentioned period.

Keywords: Russian short stories, text corpus, frequency dictionary, Russian lexicography, stylometry.

Citation: A.O. Grebennikov, N.M. Marusenko, T.G. Skrebtsova, Mapping word frequencies in fiction on sociopolitical context: the case of early 20th century Russian short stories, Terra Linguistica, 14 (1) (2023) 21–29. DOI: 10.18721/JHSS.14103

Научная статья
DOI: https://doi.org/10.18721/JHSS.14103

ЧАСТОТНЫЙ СЛОВАРЬ ХУДОЖЕСТВЕННОЙ ПРОЗЫ В КОНТЕКСТЕ СОЦИОПОЛИТИКИ (НА МАТЕРИАЛЕ «КОРПУСА РУССКОГО РАССКАЗА 1900–1930 ГГ.»)

А.О. Гребенников ✉ , **Н.М. Марусенко** , **Т.Г. Скребцова**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ a.grebennikov@spbu.ru

Аннотация. Работа выполнена на материале «Корпуса русского рассказа 1900–1930 гг.» – масштабного проекта, направленного на сбор, цифровую обработку, анализ и представление произведений русской литературы начала XX века в электронном виде. Корпус содержит рассказы нескольких тысяч авторов, как признанных мастеров художественного слова, так и практически неизвестных. Исследование проведено на материале аннотированной выборки из Корпуса, которая служит «полигоном» для проверки гипотез, касающихся стиля языка эпохи, лингвистами, литературоведами и лексикографами. Выборка разделена на три подвыборки, отражающие основные этапы русской истории начала XX века: 1) довоенный период (1900–1913), 2) военно-революционные годы (1914–1922) и 3) советский период (1923–1930). Установлено, что анализ корпусного материала показателен при прослеживании различный изменений в использовании языка, включая грамматику, лексику, синтаксические модели, коллокации и стилистику. В настоящем исследовании построенные по выборкам частотные словари используются для выявления значимых изменений в лексическом составе, которые рассматриваются в социополитическом контексте. Полученные результаты представляют интерес для специалистов в области русского языка, стилеметрии и лексикографии.

Ключевые слова: Русский рассказ, корпус текстов, частотный словарь, лексикография, стилеметрия.

Для цитирования: Гребенников А.О., Марусенко Н.М., Скребцова Т.Г. Частотный словарь художественной прозы в контексте социополитики (на материале «Корпуса русского рассказа 1900–1930 гг.») // Terra Linguistica. 2023. Т. 14. № 1. С. 21–29. DOI: 10.18721/JHSS.14103

The Russian Short Stories Corpus (1900–1930)

The present paper draws on the Russian Short Stories Corpus (1900–1930), an ongoing project aimed to collect, digitally process, and present the Russian literary heritage of the early 20th century in an electronic form, thus making it available to a wide range of users. In particular, the Corpus is supposed to become a major resource for the linguists and literary scholars, enabling them to research into the language and style of the pre-revolutionary, revolutionary and post-revolutionary prose [1, 2]. For lexicographers, it contains valuable information on the way Russian grammar, vocabulary, phraseology, and stylistics kept changing over this tumultuous period of Russian history and will be instrumental in compiling Russian dictionaries of this period.

The Corpus currently contains a few thousand stories written in Russia, and later the Soviet Union, and published in literary journals or anthologies. It seeks to include works by a maximal number of writers, not only the famous ones but also the lesser-known or almost forgotten authors, thus maintaining a well-balanced and representative collection. The whole Corpus is divided into three chronological subcorpora, their boundaries marked by significant historical events. Thus, the first subcorpus (1900–1913) refers to a

pre-war period, the second one (1914–1922) covers a series of dramatic events (WWI, the February and October revolutions, the Civil War) that resulted in an overall radical change in the Russian political landscape and social life, and the third one (1923–1930) accounts for the post-war socialist period.

Each author can be represented by a single story per period. Stories written in emigration are not included in the Corpus. Thus, the Corpus features two stories by Ivan Bunin, one written in the 1st period and the other in the 2nd one. As the writer left Russia shortly after the revolution, his 3rd period stories are not accounted for.

Besides Ivan Bunin, the Corpus, thus, contains stories by such prominent Russian authors as Leo Tolstoy, Leonid Andreev, Arkady Averchenko, Alexander Blok, Sergey Esenin, Konstantin Balmont, Andrey Belyj, Anton Chekhov, Maxim Gorky, Zinaida Gippius, Nadezhda Teffi, Alexander Kuprin, Mikhail Zoschenko, Evgeny Zamyatin, Ivan Schmelev, Valentin Kataev, Veniamin Kaverin, Mikhail Kuzmin, Isaac Babel, Mikhail Bulgakov, Yuri Olesha, Arkady Gaydar, Konsyantin Paustovsky, Andrey Platonov, Mikhail Sholokhov, Alexey Tolstoy, etc.

From the text corpus, a random sample was taken containing 310 stories by 300 authors, ca. 100 stories per period (the slight discrepancy in numbers is due to the fact that some writers feature in more than one period). This sample serves as an initial testbed enabling scholars to put forward and prove or refute the hypotheses bearing on Russian language and literature of the given period [3–5]. The present research is also based on this sample.

Word Frequency Distribution as a Window on the Sociopolitical Context

The underlying idea of any research on the Russian Short Stories Corpus (1900–1930) is that language use cannot help being affected by the sociopolitical processes, hence the division of the whole Corpus into subcorpora in accordance with the major milestones in Russian history (see above). The present paper is no exception. Focusing on the short stories vocabulary and, more specifically, on frequency-sorted word lists, it aims to find significant variation across the periods and account for it in terms of political events and social developments.

Word frequency analysis has proved instrumental in language studies, in general, and in corpus linguistics, in particular [6–8]. With respect to the project concerned, it has been used to explore the major statistical parameters of the Corpus, including the words' absolute and relative frequency, rankings, part-of-speech distribution, rank mean, keyness, lexical specificity, as well as to perform cluster analysis both for each individual period and for the whole corpus [9].

Apart from this, word frequency analysis is helpful in bringing out lexical features characteristic of a writer's individual style. Thus, a comparison of word frequency ranks drawn from the works of a few writers may provide an insight into their individual world views and priorities. It has been shown, in particular, that Ivan Bunin's stories are primarily about the rural life and the beauty of nature, whereas Anton Chekhov focused mainly on social life and human relationships. Word frequency analysis of the stories by Leonid Andreev has convincingly demonstrated his obsession with the tragic aspects of life, including loneliness and fear of death [10].

In the present research, word frequency ranks have been calculated over a variegated collection of stories by a few hundred authors. Individual differences are thus neutralized, and the results obtained may be said to reveal a certain flavour of the epoch. The frequency dictionaries under investigation have been complied using UNILEX-T software [11].

The technique is not ideal in that certain errors can be made in automatically deriving lemmas from the word forms, due to the homonymy. This is often the case with the highly inflected languages, such as Russian. Still, the ratio of such errors is quite small and can usually be neglected.

Another minor problem may arise from the polysemous words being treated as a single unit. Semantic tagging has always been a challenge to automatic processing, and the current state of the project does not provide such an option. Therefore, strictly speaking, one cannot check which of the individual word senses

suddenly got activated and brought about an increase in the overall frequency. But plausible conjectures can still be made, drawing on the previously detected dynamics of change in the stories' thematic content [5, 12, 13] and the political context in which the stories were produced. This is the case, in particular, with the words *tovarishch* ('comrade') and *krasnyj* ('red'), whose frequency drastically rose in the socialist period, obviously due to the activation of the new, ideological, senses.

In what follows, "upper zones" of the frequency lists of all the three periods are analysed, each containing content words with frequency over 100. For each period, the number of such lexical units is well over 200. Taken together, they amount to ca. 800. Table 1 provides data on some of the words discussed below.

Interestingly, the highest six ranks of all the frequency lists are filled by the same six content words (though, with varying order), namely *govorit'* ('to say'), *skazat'* ('to tell'), *odin* ('one'), *glaz* ('eye'), *ruka* ('hand, arm'), and *moch* ('can, may, be able'). Below these top ranks, the frequency distributions display quite a few noteworthy differences.

By comparing an upper zone of a later period with that/those of the earlier one(s), the following terms are identified:

1. words previously unfound in the upper zone;
2. words that, by contrast, are no longer present in the upper zone;
3. words demonstrating sharp drops and rises within the upper zone across periods.

In all the three types of cases, an attempt is made to interpret our findings in light of the relevant sociopolitical context and link them to the previously detected dynamics of change in the thematic content of the Russian short stories [12, 13].

Tracing Word Frequency Change across the Periods

1. The second (wartime) vs. the first (pre-war) period

In the wartime period (1914–1922) the words *ofitser* ('officer'), *russkij* ('Russian'), *dyakon* ('deacon') and *pisat'* ('write') made their way into the upper zone of the frequency list. This obviously resulted from the very character of the epoch. A long chain of wars and revolutions brought about, among other things, the separation of families, anxiety, distress, and sorrow, the need to keep in touch and pray. The rankings of the words *soldat* ('soldier'), *Bog* ('God') and *pis'mo* ('letter'), already present in the upper zone in the pre-war period, also went up. These facts are in accord with the increased activation of the relevant themes.

The war issues pushed down themes bearing on the regular work and study, so the words *barin* ('master'), *khozyain* ('employer'), *rabotat'* ('to work'), *rabochij* ('worker'), *student* ('student') left the upper zone.

A tougher time demanded a tougher modality, with the word *mozhno* ('one may') leaving the upper zone and the words *dolzhnyj* ('one must') and *nel'zya* ('one should not'), by contrast, entering it. Thus, permission was replaced by compulsion and prohibition.

Another conspicuous fact indicative of a difficult time is a significant drop in frequency of a wide range of terms carrying positive connotations, cf. *prazdnik* ('feast'), *dobryj* ('kind'), *svetlyj* ('bright'), *krasivyj* ('beautiful'), *vesyolyj* ('merry'), *smekh* ('laughter'), *schast'je* ('happiness'), *ulybka* ('smile'), *ulybatsya* ('to smile'), *vera* ('faith'), *tikhij* ('silent'), *tishina* ('silence'). All of these left the upper zone in the 2nd period, with only a few to return in the 3rd one (see below). Accordingly, topics bearing on love, family life, charity, magnanimity, etc. show decreasing frequencies.

The words *pit'* ('to drink') and *pjanyj* ('drunken') also went well below the upper zone, evidently due to prohibition enforced in Russia at the beginning of WWI. It continued through the turmoil of the revolutions and the Civil War until 1925.

2. The third (post-war, socialist) period vs. the preceding ones

Perhaps, the most remarkable feature of the word frequency distribution in the 3rd period is the upward movement of a vast number of concrete nouns associated, firstly, with rural life and peasantry, and secondly, with technical progress. Thus, the upper zone was enriched by such words as *ded* ('grandfather, old man'), *starukha* ('old woman'), *rebyata* ('children'), *pole* ('field'), *khleb* ('bread'), *kust* ('shrub'), *trava*

(‘grass’), *sobaka* (‘dog’), *kon*’ (‘horse’), *ptitsa* (‘bird’), *mashina* (‘machine’), *poezd* (‘train’), *vagon* (‘railway carriage’), *khod* (‘motion’). The frequency of the corresponding themes enjoyed a sharp rise, too. Abstract nouns, by contrast, yielded, many of them leaving the upper zone.

The list of the body-part names steadily featuring in the upper zone – *ruka* (‘hand, arm’), *glaz* (‘eye’), *golova* (‘head’), *litso* (‘face’), *guba* (‘lip’), *zub* (‘tooth’), *noga* (‘leg, foot’), *telo* (‘body’), *plecho* (‘shoulder’), *palets* (‘finger’), *volosy* (‘hair’) – in the 3rd period was almost doubled by the adding of *nos* (‘nose’), *ukho* (‘ear’), *yazyk* (‘tongue’), *sheya* (‘neck’), *shcheka* (‘cheek’), *boroda* (‘beard’), *bok* (‘side’), *koleno* (‘knee’). There are more numerals to be found in the top ranks, too.

The permission modality (*mozhno*) is back, with prohibition (*nel’zya*) gone and compulsion (*dolzhnyj*) remaining. Some words that left the upper zone in the 2nd period are back, too, cf. *vesyolyj* (‘merry’), *smekh* (‘laughter’), *tikhij* (‘silent’), *tishina* (‘silence’), *igrat*’ (‘play’), *razgovor* (‘talk’), *rabotat*’ (‘to work’), *rabochij* (‘worker’), which must be due to the beginning of peace. Accordingly, the words *ofitser* (‘officer’) and *soldat* (‘soldier’) left the upper zone.

Social relations in the 3rd period center primarily on work and family, hence a drop in the frequency of the words *gost*’ (‘guest’) and *znakomyj* (‘acquaintance’). Family relations, though, are also fading, cf. the falling frequency of *muzh* (‘husband’), *zhena* (‘wife’), *deti* (‘children’). *Rebyonok* (‘child’) already left the upper zone in the 2nd period and failed to re-appear. These lexical trends are corroborated by a similar dynamics of change in the stories’ thematic component.

Many words remain in the upper zone throughout all the three periods. Some of them hold a more or less stable position in frequency rankings, while others demonstrate a progressive upward or downward movement pattern.

The rising pattern is particularly characteristic of the words tovarishch (‘comrade’) and krasnyj (‘red’). The opposite trend can be observed in words referring to the family life and those denoting emotional and spiritual life aspects, cf. *lyubit*’ (‘to love’), *chuvstvovat*’ (‘to feel’), *smeyatsya* (‘to laugh’), *dusha* (‘soul’), *mysl*’ (‘thought’), *Bog* (‘God’).

Discussion

Above, the most spectacular word frequency changes have been mentioned that can be easily accounted for in terms of the relevant sociopolitical context. However, with other words, the dynamics of frequency change is at least not so understandable and may even seem counter-intuitive. Thus, the words *strashnyj* (‘dreadful’), *strakh* (‘fear’), *uzhas* (‘horror’), *drozhat*’ (‘tremble’), *umeret*’ (‘die’), *toska* (‘anguish’), *bol’noj* (‘sick’), present in the upper zone in the 1st (pre-war) period, left it in the 2nd (wartime) period, although it would look more natural the other way round.

It may also seem strange that the military terms *ruzhjo* (‘gun’) and *rota* (‘company as a military unit’), together with *krov*’ (‘blood’), were absent from the upper zone in the 2nd period but did enter it in the 3rd one. One would expect them, instead, to show higher frequency in the stories of 1914–1922. This fact, though, nicely fits with our previous finding concerning the stories themes, as there proved to be twice as many stories about the Civil War in the 3rd period as in the 2nd one [13]. Such postponed effect, in general, is typical of the decisive events affecting the very course of a nation’s history. They retain significance for many decades, being evoked in scholarship, literature, and art.

Other cases defying a rough and ready explanation are the words with a broken-line pattern of frequency dynamics, reaching a local maximum or minimum in the 2nd period. The whole lot looks rather heterogeneous and inconclusive, so they are not considered in detail. Perhaps, such patterns would become revealing if a larger upper zone of the frequency distribution were examined.

Conclusion

In the present paper, the upper zones of the word frequency distribution in early 20th century Russian short stories have been analysed. The cases well-marked by a progressive dynamics of change have been

Table 1. Frequency ranks of selected words across the three periods

LEMMA	PERIOD 1 (1900–1913)	PERIOD 2 (1914–1922)	PERIOD 3 (1923–1930)
skazat' (to say)	1	2	3
odin (one)	2	3	4
glaz (eye)	3	4	2
govorit' (to tell)	4	1	5
ruka (hand, arm)	5	5	1
moch (may, can, be able)	6	6	6
dusha (soul)	42	35	151
zhena (wife)	52	84	139
chuvstvovat' (to feel)	54	152	197
deti (children)	69	107	198
mozhno (one may)	80		95
bog (god)	89	65	211
vera (faith)	114		
soldat (soldier)	116	62	
milyj (gentle, nice)	122	176	
chuvstvo (feeling)	126		
lyubov' (love)	134	112	
strashnyj (horrible)	135		241
krasivyj (beautiful)	143		
muzh (husband)	146	166	236
veselyj (merry)	152		232
krasnyj (red)	163	110	54
uzhas (horror)	166		
ulybatsya (to smile)	173		
znakomyj (acquaintance)	179	193	
drozhat' (to tremble)	185		276
pis'mo (letter)	190	117	
tovarisch (comrade)	198	105	138
rabochij (worker)	202		112
rebyonok (infant, child)	213		
student (student)	217		
rabotat' (to work)	218		109
schastje (happiness)	228		
pyanyj (drunken)	241		
derevnya (village)	250	192	150
muzhik (peasant man)	252		
smekh (laughter)	257		234
izba (peasant hut)	266		138
prazdnik (feast)	268		
gost' (guest)	272	165	
dyakon (deacon)		83	
baba (peasant woman)		143	81
nel'zya (one should not)		147	
ofitser (officer)		169	

End of table 1

pisat' (to write)		181	
russkij (Russian)		182	
krov' (blood)			105
vagon (railway carriage)			173
rota (company as a military unit)			219
ruzhjo (gun)			278

specifically focused on. Most of them can be accounted for in terms of the relevant sociopolitical situation and the previously detected changes in the stories' thematic content. Yet, there are words whose frequency change pattern remains not quite clear. An extension of the upper zone may help as it will bring into light a larger number of similar cases.

Also, beyond our present study are words demonstrating a steady position in frequency rankings regardless of the historical context and thus representing a kind of distribution invariants. The top six ranks being filled by the same set of words is, perhaps, the brightest example, but certainly not the only one. It would be of interest to examine how far such invariance stretches by extending the temporal boundaries of the stories beyond the given timespan.

The perspectives of our future work are manifold. Firstly, the frequency-sorted word lists of our sample can be set against the frequency distributions drawn from the stories by a particular author (see [13–17]). A pilot study [18] has shown a remarkable discrepancy between the two data sets testifying to the significant impact of personal style on the literary works' vocabulary. Secondly, it would be of interest to compile word frequency list drawing on the Russian short stories of the 21st century and then compare it to the data at hand. The above-cited paper has revealed striking differences in the upper zone of the frequency distributions that have occurred over a century (Ibid). Thirdly, an extension of the present sample is being looked forward to, to make it more representative and well-balanced, thus increasing the reliability of results. This is crucial for a broad range of research on the corpus not only within linguistics, but also in literary theory and digital humanities at large.

A special direction in the future research has to do with lexicography. Along with the comprehensive dictionaries of the Russian language that have been compiled in the Russian Academy of Sciences, there is a growing interest in the language of particular historical periods. Thus, the Russian dictionaries of the 18th and 19th centuries are currently under way. It would only be logical that the focus eventually shift to the early 20th century. This period is marked by certain distinctive features of its own, e.g. acronyms and abbreviations, coined words and phrases, novel ideological word senses, shifts in lexical use, etc. The coverage it has so far received is certainly insufficient. The corpus data will be of help to lexicographers in their future work, and frequency lemma lists have been shown to be quite useful in assessing the relative frequency of individual words [19]).

Acknowledgements

The materials of the article were presented at the IV International Conference on Engineering and Applied Linguistics "Piotrovsky Readings – 2022", dedicated to the 100th anniversary of the birth of Professor R.G. Piotrovsky at the Herzen State Pedagogical University on November 22, 2022.

REFERENCES

[1] **G. Martynenko, T. Sherstinova, T. Popova, A. Melnik, Ye. Zamirajlova,** O printsipakh sozdaniya korpusa russkogo rasskaza pervoy treti XX veka [On the principles of creation of the Russian short stories

corpus of the first third of the 20th century]. Proc. of the XV Int. Conference on Computer and Cognitive Linguistics "TEL 2018", Kazan, 2018, pp. 180—197.

[2] **G. Martynenko, T. Sherstinova,** Linguistic and Stylistic Parameters for the Study of Literary Language in the Corpus of Russian Short Stories of the First Third of the 20th Century. R. Piotrowski's Readings in Language Engineering and Applied Linguistics, Proc. of the III Int. Conference on Language Engineering and Applied Linguistics (St. Petersburg, Nov., 27, 2019), CEUR Workshop Proceedings, 2552, 2020, pp. 105—120.

[3] **G. Martynenko, T. Sherstinova,** Emotional Waves of a Plot in Literary Texts: New Approaches for Investigation of the Dynamics in Digital Culture. Digital Transformation and Global Society (DTGS 2018). Communications in Computer and Information Science, 859. Springer, Cham, 2018, pp. 299—309. Available at: https://link.springer.com/chapter/10.1007/978-3-030-02846-6_24 (accessed 10.02.2023).

[4] **T. Skrebtsova,** Struktura narrativa v russkom rasskaze nachala XX veka [Narrative structure of the Russian short story in the early XX century], Proc. of the Int. Conference "Corpus Linguistics-2019", St. Petersburg University Press, St. Petersburg, 2019, pp. 426—431.

[5] **T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, M. Kirina,** 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, Mexico City, Mexico, 12-17 October 2020, LNAI 12469, 2020, pp. 134—151.

[6] **M. Oakes,** Statistics for Corpus Linguistics, Edinburgh University Press, Edinburgh, 1998.

[7] **G. Leech, P. Rayson, A. Wilson,** Word Frequencies in Written and Spoken English: based on the British National Corpus, London: Longman, 2001.

[8] **A. Baron, P. Rayson, D. Archer,** Word frequency and key word statistics in historical corpus linguistics, Anglistik: International Journal of English Studies, 20 (1), 2009, pp. 41—67.

[9] **T. Sherstinova, A. Grebennikov, T. Skrebtsova, A. Guseva, M. Gukasian, I. Egoshina, M. Turygina,** Frequency Word Lists and Their Variability (the Case of Russian Fiction in 1900–1930). 27th Conference of Open Innovations Association FRUCT, University of Trento, Italy, 2020, pp. 366—373. Available at: https://fruct.org/publications/acm27/files/She.pdf (accessed 10.02.2023).

[10] **A. Grebennikov, T. Skrebtsova,** Yazykovaya kartina mira v russkom rasskaze nachala XX veka [World through the Prism of the Early XX-century Russian Short Stories], Philosophy and the Humanities in the Information Society, 3, 2019, pp. 82—92.

[11] **Zh. Anoshkina,** Podgotovka chastotnyh slovarej i konkordansov na komp'yutere [Computer-assisted Dictionary and Concordance Making], V.V. Vinogradov Russian Language Institute of the Russian Academy of Sciences Press, Moscow, 1995.

[12] **T. Sherstinova, T. Skrebtsova,** Russian Literature Around the October Revolution: A Quantitative Exploratory Study of Literary Themes and Narrative Structure in Russian Short Stories of 1900–1930, Proc. of the Int. Workshop "Computational Linguistics" (St. Petersburg, 17-20 June, 2020), CEUR Workshop Proceedings, 2813, 2021, pp. 117—128. Available at: http://ceur-ws.org/Vol-2813/rpaper09.pdf (accessed 10.02.2023).

[13] **T. Skrebtsova,** Thematic tagging of literary fiction: the case of early 20th century Russian short stories, Proc. of the Int. Workshop "Computational Linguistics" (St. Petersburg, 17-20 June, 2020), CEUR Workshop Proceedings, 2813, 2021, pp. 265—276. Available at: http://ceur-ws.org/Vol-2813/rpaper20.pdf (accessed 10.02.2023).

[14] **A. Grebennikov, G. Martynenko,** Chastotnyy slovar rasskazov A.P. Chekhova [Frequency Dictionary of Anton Chekhov's Short Stories], St. Petersburg University Press, St. Petersburg, 1999.

[15] **A. Grebennikov, G. Martynenko,** (2003), Chastotnyy slovar rasskazov L.N. Andreeva [Frequency Dictionary of Leonid Andreev's Short Stories], St. Petersburg University Press, St. Petersburg, 2003.

[16] **A. Grebennikov, G. Martynenko,** Chastotnyy slovar rasskazov A.I. Kuprina [Frequency Dictionary of Alexander Kuprin's Short Stories], St. Petersburg University Press, St. Petersburg, 2006.

[17] **A. Grebennikov, G. Martynenko,** Frequency Chastotnyy slovar rasskazov A.I. Bunina [Dictionary of Ivan Bunin's Short Stories], St. Petersburg University Press, St. Petersburg, 2011.

[18] **A. Grebennikov, N. Marusenko,** Korpus russkogo rasskaza nachala XX veka. Primer lingvostatisticheskogo analiza [The Early XX-century Russian Short Stories Corpora. An Example of Lingvo-statistical analysis], Proc. of the 23rd Int. conf. "Internet and Modern Society" (IMS-2020), St. Petersburg, 2020, pp. 21—29.

[19] **D. Lindemann, I. San Vicente,** Building corpus-based frequency lemma lists. Procedia – Social and Behavioral Sciences, 2015, pp. 266–277.

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Alexander O. Grebennikov
Гребенников Александр Олегович
E-mail: a.grebennikov@spbu.ru
ORCID: https://orcid.org/0000-0003-2856-5049

Natalya M. Marusenko
Марусенко Наталия Михайловна
E-mail: n.marusenko@spbu.ru
ORCID: https://orcid.org/0000-0002-3347-1373

Tatyana G. Skrebtsova
Скребцова Татьяна Георгиевна
E-mail: t.skrebtsova@spbu.ru
ORCID: https://orcid.org/0000-0002-7825-1120

Submitted: 11.02.2023; Approved: 22.03.2023; Accepted: 22.03.2023.
Поступила: 11.02.2023; Одобрена: 22.03.2023; Принята: 22.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 30–40.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 30–40.

Научная статья
УДК 81.32
DOI: https://doi.org/10.18721/JHSS.14104

ИССЛЕДОВАНИЕ ВЛИЯНИЯ ПАРАМЕТРОВ МОРФОЛОГИЧЕСКОЙ СЛОЖНОСТИ НА ТРУДНОСТЬ ВОСПРИЯТИЯ МЕДИАТЕКСТА С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ СТАТИСТИЧЕСКОГО АНАЛИЗА ДАННЫХ

Т.Г. Евтушенко[1] ✉, Е.С. Клочкова[1], А.В. Лапутенко[2], Н.В. Евтушенко[3]

[1] Санкт-Петербургский политехнический университет Петра Великого, Санкт-Петербург, Российская Федерация;

[2] Национальный исследовательский Томский государственный университет, г. Томск, Российская Федерация;

[3] Институт системного программирования РАН, Москва, Российская Федерация

✉ evtushenkotg@gmail.com

Аннотация. Предлагаемая работа посвящена изучению одного из аспектов сложности, влияющих на восприятие медиатекста: параметров морфологической сложности, а также их взаимодействию с поверхностными характеристиками текста, такими как средняя длина предложения, средняя длина слова и т.п. В работе исследуется вопрос о связи количественных параметров (метрик) объективной сложности текста, которая обусловлена его языковыми характеристиками, со степенью трудности восприятия текста читателем. Для определения и подсчета метрик морфологической сложности использовался корпус из 1000 размеченных новостных текстов (общим объемом 140000 словоупотреблений) с веб-сайтов российских ВУЗов. Для каждого текста были подсчитаны следующие величины: доля слов различных частей речи, доля отдельных граммем, соотношение именности-глагольности, соотношение знаменательных и служебных частей речи, средняя длина предложения, средняя длина слова и т.д. Анализ морфологической сложности был дополнен результатами опроса представителей целевой аудитории веб-сайта ВУЗа (абитуриентов, студентов и аспирантов), которые оценили трудность 255 новостных текстов по пятибалльной шкале. Далее на основе собранных данных проводился корреляционно-регрессионный анализ для определения значимости анализируемых метрик морфологической сложности и степени их влияния на трудность восприятия текста. На основе анализа используемых полученных моделей линейной регрессии было установлено, что наиболее значимыми метриками морфологической сложности являются доля полных причастий, доля словоформ в родительном падеже, доля кратких прилагательных и доля числительных. Кроме того, проведенный анализ подтвердил вывод предыдущих исследований о значимости таких поверхностных метрик как средняя длина предложения и средняя длина словоформы. В результате анализа были предложены две формулы для расчета степени понятности новостного текста: 1) формула, основанная на трех метриках, которые чаще всего встречаются в моделях; 2) формула, основанная на модели с наиболее высокой точностью и учитывающая пять морфологических и поверхностных метрик.

Ключевые слова: сложность текста, понятность, морфологические параметры, медиатекст, корреляционно-регрессионный анализ.

Финансирование: Проект выполнен при финансовой поддержке программы стратегического академического лидерства «Приоритет 2030» Российской Федерации (Договор № 075-15-2021-1333 от 30.09.2021).

Для цитирования: Евтушенко Т.Г., Клочкова Е.С., Лапутенко А.В., Евтушенко Н.В. Исследование влияния параметров морфологической сложности на трудность восприятия медиатекста с использованием методов статистического анализа данных // Terra Linguistica. 2023. Т. 14. № 1. С. 30–40. DOI: 10.18721/JHSS.14104

Research article
DOI: https://doi.org/10.18721/JHSS.14104

STUDYING THE IMPACT OF MORPHOLOGICAL PARAMETERS ON TEXT READABILITY USING STATISTICAL ANALYSIS METHODS

T.G. Evtushenko[1] ✉, Y.S. Klochkova[1], A.V. Laputenko[2], N.V. Evtushenko[3]

[1] Peter the Great St. Petersburg Polytechnic University, St. Petersburg, Russian Federation;

[2] National Research Tomsk State University, Tomsk, Russian Federation;

[3] Institute for System Programming of the Russian Academy of Sciences, Moscow, Russian Federation

✉ evtushenkotg@gmail.com

Abstract. The paper addresses one of the important aspects of text complexity, namely the dependency of text readability on a set of morphological and text surface metrics such as the average length of words, sentences, etc. The correlation between the objective text complexity which is specified by quantitative parameters of the linguistic features and the subjective text complexity, i.e. the difficulty of text comprehension as a psychological phenomenon, is analyzed. To assess the morphological text complexity we used an annotated dataset consisting of 1000 online news texts (140000 tokens) retrieved from the websites of Russian universities. For each text unit the ratio of each part-of-speech per token is measured. Online news texts of the dataset were also assessed by a target audience of the website, i.e. applicants, undergraduate and postgraduate students. As a result, the dataset was automatically annotated based on text linguistic features and human-labelled based on experts' estimates of text readability on a 5-point scale. To assess the significance of morphological metrics and their influence on text readability, the correlation and regression analysis was carried out. To automatically classify a text as 'easy-to-read' or not 'easy-to-read', both single feature and compound models including more than one metric were constructed. In agreement with the prior research the most common metrics influencing text readability appear to be text surface characteristics. However, the proposed models also made it possible to establish the significance of morphological parameters, used both in single feature and compound models, such as the use of participles, nouns in the genitive case, adjectives and numerals, which should be taken into account in analyzing news text readability. Moreover, novel formulae for assessing readability were proposed based on the studied coefficients.

Keywords: text complexity, readability, morphological features, media text, correlation and regression analysis.

Acknowledgements: The project was implemented with the financial support of the strategic academic leadership program "Priority 2030" of the Russian Federation (project No. 075-15-2021-1333 of 30.09.2021).

Citation: T.G. Evtushenko, E.S. Klochkova, A.V. Laputenko, N.V. Evtushenko, Studying the impact of morphological parameters on text readability using statistical analysis methods, Terra Linguistica, 14 (1) (2023) 30–40. DOI: 10.18721/JHSS.14104

Введение

Предлагаемое исследование посвящено выявлению тех морфологических характеристик сложности, которые оказывают существенное влияние на трудность восприятия текста читателем.

Восприятие и понимание текста зависит от его сложности как совокупности языковых средств разных языковых уровней, которые используются автором для создания текста в соответствии с определенной коммуникативной задачей и с учетом конкретной коммуникативной ситуации.

Являясь исходно математическим понятием, а следовательно, и объектом изучения в математических дисциплинах и теории информации, сложность текста трансформировалась в трансдисциплинарную область исследования [1]. В области лингвистики исследования сложности имеют длительную историю изучения (см. далее). При этом сложность как лингвистический конструкт рассматривается с различных точек зрения: как характеристика языковой системы в целом [2] и как характеристика отдельного речевого произведения [3]. В данном исследовании речь пойдет о втором аспекте сложности.

Разработка проблемы сложности текста ведется в зарубежной лингвистике, преимущественно в американской, уже с 40-х гг. 20 века. Для измерения сложности были выработаны формулы, среди которых наиболее популярными являются формула Флеша-Кинкейда [4, 5], формула читабельности SMOG [6], автоматический индекс удобочитаемости, индекс Колман-Лиау и ряд других [3]. Все перечисленные формулы основаны на учете количественных параметров текста (метрик), которые отражают сложность единиц различных языковых уровней: морфологического (длина слова в символах и слогах), синтаксического (длина предложения, длина самого текста), лексического (доля низкочастотной лексики в тексте).

В советской лингвистике в 70-х гг. 20 века проблема сложности текста изучалась в рамках направления квантитативной лингвистики [7]. В частности, были предложены формула М.С. Мацковского [8] и формула Ю.А. Тулдавы [9]. Эти формулы для расчета сложности текста так же, как и зарубежные, основаны на учете таких количественных параметров текста, как средняя длина предложения и длина слова в слогах. Популярностью пользуется также адаптированная для русского языка формула Флеша-Кинкейда [10]. Более подробный обзор работ данного направления представлен в работе [3].

В настоящее время в исследовании сложности можно отметить следующие тенденции.

Во-первых, применяется дифференцированный подход к сложности в зависимости от его функционально-стилистической принадлежности. В частности, отдельно изучаются сложность дидактического текста [3], сложность юридических документов [11–13]. Исследователи также разрабатывают подходы к автоматизированной оценке сложности учебных текстов на русском языке как иностранном [14, 15].

Во-вторых, в современных лингвистических исследованиях широко применяются методы автоматизированной обработки текстов для решения различных задач анализа речевого материала [16–18]. Такие методы, в частности модели машинного обучения, используются и при разработке моделей сложности текста.

Так, например, в исследованиях А.Н. Лапошиной сложность текста определяется с помощью регрессионной модели, обученной на корпусе из 800 текстов из пособий по РКИ [19]. В основе моделей сложности, разрабатываемых в рамках проектов РНФ 19-18-00525 «Понятность официального русского языка: юридическая и лингвистическая проблематика» [11, 12], заложен алгоритм градиентного бустинга на основе деревьев решений для задачи классификации. В работах, посвященных анализу сложности учебных текстов, предлагаются формулы, полученные на основе линейной регрессии с регуляризацией [20–22].

Таким образом, можно отметить, что за достаточно длительную историю изучения сложности текста в лингвистике разработаны различные подходы к определению этого понятия и вы-

явлению факторов, влияющих на тот или иной уровень сложности конкретного текста. Однако многие задачи, решение которых необходимо для понимания этого явления, все еще не имеют однозначного ответа. Одной из таких задач является выявление тех лингвистических параметров, которые оказывают наибольшее воздействие на понимание текста читателем, причем особый интерес представляет определение комбинаций нескольких параметров, поскольку, как показывают предыдущие исследования, измерение метрик сложности изолированно не дает адекватного представления о сложности целого текста.

Применение методов компьютерной лингвистики и больших массивов текстов позволило, наряду с уже известными формулами читабельности, поставить вопрос об учете большего количества метрик, учитывающих влияние на сложность текста языковых явлений разных уровней.

Насколько нам известно, на данный момент при определении уровня сложности текста не учитываются или ограниченно учитываются следующие факторы: 1) взаимосвязь объективной сложности текста и трудности его восприятия читателем; 2) влияние морфологических параметров, таких как соотношение слов разных частей речи в тексте, употребление отдельных граммем и т.п.

Таким образом, предлагаемая работа направлена на определение ряда морфологических параметров и их частотных комбинаций, которые оказывают влияние на трудность восприятия текста читателем. В работе мы придерживаемся дифференцированного подхода к определению сложности и фокусируемся только на медиатекстах на русском языке.

Подходы к определению и измерению сложности текста

Обсуждая сложность текста с лингвистической точки зрения, исследователи вводят различные термины для обозначения этого понятия, что отражает его определенную двойственность. Большинство исследователей так или иначе выделяют две стороны этого явления, объективную и субъективную сложность, для обозначения которых используют разные термины. Набор объективных лингвистических характеристик, в основном формальных, присущих тому или иному произведению, обозначается как собственно «сложность» (в зарубежных источниках – complexity). Субъективная сложность, которая рассматривается как психолингвистическое явление, обозначается терминами «трудность», «понятность». Субъективная сложность предполагает учет таких психолингвистических параметров, как когнитивные способности читающих, наличие у них определенных фоновых знаний, мотивации к чтению и т.п. [2, 23]. В работах зарубежных исследователей используются также такие термины, как «читабельность», «удобочитаемость» (англ. readability), которые обозначают уровень легкости восприятия текста читателем, что соотносится с его понятностью. В данной работе, вслед за Н.С. Валгиной под понятностью текста будем понимать “возможность определить смысл, доходчивость – возможность преодолеть «препятствия», возникающие при передаче информации” [24].

Как отмечалось в предыдущих исследованиях, сложность текста можно рассматривать как переменную, значение которой вычисляется на основе числовых показателей соответствующих признаков. Измерение сложности может осуществляться посредством подсчета известных индексов удобочитаемости по формулам, представленным в литературе (см.выше), учитывающим поверхностные, или базовые (surface metrics or baseline surface features), характеристики текста, такие как длина предложения, количество слов и предложений в тексте. Кроме того, в ряде работ учитываются количество языковых единиц разных уровней. На уровне морфологии выделяют такие показатели, как доля слов разных частей речи, формы родительного и творительного падежей существительных, соотношение глаголов и существительных и т.п. [25].

Методы и материал исследования

В качестве материала исследования в работе использовался датасет, сформированный в ходе реализации проекта «Цифровые технологии в лингвистике: модель автоматической оценки рече-

вого воздействия мультимодального электронного текста». Датасет включает новостные тексты на русском языке с сайтов ведущих российских ВУЗов. Все тексты были размечены с помощью pymorphy2 и синтаксического анализатора natasha.

Исследование проводилось по алгоритму:

- отбор имеющихся в литературе базисных и морфологических метрик и выбор наиболее релевантных из них для медиатекстов;
- опрос целевой аудитории для оценки понятности медиатекста;
- статистическая обработка собранных данных;
- анализ результатов с целью выявления наиболее влиятельных метрик и их частотных сочетаний.

Опрос по оценке понятности текстов проводился среди студентов и преподавателей ВУЗов: количество студентов – 90 человек, количество текстов – 250. Респондентам предоставлялись тексты с веб-сайтов ведущих высших образовательных учреждений и анкета, в которой они должны были проставить оценки читабельности/понятности предъявленных текстов. Оценка каждого текста ставилась как средняя оценка на основании результатов прочтения каждого текста тремя экспертами по 5-балльной шкале; обработка результатов опроса проводилась по методу экспертной оценки.

Существующие модели машинного обучения, в основном, решают задачи классификации или предсказания. Большинство таких моделей не позволяют выявить вес лингвистических признаков и их комбинаций, которые могут влиять в той или иной мере на понятность текста. В то же время, классические алгоритмы регрессионного анализа позволяют создавать как описательные, так и предсказательные модели. Как отмечалось выше, целью работы является определение морфологических характеристик текста, влияющих на трудность восприятия текста, и выявление частотных комбинаций этих параметров, наиболее часто встречающихся в статистических моделях, т.е. параметров, которые с высокой частотой встречаются в полученных 155 моделях.

При составлении массива данных на основе уже имеющихся списков метрик, составленных авторами других работ, с учетом жанра текста и релевантности метрик для статистического анализа, были отобраны соответствующие морфологические параметры, пронумерованные для удобства анализа следующим образом:

(0) индекс аналитичности,
(1) индекс субстантивности,
(2) индекс местоименности,
(3) доля словоформ в родительном падеже,
(4) доля словоформ в творительном падеже,
(5) доля кратких прилагательных,
(6) доля полных причастий,
(7) доля деепричастий,
(8) доля инфинитивов,
(9) доля числительных,
(10) доля частиц,
(11) соотношение имённости-глагольности.

При проведении анализа влияния метрик на понятность текста мы исключили общепринятые формулы индексов читабельности. Использование готовых формул ограничивало проводимое исследование, вследствие чего индексы читабельности были разложены на отдельные метрики, чтобы изучить сочетаемость поверхностных характеристик непосредственно с морфологическими текстовыми параметрами. Среди поверхностных характеристик текста были выделены следующие:

(12) средняя длина слова в слогах(ASW);
(13) среднее количество букв на 100 слов;

(14) среднее количество предложений на 100 слов;
(15) доля длинных слов (слова длиннее 6 букв);
(16) средняя длина предложения в словах (ASL);
(17) среднее количество букв и цифр в слове (CbyW).

Построение формулы для определения воспринимаемости текста на основе его морфологических характеристик

Для определения наиболее существенных морфологических метрик, влияющих на восприятие текста, был выбран метод корреляционно- регрессионного анализа [26]. Анализ значимости выбранных метрик выполнялся на основе полного перебора всех моделей линейной регрессии, содержащих все возможные комбинации из 18 выбранных метрик для 255 оцененных экспертами текстов.

Ниже представлен фрагмент таблицы (табл. 1) с комбинацией наиболее значимых для решения данной задачи метрик, где RMSE — значение корня из среднеквадратической ошибки, а коэффициенты могут быть положительными и отрицательными. Знак «–» указывает на то, что при увеличении количества словоупотреблений в данной форме трудность восприятия текста возрастает. Отсутствие знака «-» перед коэффициентом предполагает положительное значение количественной характеристики и указывает на то, что при увеличении количества словоупотреблений в данной форме трудность восприятия текста снижается. Так как исходные значения отдельных метрик для 255 текстов лежали в различных числовых диапазонах, все значения предварительно были стандартизованы для приведения к единой шкале. Соответственно, числовые значения коэффициентов для каждой метрики ниже приведены в единицах стандартных отклонений для этой конкретной метрики, что позволяет сравнивать метрики между собой по величине влияния на трудность текста.

Таблица 1. Фрагмент таблицы с построенными моделями
Table 1. A fragment of a table with constructed models

Комбинация метрик	RMSE	Коэффициенты
6, 14, 17	0,64	-0,1393 0,1805 -0,0921 3,8199
6, 10, 12, 16, 17	0,62	–0,1150; –0,0983; –0,2110; –0,1408; –0,1628; 3,8199
6, 9, 14, 15	0,63	–0,1201; –0,1172; 0,1659; –0,1256; 3,8199
6, 9, 13, 16	0,63	–0,1122; –0,1124; –0,1493; –0,1536; 3,8199
6, 9, 15, 16	0,63	–0,1261; –0,1042; –0,1298; –0,1626; 3,8199
10, 12, 16, 17	0,63	–0,1015; –0,2371; –0,1527; –0,1606; 3,8199
6, 9, 14, 15	0,63	–0,1201; –0,1172; 0,1659; –0,1256; 3,8199
6, 16	0,64	–0,1384; –0,1782; 3,8199
6, 14	0,64	–0,1315; 0,1759; 3,8199
14	0,66	0,2019; 3,8199
16	0,66	–0,1999; 3,8199
12	0,66	–0,1836; 3,8199
1, 6	0,66	–0,1001; –0,1667; 3,8199
6	0,67	–0,1663; 3,8199
2	0,67	0,1563; 3,8199

Из всего множества построенных моделей были выбраны 155 моделей, каждый коэффициент регрессии в которых является статистически значимым. Каждая модель имеет в своем составе не более 5 метрик. В табл. 2 приведен список метрик, упорядоченных по количеству моделей (из

числа выбранных 155 моделей), в которых соответствующая метрика присутствует. Среди наиболее часто «используемых» метрик оказалась доля полных причастий (входит в состав 65 моделей), среднее количество предложений на 100 слов (входит в состав 58 моделей) и среднее количество букв и цифр в слове (входит в состав 55 моделей). Из исходного набора метрик в этот список не попала доля словоформ в творительном падеже и индекс аналитичности/автосемантичности.

Относительно высокая частота вхождения отдельной метрики в различные модели линейной регрессии может говорить о ее важности для анализа текстов в рамках описанной задачи при использовании более сложных моделей.

Таблица 2. Распределение частотности встречаемости той или иной метрики в моделях
Table 2. Distribution of the frequency of occurrence of a particular metric in the models

№ метрики	Метрика	Частота
6	Доля полных причастий	65
14	Среднее количество предложений на 100 слов	58
17	Среднее количество букв и цифр в слове (CbyW)	55
16	Средняя длина предложения в словах (ASL)	45
13	Среднее количество букв на 100 слов	38
3	Доля словоформ в родительном падеже	35
5	Доля кратких прилагательных	34
9	Доля числительных	31
12	Средняя длина слова в слогах (ASW)	31
15	Доля длинных слов	27
10	Доля частиц	16
8	Доля инфинитивов	14
2	Индекс местоименности	11
1	Индекс субстантивности	10
11	Соотношение имённости-глагольности	2
7	Доля деепричастий	1
0	Индекс аналитичности/автосемантичности	0
4	Доля словоформ в творительном падеже	0

Например, для модели (первая строчка в табл. 1) на основе только трех самых частотных метрик была получена формула (1):

$$y = -0,1393 \cdot x_6 + 0,1805 \cdot x_{14} - 0,0921 \cdot x_{17} + 3,8199, \quad (1)$$

где y — экспертная оценка (уровень воспринимаемости текста); x_6 — доля полных причастий; x_{14} — среднее количество предложений на 100 слов; x_{17} — среднее количество букв и цифр в слове.

Согласно данной формуле трудность восприятия текста возрастает с увеличением доли полных причастий, средней длины слова и средней длины предложения.

Коэффициенты при соответствующих параметрах позволяют судить об относительной степени их положительного или отрицательного влияния на величину целевой переменной, в данном случае на оценку понятности текста. Согласно приведенной модели при увеличении в тексте доли полных причастий воспринимаемость текста уменьшится на 0,1393, при увеличении на 1 среднего числа предложений более чем со 100 словами, воспринимаемость увеличится на 0,1805, при увеличении среднего количества букв и цифр в слове, воспринимаемость уменьшится на 0,0921. Значение корня из среднеквадратической ошибки (RMSE) для представленной модели

линейной регрессии равно 0,6376, т.е., в среднем ошибка определения уровня воспринимаемости с помощью данной модели составляет 0,6376 единиц.

Наименьшим значением ошибки RMSE (0,6166) из отобранных 155 моделей обладает модель, приведенная ниже и учитывающая значения 5 метрик. Данные метрики также часто сочетаются с другими метриками в рамках отдельных моделей (например, вторая строчка в табл. 1, что отражено в формуле (2):

$$y = -0,1150 \cdot x_6 - 0,0983 \cdot x_{10} - 0,2110 \cdot x_{12} - 0,1408 \cdot x_{16} - 0,1628 \cdot x_{17} + 3,8199, \quad (2)$$

где x_6 — доля полных причастий; x_{10} – доля частиц; x_{12} – средняя длина слова в слогах (ASW); x_{16} – средняя длина предложения в словах (ASL); x_{17} — среднее количество букв и цифр в слове.

Согласно формуле (2) трудность восприятия текста возрастает с увеличением доли полных причастий, доли частиц, средней длины слова и средней длины предложения.

Заключение

Таким образом, на основе проведенных исследований были получены следующие результаты. На основе анализа полученных моделей линейной регрессии были определены наиболее часто встречающиеся метрики, позволяющие адекватно описать уровень воспринимаемости медиатекста в зависимости от его объективных характеристик.

Наиболее встречающимися метриками в построенных моделях являются поверхностные: среднее количество предложений на 100 слов, среднее количество букв и цифр в слове (CbyW), средняя длина предложения в словах (ASL), что хорошо согласуется с результатами предыдущих исследований.

Среди морфологических метрик, существенно влияющих на восприятие: доля полных причастий, доля словоформ в родительном падеже, доля кратких прилагательных, доля числительных.

На основе первых трех наиболее частотных в моделях метрик в их взаимосвязи предложена формула (1) для определения степени понятности медиатекста.

Модель с наиболее высокой точностью из рассмотренных представлена формулой (2). Эта модель учитывает комбинацию 5 метрик: доля полных причастий, доля частиц, средняя длина слова в слогах, средняя длина предложения в словах и среднее количество символов в слове.

Использованный алгоритм отбора моделей линейной регрессии является универсальным и может быть использован для текстов другой жанрово-стилистической принадлежности.

Как видно из представленных результатов, морфологические метрики тесно связаны с длиной словоформы и длиной предложения. Такая корреляция, на наш взгляд, обусловлена увеличением количества языковых единиц и иерархических и линейных связей между ними, которые читатель обрабатывает при восприятии текста.

Таким образом, наше исследование расширяет возможности автоматической оценки читабельности медиатекстов, помещенных на сайтах ведущих ВУЗов. Тексты, ориентированные на читателя (легкочитаемые), позволят увеличить читательскую аудиторию, что может способствовать повышению популярности вуза в медийном пространстве. В дальнейшем мы предполагаем изучить группы метрик других языковых уровней с целью определения параметров, в наибольшей степени влияющих на понятность текста.

СПИСОК ИСТОЧНИКОВ

1. **Bastardas-Boada A.** From language shift to language revitalization and sustainability. A complexity approach to linguistic ecology. Barcelona: Edicions de la Universitat de Barcelona, 2019, pp. 337–349.

2. **Dahl Ö.** The growth and maintenance of linguistic complexity. Amsterdam: John Benjamins, 2004. 336 p.

3. **Солнышкина С.И., Кисельников А.С.** Сложность текста: этапы изучения в отечественном прикладном языкознании // Вестник Томского государственного университета. Филология. 2015. № 6 (38). С. 86–99. DOI: 10.17223/19986645/38/7

4. **Flesch R.** The Art of Readable Writing. Harper & Row, 1949. 237 p.

5. **Kincaid J.P., Fishburne R.P., Rogers R.L., Chissom B.S.** Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station, 1975. 40 p.

6. **McLaughlin G.H.** SMOG Grading – a New Readability Formula // Journal of Reading. 1969. 12 (8). P. 639–646.

7. **Пиотровский Р.Г., Бектаев К.Б., Пиотровская А.А.** Математическая лингвистика. М.: Высшая школа, 1977. 383 с.

8. **Мацковский М.С.** Проблемы читабельности печатного материала // Смысловое восприятие речевого сообщения в условиях массовой коммуникации. М., 1976. С. 126–142.

9. **Тулдава Ю.А.** Об измерении трудности текстов // Учен. зап. Тарт. ун-та: Труды по методике преподавания иностранных языков. 1975. Вып. 345. С. 102–120.

10. **Оборнева И.В.** Автоматизированная оценка сложности учебных текстов на основе статистических параметров: дис. ... канд. пед. наук. М., 2006. 165 с.

11. **Белов С.А., Гулида В.Б.** Язык юридических документов: сложности понимания // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН. Т. 15. Ч. 1. 2019. С. 56–103. DOI 10.30842/alp2306573715104

12. **Blinova O.V., Belov S.A.** Legal corpus «CorRIDA» and lexical complexity assessment of Russian official texts // Book of Abstracts of the 1st International Conference of the Austrian Association for Legal Linguistics "Contemporary Approaches to Legal Linguistics", University of Vienna, 2019. P. 42.

13. **Блинова О.В., Тарасов Н.А.** Сложность русских правовых текстов: методы оценки и языковые данные // Труды международной конференции «Корпусная лингвистика-2021». СПб.: Скифия-принт, 2021. С. 175–182.

14. **Лапошина А.Н.** Автоматическое определение сложности текста по РКИ // Сборник материалов международной научно-практической интернет-конференции «Актуальные вопросы описания и преподавания русского языка как иностранного/неродного». М., 2018. С.573–579

15. **Laposhina A.N., Veselovskaya T.S., Lebedeva M.U., Kupreshchenko O.F.** Automated Text Readability Assessment For Russian Second Language Learners // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Issue 17 (24). 2018. P. 403–413. DOI: 10.22363/2618-8163-2021-19-3-331-345

16. **Се Линьи, Загайнов А.И.** Моделирование характеристик персонажей и их взаимосвязей в сюжете художественного произведения методами численного фрактального анализа // Terra Linguistica. 2022. Т. 13, № 3. С. 36–47. DOI: 10.18721/JHSS.13304

17. **Митрофанова О.А., Гаврилик Д.А.** Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Т. 13, № 4. С. 22–40. DOI: 10.18721/JHSS.13402

18. **Андреев В.С**. Экспоненциальное распределение частей речи в стихотворном тексте: опыт стилеметрического анализа // Общество. Коммуникация. Образование. 2021. Т. 12, № 4. С. 94–104. DOI: 10.18721/JHSS.12407

19. **Лапошина А.Н., Лебедева М.Ю.** Текстометр: онлайн-инструмент определения уровня сложности текста по русскому языку как иностранному // Русистика. 2021. Т. 19. №3. С. 331–345. DOI: 10.22363/2618-8163-2021-19-3-331-345

20. **Solovyev V., Ivanov V., Solnyshkina M.** Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics // Journal of Intelligent & Fuzzy Systems. 2018. Vol. 34 (5). Pp. 3049–3058. DOI: 10.3233/JIFS-169489

21. **Solovyev V., Solnyshkina M., Ivanov V.** Prediction of reading difficulty in Russian academic texts // Journal of Intelligent & Fuzzy Systems. 2019. Vol. 36. Is. 5. P. 4553–4563. DOI: 10.3233/JIFS-179007

22. **Solnyshkina M., Ivanov V., Solovyev V.** Readability Formula for Russian Texts: A Modified Version // Proceedings of the 17th Mexican International Conference on Artificial Intelligence. Guadalajara. 2018. Part II. P. 132–145. DOI: 10.1007/978-3-030-04497-8_11

23. **Томина Ю.А.** Объективная оценка языковой трудности текстов (описание, повествование, рассуждение, доказательство): дис. ... канд. пед. наук. М., 1985. 226 с.

24. **Валгина Н.С.** Теория текста. Москва.: Логос, 2003. 173 с

25. **Блинова О.В., Тарасов Н.А.** Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности //Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21, дополнительный том. 2022. С. 1017–1028

26. **James G., Witten D., Hastie T., Tibshirani R.** An Introduction to Statistical Learning. In Springer Texts in Statistics. Springer New York. 2013. 426 p. DOI: 10.1007/978-1-4614-7138-7

REFERENCES

[1] **A. Bastardas-Boada,** From language shift to language revitalization and sustainability. A complexity approach to linguistic ecology. Barcelona: Edicions de la Universitat de Barcelona, 2019, pp. 337–349.

[2] **Ö. Dahl,** The growth and maintenance of linguistic complexity. Amsterdam: John Benjamins, 2004.

[3] **S.I. Solnyshkina, A.S. Kiselnikov,** Slozhnost teksta: etapy izucheniya v otechestvennom prikladnom yazykoznanii, Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya. 6 (38) (2015) 86–99. DOI: 10.17223/19986645/38/7

[4] **R. Flesch,** The Art of Readable Writing. Harper & Row, 1949.

[5] **J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom,** Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy enlisted personnel (Research Branch Report 8-75). Memphis, TN: Naval Air Station, 1975.

[6] **G.H. McLaughlin,** SMOG Grading – a New Readability Formula, Journal of Reading. 12 (8) (1969) 639–646.

[7] **R.G. Piotrovskiy, K.B. Bektaev, A.A. Piotrovskaya,** Matematicheskaya lingvistika. Vysshaya shkola, Moscow, 1977.

[8] **M.S. Matskovskiy,** Problemy chitabelnosti pechatnogo materiala, Smyslovoye vospriyatiye rechevogo soobshcheniya v usloviyakh massovoy kommunikatsii. M., 1976. Pp. 126–142.

[9] **Yu.A. Tuldava,** Ob izmerenii trudnosti tekstov, Uchen. zap. Tart. un-ta: Trudy po metodike prepodavaniya inostrannykh yazykov. 345 (1975) 102–120.

[10] **I.V. Oborneva,** Avtomatizirovannaya otsenka slozhnosti uchebnykh tekstov na osnove statisticheskikh parametrov: dis. ... kand. ped. nauk. M., 2006. 165 p.

[11] **S.A. Belov, V.B. Gulida,** Yazyk yuridicheskikh dokumentov: slozhnosti ponimaniya, Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovaniy RAN. 15 (1) (2019) 56–103. DOI 10.30842/alp2306573715104

[12] **O.V. Blinova, S.A. Belov,** Legal corpus "CorRIDA" and lexical complexity assessment of Russian official texts, Book of Abstracts of the 1st International Conference of the Austrian Association for Legal Linguistics "Contemporary Approaches to Legal Linguistics", University of Vienna, 2019. P. 42.

[13] **O.V. Blinova, N.A. Tarasov,** Slozhnost russkikh pravovykh tekstov: metody otsenki i yazykovyye dannyye [The complexity of Russian legal texts: assessment methods and language data], Proceedings of the International Conference "Corpus Linguistics-2021". SPb.: Skifiya-print, 2021. Pp. 175–182.

[14] **A.N. Laposhina,** Avtomaticheskoye opredeleniye slozhnosti teksta po RKI [Automatic determination of the complexity of the text by the RCT], Collection of materials of the international scientific and practical Internet conference "Topical issues of describing and teaching Russian as a foreign/non-native language". M., 2018. Pp.573–579.

[15] **A.N. Laposhina, T.S. Veselovskaya, M.U. Lebedeva, O.F. Kupreshchenko,** Automated Text Readability Assessment For Russian Second Language Learners, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018". Issue 17 (24) (2018) 403–413. DOI: 10.22363/2618-8163-2021-19-3-331-345

[16] **Xie Linyi, A.I. Zagaynov,** Modeling of character characteristics and their relationships in a novel plot by methods of numerical fractal analysis, Terra Linguistica, 13 (3) (2022) 36–47. DOI: 10.18721/JHSS.13304

[17] **O.A. Mitrofanova, D.A. Gavrilic,** Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts, Terra Linguistica, 13 (4) (2022) 22–40. DOI: 10.18721/JHSS.13402

[18] **V.S. Andreev,** Exponential distribution of parts of speech in verse text: experience in stylometric analysis, Society. Communication. Education, 12 (4) (2021) 94–104. DOI: 10.18721/JHSS.12407

[19] **A.N. Laposhina, M.Yu. Lebedeva,** Tekstometr: onlayn-instrument opredeleniya urovnya slozhnosti teksta po russkomu yazyku kak inostrannomu, Rusistika. 19 (3) (2021) 331–345. DOI: 10.22363/2618-8163-2021-19-3-331-345

[20] **V. Solovyev, V. Ivanov, M. Solnyshkina,** Assessment of reading difficulty levels in Russian academic texts: Approaches and metrics, Journal of Intelligent & Fuzzy Systems. 34 (5) (2018) 3049–3058. DOI: 10.3233/JIFS-169489

[21] **V. Solovyev, M. Solnyshkina, V. Ivanov,** Prediction of reading difficulty in Russian academic texts, Journal of Intelligent & Fuzzy Systems. 36 (5) (2019) 4553–4563. DOI: 10.3233/JIFS-179007

[22] **M. Solnyshkina, V. Ivanov, V. Solovyev,** Readability Formula for Russian Texts: A Modified Version, Proceedings of the 17th Mexican International Conference on Artificial Intelligence. Guadalajara. 2018. Part II. P. 132–145. DOI: 10.1007/978-3-030-04497-8_11

[23] **Yu.A. Tomina,** Obyektivnaya otsenka yazykovoĭ trudnosti tekstov [Objective assessment of the linguistic difficulty of texts] (description, narration, reasoning, proof): dis. ... Candidate of Pedagogical Sciences. M., 1985. 226 p.

[24] **N.S. Valgina,** Teoriya teksta [Text theory]. Moskva.: Logos, 2003. 173 p.

[25] **O.V. Blinova, N.A. Tarasov,** Metriki slozhnosti russkikh pravovykh tekstov: otbor, ispolzovaniye, pervichnaya otsenka effektivnosti [Metrics of complexity of Russian legal texts: selection, use, primary evaluation of effectiveness], Computational linguistics and intelligent technologies: Based on the materials of the annual international conference "Dialogue". Vyp. 21, dopolnitelnyy tom. 2022. Pp. 1017–1028.

[26] **G. James, D. Witten, T. Hastie, R. Tibshirani,** An Introduction to Statistical Learning. In Springer Texts in Statistics. Springer New York. 2013. 426 p. DOI: 10.1007/978-1-4614-7138-7

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Евтушенко Татьяна Геннадьевна
Tatiana G. Evtushenko
E-mail: evtushenkotg@gmail.com
ORCID: https://orcid.org/0000-0001-5338-3656

Клочкова Елена Сергеевна
Yelena S. Klochkova
E-mail: klochkova_es@spbstu.ru
ORCID: https://orcid.org/0000-0002-6326-8392

Лапутенко Андрей Владимирович
Andrey V. Laputenko
E-mail: laputenko.av@gmail.com

Евтушенко Нина Владимировна
Nina V. Evtushenko
E-mail: evtushenko@ispras.ru
ORCID: https://orcid.org/0000-0002-4006-1161

Поступила: 27.01.2023; Одобрена: 06.03.2023; Принята: 17.03.2023.
Submitted: 27.01.2023; Approved: 06.03.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 41–56.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 41–56.

Научная статья
УДК 8'33
DOI: https://doi.org/10.18721/JHSS.14105

МАШИННЫЙ ПЕРЕВОД В ЭПОХУ ЦИФРОВИЗАЦИИ: НОВЫЕ ПРАКТИКИ, ПРОЦЕДУРЫ И РЕСУРСЫ

О.Н. Камшилова ✉ , **Л.Н. Беляева**

Российский государственный педагогический университет им. А.И. Герцена, Санкт-Петербург, Российская Федерация

✉ onkamshilova@gmail.com

Аннотация. В российской лингвистике к условиям цифровизации традиционно относят применение математических и компьютерных методов для решения различных задач, в основном связанных с проблемами обработки текста в различных автоматизированных системах. В статье анализируется влияние процесса цифровизации на назначение и использование систем машинного перевода (МП) в современных условиях, описываются новые практики использования продуктов МП как в профессиональной переводческой деятельности, так и в рамках решения частных задач пользователей таких систем. Отмечаются объективные преимущества и недостатки МП с точки зрения практикующих переводчиков-профессионалов и простых пользователей. Рассматриваются новые условия работы переводчиков, их новые роли и навыки, определяемые влиянием цифровизации на работу с текстом. Специальное внимание уделяется постредактированию продуктов МП как новой области профессиональной деятельности переводчика, необходимой для обеспечения качественного перевода, извлечению корректной информации. Определяется объем необходимого и достаточного постредактирования при решении частных задач пользователями-непрофессионалами. Анализируются доступные процедуры и лингвистические ресурсы, способные оптимизировать работу с системами МП.

Ключевые слова: цифровизация, машинный перевод (МП), системы МП, постредактирование, переводческие практики, лингвистические ресурсы.

Для цитирования: Камшилова О.Н., Беляева Л.Н. Машинный перевод в эпоху цифровизации: новые практики, процедуры и ресурсы // Terra Linguistica. 2023. Т. 14. № 1. С. 41–56. DOI: 10.18721/JHSS.14105

Research article
DOI: https://doi.org/10.18721/JHSS.14105

MACHINE TRANSLATION IN THE AGE OF DIGITALIZATION: NEW PRACTICES, PROCEDURES AND RESOURCES

O.N. Kamshilova ✉ , **L.N. Beliaeva**

Herzen State Pedagogical University of Russia,
St. Petersburg, Russian Federation

✉ onkamshilova@gmail.com

Abstract. In Russian linguistics digitalization is traditionally associated with the use of mathematical and computer methods applied mainly to text processing problems in various automated systems. The article analyzes the impact of digitalization on the use and purpose of machine translation (MT) systems in modern conditions. It describes new practices of using MT products both by professional translators and by a general MT system user for their individual purpose. It highlights the objective advantages and disadvantages of MT application from the point of view of practicing professional translators and ordinary users as well. The article considers a translator's new working conditions, their new roles and skills determined by the impact of digitalization on working with text. It pays special attention to post-editing MT products as a translator's new professional activity, which is needed to ensure high-quality translation and to extract correct information. It also describes the necessary and sufficient post-editing procedures to be performed by non-professional users while pursuing their own goals through MT application. Finally, the research focuses on the analysis of procedures and available linguistic resources that can optimize working with MT systems.

Keywords: digitalization, machine translation (MT), MT systems, post-editing, translation practices, linguistic resources.

Citation: O.N. Kamshilova, L.N. Beliaeva, Machine translation in the age of digitalization: new practices, procedures and resources, Terra Linguistica, 14 (1) (2023) 41–56. DOI: 10.18721/JHSS.14105

Введение

В современном цифроориентированном мире, благодаря изменению геополитической и социокультурной ситуаций, возникновению все новых возможностей представления информации в цифровой среде, разработке телекоммуникационных систем и технологий, новых средств связи, систем автоматического анализа и обработки информации, кардинально меняется как само представление о тексте, так и представление о способах и возможностях работы с текстами на естественных и искусственных языках.

Процесс цифровизации (что бы конкретно не понималось под этим процессом, ср., например, [1]) и особенности его реализации определяют необходимость учета целого комплекса составляющих: технических, связанных с выбором соответствующих технологий, а также собственно лингвистических, определяющих особенности лингвистического обеспечения всего применяемого комплекса методов и технологий.

В российской лингвистике к условиям цифровизации традиционно относят применение математических и компьютерных методов для решения различных задач, в основном связанных с проблемами обработки текста в различных автоматизированных системах или, вообще, к приложению лингвистических знаний в других сферах деятельности. В связи с задачами настоящей статьи выделим особо сферу машинного перевода (МП) и, в частности, работу с продуктами МП.

Технические составляющие процесса цифровизации ограничиваются и определяются учетом оснащенности различных учреждений компьютерами, способными поддерживать процесс диа-

лога и дистанционного обмена данными. Собственно *лингвистические составляющие* включают использование специальных предметно-ориентированных словарных систем и баз данных, локализацию программного обеспечения специализированных систем поддержки обучения, формирование специализированных корпусов текстов, применение систем МП в процессе локализации и лексикографирования.

Широкое развитие систем МП, их полная доступность делают бессмысленной борьбу с их использованием как в работе переводчиков и специалистов в различных областях знаний, так и при обучении иностранным языкам и при подготовке переводчиков. Более целесообразным и актуальным представляется включение работы с системами МП и постредактирования продуктов МП в требования к профессиональной подготовке (ср. [2–3]). Задачей настоящего исследования является анализ лингвистических процедур и доступных ресурсов работы с системами МП, способных обеспечить техническую и лингвистическую составляющие функционала специалистов по работе с текстом.

Методология

В основе исследования лежит критический анализ назначения и использования систем МП в современных условиях, обзор специальных условий, необходимых как для работы с системами МП, так и для работы по редактированию (постредактированию) результатов МП, сравнительный анализ ресурсов и процедур работы с системами МП, обеспечивающих успешное использование МП для решения профессиональных и пользовательских задач разного рода (исследовательских, образовательных, информационных).

Результаты исследования

Назначение и использование систем МП в современных практиках перевода

Основным назначением систем машинного перевода (МП) является оперативный перевод специальных (научных, научно-технических, информационных) текстов. Качество результатов работы такой системы определяется требованиями его потребителей: специалистов в конкретной области знаний и переводчиков. Целесообразность использования систем МП установлена давно и не требует новых доказательств. Применение систем МП в рамках научного и технического перевода обеспечивает, во-первых, выигрыш в продуктивности за счет значительного сокращения временных затрат; во-вторых, выигрыш в качестве за счет согласованности и корректности перевода терминологии при использовании предметно-ориентированных систем; в-третьих, выигрыш в удобстве организации работ за счет возможности контроля над переводом потоков текстов и возможности разделения работы между несколькими переводчиками.

Очевидные успехи в области разработки систем МП привели к тому, что сегодня можно наблюдать серьезные изменения в их назначении и использовании: если *исторически* машинный перевод был ориентирован на специальные тексты, а пользователями этих систем были переводчики или специалисты отдельных предметных областей, то сегодня в практике МП складывается ряд *новых направлений*. В частности, перевод литературных произведений и субтитров, что не удивительно, поскольку применение МП в этой сфере дает 36% увеличения производительности труда переводчиков в соотношении «количество переведенных слов на единицу времени» [4–5]. Пользователем систем МП в данном случае по-прежнему остается переводчик, однако приходится констатировать низкое качество перевода, отсутствие редактирования результатов МП, что можно объяснить большим спросом на массовую продукцию (массовая литература, кино и сериалы на многочисленных частных каналах и т.п.), в результате потребитель получает некачественные переводы, демонстрирующие пренебрежение элементарными правилами перевода вроде всем известных «ложных друзей» переводчика:

Алые губы подчеркивали пронзительную сексуальность и безупречную красоту актрисы, а в сочетании с роскошными драгоценностями, ***меховыми манто*** *и открывающими плечи* ***сатиновыми*** *платьями заявляли всему миру о ее славе, успехе и богатстве.*[1]

Еще одно направление и канал использования и распространения некачественного перевода – новостные агрегаторы, сетевые издания и каналы на платформах Telegram, YouTube и т.п, типа сайта AKKet.com:

Многие россияне воспользовались возможностью и купили товары от компании IKEA, получив ***такие*** *в пункте выдачи заказов или с доставкой на дом. Получать* ***такие*** *в* ***свои*** *руки покупатели будут до самого конца сентября <...> В рамках последней распродажи многие россияне купили себе* ***различные самые разные*** *товары, которые были доступны для покупки* ***до середине*** *августа <...> При этом магазины в какой-то момент также заработают, потому что иначе смысла сохранять* ***такие*** *и поддерживать их в* ***актуальном*** *состоянии* ***бы попросту не было****.*[2]

Об использовании МП в данном тексте свидетельствуют не только ошибки в предложном управлении, лексической сочетаемости, порядке слов, но и анафорическое употребление местоимения *такие* (ср. *such*, *these*), в совершенно не характерной для русского языка конструкции и, напротив, естественной для английского. Подобные неотредактированные тексты можно встретить на многих официальных каналах иностранных новостных агентств, публикующих материалы на русском языке. Все это свидетельствует о том, что редактирование переводных материалов приносится в жертву скорости публикации.

При этом нельзя не отметить, что качество результатов систем нейронного машинного перевода в последнее время повысилось кардинально, выявляя в рамках переводческих техник новое противопоставление: *machine translation* ↔ *translation from scratch*, т.е. противопоставление полного или частичного машинного перевода переводу ручному от начала до конца. При этом последний становится «штучным товаром», предполагая высочайшую квалификацию переводчика, а первый одинаково популярен как у переводческих компаний, выполняющих огромный объем работы, так и у широкого круга лиц – от переводчиков-фрилансеров до специалистов-исследователей, студентов, решающих академические задачи, и просто пользователей сети, предлагающей доступные инструменты типа Гугл-переводчика. Обращение широкого круга пользователей к применению результатов МП определяет третье направление в изменении характера практик МП.

Анализ материалов научных журналов, конференций, студенческих исследовательских работ позволяет заключить следующее: системы МП широко используются специалистами в различных областях знаний для анализа и перевода статей, извлекаемых из множества источников, в первую очередь из материалов конференций, представленных в системе Интернет, а также для перевода заглавий, аннотаций, ключевых слов и собственно текстов, написанных на родном языке автора, для публикации в отечественных и зарубежных журналах:

In article the heuristic potential categories "rationality" is considered, the basic approaches to its studying are shown. The author allocates initial characteristics and genetic roots of rationality, analyzes the reasons of a crisis state of idea "ratio" in modern humanitarian knowledge. The further prospects of theoretical judgement of a phenomenon of "rationality" are discussed.[3]

При этом «новые пользователи» демонстрируют следующие типичные недостатки в работе с результатами МП:

- не знают процедур работы с такими системами;
- не осознают необходимости постредактирования и возможности минимизации объема постредактирования результатов МП;

[1] Rachel Felder. RED LIPSTICK – An Ode to a Beauty Icon, перевод «Секретное оружие : история красной помады / Рейчел Фелдер ; [перевод с английского А. А. Джапаридзе»: Эксмо; Москва; 2021, Научный редактор Анна Жуковская

[2] https://akket.com – публикация от 08.09.22

[3] Аннотация к научной статье, поданной в рецензируемый журнал

• не понимают связи объемов и методов постредактирования с полнотой и точностью автоматического словаря соответствующей системы.

При всех определенных успехах современных систем МП отношение к ним весьма неоднозначное как у переводчиков-профессионалов, так и в академической среде.

Более 38% практикующих переводчиков продолжают опасаться использования МП как возможного конкурента, который сократит объем их работы или вообще заменит их [6]. Кроме того, объем ежедневного постредактирования, выполняемого переводчиками, в среднем составляет приблизительно 5 000 слов, этот процесс вызывает неприятие переводчиков и отрицательное отношение к результатам МП в целом. Новые характеристики самого процесса перевода и требований к его качеству и скорости получения результатов определяются еще и тем, что он включен в так называемую индустрию локализации (*Localization Industry*) [7–8], которая охватывает не только собственно перевод, но и адаптацию его результата к культуре принимающего языка, решение маркетинговых и технологических задач. При этом следует учитывать, что результаты обработки текстов на разных языках являются базой решения различных научных и практических задач, поэтому использование систем МП является важной частью перевода как технологического процесса.

В академической сфере борьба с использованием МП при подготовке студентами материалов на иностранном языке, в т.ч. на русском, если речь идет об обучении иностранных студентов и аспирантов, носит принципиальный характер. Преподаватели иностранных языков рассматривают использование таких систем студентами как совершенно недопустимое и неэтичное. Почти половина тех, кто изучает или преподает романские языки (испанский, французский, итальянский и португальский), считает применение МП непродуктивным, более того, рассматривают это как мошенничество [9]. При этом нельзя отрицать, что практичность, простота использования и бесплатный доступ к сетевым системам МП сделали эти инструменты очень популярными среди изучающих язык. Предваривший наше исследование опрос студентов и аспирантов филологического факультета РГПУ им. А.И. Герцена показал, что большинство использующих МП для перевода на английский язык предпочитают систему Google Translate и оценивают результат работы с этим инструментом достаточно высоко. Поэтому сегодня в академической среде наряду с этическим запретом к использованию МП наблюдается новая тенденция – поиск эффективных способов использования таких систем в обучении языкам [10].

Наш собственный опыт использования систем МП в обучении позволяет утверждать, что оценка эффективности и целесообразности использования МП зависят от типа изучаемого языка, а приемлемость и процедуры использования систем МП преподаватель должен оценивать сам, исходя из возможностей образовательной среды и уровня подготовки студентов.

Постредактирование как новый вид деятельности переводчика

Переводчик должен обладать знаниями в сфере владения терминологией предметной области, входного языка и языка перевода, межкультурных отношений, поиска информации, технологических инструментов и программ, а также в сфере поставки услуг переводчиков и др. Постредактирование как особый вид деятельности переводчика предполагает умения выполнять корректуру и редактировать перевод, а также умение устанавливать и контролировать стандарты качества. Однако, общая структура цифровизации сегодня предполагает изменение не только деятельности постредактора, но и процесса редактирования в целом.

Традиционно редактор рассматривается как специалист, основной задачей которого является *редактирование*, т.е. установление меры ценности готовящегося к изданию произведения и приведение его к стандартной форме, соответствующей стилистическим требованиям (ср. [11]). Последовательность работы редактора в той форме, в которой она определяется полученными в вузовском обучении знаниями, можно описать следующим образом: определение способа изложения (функционально-смыслового типа речи: описания, повествования, рассуждения) и его разновидности по конкретным признакам с опорой на их знание; установка

соответствия требованиям к способу построения, выбору языковых средств, поиск возможных недостатков и типичных ошибок [12]. Под литературным редактированием понимается «совершенствование редактором формы литературного произведения (композиции, языка, стилистических качеств)» [13, с. 9].

Активно развивающийся практически во всех сферах деятельности процесс цифровизации изменил традиционный формат работы с текстом, делая обязательным освоение новых платформ, работу с интернет-версиями изданий и принятие форматов мультимедийной редакции. В современных условиях работа редактора предполагает учет этих новых возможностей, платформ и технологий, поскольку аудитория перешла в киберпространство, т.е. можно утверждать, что происходят изменения профессиональных компетенций редактора под воздействием как внешних, так и внутренних факторов [14]. Новый подход к созданию и ведению сайтов, к созданию журналистских текстов потребовал от редактора освоения новых ролей и навыков, не входивших ранее в сферу его базовых компетенций. У общей профессии редактора появились новые варианты: администратор сайта, контент-менеджер, аккаунт-менеджер, seo- и smm-специалисты (*social media marketing – smm*), редактор пользовательского контента, мультимедийный редактор, редактор-модератор (фасилитатор), редактор-рерайтер и множество других. Редактор-модератор организует коммуникацию с аудиторией, целью его работы является повышение уровня доверия к определённому СМИ. В функции редактора-рерайтера входит поддержание сайта конкретного СМИ в актуальном состоянии, постоянное обновление контента, оптимальное мультимедийное наполнение, а также интерактивное взаимодействие с аудиторией [15], см. также Программу переподготовки редакторов на факультете журналистики МГУ[4]. Следовательно, подготовка переводчиков, равно как и лингвистов (в широком смысле, включая подготовку многих профессий, предполагающих работу с текстом [16]) должна кардинально измениться.

Независимо от полноты автоматического словаря системы и его привязки к предметной области, результат МП требует постредактирования на уровне синтаксической структуры всего предложения, на лексическом уровне для уточнения и/или изменения перево¬дов отдельных слов и словосочетаний, введения переводов для незарегистрированных в словаре единиц, а также для изменения морфологических характеристик рода, числа, падежа, уточнения форм времени и залога, введения корректной пунктуации. При оценке трудоемкости этого процесса внесение стилистических изменений обычно не рассматривается. Проведенные исследования [17–18] показали, что отказ от включения МП в процесс перевода с последующим трудоемким редактированием больше свойственен профессиональным переводчикам, чем тем, кто еще только получает эту профессию. Возможно, это связано еще и с уровнем компьютерной грамотности опрошенных, а также с небольшим опытом перевода. Многолетний опыт собственной работы авторов показывает, что работа с постредактированием результатов МП оставляет простор для решения творческих и лингвистических задач. Обучение постредактированию результатов МП должно составлять обязательную часть подготовки современных специалистов и не только переводчиков.

Постредактирование результатов МП в решении частных задач

Использование систем МП в исследовательской работе требует особого подхода к получаемому результату и к объему необходимого и достаточного постредактирования, при этом особое внимание должно уделяться переводу слов, которые не зарегистрированы в автоматическом словаре используемой системы, являются неологизмами и могут включаться в словосочетания, чаще всего терминологические, называющие новые реалии – именованные сущности. Постредактирование на лексическом уровне требует уточнения и изменения переводов конкретных лексических единиц, на синтаксическом – преобразования структуры предложения. Например, в случаях перевода с английского языка на русский необходимы проверка согласования по роду, числу

[4] Программа профессиональной переподготовки «Редактор текстов для СМИ». Факультет журналистики МГУ им. М.В. Ломоносова (сайт). [Электронный ресурс]. URL: http://www.journ.msu. ru/education/extra/editor/

и падежу, уточнение места подлежащего, иногда полная перестройка предложения или переход к непрямой структуре типа *we have ^мы имеем *у нас есть*.

Если учесть, что при переводе научного или технического текста на выбор переводного эквивалента конкретного термина профессионалом-переводчиком затрачивается до 75% времени, необходимого для перевода текста в целом [19–21], то трудоемкий процесс обращения с иноязычной терминологией, формирования и описания переводных эквивалентов новых, вводимых в научный обиход терминов, оказывается исключен из процесса обучения студентов и аспирантов академическому письму, что при достаточно высоком уровне владения иностранным языком, позволяющем специалисту в определенной предметной области читать и понимать иноязычный научный текст даже при наличии беспереводных терминологических единиц, приводит к созданию научных текстов на безумной смеси языков, новом научном воляпюке. Вряд ли пример приведенного ниже фрагмента научного текста можно рассматривать как корректный текст на русском языке:

Это справедливо, например, для энвайронментализма, который, в свою очередь, был тесно связан с другими арт-практиками, в частности, ленд-артом, экологическим искусством, арте повера, био-артом, art&science.

Постредактирование результатов МП и получение окончательного варианта перевода текста требуют обращения к словарным и энциклопедическим базам данных. Сложности работы на этом уровне определяются тем, что использование систем машинного перевода специалистами в различных областях знаний, особенно использование предметно-ориентированных практических систем, действительно дает возможность понимания общего содержания текста, но не вводит непереведенные системой лексические единицы в систему лексики языка перевода, русского языка в частности, что определяет объем заимствований при создании текстов на русском языке в соответствующей предметной области. Тем самым создается база для неоправданного введения в русский язык массы новых слов, в частности англицизмов. Поскольку эти слова встречаются недостаточно устойчиво, их значение определено нечетко, а сами слова не введены в терминологические базы данных, их можно рассматривать как молодые, не полностью освоенные лексикой русского языка – лексические недопески – то ли терминоиды, то ли сленгизмы [22], то ли единицы общеупотребительной лексики.

Совершенно очевидно, что перевод стандартизированной терминологии и поиск эквивалентов для новых терминологических единиц остаются одной из самых трудоемких задач, решаемых и в процессе перевода, и в постредактировании результатов МП.

Ресурсы и процедуры для оптимизации работы с системами МП

В общепринятом сценарии постредактирования результатов машинного перевода человек-постредактор исправляет перевод, созданный системой МП. Так, например, множество систем переводческой памяти (Trados, MemoQ, OmegaT, и др.) обеспечивают доступ к системам МП, результаты работы которых могут быть объединены в совокупность переводов систем переводческой памяти. Следует иметь в виду, что в больших переводческих проектах часто возникает необходимость разделения текста большого объема между несколькими переводчиками. В этих условиях именно система МП, настроенная на конкретную предметную область, позволяет унифицировать перевод терминологии.

Недавно появилась возможность активного машинного обучения и взаимодействия [23], при котором система МП переупорядочивает сегменты исходного языка, которые должны переводиться и подвергаться постредактированию, чтобы максимизировать продуктивность и эффект от накопленного опыта. Вместо представления текста в первоначальной последовательности, система сортирует сегменты согласно степени доверия к полученным ранее результатам МП, так что она может научиться более корректно решать задачу перевода на основе исправлений, сделанных человеком. Это сопровождается новыми концептуализациями последовательности выполнения

работ по переводу и постредактированию, которые связывают эти новые возможности с инновационным использованием краудсорсинга.

Чтобы полностью использовать потенциал краудсорсинга, необходимо найти новые способы для того, чтобы делить связные тексты на фрагменты, которые независимо друг от друга несколько переводчиков могут переводить и редактировать одновременно. Под действием увеличившегося спроса на производительность перевода и на то, чтобы цикл перевода стал короче, известные инструментальные средства перевода (Wordbee, Trados, MATECAT) предлагают объединенные функциональные возможности.

Некоторые компании (Unbabel1, MotaWord2), которые используют технологии LSP, позволяющие подсчитывать и ограничивать траффик, регулировать скорость закачки и приоритеты, фильтровать и распределять контент, ищут возможности экспериментировать с более динамичным подходом к совместному переводу, при котором документ разбивается на меньшие единицы. Компания MotaWord, например, объявила, что разработала «самую быструю в мире платформу для ручного перевода», основанную на совместной облачной платформе, «эффективно координируемой через интеллектуальный внутренний интерфейс», в котором участвует более 9 000 переводчиков [24, с.11]. Это возможно только в том случае, если большие документы раздроблены на маленькие сегменты, чтобы постредактировать ограниченное количество меньших фрагментов текста с введением в действие «толпы» (crowd) переводчиков.

Однако, неясно, как переводчики справляются с ситуацией, в которой меньшие сегменты, возможно выбранные из разных частей одного и того же документа, представлены в отрыве от контекста. Влияние того, что переводчики переводят сегменты в порядке, не соответствующем исходному, на процесс перевода ранее не исследовалось.

Для совершенствования работы постредактора и оптимизации ее результата современные исследования процесса постредактирования предлагают методы слежения за положением глаз или зрачков постредактора и регистрацию сигналов от нажатий человеком клавиш на клавиатуре компьютера, а также метрики, принятые в области исследования процесса перевода.

Практическая работа переводчика с системой машинного перевода предусматривает:

- ручное предредактирование текста – подготовку исходного текста (массива текстов) к переводу;
- ручное постредактирование переводов – редактирование результатов работы системы МП;
- ведение собственного (пользовательского) словаря, фиксирующего результаты МП и определяющего настройку системы МП на задачи конкретного переводчика.

Предредактирование текста необходимо для установления единства используемой терминологии, например, в системах извлечения данных (*data mining systems*), при работе с которыми некорректные результаты часто возникают из-за расхождения между данными, извлекаемыми из текста, и номинацией соответствующих объектов в словарном обеспечении (базах данных или онтологиях). В задачи предредактирования должно включаться исправление ошибок в текстах на глобальном английском языке, в частности, в излишней пассивизации текста, и упрощение текста в связи с решением задач перевода и инженерии знаний.

Постредактирование результатов МП и получение окончательного варианта перевода текста требует обращения к словарным и энциклопедическим базам данных, корпусам текстов, заранее выбранным переводчиком. При решении вопроса о выборе перевода конкретной терминологической единицы необходимо привлечение миниконкорданса. В результате такой работы на этапе собственно перевода должен формироваться пользовательский словарь, характеризующий терминологические особенности конкретного текста. Этот словарь на этапе поддержки системы добавляется в ее лингвистические ресурсы.

В процессе перевода в режиме реального времени текст может предварительно обрабатываться, при этом:

• получение частотного словаря и миниконкорданса по конкретному тексту помогает выявить основную терминологию и установить ее контекст,

• предварительное редактирование текста позволяет снять его стилистические несообразности, устранить сверхдлинные предложения и т.д.

• использование системы машинного перевода, выбранной и настроенной на необходимую предметную область, дает вариант перевода, который требует анализа и постредактирования.

Собственно процесс редактирования перевода может в этом случае осуществляться либо в стандартном двухоконном интерфейсе Windows, когда исходный текст и перевод представлены в двух параллельных окнах, что позволяет редактировать и заново переводить оба текста, либо в режиме работы только с текстом перевода. Опыт практической работы с системами МП позволяет рекомендовать именно первый вариант как наиболее удобный.

Таким образом, после завершения перевода конкретного текста должна происходить перенастройка лингвистических ресурсов: пополняется корпус параллельных текстов за счет исходного текста и его перевода, формируется и/или пополняется пользовательский словарь, включающий терминологию, выявленную и проверенную переводчиком, пополняется база словарей.

Примером специально разработанной для решения исследовательских задач в области характеристик результатов систем МП является электронная платформа E-NBU, созданная в Лаборатории языковых технологий Нового болгарского университета (New Bulgarian University). Эта платформа изначально разрабатывалась как инструмент для преподавания языка и представляла собой генератор онлайновых упражнений, извлекаемых из аннотированных корпусов с экспортом в систему Moodle или другие образовательные платформы. Платформа E-NBU в настоящее время расширена дополнительными модулями и функциональными возможностями, позволяющими проводить исследования в области перевода и анализа ошибок, а также поддерживать лексикографические проекты.

Другой тип электронной платформы – это достаточно давно разрабатываемая в РГПУ им. А.И. Герцена модель автоматизированного рабочего места (АРМ) преподавателя и переводчика [25–26]. АРМ преподавателя, включает

1) среду для создания, организации и ведения электронных текстовых архивов и извлечения корпусов текстов;

2) модули для проведения лингвистического анализа: лемматизатор, анализатор частей речи, приписывающий словоформам или словосочетаниям соответствующие теги; анализатор терминов; морфологический анализатор, синтаксический анализатор; анализатор многокомпонентных лексических единиц (включая сложные термины, аналитические формы, фразеологические единицы); средство выравнивания параллельных текстов; конкордансер;

3) лингвистическую базу данных, позволяющую работать с корпусом без потери информации;

4) модули, предназначенные для создания и редактирования онлайновых упражнений.

Специальные средства для ведения электронного текстового архива позволяют формировать различные метаданные, которые могут, по отдельности или в комбинациях, составлять основу для построения корпусов текстов. После проведения лингвистического анализа, необходимого для создания корпуса текстов, из него могут извлекаться вторичные («виртуальные») корпусы: списки предложений, содержащих определенную единицу – лемму, словоформу, словосочетание, тег или комбинацию тегов. Применяемая архитектура (см. рис. 1) позволяет параллельно использовать несколько систем предварительной обработки и сравнивать их результаты, что превращает такое АРМ в удобную среду для проведения экспериментов и исследований [27], которая может подключаться к системе Moodle.

Большое число лингвистических ресурсов сегодня представляет собой базы и банки данных, а также корпусы текстов, предназначенные в том числе для обучения языкам и подготовки переводчиков. К числу таких ресурсов относятся:

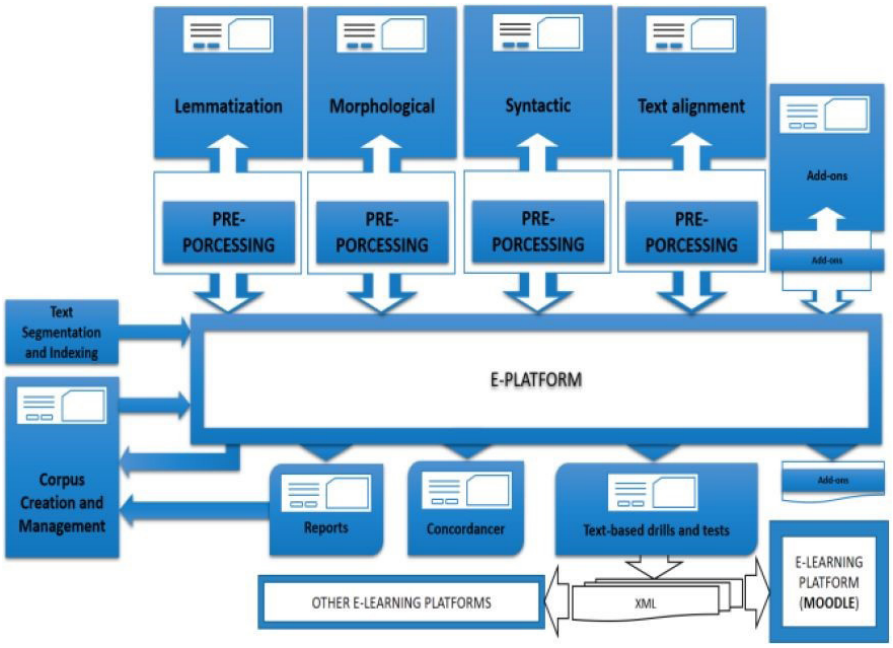

Рис. 1. Архитектура платформы E-Platform [27, с.100]

Fig. 1. Architecture of the E-Platform [27, p. 100]

• Ресурс *TAUS Data Cloud* (https://www.taus.net/data/taus-data-cloud) – крупнейшее хранилище переводческой памяти, объем которого превышает 79 миллиардов слов в более чем 2200 языковых пар. Этот ресурс в основном используется для обучения системам МТ как в промышленности, так и в научных и учебных кругах. Пользователи могут извлекать данные, если у них достаточно кредитов, которые можно либо купить, либо заработать, загрузив собственные переводческие материалы. Данные доступны для поиска на основе относительно небольшого набора меток: язык источника и язык перевода, предметная область, тип контента, владелец данных и поставщик данных.

• Ресурс *My Memory* (http://mymemory.translated.net), управляемый компанией LSP Translated, включает в себя программное обеспечение для управления терминологическими данными Европейского союза и Организации Объединенных Наций, а также данные, полученные с многоязычных веб-сайтов. Его основными пользователями являются те, чья работы связана с лингвистическими технологиями, те, кто занимается обработкой текстов на естественных языках, и переводчики. Загрузка массивов переводческой памяти является бесплатной. Поиск на основе метаданных ограничен языковыми парами и предметными областями.

• Ресурс *European Parliament Proceedings Parallel Corpus* (Параллельный корпус слушаний в Европейском парламенте [28]) характеризуется большим размером, многоязычностью и выравниванием по предложениям. Однако в него включен ограниченный круг предметных областей и типов текста. Метаданные очень скудны (языки и направления перевода), что ограничивает его полезность для учебных целей.

• Европейская комиссия предоставляет свободный доступ к ряду своих крупных многоязычных систем переводческой памяти и корпусов из областей политики, права и экономики (https://ec.europa.eu/jrc/en/language-technologies). Эти ресурсы часто используются для обучения систем МП и для поддержки систем переводческой памяти в конкретных предметных областях. Метаданные также скудны.

• Ресурс *OPUS* [29] представляет собой бесплатную, постоянно растущую коллекцию уже существующих, автоматически обогащаемых (например, тегами частей речи) корпусов и систем

переводческой памяти, извлеченных из Интернета. Его основными преимуществами являются большой размер, множество языковых пар, текстовое разнообразие с точки зрения предметных областей и типов текстов, разнообразие способов перевода (письменный перевод, локализация и субтитры), а также его открытый характер. С другой стороны, метаданных в нем мало: они предоставляют различные, но небольшие наборы меток в зависимости от соответствующего корпуса, доступ к которому осуществляется через интерфейс OPUS, например, корпус EuroParl. Благодаря своему размеру, вариативности и бесплатной доступности он может быть полезным для учебных целей.

• Ресурс *Translational English Corpus* (http://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-englishcorpus-tec/) включает в себя несколько субкорпусов, состоящих из различных видов текстов, и сообщает множество метаданных об экстралингвистических параметрах текстов, включая данные о переводчиках, эти тексты создавших. В отличие от других ресурсов, этот корпус является не параллельным, а одноязычным сопоставимым корпусом, противопоставляющим оригинальный текст на английском языке английскому языку перевода. Использование его данных оказало влияние на теорию перевода, особенно в том, что касается исследований переводческих и стилистических вариаций. Несомненна его значимость и для развития навыка постредактирования.

• Ресурс *MeLLANGE Learner Translator Corpus* (http://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-englishcorpus-tec/) является многоязычным учебным корпусом и включает не профессиональные переводы, а переводы, выполненные переводчиками-стажерами. Несмотря на сравнительно небольшой размер, он предоставляет множество специфичных для перевода метаданных на различных уровнях, включая информацию о профилях переводчиков и процессах перевода (например, затраченное время и тип используемых средств перевода). Он лучше всего подходит для изучения дидактических аспектов перевода, качества перевода и предоставляет обширный материал для работы с постредактированием.

• Ресурс *Human Language Project* [30] представляет собой попытку создания универсального корпуса для всех языков мира для изучения языковых универсалий и документации на естественных языках.

К сожалению, не существует универсального хранилища параллельных переводческих данных, подходящего для максимально широкого спектра задач, связанных с обучением языкам, переводу и постредактированию.

Заключение

МП сегодня становится регулярной и неотъемлемой частью переводческих практик. Кроме того, доступные в сети системы МП активно используются для решения частных задач – информационных, исследовательских, образовательных.

В самой практике МП намечаются новые направления его применения: от перевода сугубо специальных текстов научно-технического характера к переводу текстов художественных, рекламных, новостных.

Новые условия работы переводчика и лингвиста, а также требования к их профессиональным компетенциям, продиктованные влиянием цифровизации на работу с текстом, делают работу с системами МП неизбежной. Не пользовательские, но профессиональные навыки обращения с продуктами МП, их постредактированием должны стать основой работы специалиста.

Ресурсы и процедуры работы с системами МП, проанализированные в этом исследовании, способны обеспечить техническую и лингвистическую составляющие процесса использования систем МП, прежде всего для переводчиков и редакторов, а также исследователей, работающих с иноязычными источниками и участвующих в международном научном обмене.

Сегодня спектр систем МП постоянно расширяется и совершенствуется, и с точки зрения доступных пар языков, и по качеству результатов. Поэтому позиция неприятия результатов их работы и неучета их использования непродуктивна. Простое запрещение применения МП неэффективно, поскольку пользователи продолжают обращаться к системам МП независимо от вводимых запретов.

Материалы статьи были представлены на IV Международной конференции по инженерной и прикладной лингвистике «Пиотровские Чтения – 2022», посвященной 100-летию со дня рождения профессора Р.Г. Пиотровского в РГПУ им. А.И. Герцена 22 ноября 2022 г.

СПИСОК ИСТОЧНИКОВ

1. **Стариченко Б.Е.** Цифровизация образования: иллюзии и ожидания // Педагогическое образование в России. 2020. № 3. С. 49–586. URL: https://cyberleninka.ru/article/n/tsifrovizatsiya-obrazovaniya-illyuzii-i-ozhidaniya?ysclid=ldvzz4y1cs241893073

2. **Абросимова Н.А., Щелокова Е.А.** Пред- и постредактирование машинного перевода медицинских текстов // Мир науки, культуры, образования. 2022. № 4 (95). С. 178–181. URL: https://cyberleninka.ru/article/n/pred-i-postredaktirovanie-mashinnogo-perevoda-meditsinskih-tekstov?ysclid=ldw44lkycw553753824

3. **Нечаева Н.В., Светова С.Ю.** Постредактирование машинного перевода как актуальное направление подготовки переводчиков в вузах // Вопросы методики преподавания в вузе. 2018. Т. 7. № 25. С. 64–73. URL: https://cyberleninka.ru/article/n/postredaktirovanie-mashinnogo-perevoda-kak-aktualnoe-napravlenie-podgotovki-perevodchikov-v-vuzah/viewer

4. **Toral A., Castilho S., Hu K., Way A.** Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation, 2018. URL: https://arxiv.org/abs/1808.10432

5. **Toral A., Oliver A., Ribas-Bellestín P.** Machine Translation of Novels in the Age of Transformer // Maschinelle Übersetzung für Übersetzungsprofis, edited by Jörg Porsiel. Berlin: BDÜ Fachverlag, 2020. Pp. 276–296. URL: https://www.researchgate.net/publication/346557116

6. **de Almeida G., O'Brien S.** Analysing Post-Editing Performance: Correlations with Years of Translation Experience // Proceedings of the 14th Annual conference of the European Association for Machine Translation, May 27-28, 2010, Saint Raphael, France – European Association for Machine Translation, 2010. URL: https://www.aclweb.org/anthology/2010.eamt-1.19

7. **Ачкасов А.В.** Индустрия локализации и подготовка переводчиков // Материалы VIII Международной научной конференции «Индустрия перевода». Пермь, 2016. С. 13–19. URL: https://elibrary.ru/item.asp?id=26181052&ysclid=ldwxzskzkc395978920

8. **Беляева Л.Н.** Лингвистические технологии в современном сетевом пространстве: language worker в индустрии локализации. СПб.: Книжный дом, 2016. 134 с. URL: https://search.rsl.ru/ru/record/01008793133?ysclid=ldw1rssz1875247094

9. **Koponen M., Sulubacak U., Vitikainen K., Tiedemann J.** MT for subtitling: User evaluation of post-editing productivity // EAMT-2020, Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. Lisboa, Portugal, 2020. Pp.115–124. URL: https://www.researchgate.net/publication/348688005

10. **Alves N., Marques M., Guimaraes P., Almeida J.A, Canario R.** Education and Training Courses in Portugal: Another kind of schooling? 2021. URL: https://www.researchgate.net/publication/351050875

11. **Стефанов С.И.** Реклама и полиграфия: опыт словаря-справочника. М.: Гелла принт, 2004. 340 с. URL: https://search.rsl.ru/ru/record/01002486074?ysclid=ldw3y0ofmq501451819

12. **Колесникова О.И.** Обучение студентов литературному редактированию в контексте развития художественно-языковой компетенции // Ярославский педагогический вестник. 2016. № 2. С. 60–63. URL: http://vestnik.yspu.org/releases/2016_2/14.pdf

13. **Накорякова К.М.** Справочник по литературному редактированию для работников массовой информации. М.: Флинта: Наука, 2010. 200 с. URL: https://search.rsl.ru/ru/record/01004396714?ysclid=ldw3w40ial722575012

14. **Хлопунова О.В., Цаканян А.А.** Профессиональные компетенции современного редактора: лидер или менеджер // Вестник Волжского университета имени В.Н. Татищева. 2019. № 4 (1).

С. 1–10. URL: https://cyberleninka.ru/article/n/professionalnye-kompetentsii-sovremennogo-redaktora-lider-ili-menedzher?ysclid=ldw1g8crls909581657

15. **Фролова В.И.** О меняющейся роли редактора в эпоху обновления интернет-коммуникации // Вестник Волжского университета им. В.Н. Татищева. 2015. № 4 (19). С. 51–56. URL: https://cyberleninka.ru/article/n/o-menyayuscheysya-roli-redaktora-v-epohu-obnovleniya-internet-kommunikatsii?ysclid=ldw1l592hr40120245

16. **Beliaeva L.N., Kamshilova O.N.** Problems and Perspectives of Language Worker Professional Training // International Journal of Open Information Technologies. 2018. Vol. 6. № 3. Pp. 35–42. URL: http://injoit.org/index.php/j1/article/view/641

17. **Moorkens J., O'Brien S.** Post-Editing Evaluations: Trade-offs between Novice and Professional Participants // EAMT 2015. Proceedings of the 18th Annual Conference of the European Association for Machine Translation. Antalya, Turkey, May 11-13, 2015. Pp. 75–81. URL: https://www.researchgate.net/publication/275031846

18. **Zaretskaya A.** The Use of Machine Translation among Professional Translators // Proceedings of the EXPERT Scientific and Technological Workshop, Malaga, 26th and 27th June 2015: Editions Tradulex, Geneva, 2015. Pp. 1–13. URL: https://www.researchgate.net/publication/283667234

19. **Кудашев И.С.** Проектирование переводческих словарей специальной лексики. Helsinki: Univ. of Helsinki, Dep. of transl. studies, 2007. 443 с. URL: https://helda.helsinki.fi/bitstream/handle/10138/19272/proektir.pdf

20. **Vasiljevs A., Pinnis M., Gornostay T.** Service model for semi-automatic generation of multilingual terminology resources // Terminology and Knowledge Engineering. 2014. Pp. 67–76. URL: https://www.researchgate.net/publication/266565145

21. **Vieira L.N.** Post-Editing of Machine Translation // M. O'Hagan (Ed.), The Routledge Handbook of Translation and Technology. Routledge, 2019. Pp. 319–335. URL: http://www.routledge.com/9781138232846

22. **Фельде О.В.** Эффективное речевое общение (базовые компетенции). Словарь-справочник под редакцией А.П. Сковородникова. Красноярск: Сибирский федеральный университет, 2012. URL: http://elib.sfu-kras.ru/handle/2311/63722

23. **Ortiz-Martinez D.** Online Learning for Statistical Machine Translation // Computational Linguistics. 42 (1). 2016. Pp. 121–161. URL: http://dx.doi.org/10.1162/COLI_a_00244

24. **Báez C.T., Schaeffer M., Carl M.** Experiments in Non-Coherent Post-editing // The Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT). Varna, Bulgaria, Sept 7, 2017. Pp. 11–20. URL: https://www.researchgate.net/publication/322031657

25. **Беляева Л.Н., Джепа Т.Л., Зак Г.Н., Камшилова О.Н., Нымм В.Р., Разумова В.В.** Автоматизированное рабочее место филолога в структуре образовательного пространства современного вуза / коллективная монография. СПб.: Книжный дом, 2013. 127 с. URL: https://search.rsl.ru/ru/record/01006589603?ysclid=ldw33c4dsz595025912

26. **Беляева Л.Н., Джепа Т.Л.** Автоматизированное рабочее место переводчика: лингвистические ресурсы и технологии // Структурная и прикладная лингвистика. 2012. № 9. С. 109–128. URL: https://www.elibrary.ru/item.asp?id=20351467&ysclid=ldw3as3jgx839950183

27. **Stambolieva M.** Corpus Linguistics, Translation and Error Analysis // Proceedings of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019)/ Varna, Bulgaria. 2019. Pp. 98–104. URL: https://www.researchgate.net/publication/337854111

28. **Koehn P.** Shared task: Statistical machine translation for European languages // ACL Workshop on Parallel Texts. 2005. Pp. 79–86. URL: https://www.researchgate.net/publication/234795504

29. **Tiedemann J.** Parallel Data, Tools and Interfaces in OPUS // Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). 2012. Pp. 2214–2218. URL: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

30. **Abney S., Bird S.** The Human language Project: building a universal corpus of world's languages. In: Proceedings of the 48th ACL. 2010. Pp. 88–97. URL: https://www.researchgate.net/publication/220873435

REFERENCES

[1] **B.Ye. Starichenko,** Tsifrovizatsiya obrazovaniya: illyuzii i ozhidaniya [Digitalization of education: illusions and expectations], Pedagogicheskoye obrazovaniye v Rossii, 3 (2020) 49–586. Available at: https://cyberleninka.ru/article/n/tsifrovizatsiya-obrazovaniya-illyuzii-i-ozhidaniya?ysclid=ldvzz4y1cs241893073

[2] **N.A. Abrosimova, Ye.A. Shchelokova,** Pred- i postredaktirovaniye mashinnogo perevoda meditsinskikh tekstov [Pre- and post-editing of machine translation of medical texts], Mir nauki, kultury, obrazovaniya. 4 (95) (2022) 178–181. Available at: https://cyberleninka.ru/article/n/pred-i-postredaktirovanie-mashinnogo-perevoda-meditsinskih-tekstov?ysclid=ldw44lkycw553753824

[3] **N.V. Nechayeva, S.Yu. Svetova,** Postredaktirovaniye mashinnogo perevoda kak aktualnoye napravleniye podgotovki perevodchikov v vuzakh [Post-editing machine translation as a new activity for teaching translation at universities], Voprosy metodiki prepodavaniya v vuze. 7 (25) (2018) 64–73. Available at: https://cyberleninka.ru/article/n/postredaktirovanie-mashinnogo-perevoda-kak-aktualnoe-napravlenie-podgotovki-perevodchikov-v-vuzah/viewer

[4] **A. Toral, S. Castilho, K. Hu, A. Way,** Attaining the Unattainable? Reassessing Claims of Human Parity in Neural Machine Translation, (2018). Available at: https://arxiv.org/abs/1808.10432

[5] **A. Toral, P. Oliver,** Ribas-Bellestín, Machine Translation of Novels in the Age of Transformer, Maschinelle Übersetzung für Übersetzungsprofis, edited by Jörg Porsiel. Berlin: BDÜ Fachverlag, 2020. Pp. 276–296. Available at: https://www.researchgate.net/publication/346557116

[6] **G. de Almeida, S. O'Brien,** Analysing Post-Editing Performance: Correlations with Years of Translation Experience, Proc. of the 14th Annual conference of the European Association for Machine Translation, May 27-28, 2010, Saint Raphael, France, 2010. Available at: https://www.aclweb.org/anthology/2010.eamt-1.19

[7] **A.V. Achkasov,** Industriya lokalizatsii i podgotovka perevodchikov [Localization industry and translator training], Materialy VIII Mezhdunarodnoy nauchnoy konferentsii «Industriya perevoda» ["Translation Industry"]. Perm, 2016, pp. 13–19. Available at: https://elibrary.ru/item.asp?id=26181052&ysclid=ldwxzskzkc395978920

[8] **L.N. Belyayeva,** Lingvisticheskiye tekhnologii v sovremennom setevom prostranstve: language worker v industrii lokalizatsii [Linguistic technologies in modern network space: language worker in localization industry], Knizhnyy dom, SPb, 2016. Available at: https://search.rsl.ru/ru/record/01008793133?ysclid=ldw1rssz1875247094

[9] **M. Koponen, U. Sulubacak, K. Vitikainen, J. Tiedemann,** MT for subtitling: User evaluation of post-editing productivity, EAMT-2020, Proc. of the 22nd Annual Conference of the European Association for Machine Translation. Lisboa, Portugal, 2020, pp. 115–124. Available at: https://www.researchgate.net/publication/348688005

[10] **N. Alves, M. Marques, P. Guimaraes, J.A. Almeida, R. Canario,** Education and Training Courses in Portugal: Another kind of schooling? 2021. Available at: https://www.researchgate.net/publication/351050875

[11] **S.I. Stefanov,** Reklama i poligrafiya: opyt slovarya-spravochnika [Advertising and printing: the experience of a reference dictionary], Gella print, M., 2004. Available at: https://search.rsl.ru/ru/record/01002486074?ysclid=ldw3y0ofmq501451819

[12] **O.I. Kolesnikova,** Obucheniye studentov literaturnomu redaktirovaniyu v kontekste razvitiya khudozhestvenno-yazykovoy kompetentsii [Students' Training to the Literary Editing in the Context of the Artistic-Language Competence Development], Yaroslavskiy pedagogicheskiy vestnik. 2 (2016) 60–63. Available at: http://vestnik.yspu.org/releases/2016_2/14.pdf

[13] **K.M. Nakoryakova,** Spravochnik po literaturnomu redaktirovaniyu dlya rabotnikov massovoy informatsii [A Handbook of Literary Editing for Media Professionals], Flinta: Nauka, M., 2010. Available at: https://search.rsl.ru/ru/record/01004396714?ysclid=ldw3w40ial722575012

[14] **O.V. Khlopunova, A.A. Tsakanyan,** Professionalnyye kompetentsii sovremennogo redaktora: lider ili menedzher [Professional competencies of a modern editor: leader or manager], Vestnik Volzhskogo universiteta imeni V.N. Tatishcheva. 4 (1) (2019) 1–10. Available at: https://cyberleninka.ru/article/n/professionalnye-kompetentsii-sovremennogo-redaktora-lider-ili-menedzher?ysclid=ldw1g8crls909581657

[15] V.I. **Frolova,** O menyayushcheysya roli redaktora v epokhu obnovleniya internet-kommunikatsii [On the changing role of editor in the age of internet communication renewal], Vestnik Volzhskogo universi-

teta im. V. N. Tatishcheva. 4 (19) (2015) 51–56. Available at: https://cyberleninka.ru/article/n/o-menyayuscheysya-roli-redaktora-v-epohu-obnovleniya-internet-kommunikatsii?ysclid=ldw1l592hr40120245

[16] **L.N. Beliaeva, O.N. Kamshilova,** Problems and Perspectives of Language Worker Professional Training, International Journal of Open Information Technologies. 6 (3) (2018) 35–42. Available at: http://injoit.org/index.php/j1/article/view/641

[17] **J. Moorkens, S. O'Brien,** Post-Editing Evaluations: Trade-offs between Novice and Professional Participants, EAMT 2015. Proc. of the 18th Annual Conference of the European Association for Machine Translation. Antalya, Turkey, 2015, pp. 75–81. Available at: https://www.researchgate.net/publication/275031846

[18] **A. Zaretskaya,** The Use of Machine Translation among Professional Translators, Proc. of the EXPERT Scientific and Technological Workshop, Malaga: Editions Tradulex, Geneva, 2015, pp. 1–13. Available at: https://www.researchgate.net/publication/283667234

[19] **I.S. Kudashev,** Proyektirovaniye perevodcheskikh slovarey spetsialnoy leksiki [Designing LSP Dictionaries for Translators], Univ. of Helsinki, Dep. of transl. studies, Helsinki, 2007. Available at: https://helda.helsinki.fi/bitstream/handle/10138/19272/proektir.pdf

[20] **A. Vasiljevs, M. Pinnis, T. Gornostay,** Service model for semi-automatic generation of multilingual terminology resources, Terminology and Knowledge Engineering, 2014, pp. 67–76. Available at: https://www.researchgate.net/publication/266565145

[21] **L.N. Vieira,** Post-Editing of Machine Translation, The Routledge Handbook of Translation and Technology. Routledge, 2019, pp. 319–335. Available at: http://www.routledge.com/9781138232846

[22] **O.V. Felde,** Effektivnoye rechevoye obshcheniye (bazovyye kompetentsii). Slovar-spravochnik pod redaktsiyey A.P. Skovorodnikova [Effective verbal communication (basic competencies): a reference dictionary, A.P. Skovorodnikov – ed.], Krasnoyarsk, 2012. Available at: http://elib.sfu-kras.ru/handle/2311/63722

[23] **D. Ortiz-Martinez,** Online Learning for Statistical Machine Translation, Computational Linguistics. 42(1) (2016) 121–161. Available at: http://dx.doi.org/10.1162/COLI_a_00244

[24] **C.T. Báez, M. Schaeffer, M. Carl,** Experiments in Non-Coherent Post-editing, Proc. of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT). Varna, Bulgaria, 2017, pp. 11–20. Available at: https://www.researchgate.net/publication/322031657

[25] Avtomatizirovannoye rabocheye mesto filologa v strukture obrazovatelnogo prostranstva sovremennogo vuza [Automated workplace of a philologist in the structure of university educational space] / L.N. Belyayeva, T.L. Dzhepa, G.N. Zak, O.N. Kamshilova, V.R. Nymm, V.V. Razumova, Knizhnyy dom, SPb., 2013. Available at: https://search.rsl.ru/ru/record/01006589603?ysclid=ldw33c4dsz595025912

[26] **L.N. Belyayeva, T.L. Dzhepa,** Avtomatizirovannoye rabocheye mesto perevodchika: lingvisticheskiye resursy i tekhnologii [Translator's automated workplace: linguistic resources and technologies], Strukturnaya i prikladnaya lingvistika. 9 (2012) 109–128. Available at: https://www.elibrary.ru/item.asp?id=20351467&ysclid=ldw3as3jgx839950183

[27] **M. Stambolieva,** Corpus Linguistics, Translation and Error Analysis, Proc. of the Human-Informed Translation and Interpreting Technology Workshop (HiT-IT 2019) Varna, Bulgaria, 2019, pp. 98–104. Available at: https://www.researchgate.net/publication/337854111

[28] **P. Koehn,** Shared task: Statistical machine translation for European languages, ACL Workshop on Parallel Texts, 2005, pp. 79–86. Available at: https://www.researchgate.net/publication/234795504

[29] **J. Tiedemann, Parallel Data,** Tools and Interfaces in OPUS, Proc. of the Eighth International Conference on Language Resources and Evaluation (LREC'12), 2012, pp. 2214–2218. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/463_Paper.pdf

[30] **S. Abney, S. Bird,** The Human language Project: building a universal corpus of world's languages, Proc. of the 48th ACL, 2010, pp. 88–97. Available at: https://www.researchgate.net/publication/220873435

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Камшилова Ольга Николаевна
Olga N. Kamshilova
E-mail: onkamshilova@gmail.com
ORCID: https://orcid.org/0000-0002-1488-2206

Беляева Лариса Николаевна
Larisa N. Beliaeva
E-mail: lauranbel@gmail.com
ORCID: https://orcid.org/0000-0002-8622-4595

Поступила: 16.01.2023; Одобрена: 27.02.2023; Принята: 17.03.2023.
Submitted: 16.01.2023; Approved: 27.02.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 57–69.
Terra Linguistica. 2023, Том 14, № 1. С. 57–69.

Review article
UDC 81'32
DOI: https://doi.org/10.18721/JHSS.14106

LEARNER CORPORA: RELEVANT INFORMATION AND AN OVERVIEW OF THE EXISTING FRAMEWORKS

M.V. Khokhlova ✉

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ m.khokhlova@spbu.ru

Abstract. In the modern world, there is a constant interest in foreign languages. Therefore, the question of learning about the language used by non-native speakers of a certain language, as well as describing their mistakes is a highly relevant matter. Learner corpora differ not only according to the languages they focus on, but also in relation to a number of their properties. The purpose of the study is to present a review the learner corpora available for different languages, as well as to compare the approaches that exist for their annotation. The paper considers the origins of learner corpus research, focuses on the main the stages of a project, types of learner corpora (which may differ in their tasks, students' mother tongue, language proficiency, text genre, data type, etc.), linguistic and metatextual information that accompany texts and provides a classification of errors. The paper gives a brief overview of annotation tools and corpus platforms that can be used for building a learner corpus.

Keywords: learner corpora, typology, errors, annotation, second language acquisition.

Acknowledgements: The study was carried out with the financial support of St. Petersburg State University (project No. 92563238).

Citation: M.V. Khokhlova, Learner corpora: relevant information and an overview of the existing frameworks, Terra Linguistica, 14 (1) (2023) 57–69. DOI: 10.18721/JHSS.14106

Обзорная статья
DOI: https://doi.org/10.18721/JHSS.14106

КОРПУСА УЧЕБНЫХ ТЕКСТОВ: ДАННЫЕ И ОБЗОР СУЩЕСТВУЮЩИХ ПОДХОДОВ

М.В. Хохлова ✉

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ m.khokhlova@spbu.ru

Аннотация. В современном мире не угасает интерес к иностранным языкам. Поэтому вопрос их изучения в качестве неродного, а также описание ошибок, которые допускают обучающиеся, не теряет своей актуальности. Учебные корпуса различаются не только в зависимости от языкового материала, но и по ряду своих характеристик. Целью статьи является обзор корпусов учебных текстов разных языков, а также сравнение подходов, которые существуют для их разметки (прежде всего, метатекстовой). В работе рассматриваются основные этапы разработки проектов, типы учебных корпусов (которые могут отличаться по своим задачам, по родному языку студентов, уровню владения языком, жанру текстов, типу данных и т.д.), лингвистическая и метатекстовая информация, которая сопровождает тексты, а также приводится классификация ошибок. В статье дается краткий обзор инструментов для разметки и платформ, которые можно использовать для создания учебного корпуса.

Ключевые слова: корпусы учебных текстов, типология, ошибки, разметка, усвоение второго языка.

Финансирование: Исследование выполнено при финансовой поддержке Санкт-Петербургского государственного университета (проект № 92563238).

Для цитирования: Хохлова М.В. Корпуса учебных текстов: данные и обзор существующих подходов // Terra Linguistica. 2023. Т. 14. № 1. С. 57–69. DOI: 10.18721/JHSS.14106

Introduction

Now in the 21st century, we can observe the processes of population migration. Open borders allow people to travel, study and work in different countries. The number of non-native speakers living in various countries has increased significantly over the past few years. The language of a non-native speaker has specific unusual characteristics both in vocabulary and grammar, and hence such a system deserves to be studied. The target audience of such corpora can be not only teachers or students, but also linguists who analyze second language acquisition through corpus data. This can help to create specialized tutorials, develop methods, and also to describe mechanisms of error production. Empirical linguistic evidence from learner corpora is hence the most valuable source of examples and can contribute to the understanding of the processes emerging during foreign or second language acquisition.

Overview

The origins of learner research projects can be traced back to the 1980s when computer technologies facilitated the processes of storage, retrieval and processing of large amounts of texts. This resulted in the launch of electronic collections of written and spoken linguistic data that represented the language of foreign and second language students. The work by Rosen et al. [1] proved to be one of the most profound and up-to-date monographs that are totally focused on learner corpora. Let us turn to some projects that focus on texts produced by learners. An excellent review of the projects can be found on the website of CLARIN initiative [2].

English was the first target language to have a learner corpus. The project entitled the *International Corpus of Learner English* (ICLE) [3] is often referred to as the first learner corpus that was based upon the principles of corpus linguistics [4–5]. Granger emphasizes that "the release of a learner corpus such as the ICLE marks the beginning of a new stage in the evolution of learner corpus research" [6, p. 544]. However, there were other projects focused on the analysis of learner language that foreshadowed ICLE in 1980s and 1990s. ICLE was a large-scale corpus with data collected from respondents with various L1 backgrounds. The initial corpus was about 2 mln words, while the second version was 3.7 mln words. It comprised essays written by advanced learners of English, which were university students. This corpus had a tremendous influence and launched the development of similar projects, namely, the LOCNESS (Louvain Corpus of Native English Essays), LINDSEI (Louvain International Database of Spoken English Interlanguage) LOCNEC (Louvain Corpus of Native English Conversation).

Publishing houses and testing organizations are also interested in such resources and build their own corpora. Cambridge University Press built the *Cambridge English Corpus* (CEC) [7] that includes two parts: the first subcorpus (1.8 bln words) compiles texts by native speakers (British and American), while the second subcorpus (55 mln words) focuses on how non-native speakers use English. The latter is also known as the Cambridge Learner Corpus (CLC) and comprises written data produced by more than 200,000 L2 English learners from 173 countries in such language exams as Cambridge English (all levels), CELS, IELTS and others. The corpus data is error-annotated which makes it possible to compute frequencies of different types of errors, to see the contextualized usage of the word and possible mistakes, to see difficult cases, and to compile dictionary entries and other material for language learners. Longman collected several corpora combined in the Longman Corpus Network. The Longman Learner Corpus (10 mln words), being its part, was compiled for the production of the Longman Active Study Dictionary.

Speaking about Russian learner corpora, we should name the *Russian Learner Corpus of Academic Writing* (RULEC) that was the earliest project in this direction for Russian. Its detailed description is made in [8–9]. Nowadays it is a subcorpus within the *Russian Learner Corpus* [10] that was created at the HSE Laboratory for Corpus Research (Moscow). The corpus represents the so-called "non-standard Russian". It contains samples of oral and written speech of two groups of Russian speakers: those who study Russian as a foreign language and heritage speakers. The latter includes people for whom Russian is not the main language, while they began to learn it as their first language in childhood (for example, emigrants). Along with lexical and grammatical features, error annotation is also available in the corpus. It includes spelling, morphological, syntactic, and lexical errors, as well as errors in constructions. Search results contain not only original versions but also the corrected ones. Metatextual features indicate the author's dominant language (American English, German, French, Italian, Norwegian, Dutch, Finnish, Swedish, etc.) and the level of Russian language proficiency according to the CEFR and ACTFL scales. When searching, one can specify a subcorpus taking into account the required characteristics. The corpus can be used to study the assimilation of the Russian language and in the teaching of Russian as a foreign language.

Typology

Nowadays, there are 190 learner corpora registered by the Centre for English Corpus Linguistics at Catholic University of Louvain [11]. Beyond these resources, there are also various own learner corpora, which were developed by universities or research groups. All these corpora can differ according to their properties and their volume. The issue of corpus volume is extremely important (see, for example, the discussion about the dependency of collocations on corpus volume [12] while a small amount of data does not provide sufficient evidence for frequencies and hence hinders the use of statistical tests. Nevertheless, it is difficult to say if a corpus is big or small because there is no agreement about how much data is enough. The only thing that matters, in this case, is the quality of a corpus. If it is poor, the size of a corpus will have no importance. Moreover, if a corpus was collected according to perfect and strict design criteria, even a small corpus will be of great value.

In second language acquisition (SLA), one can distinguish between second and foreign language. The former means that the language is acquired in its natural environment, for example, English as a Second Language (ESL) implies learning English in English-speaking countries (such as the United Kingdom). Foreign language acquisition (FLA) deals with studying languages in a context where it is not generally spoken. If we apply this paradigm to learner corpora, hence there can be corpora focusing on SLA and FLA tasks. However, some authors use these terms as synonyms to describe learning a second (nonnative) language.

Adopting classification schemes used by Tono [13], Granger [14–15], Rosen et al. [1], we can define learner corpora thoroughly by the following characteristics:

1. Language-related criteria

Medium. Learner corpora can cover written or oral data. The former is the dominant source, while spoken learner corpora are still rare. Nowadays, ambitious projects contain audio and video fragments known as multimedia (multimodal) learner corpora. Among them, we can name the Multimedia Adult ESL Learner Corpus [16], the PAROLE corpus [17], and the TAITO corpus that contains videos of partially transcribed discussions [18]. Granger [14] mentions the Telekorp project [18], "which results from five years of computer-mediated communication between learners of German in the US and learners of English in Germany" [14, p. 261].

Genre. Learner corpora tend to represent one kind of texts (mostly, essays). It is time and labor-consuming to collect many genres. Nevertheless, the existing projects try to overcome this drawback. For example, the BELC (Barcelona English Language Corpus) contains speech recordings across four tasks (written composition, oral narrative, oral interview, and role play) [20]. The ICLFI (International Corpus of Learner Finnish) comprises both non-fictional (e.g., essays, argumentative texts) and fictional texts (e.g., narratives, letters) [21–22].

2. Task-related criteria

Time of collection. Texts can be collected ad hoc and only once or over a period of time and thus, we can speak about cross-sectional vs. longitudinal data. Cross-sectional corpora represent data from different types of learners at a single point in time, while longitudinal corpora focus on the same learners during certain time periods. A particular mixture between them is quasi-longitudinal corpora that represent data collected simultaneously from learners with different proficiency levels. The *Longitudinal Database of Learner English* focuses on collecting longitudinal learner data from the same students over several years [23].

Task. Here we can differentiate between spontaneous and prepared texts, i.e., the ones generated in the classroom and those written at home. The use of references can also be limited.

Pedagogical use. The majority of learner corpora are corpora for delayed pedagogical use. It means that they are built on texts from a given sample of students and then will be processed and used for other (next) groups of learners and not for the ones who produce the texts. The opposite example is corpora for immediate pedagogical use (the same students can benefit from their "own" corpora).

3. Learner-related criteria

First language (L1, mother tongue). Learner corpora can contain data from learners with the same mother tongue or with several different L1 backgrounds. The ESF (European Science Foundation Second Language) Database comprises data collected by research groups from the Netherlands, Great Britain, France, Germany and Sweden [24–25]. The target languages are Dutch, English, French, German and Swedish. For each target language, two source languages were selected: Punjabi and Italian for English, Italian and Turkish for German, Turkish and Arabic for Dutch, Arabic and Spanish for French, Spanish and Finnish for Swedish. SweLL (Swedish Learner Language Corpus) presents data collected from learners who speak 64 languages from different language families [26]. The ten most frequent mother tongue backgrounds are English, Persian, German, Chinese, Russian, Arabic, Spanish, Thai, Somali and Vietnamese. The ICLE covers 11 different native languages.

Target language (L2). Usually, learner corpora focus on one language. The majority of them deal with English, but there are also corpora for other languages such as German, French, Dutch, Czech etc. The bilingual part of the CHILDES database represents data collected from children learning two or more languages [27]. The Multilingual Learner Corpus represents data from speakers of Brazilian Portuguese who learn different languages (English, German and Spanish) [28].

Proficiency in the target language. According to this feature, we can distinguish between corpora with texts collected from students at the same level of language knowledge and those with texts from speakers at various levels.

4. Data-related criteria

Owner. Collection of data for corpora can be initiated by publishing houses, companies or universities. Hence we can speak about commercial (the Longman Learners' Corpus or the Cambridge Learner Corpus) or academic corpora (the Louvain International Database of Spoken English Interlanguage (LINDSEI)). The former "tend to be much larger and have a wider range of mother tongue backgrounds" [14, p. 260].

Accessibility. This distinction is related to the one mentioned above. Depending on their funding, corpora can be freely accessible online or aimed for limited (in-house or commercial) use.

Annotation. As a rule, corpora should be annotated (at least POS-tagging), but some learner corpora represent only raw data without added linguistic information.

Levels of learner corpus design

Text collection is the first step in building corpora. The collection of texts can be organized in different ways, and it reflects the specifics of learner corpora and differs from the procedure for standard corpora. Many written learner corpora deal with one genre, e.g., essays produced by students, as it is relatively easy to collect them after exams or exercises. The compilers of a corpus should elaborate particular guidelines that include instructions for text collectors, the choice of text topics and their types, metadata description, and consent for learners about the use of texts. The next question that should be addressed with attention is the initial form of texts, i.e. whether they are electronic or written on paper, or where they are produced, i.e., at home or in a class. This is not as straightforward issue as it may seem. Electronic texts can be influenced by spell-checkers and texts written at home can have fewer errors than the ones produced in a class. The processing of hand-written texts raises the question of their OCR recognition or transcription and is time-consuming. This step usually involves an orthographic form but in case of spoken corpora texts can be supplied with phonemic or phonetic transcriptions.

Once the text is preprocessed and converted into an appropriate format, it can be annotated. The next step of corpus building deals with adding special data to texts, i.e., their mark-up. Any corpus usually has an annotation, which can be either textual or linguistic. In the case of learner corpora, their building process resembles one for standard corpora, but there are some peculiarities. Learner language abounds with errors and hence differs from the standard language, so it is necessary to take into account these discrepancies. Accordingly, learner corpora require a special type of markup, i.e., error annotation. Next, we will look at some of the principles that underlie various types of mark-up.

The first type of annotation implies specifying textual characteristics that describe texts. Standard textual annotation identifies sentences, paragraphs, sections, headings, and other features dealing with the structure of a document. It can also be helpful for learner corpora. In terms of learner corpora, one can pay attention to such unusual features as corrections (insertions or deletions) made by learners in handwritten texts. The next step involves grammatical annotation, which results in tokenization, lemmatization, POS-tagging, and determining other grammatical features. The most elaborate strategy implies syntactic parsing, semantic or discourse annotation. However, learner corpora need another type of specific annotation, namely, error annotation, that can be performed in most cases manually. This layer of annotation deals with the detection of errors, their description (categorization) and correction.

Metatextual annotation

Metatextual annotation is highly important for second language analysis as it helps to build subcorpora based on relevant features and hence to investigate linguistic phenomena inherent to the students of a particular proficiency level, age, education or social class. Metadata deal with texts as a whole and imply information about texts themselves (title, year of publication, medium, register etc) or their authors (in this case it describes sociolinguistic features), being one of the most important types of annotation in case of a learner corpus [13–15]. The usefulness of a learner corpus depends on this kind of annotation as non-properly described data fail to contribute to confirmation or rejection of linguistic hypotheses. Metatextual markup enables a user to define and select subcorpora, i.e. find those texts that meet the specified parameters.

There are different approaches to textual annotation that can focus on describing authors or texts. Granger rightly points out that "extra care has to be taken in collecting the data for learner corpora given the large number of variables affecting the learning/acquisition process" [6, p. 538]. We can proceed from the idea that the following positions can be reflected in metadata that describes the authors of texts for learner corpora: 1) the learner's first language; 2) education; 3) gender; 4) age; 5) other languages; 6) the duration of language learning. Below we will dwell on a number of projects (this list was inspired by [1]) and metadata description used by them taking into account that they are a few of many.

ICLE, on the one hand, follows the design criteria introduced by [4] and, on the other hand, tries to describe characteristics of learner texts. Granger [14] distinguishes between learner variables which concern a student and task variables that characterize the language situation. In its turn, each type can be described in terms of general variables (can apply to any corpus) and L2-specific variables. General learner features are age, gender, region and mother tongue, while L2-specific are learning context, proficiency level, amount of L2 exposure and knowledge of other foreign languages. General task variables are represented by medium, field, genre (text type), whereas task type (activities that learners are involved in: conversation, role-play, interview, essays etc.) and conditions belong to L2-specific task characteristics that can influence leaner's text generation (time limit, topic, mother tongue of interviewer etc.). The list of features used in ICLE can describe a text quite exhaustively, nevertheless the authors name one variable that plays a crucial role for learner corpora but is difficult to be recorded, that is "the teaching methodology and pedagogical materials to which the learners have been exposed" [3, p. 4].

CzeSL corpora have 15 items about the author of the text and 15 items about the text itself [1, p. 54]. While building a learner corpus for Russian, the authors [8] described 8 metadata items that were grouped into two categories, namely, author- and text-related features. The former included six subcategories, for example, language background (L2 learner or heritage speaker), dominant language (American English, German, Korean, etc.), and proficiency level according to CEFR (ACTFL). The latter contains three subcategories: mode (written or oral), genre (answers, essays, blogs, letters, stories, descriptions, etc.), and time limit (limited or unlimited).

The whole range of codes for describing texts implies the following positions [1, p. 56]:

- text id;
- date of the text collection;
- medium of the text (manuscript or electronic);
- time limit in minutes;
- permitted resources (yes or none, dictionary, textbook, other);
- part of exam (yes or n/a, interim, final);
- size limit in words;
- title of the essay;
- type of the topic (general or specific);
- activity before writing the text (exercise, discussion, visual, vocabulary, other or none);
- assigned topic (multiple choice, specified, free, or other);

- assigned genre (free or specified);
- predominant genre in the text (informative, descriptive, argumentative, or narrative);
- text length in words;
- range of text length in words.

RULEC/ RLC uses the following metadata categories: name (pseudonym), gender, language background and language experience of the student (either for L2 or HL), language proficiency level, time stamp (week and academic year when the paper was written), time limit under which the paper was written (timed or non-timed), text type (one paragraph or a long research paper), text function (e.g. narration, argumentation), and indication if the paper was written individually or in a group.

The author metadata used in CLC includes the following characteristics: age, gender, first language, nationality, exam, CEFR and ALTE levels, year, educational level, and years of English study. CLC focuses on language exams and hence here we find a range of features describing textual characteristics (exam level, date, format, style, register). It pays much attention to exam scripts and metadata and includes the binary feature of whether the exam was passed. Annotation differentiates between CEFR level student performance and CEFR level exam.

Next, we also examined four other corpora, an overview of which is given in [1], namely: the ASK corpus of Norwegian [29], "Learner corpus for Portuguese" (*COPLE2*) [30], "Croatian Learner Text Corpus (*CroLTec*) [31] and "Fehlerannotiertes Lernerkorpus des Deutschen als Fremdsprache" (*Falko*) [32].

Table 1 shows the combined metafeatures that are used for describing authors in different corpora (learner-related criteria). Some metafeatures are obvious and easy to collect (such as gender or age that can be given in the questionnaire), while others are not as clear. For example, proficiency level should be additionally estimated by data collectors or teachers via language tests or the same length for the duration of study can lead to different language knowledge (due to study breaks). As we can see, with regard to meta-features, there is a core that is the same for all the corpora we have considered. Something may be different or may be implemented in a slightly different way. For instance, the "Nationality" field in some corpora can be denoted as "Country of origin" in others. Falko has the largest list of metafeatures dealing with mother tongues. The same holds true for ICLE: it indicates up to three foreign languages and the same number of mother tongues. Some corpora evaluate period spent in the country of the studied language in years, whereas ICLE counts it in months.

Some characteristics indicated as metafeatures and presented in the table serve as bases that can be chosen for dividing corpora according to their types (see section on learner corpora typology).

Error annotation

Error annotation constitutes an important part for a learner corpus and is far from straightforward as it can be difficult to make a guess about the author's intention. This stage implies the following three steps: error detection, classification and correction. Error classification usually involves a taxonomy, which predefines certain types of errors (e.g., lexical, morphological, syntactic ones). An error can be corrected on the next step and it means that the annotation will hence mark two variants for the given item (erroneous and correct ones). So we can speak about two levels of error annotation, the first one implies labeling according error categories, while the second type deals with error correction.

The question of whether it is possible to be sure that the error has been correctly detected and (if necessary) corrected deserves special attention. On the one hand, there are obvious cases, on the other hand, some examples are ambiguous and can suggest several possible correct versions or the annotator even fail to understand the author's intention.

Errors may vary depending on the level of language proficiency and include a wide range from orthographic and to stylistic ones. Rosen et al. [1] propose an elaborate error annotation scheme including two levels. The first one uses two-tier approach and allows the annotators to make their corrections without classifying the errors. The second one is based on standard linguistic categorization. The error tagset

Table 1. Metadata about authors

	id/name	sex	age	age category	first language	first language group	other languages	ACTFL	ALTE	CEF	language in the family	years in the country	program of study	duration of study	hours of study per week	textbooks	bilingual	language background	stay in the country	origin / nationality	education	Exam
CzeSL[1]	+	+	+	+	+	+	+	–		+	+	+	+	+	+	+	+	*				
RULEC /RLC	+	+			+			+		+			+					+				
ASK	+	+	+		+							+			+				+	+	+	
CLC		+	+		+				+	+				+							+	+
COPLE2	+	+	+		+		+			+	+	+		+						+	+	
CroLTec		+	+		+		+			+	+	+	+				+			+		
Falko	+	+	+		+		+					+	+	+					+		+	+
ICLE		+	+		+		+						+	+						+		

[1] CzeSL are aimed only at L2 learners.

includes: 1) incorrect words; 2) foreign words; 3) tags for incorrect inflections; 4) wrong word boundaries; 5) stylistic errors; 6) miscellaneous errors. The second type of classification involves errors in the following categories:

1) agreement;
2) valency;
3) pronominal reference;
4) analytical verb forms or compound predicates;
5) reflexive expressions;
6) negation;
7) redundant or missing items;
8) wrong word order;
9) lexicon or phraseology;
10) grammar category;
11) style.

RLC adopted the following error classification [10]:

1) orthographic errors;
2) morphologic errors;
3) syntactic errors;
4) errors in constructions;
5) lexical errors;
6) additional features.

The last group is used for tagging miscellaneous errors such as calques from other languages (interference), missed words, word, morpheme or letter substitution or other types of errors, especially of those that cannot be identified properly.

Since error correction and classification is the core part of learner corpora, it is then crucial to do it with minimal errors. The question of how many annotators are sufficient for this task is discussed in [33]. Obviously, we need to have more than one or at least two experts; however their agreement can be not so high. The evaluation of the manual annotation and its consistency deserves special attention. Many authors use the metric κ (kappa) from [34] that varies within the interval $[-1, 1]$: $\kappa = -1$ means perfect disagreement, $\kappa = 1$ shows perfect agreement, and $\kappa = 0$ suggests that the agreement is equal to chance.

Linguistic annotation

Nowadays linguistic annotation is mostly done automatically. Tools for automated analysis are integrated into corpus systems or available as separate programs. The results yields high accuracy but should be treated with caution as there are errors in lemmatization or tags. Nevertheless the importance of linguistic annotation for corpora cannot be overestimated. Linguistic data of various types makes it possible to search grammatical categories, parts of speech, sentence structure, etc.

Automatic linguistic processing usually includes sentence splitting, tokenization and morphological analysis. At this stage, the system will mark words that can potentially contain an error, since they are absent in the morphological dictionary of the system. Of course, it should be remembered that the system may not contain some word, which at the same time exists in the language and which was used by the student. Thus, taggers can help annotators to find errors.

There is a large number of different systems that can be either language-specific or not. Below we will dwell on software available for Russian. Russian is an inflectional language and thus morphological forms play a key role. There are a number of well-recommended analyzers that can be used for annotating Russian texts. The morphological annotation of the RLC was carried out with the MyStem [35–36]. Below we show the example of this annotation performed by the program for the sentence *Segodnja my pishem sochineniye o semje* 'Today we are writing an essay about a family'.

Сегодня{сегодня=ADV=}
мы{мы=SPRO,pl,1p=nom}
пишем{писать=V,ipf,tran=inpraes,pl,indic,1p}
сочинение{сочинение=S,n,inan=(acc,sg|nom,sg)}
о{о=PR=}
семье{семья=S,f,inan=(abl,sg|dat,sg)}

The sentence was split into the tokens, each of them being on a separate line: *Segodnja*, *my*, *pishem*, *sochineniye*, *o* and *semje*. Lemmata, parts of speech and grammatical information are then shown in braces. In case of ambiguity, the system generates several options for parsing, which are separated by vertical bars |. For example, *sochineniye* are treated automatically either as an accusative or a nominative form.

UDPipe is yet another example of a system that annotates texts according the following levels: tokenization, lemmatization, and morphological and syntactic parsing [37–38]. The UDPipe output is prodeuced in the CoNLL-U format [39]:

Along with tokens, lemmata, parts of speech and morphological features, this format indicates a syntactic head for the current token, which is either a value of ID or zero (if the token is the root of the sentence, for example, *pishem*), and a dependency relation to the head (for example, "object" for *sochineniye*). Additional dependency relation can be applied when sentences involve coordinate structures. The last field stores miscellaneous information that is not given in other columns, such as position in the sentence with respect to punctuation or any specific annotation.

Another example of a morphological analyzer for Russian is pymorphy2 [40]. The output format shares the same features with the above described examples. The annotation provides tokens, morphological tags, grammemes and lemmata (defined as "normal_form"). The field "score" shows the probability that tags are assigned correctly (1.0 correspond to a perfect result). The analyzer can predict lemmata and morphological features for words that are absent in the dictionary.

As it has been already mentioned, the language of learner texts is specific; therefore, automatic morphological annotation, although being important, nevertheless requires manual or semi-automatic verification and further correction.

Corpus platforms

Once texts are collected and processed, we should answer the following question: how can they be stored and accessed? Most projects deal with the XML format, which is the most suitable for describing corpus data, and use TEI guidelines to represent metadata. It is also crucial to choose a suitable platform or corpus manager that allows one to work with annotated texts and search in them. On the one hand, there are well-known systems, and on the other hand, it is possible to build a new system for a project. RLC uses the platform powered by Django [8]. It keeps texts in a MySQL database that has separate tables for each of the text layers (metadata, sentence, morphological and error annotation ones). The system enables online upload of external new texts and then automatically processes them by the MyStem tagger.

Sketch Engine [41] is widely used by many corpora, among them CLC, Arabic Learner Corpus, Estonian Corpus for Learners and Guangwai-Lancaster Chinese Learner Corpus. Fig. 1 shows an example of text types available for the Open Cambridge Learner Corpus (an uncoded subset of CLC) in Sketch Engine. Based on these attribute, a user can build an appropriate subcorpus for his (her) tasks.

Among other platforms used for corpus building and analytics we can name KonText [42], IMS Open Corpus Workbench (CWB) [43], WordSmith [44], AntConc [45] and LancsBox [46].

Conclusion

In our work, we tried to provide an overview of some learner corpora. As one can see, these resources are diverse and the language of non-native speakers deserves to be studied more profoundly. We also

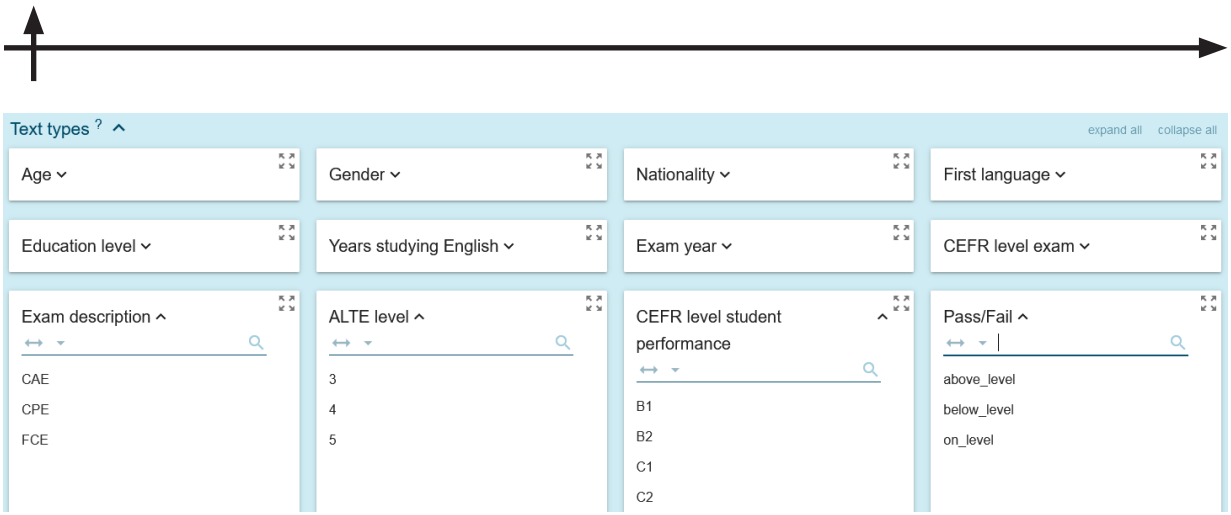

Fig. 1. Example of text metafeatures in Sketch Engine

sketched the pipeline that can be used for building a learner corpus and focused on issues related to its annotation. Error annotation requires special attention as it is the main part of such a corpus and provides empirical evidence of learner performance helping to reveal real problems the language learner encounter and not the ones that are described in dictionaries and grammars.

REFERENCES

[1] **A. Rosen, J. Hana, B. Vidová Hladká, T. Jelínek, S. Škodová, B. Štindlová,** Compiling and annotating a learner corpus for a morphologically rich language: CzeSL, a corpus of non-native Czech. Praha, 2021.

[2] CLARIN. Available at: https://www.clarin.eu/resource-families/L2-corpora (accessed 10.02.2023).

[3] **S. Granger, M. Dupont, F. Meunier, H. Naets, M. Paquot,** The International Corpus of Learner English. Version 3. Louvain-la-Neuve: Presses universitaires de Louvain, 2020.

[4] **S. Atkins, J. Clear, N. Ostler,** Corpus Design Criteria. Literary & Linguistic Computing, 7 (1) (1992) 1–16.

[5] **A. McEnery, V. Brezina, D. Gablasova, J.V. Banerjee,** Corpus Linguistics, learner corpora and SLA: employing technology to analyze language use. In: Annual Review of Applied Linguistics. 39 (2019) 74–92.

[6] **S. Granger,** The International Corpus of Learner English: A New Resource for Foreign Language Learning and Teaching and Second Language Acquisition Research. TESOL Quarterly. 37 (3) (2003) 538–546.

[7] **D. Nicholls,** The Cambridge Learner Corpus – error coding and analysis for lexicography and ELT, in Archer, D, Rayson, P, Wilson, A and McEnery, T (Eds), Proceedings of the Corpus Linguistics 2003 Conference, Lancaster University, 2003, pp. 572–581.

[8] **E. Rakhilina, A. Vyrenkova, E. Mustakimova, A. Ladygina, I. Smirnov,** Building a learner corpus for Russian. In: Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition at SLTC, 2016. Available at: http://aclweb.org/anthology/W16-65 (accessed 10.02.2023).

[9] **O. Kisselev,** Russian Learner Corpora Research: State of the Art and Call for Action. In Bakhtiniana, São Paulo, 18 (1) (2023) 8–29.

[10] RLC. Available at: http://web-corpora.net/RLC (accessed 10.02.2023).

[11] CECL. Learner corpora around the world. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html (accessed 10.02.2023).

[12] **M. Khokhlova, V. Benko,** Size of corpora and collocations: the case of Russian. In Slovenščina 2.0, 8 (2) (2020) 58–77.

[13] **Y. Tono,** Learner corpora: Design, development and applications. In Proceedings of the 2003 Corpus Linguistics Conference, D. Archer, P. Rayson, A. Wilson & T. McEnery (eds). UCREL: Lancaster University, 2003, pp. 800–809.

[14] **S. Granger,** Learner corpora. In Lüdeling, A. & Kytö, M. (eds.) Corpus Linguistics. An International Handbook. Volume 1. Berlin & New York: Walter de 5 Gruyter (2008a) pp. 259–275.

[15] **S. Granger,** Learner Corpora in Foreign Language Education. In Van Deusen-Scholl N. and Hornberger N.H. (ed.) Encyclopedia of Language and Education. Volume 4. Second and Foreign Language Education. Springer (2008b) pp. 337–351.

[16] **S. Reder, K. Harris, K. Setzler,** The Multimedia Adult Learner Corpus. In TESOL Quarterly (2003), 37 (3) (2003) 546–557. Available at: https://www.researchgate.net/publication/251678834_The_Multimedia_Adult_Learner_Corpus (accessed: 10.02.2023).

[17] **H. Hilton,** Annotation and analyses of temporal aspects of spoken fluency. CALICO Journal, 26 (2009) 644–661.

[18] TAITO. Available at: http://urn.fi/urn:nbn:fi:lb-2014073035 (accessed: 10.02.2023).

[19] **J.A. Belz, N. Vyatkina,** Learner Corpus Research and the Development of L2 Pragmatic Competence in Networked Intercultural Language Study: The Case of German Modal Particles, Canadian Modern Language Review/Revue canadienne des langues vivantes 62.1 (2005) 17–48.

[20] **C. Muñoz,** (ed.) Age and the Rate of Foreign Language Learning. Clevedon: Multilingual Matters. (2006).

[21] **J.H. Jantunen, S. Brunni,** Morphology, lexical priming and second language acquisition: A corpus-study on learner Finnish. In S. Granger, G. Gilquin and F. Meunier (eds.) Twenty Years of Learner Corpus Research. Louvain-la-Neuve. Presses universitaires de Louvain, 2013, pp. 235–245.

[22] **S. Brunni, L.-M. Lehto, J.H. Jantunen, V. Airaksinen,** How to annotate morphologically rich learner language: principles, problems and solutions. Bergen Language and Linguistics Studies, 6 (2015) 133–152.

[23] **F. Meunier,** Introduction to the LONGDALE project. In E. Castello K. Ackerley, & F. Coccetta (Eds.), Studies in learner corpus linguistics: Research and applications for foreign language teaching and assessment Berlin: Peter Lang Publishing, 2016, pp. 123–126.

[24] ESF. In: The ESF Database. Available at: https://www.mpi.nl/world/tg/lapp/esf/esf.html (accessed: 10.02.2023).

[25] **H. Feldweg,** The European Science Foundation Second Language Database. Nijmegen: Max-Planck-Institute for Psycholinguistics, 1991.

[26] **E. Volodina, I. Pilán, I. Enström, L. Llozhi, P. Lundkvist, G. Sundberg, M. Sandell,** SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. In LREC Proceedings 2016, 2016, pp. 206–212. Available at: https://aclanthology.org/L16-1031 (accessed: 10.02.2023).

[27] CHILDES. Available at: https://childes.talkbank.org/ (accessed: 10.02.2023).

[28] **S. Tagnin,** A multilingual learner corpus in Brazil. In: Archer, D., Rayson, P., Wilson A., McEnery T. (eds.) Proceedings of the Corpus Linguistics 2003 conference. UCREL, Lancaster University, 2003, pp. 940–945.

[29] **K. Tenfjord, J.E. Hagen, H. Johansen,** Norsk andresprkåskorpus (ASK) – design og metodiske forutsetninger. NOA norsk som andrespråk 25 (1) (2009) 52–81. Available: https://w2.uib.no/filearchive/tenfjord-hagen-og-johansen.-2009.-noa..pdf (accessed: 10.02.2023).

[30] **A. Mendes, S. Antunes, M. Janssen, A. Gonçalves,** The COPLE2 Corpus: A Learner Corpus for Portuguese. In: Proceedings of the Tenth Language Resources and Evaluation Conference – LREC'16, 23-28 May 2016, Portoroz, Slovenia, 2016, pp. 3207–3214.

[31] **N. Mikelić Preradović, M. Berać, D. Boras,** Learner Corpus of Croatian as a Second and Foreign Language. Multidisciplinary Approaches to Multilingualism. Ur. Cergol Kovačević, Kristina i Udier, Sanda Lucija. Peter Lang. Frankfurt am Main, Njemačka, 2015, pp. 107–126.

[32] **M. Reznicek, A. Lüdeling, C. Krummes, F. Schwantuschke, M. Walter, K. Schmidt, H. Hirschmann, T. Andreas,** Das Falko-Handbuch. Korpusaufbau und Annotationen Version 2.01, 2012. Available at: https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko (accessed: 10.02.2023).

[33] **R. Snow, B. O'connor, D. Jurafsky, A.Y. Ng,** Cheap and fast---but is it good?: evaluating non-expert annotations for natural language tasks. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, pp. 254–263.

[34] **J. Cohen,** A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20 (1) (1960) 37–46.

[35] **I. Segalovich,** A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. In Proceedings of the MLMTA 2003. USA, 2003, pp. 273–280.

[36] **I. Segalovich, V. Titov,** Mystem, 1997. Available at: https://yandex.ru/dev/mystem/ (accessed: 10.02.2023).

[37] **M. Straka, J. Straková,** Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. In: Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies, 2017, pp. 88–99.

[38] UDPipe. Available at: https://ufal.mff.cuni.cz/udpipe/1 (accessed: 10.02.2023).

[39] CoNLL-U Format. Available at: http://universaldependencies.org/docs/format.html (accessed: 10.02.2023).

[40] **M. Korobov,** Morphological Analyzer and Generator for Russian and Ukrainian Languages. In: Analysis of Images, Social Networks and Texts, 2015, pp. 320–332.

[41] **A. Kilgarriff, V. Baisa, J. Bušta, M. Jakubíček, M. Kovář, J. Michelfeit, P. Rychlý, V. Suchomel,** The Sketch Engine: ten years on. Lexicography, 1 (1) (2014) 7–36.

[42] **T. Machálek,** KonText – a modern, customizable corpus query interface. Abstract of a talk presented at the conference Corpus Linguistics 2017, Birmingham, 2017. Available at: https://www.birmingham.ac.uk/Documents/collegeartslaw/corpus/conference-archives/2017/general/paper341.pdf (accessed: 10.02.2023).

[43] **S. Evert, A. Hardie,** Twenty-first century Corpus Workbench: Updating a query architecture for the new millennium. In Proceedings of the Corpus Linguistics 2011 conference, University of Birmingham, UK, 2011.

[44] **M. Scott,** WordSmith Tools version 8, Stroud: Lexical Analysis Software. Available at: https://lexically.net/wordsmith/ (accessed: 10.02.2023).

[45] **L. Anthony,** AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University, 2020. Available at: https://www.laurenceanthony.net/software (accessed: 10.02.2023).

[46] **V. Brezina, P. Weill-Tessier, A. McEnery,** #LancsBox v. 5.x. [software], 2020. Available at: http://corpora.lancs.ac.uk/lancsbox (accessed: 10.02.2023).

СВЕДЕНИЯ ОБ АВТОРЕ / INFORMATION ABOUT AUTHOR

Maria V. Khokhlova
Хохлова Мария Владимировна
E-mail: m.khokhlova@spbu.ru
ORCID: https://orcid.org/0000-0001-9085-0284

Submitted: 20.01.2023; Approved: 10.03.2023; Accepted: 17.03.2023.
Поступила: 20.01.2023; Одобрена: 10.03.2023; Принята: 17.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 70–87.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 70–87.

Научная статья
УДК 81'32, 81'33
DOI: https://doi.org/10.18721/JHSS.14107

ДИНАМИЧЕСКОЕ ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ РУССКОЯЗЫЧНОГО КОРПУСА ЮРИДИЧЕСКИХ ДОКУМЕНТОВ

О.А. Митрофанова ✉ , **М.М. Атугодаге**

Санкт-Петербургский государственный университет,
Санкт-Петербург, Российская Федерация

✉ o.mitrofanova@spbu.ru

Аннотация. Статья посвящена анализу результатов динамического тематического моделирования законодательных актов Российской Федерации, указов высших должностных лиц и постановлений Верховного и Конституционного Судов за 2008–2022 годы, входящих в исследовательский корпус русскоязычных юридических документов. В статье описаны процедуры формирования и предобработки корпуса, эксперименты по обучению тематических моделей на данном корпусе. Рассматривается как стандартная тематическая модель, так и динамическая тематическая модель, учитывающая изменение тем корпуса во времени. После обучения моделей в различных условиях были определен набор оптимальных параметров обучения. В качестве основного инструмента тематического моделирования использовалась библиотека BERTopic на языке программирования Python, комбинирующая алгоритмы построения тематических моделей и нейросетевые контекстуализированные модели распределенных векторных вложений. Исследовательские данные могут представлять интерес не только для специалистов в области компьютерной лингвистам, но и для социологов, политологов, юристов, работающих с законодательными документами.

Ключевые слова: тематическое моделирование, динамическая тематическая модель, BERTopic, корпус русскоязычных юридических документов, Российская Газета.

Финансирование: НИП СПбГУ № 75254082 «Моделирование коммуникативного поведения жителей российского мегаполиса в социально-речевом и прагматическом аспектах с привлечением методов искусственного интеллекта», грант РНФ № 21-78-10148 «Моделирование значения слова в индивидуальном языковом сознании на основе дистрибутивной семантики».

Для цитирования: Митрофанова О.А., Атугодаге М.М. Динамическое тематическое моделирование русскоязычного корпуса юридических документов // Terra Linguistica. 2023. Т. 14. № 1. С. 70–87. DOI: 10.18721/JHSS.14107

Research article
DOI: https://doi.org/10.18721/JHSS.14107

DYNAMIC TOPIC MODELLING OF THE RUSSIAN LEGAL TEXT CORPUS

O.A. Mitrofanova ✉ , **M.M. Athugodage**

St. Petersburg State University,
St. Petersburg, Russian Federation

✉ o.mitrofanova@spbu.ru

Abstract. The article is devoted to the dynamic topic modelling analysis of legislative acts, decrees of senior officials and resolutions of the Supreme and Constitutional Courts dated 2008–2022, included into the research corpus of Russian legal documents. The article describes the procedures of corpus construction and preprocessing, training of topic models on this corpus. We consider both standard topic model and a dynamic topic model that takes into account changes in topics over time. After training the models in various conditions, a set of optimal training parameters was determined. The BERTopic library was used as the main tool for topic modelling, combining algorithms for constructing topic models and contextualized neural network models of distributed vectors. The research data may be of interest both for specialists in the field of computational linguistics as well as for sociologists, political scientists, lawyers working with legislative documents.

Keywords: topic modelling, dynamic topic model, BERTopic, Russian corpus of legal documents, Russian gazette.

Acknowledgements: Research Program of St. Petersburg State University No. 75254082 "Modeling the communicative behavior of residents of a Russian metropolis in the socio-speech and pragmatic aspects with the use of artificial intelligence methods", RSF grant No. 21-78-10148 "Modeling the meaning of a word in individual linguistic consciousness based on distributive semantics."

Citation: O.A. Mitrofanova, M.M. Athugodage, Dynamic topic modelling of the russian legal text corpus, Terra Linguistica, 14 (1) (2023) 70–87. DOI: 10.18721/JHSS.14107

Введение

Тематическое моделирование – это способ построения семантической модели коллекции текстовых документов, который определяет, к какой теме относится каждый из документов. Результатом тематического моделирования является установление соответствия между текстами корпуса и скрытыми факторами, темами (кластерами слов-тематизаторов), при этом каждый документ соотносится с некоторой вероятностью с одной или нескольким темам, а сами темы могут пересекаться: определенное слово может быть с разными вероятностями отнесено к нескольким темам [1, 2 и т.д.]. Тематические модели способствуют повышению результативности процедур извлечения информации из естественно-языковых текстов, таких как, например, автоматическая рубрикация, кластеризация и классификация документов, сентимент-анализ и т.д., а также вносят весомый вклад в обучение систем искусственного интеллекта [3–6]. Сфера применения тематических моделей широка, она охватывает корпусы текстов разных типов и жанров: новостные корпусы [7–9], корпусы социальных сетей [10–14], корпусы медицинских текстов [15], корпусы по финансам и банковскому делу [16, 17], корпусы текстов по разным областям научного знания [18], художественные корпусы [19–25] и т.д. Тематическое моделирование юридических документов – это новая исследовательская область, где удается получать ценные результаты [26]. Наше исследование призвано решить задачу изучения динамики тем в законодательных текстах, поэтому в фокусе нашего внимания находится такая модификация алгоритмов тематического моделирования как динамическое тематическое моделирование.

Автоматическая обработка текстов юридических документов представляет большой интерес как для лингвистов, так и для юристов, правоведов, социологов. В фокусе внимания исследователей находится юридическая терминология [27–29], необходимость ее гармонизации [30–32], машинный перевод [33], неоднозначность [34], сложность юридических текстов [35–38] и другие вопросы. В связи с нуждами прикладных исследований формируются специализированные корпусы юридических документов: многоязычные корпусы – Europarl (a Parallel Corpus for Statistical Machine Translation)[1], Параллельный корпус документов ООН (United Nations Parallel Corpus)[2] и т.д., для русского языка – корпус законов CorCodex, корпус решений конституционного суда CorDec, корпус локальных актов CorRIDA[3] и т.д. Однако для цели нашего исследования требуется особенный корпусной ресурс.

В ходе динамического тематического моделирования восстанавливается структура тем корпуса, документы в котором имеют хронологическую метаразметку и распределены по сегментам, соответствующим периодам времени (чаще всего, по месяцам, годам или десятилетиям). Результатом является выделение нишевых тем, характеризующих изучаемые хронологические промежутки [39, 40]. В случае применения динамического тематического моделирования оказывается возможным проведение анализа того, как формируется повестка дня в законодательных актах (как федеральных, так и региональных), указах Президента Российской Федерации и постановлениях Верховного Суда и Конституционного Суда Российской Федерации и других юридических документах России.

Изменения, происходящие в российской юриспруденции и влияющие на уровень правовой культуры в России, могут свидетельствовать о «юридическом богатстве общества», явно или косвенно представлять состояние его правовой жизни и характеризовать коллективный юридический опыт [41] в тот или иной хронологический период. Законодательные тексты своевременно отражают нововведения в общественно-политической и социальной сфере, тем самым, являются маркерами значимых событий в государстве и обществе. Представляет исследовательский интерес взаимосвязь между содержанием общественных дискуссий и их результатом в виде законодательных актов: эта взаимосвязь определима в ходе лингвистической экспертизы текстов, включающей также и процедуры тематического моделирования. Юридические документы являются источником данных о политической ситуации в стране, в том числе, об изменениях политического строя, уровня политических и гражданских свобод и т.п., о кадровых изменениях в составе руководящих органов власти, фиксируемых указом Президента или Председателя Правительства. Тем самым, исследование динамики тем в корпусе русскоязычных юридических документов может показать, как развивалась с течением времени различные аспекты деятельности государства.

Исследовательский корпус русскоязычных юридических документов

Для проведения процедур динамического тематического моделирования русскоязычных юридических текстов был использован исследовательский корпус сопоставимых документов[4], исходное предназначение которого связано с решением задачи упрощения юридических текстов. Корпус содержит юридический документ, опубликованный на сайте Российской Газеты, и его упрощенный вариант, или комментарий, написанный специалистами из Российской Газеты. Российская Газета[5] – это официальный печатный орган Правительства Российской Федерации; законы вступают в силу только после публикации в Парламентской газете, Российской газете, Собрании законодательства Российской Федерации или на Официальном интернет-портале правовой информации [42]. Интернет-портал «Российской газеты» RG.RU, который использовался при сборе документов, существует с 1999 г. и также наделён официальным статусом.

[1] https://www.statmt.org/europarl/
[2] https://conferences.unite.un.org/UNCorpus
[3] https://www.plaindocument.org/corpora
[4] https://www.kaggle.com/datasets/athugodage/russian-legal-text-parallel-corpus
[5] https://rg.ru/doc

Параллельный корпус содержит не все документы, опубликованные в Российской Газете, а только лишь те, к которым прилагался комментарий (или упрощенный текст). С одной стороны, это ограничивает рамки нашего исследования, так как нам придется работать не со всеми выпущенными законами, а лишь некоторой частью, причем довольно малой. Для сравнения, в среднем Российская Газета публикует около 1,5–3 тыс. документов в год, число комментариев имеет предельное значение – 465). С другой стороны, редакция Российской газеты производит отбор текстов для комментирования и публикует комментарий только к общественно важным документам. В результате, сам источник задает критерии отбора текстов по степени важности для общества. У нас есть возможность не тратить усилия на обработку полного массива юридических текстов, а сосредоточить исследовательское внимание на наиболее важных.

В экспериментах, описываемых в настоящей статье, использовался только сегмент корпуса, содержащий полные тексты, поэтому характеристики сегмента корпуса, содержащего комментарии, в данной статье не приводится.

Корпус содержит около 3 тыс. статей, датированных от 31 декабря 2008 г. по 28 ноября 2022 г. (это дата публикации, и, соответственно, вступления в силу). Дата публикация отмечена в отдельном столбце в следующем формате: «31 декабря 2008». В нашем корпусе больше всего статей за 2018 г. (465 статей), меньше всего – за 2008 г. (1 статья). Распределение документов исследовательского корпуса по годам представлено на рис. 1. Мы приняли решение сохранить хронологические рамки корпуса и оставить в экспериментальном материале единственную статью за 2008 г., поскольку сегментация на периоды времени при построении динамической тематической модели производится не по годам, а по дням, это позволяет соблюсти требование сбалансированности данных.

Как можно увидеть из графика на рис. 1 выше, за последнюю пятилетку было опубликовано больше документов (1664 документов), чем за предыдущее десятилетие (1299 документов). Это можно связать с тем, что интернет-портал развивается и публикует все больше комментариев к законам. Следует заметить, что до 2008 г. Российская Газета вообще не публиковала комментарии.

В среднем размер документа составляет около 400–600 токенов. Самый крупный документ содержит чуть более 70 тыс. токенов. Есть документы с количеством токенов около 100 –

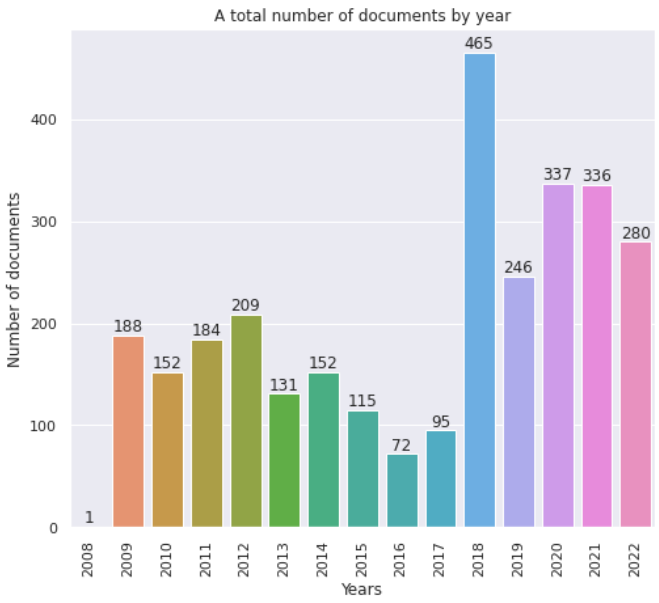

Рис. 1. Распределение документов исследовательского корпуса по годам

Fig. 1. Distribution of documents in the research corpus by year

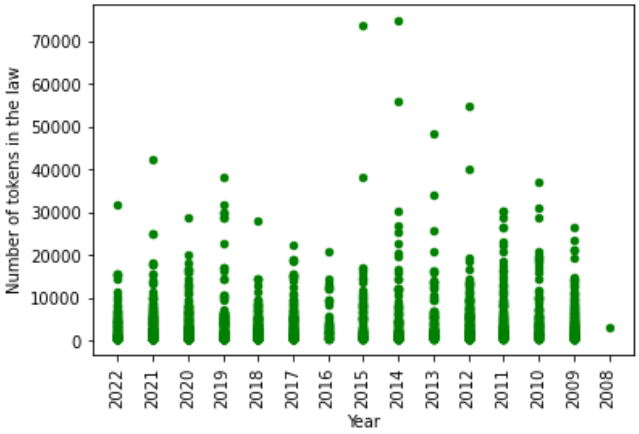

Рис. 2. Распределение документов исследовательского корпуса по годам с учетом объема текстов в токенах
Fig. 2. Distribution of documents in the research corpus by year taking into account text size in tokens

в основном, это поправки в закон и краткие указы Президента РФ или Председателя Правительства. График на рис. 2 показывает распределение документов по годам и их объем в токенах.

Обучение тематических моделей требует предварительной обработки документов корпуса. Для решения обсуждаемых в статье исследовательских задач в текстах корпуса необходимо провести следующие процедуры:

1) упрощенная токенизация с использованием функции .split() в Python;

2) удаление стоп-слов с помощью словаря, составленного на основе списков служебных слов, оборотов, аббревиатур, латинских и числовых обозначений [19, 20, 22];

3) лемматизация с использованием библиотеки для морфологического анализа pymorphy2;[6]

4) приведение дат в формат, соответствующий требованиям к обучению модели: «31 декабря 2008» → «2008-12-31»;

5) сортировка статей по дате (первоначальная сортировка производилась по годам).

Как показали эксперименты, качество предобработки корпуса позволяет повысить качество обучения тематических моделей и содержательно улучшить результат.

Особенности процедур стандартного и динамического тематического моделирования с помощью библиотеки BERTopic

Известные способы тематического моделирования включают в себя группу алгебраических моделей (LSA (латентно-семантический анализ), NMF (неотрицательная матричная факторизация) и т.д.) и вероятностных моделей (pLSA (вероятностный латентно-семантический анализ), LDA (латентное размещение Дирихле), LPA (латентное размещение Патинко), HTMM (скрытая тематическая марковская модель) и т.д.). На практике используются их мультимодальные расширения за счет введения дополнительных параметров корпуса: авторство в ATM (автор-тематической модели), фактор адресата в ARTM (модели автор-получатель), связи между темами в HTM (иерархической тематической модели), наличие заранее заданных ключевых слов-тематизаторов в GuidedLDA (управляемом латентном размещении Дирихле), учет конструкций в составе тем в n-граммных тематических моделях, возможность обобщения состава тем с помощью меток и т.д. [1, 3, 5 и т.д.]. В последние годы появился новый класс тематических моделей, комбинирующих вероятностные процессы и модели распределенных векторов, например, LDA2Vec, Top2Vec, ETM (Embedded topic model), CTM (Contextualized topic model), BERTopic и т.д. [43–47]. Стандартные и комбинированные тематические модели могут использоваться в модификации DTM (динамическая тематическая модель) [39–40].

[6] https://pymorphy2.readthedocs.io/

Преимущество комбинированных тематических моделей заключается в том, что они позволяют улучшить качество семантического представления текста и сократить потери, связанные с использованием технологии представления корпуса в виде мешка слов (bag-of-words). Использование контекстуализированных моделей в комбинированных тематических моделях имеет ряд особенностей: по сравнению с моделями типа word2vec модели BERT (Bidirectional Encoder Representations from Transformers) сохраняют векторные представления слов с учетом контекста и поэтому чувствительны, например, к полисемии.

В комбинированной модели BERTopic, используемой в нашем исследовании, к распределенным векторным вложениям BERT применяются процедуры кластеризации со снижением размерности и ранжированием слов-тематизаторов для формирования тем [47]. Тематическое моделирование в BERTopic проходит в три этапа. На первом этапе каждый документ корпуса преобразуется в векторное представление с использованием предварительно обученной языковой модели BERT. На втором этапе проводится снижение размерности векторов методом UMAP и кластеризация результирующих векторных вложений методом HDBSCAN. На третьем этапе из кластеров документов извлекаются специфичные для них ключевые выражения (n-граммы) с использованием измененного варианта метрики c-TF-IDF. Эти слова и словосочетания ранжируются методом MMR и рассматриваются в качестве кандидатов в тематизаторы. Формула для расчета значений c-TF-IDF представлена ниже:

$$W_{t,c} = tf_{t,c} \cdot \log\left(1 + \frac{A}{tf_t}\right),$$

где $tf_{t,c}$ – частота термина t в классе c. Класс c понимается как набор документов, объединенных в единое целое. Чтобы оценить информативность термина для класса, обратная частота документа заменяется обратной частотой класса, которая рассчитывается как логарифм отношения среднего количества слов в классе A и частоты употребления термина t во всех классах. Чтобы вывести только положительные значения, к логарифмируемому выражению добавляется единица.

Данный подход, позволяющий находить важные слова для кластеров текстов, а не только для отдельных документов, лежит в основе стандартной тематической модели BERTopic, которая описывает темы, характерные для корпуса в целом. Например, документы, связанные с вакцинацией, будут объединены в глобальную тему «вакцинация». Эта тема может со временем развиваться или исчезать: возможная тема для 2011 г. – «вакцинация от гриппа», а для 2021 г. – «вакцинация от ковида». При переходе от стандартного варианта BERTopic к динамическому тематическому моделированию на третьем этапе построения модели к варианту c-TF-IDF добавляются метки времени i, формирующие дополнительную модальность.

$$W_{t,c,i} = tf_{t,c,i} \cdot \log\left(1 + \frac{A}{tf_t}\right).$$

В нашей работе мы использовали архитектуру и модели из библиотеки BERTopic[7] для тематического моделирования корпуса юридических текстов в стандартном и динамическом вариантах, которые рассматриваются в последующих разделах.

Построение стандартной тематической модели исследовательского корпуса юридических документов

В начале работы с библиотекой BERTopic нужно настроить модель векторизации и инициализировать ее класс. При настройке мы использовали стоп-словарь [19, 20, 22], а также определяли минимальную частоту вхождения слов в корпус (1) и максимальную долю документов корпуса,

[7] https://github.com/MaartenGr/BERTopic

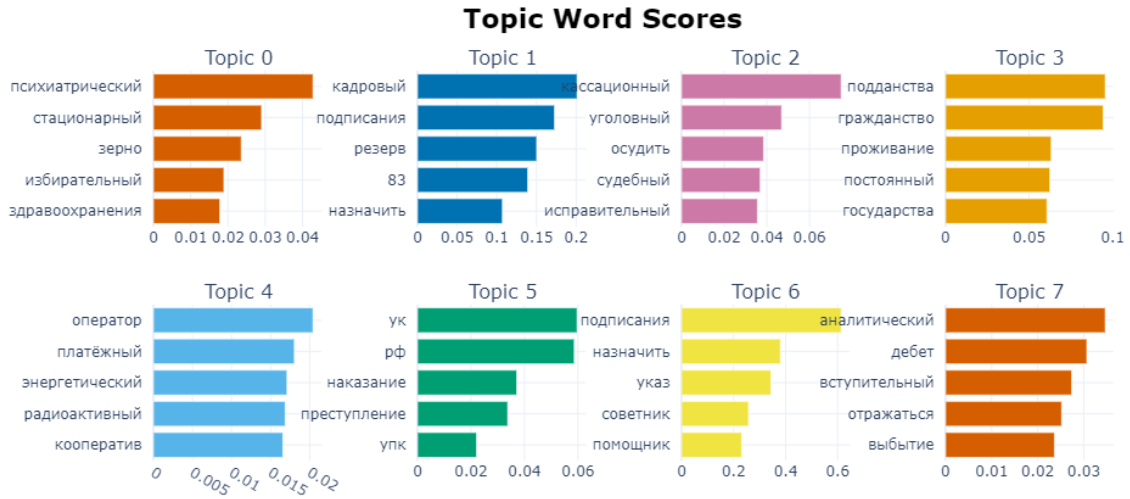

Рис. 3. Первые восемь тем из стандартной модели BERTopic

Fig. 3. The first eight topics from standard BERTopic model

включающих слова (0.6). Если увеличить значение второго параметра (0.8 или 0.95), то в числе слов-тематизаторов будут широко распространенные термины и сокращения, не дающие должного понимания содержательной стороны темы: *гк*, *рф*, *статья* и т.п.

Далее следует задать параметры обучения самой модели BERTopic и инициализировать ее класс. В результате серии экспериментов были подобраны оптимальные параметры обучения:

- минимальный размер темы – 10 слов-тематизаторов,
- первые *n* слов-тематизаторов – 10,
- размер *n*-грамм, выделяемых в корпусе: от 3 до 5.

В результате обучения стандартной тематической модели BERTopic были получены результаты, проиллюстрированные на рис. 3–7. На рис. 1 отображены 8 наиболее значимых тем (для каждой из них при визуализации выводится пять слов-тематизаторов) из 133 тем, предсказанных моделью.

Функционал библиотеки BERTopic позволяет нарисовать «карту расстояний» между темами (по значению), см. рис. 4–7.

На рис. 4–7 темы обозначены серыми кругами. Семантически близкие темы расположены близко друг к другу. Так, например, в правом нижнем углу рисунка расположена группа тем, связанных с налогообложением (рис. 4), внизу слева – группа тем, связанных с уголовными преступлениями (рис. 5), вверху по центру – группа тем, связанных с туристическим обслуживанием (рис. 6), вверху справа – группа тем, связанных с пенсионным законодательством (рис. 7). Выделяются более специфические темы, связанные с правами военнослужащих, пандемией COVID-19, топонимами – названиями регионов, числовыми обозначениями и т.д.

Динамическое тематическое моделирование корпуса русскоязычных юридических документов

Для обучения динамической тематической модели необходимо предварительно задать точки во времени, относительно которых будут оцениваться изменения в темах. В наших экспериментах в качестве данных точек использовались дни публикации текстов законодательных актов. Таким образом, на вход алгоритма подавался список дат вида: [‘2008-12-31’, ‘2009-01-15’, ‘2009-01-16’, ... ‘2022-11-28’]. Такие метки как ‘2009-01-01’ отсутствуют, поскольку в этот момент законы не публиковались. Были проведены эксперименты по динамическому тематическому моделированию с изменением числа тем.

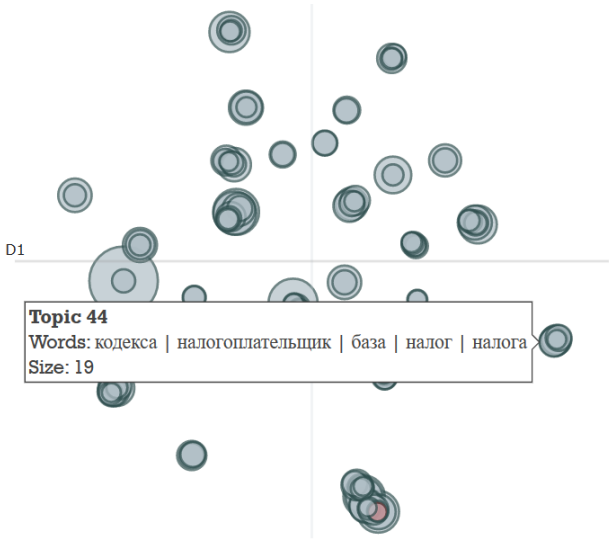

Рис. 4. «Карта расстояний» стандартной модели BERTopic (тема 44)
Fig. 4. The «distance map» for standard BERTopic model (topic 44)

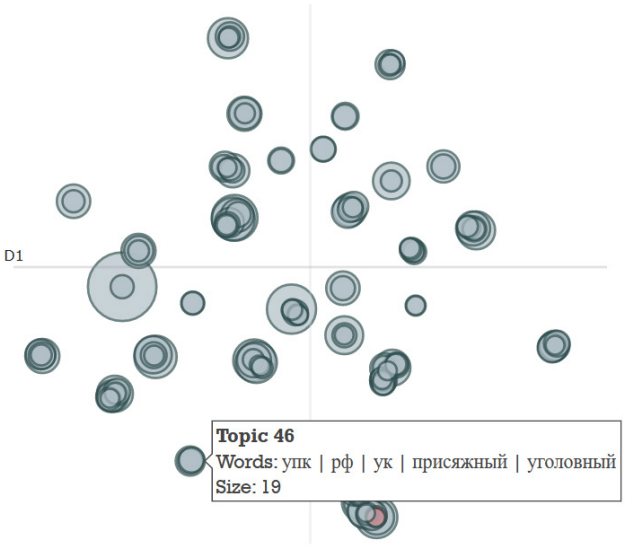

Рис. 5. «Карта расстояний» стандартной модели BERTopic (тема 46)
Fig. 5. The «distance map» for standard BERTopic model (topic 46)

На рис. 8 представлена визуализация изменений 13 наиболее значимых тем. При интерактивной работе с графиком можно отслеживать изменение состава темы: в определенный момент времени слова-тематизаторы могут быть стандартными (глобальными), либо специфическими (локальными). Как видно из графика, в начале 2018 г. резко увеличилось число документов, связанных с темами 1 и 9, которые относятся к кадровым изменениям. В начале 2018 г. прошли Президентские выборы и, в связи с этим, множество назначений на руководящие посты. В юридических документах с 2020 г. фигурирует тема борьбы с новой коронавирусной инфекцией COVID-19. На рис. 8 видно, что глобальный состав темы 2 определяется словами *коронавирусный*, *инфекция*, *covid* и т.д., а локальные темы отличаются друг от друга обозначениями регионов (*Забайкальский край*, *Ростовская область*, *Ямало-Ненецкий автономный округ* и т.д.)

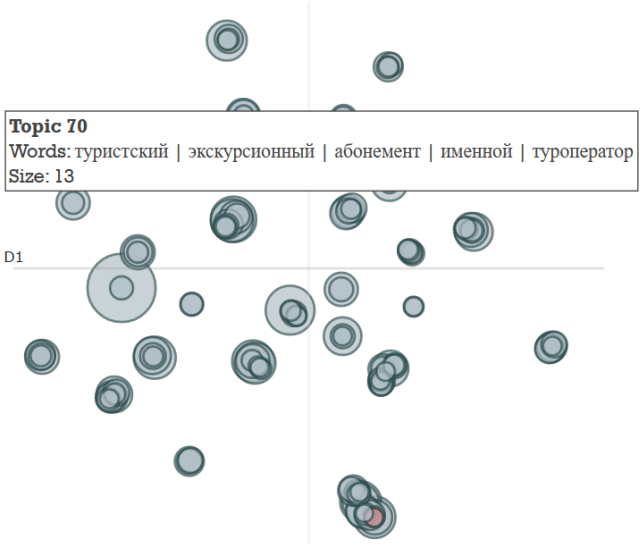

Рис. 6. «Карта расстояний» стандартной модели BERTopic (тема 70)

Fig. 6. The «distance map» for standard BERTopic model (topic 70)

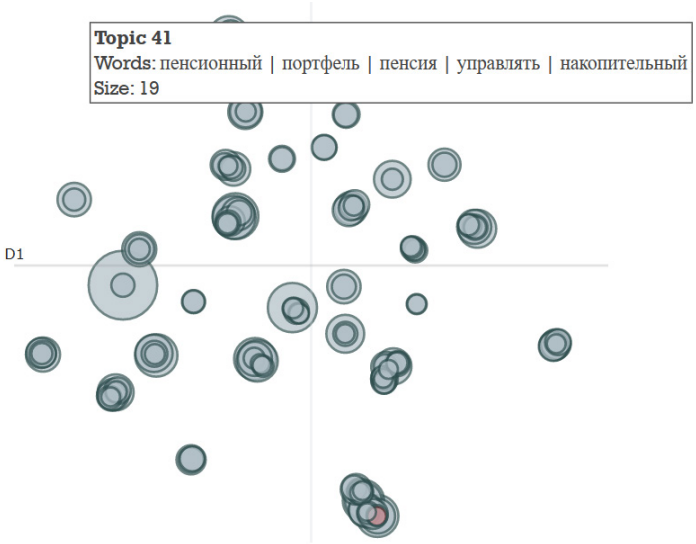

Рис. 7. «Карта расстояний» стандартной модели BERTopic (тема 41)

Fig. 7. The «distance map» for standard BERTopic model (topic 41)

На рис. 9 визуализирована динамика первых 20 значимых тем на отрезке от февраля 2021 г. по ноябрь 2022 г.

Желтым цветом отмечена динамика темы со словами-тематизаторами *подданства*, *гражданство*, *проживание*, *постоянный*, *государства* и т.д., график имеет пик, соответствующий июню 2021 г., и спад в осенне-зимний период. Суть в том, что в весенне-летний период появилось много документов по данной теме, и их число снизилось к концу 2021 г. Для рассматриваемой темы на пике ее популярности в июне 2021 г. характерны глобальные слова-тематизаторы, а вот 20 декабря 2021 г. ее состав изменился, в ней появились слова, обозначающие холодное оружие: *клинковый*, *холодное*, *кортики*, *кортиков*, *оружие* и т.д.

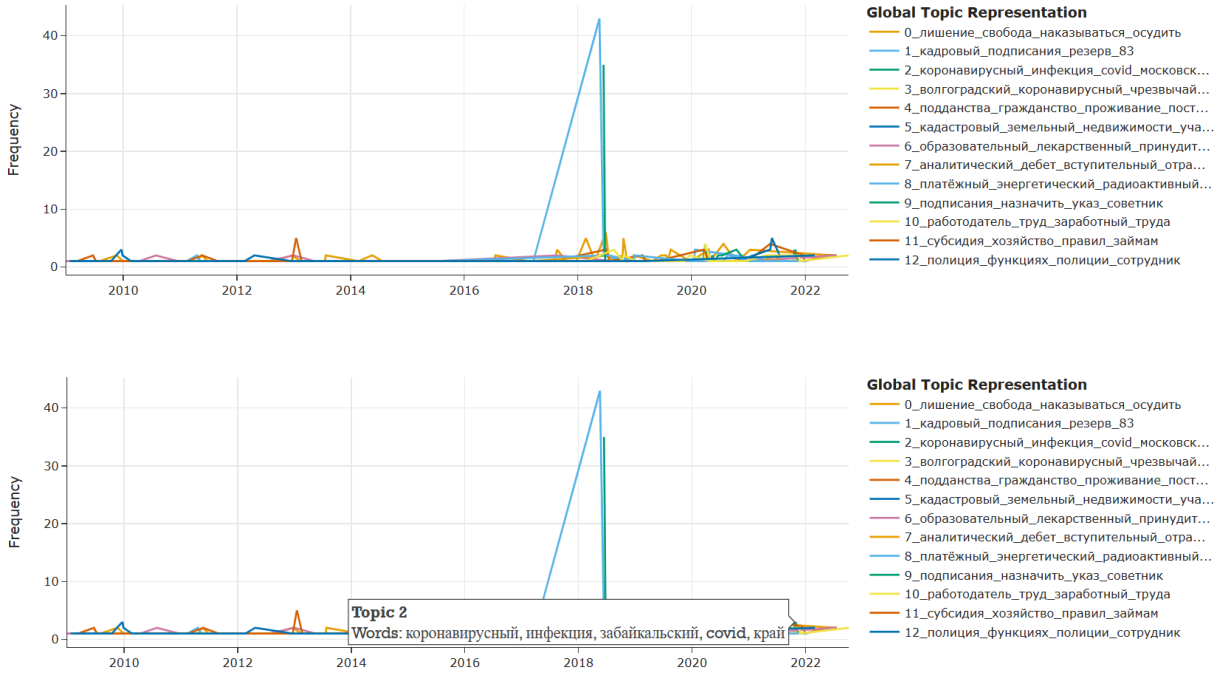

Рис. 8. Значимые темы динамической модели BERTopic

Fig. 8. Significant topics of dynamic BERTopic model

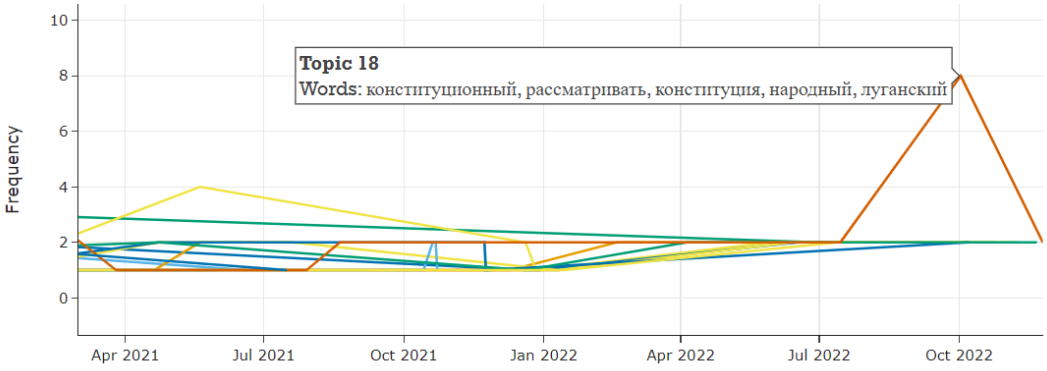

Рис. 9. Значимые темы динамической модели BERTopic (тема 18)

Fig. 9. Significant topics of dynamic BERTopic model (topic 18)

К концу выбранного нами периода, в октябре 2022 г., популярной темой в документах стала тема, связанная с конституционными изменениями в РФ и принятием в состав РФ новых регионов (Донецкая Народная Республика (ДНР), Луганская Народная Республика (ЛНР), Запорожская и Херсонская области). Данная тема на графике обозначена номером 18 и выделена красным цветом (рис. 9). В тот же период становится актуальной тема, связанная с указом об объявлении частичной мобилизации в РФ: 5 октября 2022 г. локальными ключевыми словами этой темы были: *задач*, *работодатель*, *мобилизация*, *контракт*, *добровольный*.

Рассмотрим картину динамики тем в исследовательском корпусе юридических документов за период от 31 декабря 2008 г. до конца 2011 г., см. график на рис. 10.

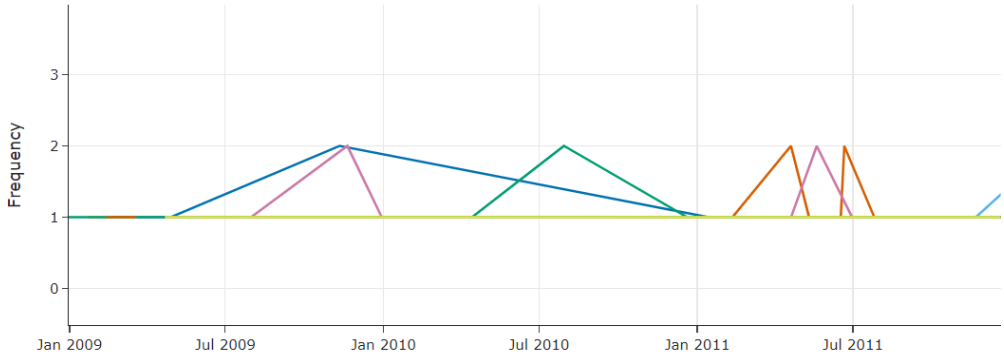

Рис. 10. Значимые темы динамической модели BERTopic (2008–2011 г.)
Fig. 10. Significant topics of dynamic BERTopic model (2008–2011)

Зеленым отмечен график темы со словами-тематизаторами, соотносимыми с правами на интеллектуальную собственность: *патентный*, *поверить*, *поверенного*, *интеллектуальный*, *квалификационный* и т.д. Синим помечен график глобальной темы правосудия ук, рф, наказание, преступление и т.д. Розовым показан график темы взяточничества: в ноябре 2009 г. она имела следующие локальные слова-тематизаторы: *лишение*, *крупный*, *такового*, *наказываться*, 1854 и т.д., а в мае 2011 г. – *взятка*, *лишение*, *штраф*, *взятки*, *наказываться* и т.д.

Интерфейс библиотеки BERTopic позволяет пользователю выбрать отдельную глобальную тему и проследить ее развитие за весь период времени, охваченный корпусом. На рис. 11 представлены изменения темы правосудия со словами-тематизаторами *ук*, *рф*, *наказание*, *преступление* и т.д. Как можно заметить, пики популярности уголовной тематики приходятся на конец 2013 г. и начало 2018 г.

В библиотеке BERTopic есть возможность проводить сравнительный анализ нескольких тем одновременно, см. рис. 12. На графике синим и красным цветами обозначены изменения в упоминании двух смежных тем, связанных с Конституцией РФ. До 2016 г. конституционная тематика не встречается в корпусе вообще. Пик популярности приходится на конец 2022 г., что может быть связано с присоединением к РФ новых территорий. Характерно, что летом 2021 г., когда вносились поправки в Конституцию РФ, популярность тем была минимальной.

Более подробно ознакомиться с результатами экспериментов можно источниках Jupyter Notebook[8] и GitHub[9].

Заключение

В данной статье представлены результаты экспериментального исследования динамики основных тем в корпусе русскоязычных юридических документов за 2008–2022 гг. с использованием методов тематического моделирования. Были построены стандартная и динамическая тематические модели корпуса с помощью библиотеки BERTopic, интегрирующей использование нейросетевых моделей распределенных векторных вложений типа трансформер (BERT), алгоритмы кластеризации и снижения размерности. В результате экспериментов были оценены возможности инструмента BERTopic в исследовании юридических текстов, осуществлена интерпретация полученных данных. Созданные тематические модели позволили выявить наиболее значимые события в жизни государства и общества, отраженные в юридических документах, помогли найти соответствующие им слова-тематизаторы, и, тем самыми, представить эти события в развитии с течением времени.

[8] https://drive.google.com/file/d/1MHGglLDtv4-HMSAgLHRFkHb0Y_Pey2Cn/view
[9] legal_dtm/BERTopic_DTM_legal_docs.ipynb at main · Athugodage/legal_dtm (github.com)

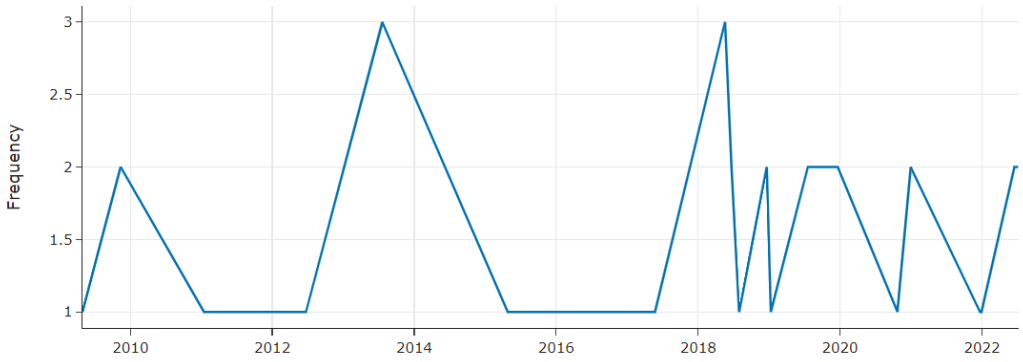

Рис. 11. Динамика темы *ук*, *рф*, *наказание*, *преступление* и т.д.

Fig. 11. Dynamic changes in the topic *Criminal code*, *Russian Federation*, *punishment*, *crime* etc.

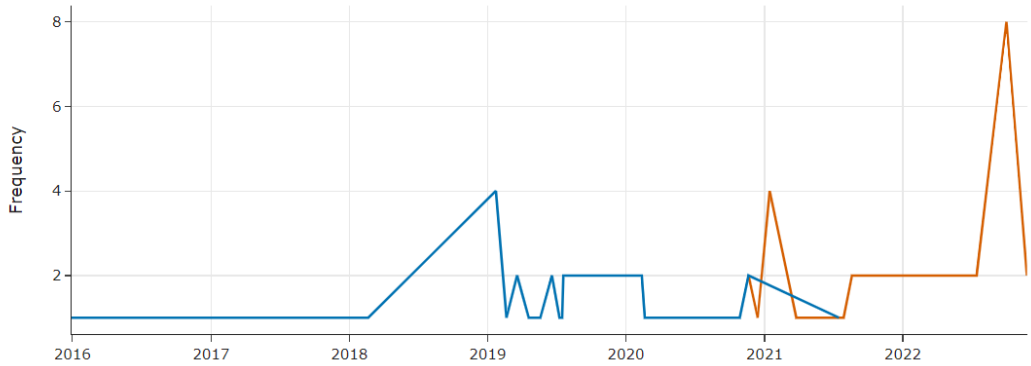

Рис. 12. Динамика двух смежных тем, связанных с Конституцией РФ

Fig. 12. Dynamic changes of two topics related to the Constitution of the Russian Federation

Данное исследование может быть полезно как компьютерным лингвистам, так и юристам, политологам, социологам и специалистам, исследующие историю развития российского законодательства в первой четверти XXI в. Проведенный анализ материала расширяет наши знания о созданном корпусе русскоязычных юридических документов, который предназначен для обучения алгоритмов упрощения текстов.

Результаты нашей работы позволят облегчить поиск релевантных документов и повысят доступность юридической информации для неспециалистов.

СПИСОК ИСТОЧНИКОВ

1. **Daud A., Li J., Zhou L., Muhammad F.** Knowledge discovery through directed probabilistic topic models: a survey // Proceedings of Frontiers of Computer Science in China, 2010. P. 280–301. URL: https://www.researchgate.net/publication/215904200_Latent_Dirichlet_allocation_LDA_and_topic_modeling_models_applications_future_challenges_a_survey

2. **Blei D.M., Ng A.Y., Jordan M.I.** Latent Dirichlet Allocation // Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. URL: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

3. **Милкова М.А.** Тематические модели как инструмент «дальнего чтения» // Цифровая экономика. № 1(5). 2019. С. 57–70. URL: http://digital-economy.ru/images/easyblog_articles/520/DE-2019-01-06.pdf?ysclid=lefaetif1z601986150

4. **Николенко С., Кадурин Е., Архангельская Е.** Глубокое обучение. Погружение в мир нейронных сетей. СПб., 2018. URL: https://b-ok.cc/book/4987601/95075c

5. **Кирина М.А.** Сравнение тематических моделей на основе LDA, STM и NMF для качественного анализа русской художественной прозы малой формы // Вестник НГУ. Серия: Лингвистика и межкультурная коммуникация. Вып. 20(2). 2022. С. 93–109. URL: https://lingngu.elpub.ru/jour/article/view/384

6. **Воронцов К.В.** Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM 2023. URL: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

7. **Карпович С.Н.** Русскоязычный корпус текстов СКТМ-ру для построения тематических моделей // Международная конференция «Корпусная лингвистика-2015». СПб., 2015.

8. **Shavrina T., Shapovalova O.** To the Methodology of Corpus Construction for Machine Learning: «TAIGA» Syntax Tree Corpus and Parser // Proceedings of the International Conference «Corpus Linguistics – 2017». Saint-Petersburg, 2019.

9. **Mitrofanova O., Kriukova A., Shulginov V., Shulginov V.** E-hypertext Media Topic Model with Automatic Label Assignment // Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science, vol. 1357. Springer, 2021. P. 102–114. URL: https://doi.org/10.1007/978-3-030-71214-3_9

10. **Кольцов С.Н., Кольцова О.Ю., Митрофанова О.А., Шиморина А.С.** Интерпретация семантических связей в текстах русскоязычного сегмента Живого Журнала на основе тематической модели LDA // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Материалы XVII Всероссийской объединенной конференции «Интернет и современное общество» IMS–2014, СанктПетербург, 19–20 ноября 2014 г. СПб., 2014. С. 135–142.

11. **Bodrunova S., Blekanov I.S., Kukarkin M.** Topics in the Russian Twitter and Relations between their Interpretability and Sentiment // 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019. P. 549–554.

12. **Mamaev I.D, Mitrofanova, O.A.** Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus // A. Filchenkov, J. Kauttonen, L. Pivovarova (Eds.). Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings. Communications in Computer and Information Science. Vol. 1292. Springer, 2020. P. 17–33. URL: https://doi.org/10.1007/978-3-030-59082-6_2

13. **Mitrofanova O., Sampetova V., Mamaev I., Moskvina A., Sukharev K.** Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling // CEUR Workshop Proceedings, 2813. 2021. P. 101–116.

14. **Nikolenko S.I., Koltcov S., Koltsova O.** Topic modelling for qualitative studies // Journal of Information Science. Vol. 43. 2017.

15. **Khawaji K., Almubark I., Almalki A., Taylor B.** Similarity Matching for Workflows in Medical Domain Using Topic Modeling // 2018 IEEE World Congress on Services (SERVICES). San Francisco, CA, USA, 2018.

16. **Шишкина В.С.** Тематическое моделирование финансовых потоков корпоративных клиентов банка по транзакционным данным. М., 2019.

17. **Dowling M., Piepenbrink A., Saqib A., Helmi H.** Machine learning in finance: A topic modeling approach. 2019. URL: https://arxiv.org/ftp/arxiv/papers/1911/1911.12637.pdf

18. **Vorontsov K.V., Voronov S.O.** Automatic Filtering of Russian Scientific Content using Machine Learning and Topic Modeling // International Conference on Computational Linguistics and Intellectual Technologies «Dialogue–2015». Moscow, 2015. URL: http://www.dialog-21.ru/media/2135/vorontsov.pdf

19. **Митрофанова О.А.** Вероятностное моделирование тематики русскоязычных корпусов текстов с использованием компьютерного инструмента GenSim // Труды международной конференции «Корпусная лингвистика–2015». СПб., 2015.

20. **Митрофанова О.А.** Исследование структурной организации художественного произведения с помощью тематического моделирования: опыт работы с текстом романа «Мастер и Маргарита» М.А. Булгакова // Труды международной конференции «Корпусная лингвистика–2019». СПб: Издательство Санкт-Петербургского университета, 2019.

21. **Скоринкин Д.А.** Семантическая разметка художественных текстов для количественных исследований в филологии (на примере романа «Война и мир» Л.Н. Толстого). Дис. ... канд. филол наук. М., 2019.

22. **Mitrofanova O.A., Sedova A.G.** Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose) // Information Technology and Computational Linguistics (ITCL 2017). Association for Computing Machinery, 2017.

23. **Rhody L.M.** Topic Modeling and Figurative Language // Journal of Digital Humanities. Vol. 2(1). Winter 2012. URL: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

24. **Sherstinova T., Mitrofanova O., Skrebtsova T., Zamiraylova E., Kirina T.** Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction // L. Martínez-Villaseñor, H. Ponce, O. Herrera-Alcántara, F.A. Castro-Espinoza (Eds.). Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Proceedings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 12469. Springer, 2020. P. 134–151. URL: https://doi.org/10.1007/978-3-030-60887-3_13

25. **Zamiraylova E., Mitrofanova O.** Dynamic topic modeling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization // R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL-2019: Proceedings of the III International Conference RWTH Aahen University. CEUR Workshop Proceedings. Vol. 2552. 2019. P. 321–339.

26. **Badenes-Olmedo C., Redondo-Garcia J.-L., Corcho O.** Legal document retrieval across languages: topic hierarchies based on synsets // arXiv. 2019. URL: https://arxiv.org/abs/1911.12637v1

27. **Дмитриева А.В.** «Искусство юридического письма»: количественный анализ решений Конституционного Суда Российской Федерации. Сравнительное конституционное обозрение. 2017. 118 (3). С. 125–133.

28. **Mattila H.E.S.** Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas. Routledge, 2013.

29. **Tiersma P.M.** Legal Language. Chicago, London: University of Chicago Press, 1999.

30. **Туранин В.Ю.** Юридическая терминология в современном российском законодательстве (теоретико-правовое исследование). Дис ... д-ра юрид. наук. Белгород: БГУ, 2017.

31. **Виландеберк А.А.** Принципы и методы гармонизации терминологии на основе корпуса специальных параллельных текстов: на материале документов ООН. Автореф. дис. ... канд филол. наук. СПб, РГПУ им. А.И.Герцена, 2005. URL: https://new-disser.ru/_avtoreferats/01002771726.pdf/f?ysclid=leciciqi30122006496

32. **Виландеберк А.А.** Корпус параллельных правовых документов как составная часть АРМ юриста-переводчика // Труды Международной научной конференции «Корпусная лингвистика 2004». СПб., 2004.

33. **Mirzagitova A.** Realisation of statistical machine translation based on a parallel Tatar-Russian corpus of legal texts Proceedings of the International Conference «Turkic Languages Processing: TurkLang – 2015». Kazan, 2015. P. 39–49.

34. **Блинова О.В., Белов С.А.** Языковая неоднозначность и неопределённость в русских правовых текстах. Вестник Санкт-Петербургского университета. Право. 2020. № 11(4). С. 774–812.

35. **Блинова О.В.** Оценка сложности русских правовых текстов: архитектура модели // Мир русского слова. 2022. № 2. С. 4–13.

36. **Блинова О.В., Тарасов Н.А.** Метрики сложности русских правовых текстов: отбор, использование, первичная оценка эффективности // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 21. М.: Изд-во РГГУ, 2022. С. 1017–1028.

37. **Кучаков Р.К., Савельев Д.А.** Сложность правовых актов в России: Лексическое и синтаксическое качество текстов / Под ред. Д. Скугаревского. СПб.: ИПП ЕУСПб, 2018.

38. **Кнутов А.В., Плаксин С.М., Григорьева Н.Л., Синятуллин Р.Х., Чаплинский А.В., Успенская А.М.** Сложность российских законов. Опыт синтаксического анализа. М.: Изд. дом Высшей школы экономики, 2020. 311 с.

39. **Greene D., Cross J.** Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach // Political Analysis. 2016. Vol. 25. P. 77–94. // arXiv. URL: https://arxiv.org/pdf/1607.03055.pdf

40. **Blei D.M., Lafferty J.D.** Dynamic topic models // Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, 2006. URL: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf

41. **Усманова Е.Ф.** Правовая культура российского общества в современных условиях // Мир науки и образования. 2016. № 3(7). URL: https://cyberleninka.ru/article/n/pravovaya-kultura-rossiysa kogo-obschestva-v-sovremennyh-usloviyah?ysclid=lefhwff9kf730175397

42. Федеральный закон от 14.06.1994 N 5-ФЗ (ред. от 01.05.2019) «О порядке опубликования и вступления в силу федеральных конституционных законов, федеральных законов, актов палат Федерального Собрания». URL: http://www.kremlin.ru/acts/bank/6332

43. **Dieng A.B., Ruiz F.J.R., Blei D.M.** Topic Modeling in Embedding Spaces // arXiv. 2019. URL: https://arxiv.org/abs/1907.04907

44. **Moody C.E.** Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2Vec // arXiv. 2016. URL: https://arxiv.org/abs/1605.02019

45. **Angelov D.** Top2Vec: Distributed Representations of Topics // arXiv. 2020. URL: https://arxiv.org/abs/2008.09470

46. **Bianchi F., Terragni S., Hovy D.** Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Volume 2: Short Papers. ACL, 2021. P. 759–766. URL: https://aclanthology.org/2021.acl-short.96/

47. Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv. 2022. URL: https://arxiv.org/pdf/2203.05794.pdf

REFERENCES

[1] **A. Daud, J. Li, L. Zhou, F. Muhammad,** Knowledge discovery through directed probabilistic topic models: a survey, Proceedings of Frontiers of Computer Science in China, 2010. Pp. 280–301. Available at: https://www.researchgate.net/publication/215904200_Latent_Dirichlet_allocation_LDA_and_topic_modeling_models_applications_future_challenges_a_survey

[2] **D.M. Blei, A.Y. Ng, M.I. Jordan,** Latent Dirichlet Allocation, Journal of Machine Learning Research. Vol. 3 (4–5). January 2003. Available at: http://jmlr.csail.mit.edu/papers/volume3/blei03a/blei03a.pdf

[3] **M.A. Milkova,** Topic models as a tool for "distant reading", Digital Economy. 1 (5) 2019 57–70. Available at: http://digital-economy.ru/images/easyblog_articles/520/DE-2019-01-06.pdf?ysclid=lefaetif1z601986150

[4] **S. Nikolenko, E. Kadurin, E. Arkhangelskaya,** Deep learning. Dive into the world of neural networks. SPb., 2018. Available at: https://b-ok.cc/book/4987601/95075c

[5] **M.A. Kirina,** Comparison of thematic models based on LDA, STM and NMF for a qualitative analysis of Russian short fiction. Vestnik NGU. Series: Linguistics and intercultural communication. 20 (2) 2022 93–109. Available at: https://lingngu.elpub.ru/jour/article/view/384

[6] **K.V. Vorontsov,** Probabilistic topic modeling: ARTM regularization theory and BigARTM open source library. 2023. Available at: http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf

[7.] **S.N. Karpovich,** Russian-language corpus of SKTM-ru texts for building thematic models, International Conference "Corpus Linguistics – 2015". SPb., 2015.

[8] **T. Shavrina, O. Shapovalova,** To the Methodology of Corpus Construction for Machine Learning: "TAIGA" Syntax Tree Corpus and Parser, Proceedings of the International Conference "Corpus Linguistics – 2017". Saint-Petersburg, 2019.

[9] **O. Mitrofanova, A. Kriukova, V. Shulginov, V. Shulginov,** E-hypertext Media Topic Model with Automatic Label Assignment, Recent Trends in Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Revised Supplementary Proceedings. Communications in Computer and Information Science, vol. 1357. Springer, 2021. Pp. 102–114. Available at: https://doi.org/10.1007/978-3-030-71214-3_9

[10] **S.N. Koltsov, O.Yu. Koltsova, O.A. Mitrofanova, A.S. Shimorina,** Interpretation of semantic links in the texts of the Russian-language segment of LiveJournal based on the LDA thematic model, Technologies of the information society in science, education and culture: a collection of scientific articles. Proceedings of the XVII All-Russian Joint Conference "Internet and Modern Society" IMS-2014, St. Petersburg, November 19-20, 2014. St. Petersburg, 2014. Pp. 135–142.

[11] **S. Bodrunova, I.S. Blekanov, M. Kukarkin,** Topics in the Russian Twitter and Relations between their Interpretability and Sentiment, 2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS). 2019. Pp. 549–554.

[12] **I.D. Mamaev, O.A. Mitrofanova,** Automatic Detection of Hidden Communities in the Texts of Russian Social Network Corpus, A. Filchenkov, J. Kauttonen, L. Pivovarova (Eds.). Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings. Communications in Computer and Information Science. Vol. 1292. Springer, 2020. Pp. 17–33. Available at: https://doi.org/10.1007/978-3-030-59082-6_2

[13] **O. Mitrofanova, V. Sampetova, I. Mamaev, A. Moskvina, K. Sukharev,** Topic modelling of the Russian corpus of Pikabu posts: Author-topic distribution and topic labelling, CEUR Workshop Proceedings, 2813. 2021. Pp. 101–116.

[14] **S.I. Nikolenko, S. Koltcov, O. Koltsova,** Topic modelling for qualitative studies, Journal of Information Science, 43 (2017).

[15] **K. Khawaji, I. Almubark, A. Almalki, B. Taylor,** Similarity Matching for Workflows in Medical Domain Using Topic Modeling, 2018 IEEE World Congress on Services (SERVICES). San Francisco, CA, USA, 2018.

[16] **V.S. Shishkina,** Tematicheskoye modelirovaniye finansovykh potokov korporativnykh kliyentov banka po tranzaktsionnym dannym [Thematic modeling of financial flows of corporate clients of the bank based on transactional data]. M., 2019.

[17] **M. Dowling, A. Piepenbrink, A. Saqib, H. Helmi,** Machine learning in finance: A topic modeling approach, 2019. Available at: https://arxiv.org/ftp/arxiv/papers/1911/1911.12637.pdf

[18] **K.V. Vorontsov, S.O. Voronov,** Automatic Filtering of Russian Scientific Content using Machine Learning and Topic Modeling, International Conference on Computational Linguistics and Intellectual Technologies "Dialogue–2015". Moscow, 2015. Available at: http://www.dialog-21.ru/media/2135/vorontsov.pdf

[19] **O.A. Mitrofanova,** Probabilistic modeling of the topics of Russian text corpora using the GenSim computer tool, Proceedings of the international conference "Corpus Linguistics – 2015". SPb., 2015.

[20] **O.A. Mitrofanova,** The study of the structural organization of a work of art using thematic modeling: experience with the text of the novel "The Master and Margarita" M.A. Bulgakova, Proceedings of the International Conference "Corpus Linguistics – 2019". St. Petersburg: St. Petersburg University Press, 2019.

[21] **D.A. Skorinkin,** Semantic markup of literary texts for quantitative research in philology (on the example of the novel "War and Peace" by L.N. Tolstoy). PhD Thesis. M., 2019.

[22] **O.A. Mitrofanova, A.G. Sedova,** Topic Modelling in Parallel and Comparable Fiction Texts (the case study of English and Russian prose), Information Technology and Computational Linguistics (ITCL 2017). Association for Computing Machinery, 2017.

[23] **L.M. Rhody,** Topic Modeling and Figurative Language, Journal of Digital Humanities. Vol. 2(1). Winter 2012. Available at: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/

[24] **T. Sherstinova, O. Mitrofanova, T. Skrebtsova, E. Zamiraylova, T. Kirina,** Topic modelling with NMF vs. expert topic annotation: the case study of Russian fiction, L. Martínez-Villaseñor, H. Ponce, O. Herrera-Alcántara, F.A. Castro-Espinoza (Eds.). Advances in Computational Intelligence. 19th Mexican International Conference on Artificial Intelligence, MICAI 2020, Proceedings. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Vol. 12469. Springer, 2020. Pp. 134–151. Available at: https://doi.org/10.1007/978-3-030-60887-3_13

[25] **E. Zamiraylova, O. Mitrofanova,** Dynamic topic modeling of Russian fiction prose of the first third of the XXth century by means of non-negative matrix factorization, R. Piotrowski's Readings in Language Engineering and Applied Linguistics. PRLEAL–2019: Proceedings of the III International Conference RWTH Aahen University. CEUR Workshop Proceedings. Vol. 2552. 2019. Pp. 321–339.

[26] **C. Badenes-Olmedo, J.-L. Redondo-Garcia, O. Corcho,** Legal document retrieval across languages: topic hierarchies based on synsets, arXiv. 2019. Available at: https://arxiv.org/abs/1911.12637v1

[27] **A.V. Dmitrieva,** "The art of legal writing": a quantitative analysis of the decisions of the Constitutional Court of the Russian Federation. Comparative constitutional review. 118 (3) (2017) 125–133.

[28] **H.E.S. Mattila,** Comparative Legal Linguistics: Language of Law, Latin and Modern Lingua Francas. Routledge, 2013.

[29] **P.M. Tiersma,** Legal Language. Chicago, London: University of Chicago Press, 1999.

[30] **V.Yu. Turanin,** Yuridicheskaya terminologiya v sovremennom rossiyskom zakonodatelstve (teoretiko-pravovoye issledovaniye). Dis ... d-ra yurid. nauk. Belgorod: BGU, 2017.

[31] **A.A. Vilandeberk,** Principles and methods of terminology harmonization based on the corpus of special parallel texts based on UN documents. PhD Thesis Abstract. St. Petersburg, Russian State Pedagogical University, 2005. Available at: https://new-disser.ru/_avtoreferats/01002771726.pdf?ysclid=leciciqi30122006496

[32] **A.A. Vilandeberk,** Corpus of Parallel Legal Documents as a Part of AWP of a Lawyer-Translator, Proceedings of the International Scientific Conference "Corpus Linguistics 2004". SPb., 2004.

[33] **A. Mirzagitova,** Realisation of statistical machine translation based on a parallel Tatar-Russian corpus of legal texts Proceedings of the International Conference "Turkic Languages Processing: TurkLang – 2015". Kazan, 2015. Pp. 39–49.

[34] **O.V. Blinova, S.A. Belov,** Linguistic ambiguity and uncertainty in Russian legal texts. Bulletin of St. Petersburg University. Right. 11 (4) (2020) 774–812.

[35] **O.V. Blinova,** Assessing the complexity of Russian legal texts: the architecture of the model, The World of the Russian Word, 2 (2022) 4–13.

[36] **O.V. Blinova, N.A. Tarasov,** Metrics of complexity of Russian legal texts: selection, use, primary evaluation of effectiveness, Computational Linguistics and Intelligent Technologies: Based on the materials of the annual international conference "Dialogue". Issue. 21. Moscow: RGGU Press, 2022. Pp. 1017–1028.

[37] **R.K. Kuchakov, D.A. Saveliev,** Complexity of Legal Acts in Russia: Lexical and Syntactic Quality of Texts / Ed. D. Skugarevsky. St. Petersburg: IPP EUSPb, 2018.

[38] **A.V. Knutov, S.M. Plaksin, N.L. Grigor'yeva, R.H. Sinyatullin, A.V. Chaplinskiy, A.M. Uspenskaya,** Complexity of Russian Laws. Parsing experience. M.: Ed. house of the Higher School of Economics, 2020.

[39] **D. Greene, J. Cross,** Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach, Political Analysis. 25 (2016) 77–94. arXiv. Available at: https://arxiv.org/pdf/1607.03055.pdf

[40] **D.M. Blei, J.D. Lafferty,** Dynamic topic models, Proceedings of the 23rd International Conference on Machine Learning. Pittsburgh, 2006. Available at: http://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf

[41] **E.F. Usmanova,** Legal culture of the Russian society in modern conditions, World of Science and Education. 3 (7) (2016). Available at: https://cyberleninka.ru/article/n/pravovaya-kultura-rossiyskogo-obschestva-v-sovremennyh-usloviyah?ysclid=lefhwff9kf730175397

[42] Federal Law № 5-FZ of June 14, 1994 (as amended on May 1, 2019) "On the Procedure for Publication and Entry into Force of Federal Constitutional Laws, Federal Laws, and Acts of the Chambers of the Federal Assembly." Available at: http://www.kremlin.ru/acts/bank/6332

[43] **A.B. Dieng, F.J.R. Ruiz, D.M. Blei,** Topic Modeling in Embedding Spaces, arXiv. 2019. Available at: https://arxiv.org/abs/1907.04907

[44] **C.E. Moody,** Mixing Dirichlet Topic Models and Word Embeddings to Make Lda2Vec, arXiv. 2016. Available at: https://arxiv.org/abs/1605.02019

[45] **D. Angelov,** Top2Vec: Distributed Representations of Topics, arXiv. 2020. Available at: https://arxiv.org/abs/2008.09470

[46] **F. Bianchi, S. Terragni, D. Hovy,** Pre-training is a Hot Topic: Contextualized Document Embeddings Improve Topic Coherence, Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Vol. 2: Short Papers. ACL, 2021. Pp. 759–766. Available at: https://aclanthology.org/2021.acl-short.96/

[47] **M. Grootendorst,** BERTopic: Neural topic modeling with a class-based TF-IDF procedure, arXiv. 2022. Available at: https://arxiv.org/pdf/2203.05794.pdf

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Митрофанова Ольга Александровна
Olga A. Mitrofanova
E-mail: o.mitrofanova@spbu.ru
ORCID: https://orcid.org/0000-0002-3008-5514

Атугодаге Марк Махешевич
Mark M. Athugodage
E-mail: m.athugodage@yahoo.com

Поступила: 22.01.2023; Одобрена: 16.03.2023; Принята: 17.03.2023.
Submitted: 22.01.2023; Approved: 16.03.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 88–97.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 88–97.

Научная статья
УДК 808.1
DOI: https://doi.org/10.18721/JHSS.14108

О СМЕНЕ ПАРАДИГМЫ АВТОРСКОГО ИНВАРИАНТА

А.А. Рогов, Н.Д. Москин, А.А. Лебедев

Петрозаводский государственный университет,
г. Петрозаводск, Российская Федерация

moskin@petrsu.ru

Аннотация. Одной из актуальных задач в филологии является атрибуция текстов. Количественный показатель, с помощью которого можно различить произведения разных авторов, нужно назвать авторским инвариантом. В работе описывается ряд исследований (методика Г. Хетсо, методика оценки парной связи грамматических классов, методика «деревьев решений», Дельта-метод), результаты которых подтверждают, что изначальное определение авторского инварианта следует скорректировать. В частности, это касается временного промежутка, на котором параметр атрибуции должен сохранять «постоянное значение». Он не обязательно совпадает со всем периодом творчества писателя. Также из-за отсутствия универсального критерия, однозначно отличающего конкретного писателя от других, следует использовать совокупность характеристик авторских инвариантов на разных уровнях языка. Проведенный анализ показывает, что термин «авторский инвариант» следует разбить на две категории – «глобальный авторский инвариант» и «локальный авторский инвариант» – которые могут последовательно изучаться независимо друг от друга.

Ключевые слова: литературный текст, лингвостатистический параметр, авторский инвариант, классификация, интеллектуальный анализ данных.

Для цитирования: Рогов А.А., Москин Н.Д., Лебедев А.А. О смене парадигмы авторского инварианта // Terra Linguistica. 2023. Т. 14. № 1. С. 88–97. DOI: 10.18721/JHSS.14108

Research article
DOI: https://doi.org/10.18721/JHSS.14108

ON THE PARADIGM SHIFT OF THE AUTHOR'S INVARIANT

A.A. Rogov , N.D. Moskin , A.A. Lebedev

Petrozavodsk State University,
Petrozavodsk, Russian Federation

moskin@petrsu.ru

Abstract. One of the urgent tasks in philology is the attribution of texts. The quantitative indicator by which one can distinguish between the works of different authors should be called the author's invariant. The paper describes a number of studies (the method of G. Kjetsaa, the method of evaluating the pair connection of grammatical classes, the method of "decision trees", the Delta method), the results of which confirm that the initial definition of the author's invariant should be corrected. In particular, this applies to the time interval during which the attribution parameter should keep “constant value”. It does not necessarily coincide with the entire period of the writer's work. Also due to the lack of a universal criterion that uniquely distinguishes a particular writer from others, one should use a set of characteristics of author's invariants at different levels of the language. The performed analysis shows that the term "author's invariant" should be divided into two categories – "global author's invariant" and "local author's invariant" – which can be consistently studied independently of each other.

Keywords: literary text, linguostatistical parameter, author's invariant, classification, data mining.

Citation: A.A. Rogov, N.D. Moskin, A.A. Lebedev, On the paradigm shift of the author's invariant, Terra Linguistica, 14 (1) (2023) 88–97. DOI: 10.18721/JHSS.14108

Введение

Вопрос об определении авторства текста – один из наиболее актуальных в филологии. Псевдонимные и анонимные произведения существовали еще в далеком прошлом; даже в эпоху античности появлялись и литературные подделки, и стилизации, и банальная путаница. Однако вопрос атрибуции подобных текстов не был устранен и после появления книгопечатания. В связи с этим важную роль могло бы сыграть формирование некоего метода, позволяющего при помощи современных информационных технологий решать данные вопросы атрибуции с высокой степенью достоверности.

Но хотя сегодня существует большое число различных способов, позволяющих определить авторство произведений, универсального метода, который был бы признан всеми исследователями, не выявлено. Очевидно, что основой изучения авторского стиля становится анализ языка конкретных литературных произведений определенного автора в контексте общеязыковых закономерностей, присущих тому или иному этапу развития национального языка и конкретному литературному стилю. Однако закономерной является постановка вопроса о существовании подобного метода как такового – можно ли пронаблюдать некоторые лингвистические универсалии, которые прослеживаются на протяжении всего творчества автора и могут быть эффективным инструментом для определения индивидуально-авторского стиля?

В отечественной лингвистике подобного рода исследования, проводимые на стыке филологии и математики, имеют давнюю историю. Одним из первых авторов, которые в практике анализа русскоязычных текстов стали опираться на такие универсальные компоненты, был Н.А. Морозов [1]. Еще в начале XX века он в своей работе «Лингвистические спектры, как средство для отличения плагиатов от истинных произведений того или другого известного автора и для определения их эпохи)» ввел понятие «лингвистический спектр» (график частоты встречаемости определенных слов или грамматических категорий в творчестве определенного автора), и на базе текстов А.С. Пушкина, Н.В. Гоголя, А.Н. Толстого, И.С. Тургенева, Н.М. Карамзина, М.Н. Загоскина

определил значимость служебных частиц в творчестве данных авторов. Н.А. Морозов считал, что такое исследование творчества позволяет «узнавать авторов», метафорически представляя свою идею следующим образом: «подобно тому, как каждый автор, всегда оставаясь человеком, имеет свою индивидуальную физиономию, так и его язык, все время оставаясь русским, или английским, или французским, обнаруживает свои особенные черты <...>» [1].

Однако более востребованным в работах на стыке математики и лингвистики стало понятие «авторский инвариант», которое было заявлено в работе В.П. Фоменко и Т.Г. Фоменко «Авторский инвариант русских литературных текстов» [2]. Исследователи под авторским инвариантом понимают числовую характеристику, соответствующую ряду требований:

1) это должен быть «бессознательный параметр», который укоренен столь глубоко, что автор о нем не задумывается, а если и пытается держать его под контролем, то не в состоянии проделывать это в течение долгого времени (вследствие чего тот возвращается в исходное состояние);

2) для текстов конкретного автора такой параметр должен быть постоянным (слабо отклоняться от среднего значения) на всех его произведениях;

3) с помощью данного параметра должны различаться разные группы авторов (т. е. он должен позволить различить писателей).

Только комбинация из трех вышеуказанных условий позволяет говорить о формировании авторского инварианта. Среди множества параметров в качестве авторского инварианта исследователями была выбрана частота употребления служебных слов, которая оказывается стабильным показателем на протяжении всего творчества писателя. Было отмечено, что именно частотность употребления предлогов, союзов и частиц соответствует всем вышеозначенным критериям (писатель, формирующий текст, специально не задумывается над служебными частями речи; параметр прослеживается во всех произведениях автора; параметр позволяет разграничивать авторов). На основании данного критерия авторами было поставлено под сомнение авторство «Тихого Дона» М.А. Шолохова.

Следует особо отметить, что вопрос об авторстве «Тихого Дона» является одним из наиболее сложных в аспекте изучаемой проблемы. Эту тему затрагивали и специалисты-литературоведы, и языковеды, в том числе и зарубежные, причем, как кажется, точка в этой дискуссии не поставлена до сих пор: существуют исследования, как указывающие на стилистическое единство всех частей «Тихого Дона», так и на значительные отличия (опираясь на язык романа, его хронологию, историческое содержание и т.п.). Не ставя перед собой задачу дать полноценный обзор данной конкретной проблемы, сошлемся на выводы, которые сделали В.П. Фоменко и Т.Г. Фоменко. Исследователи отмечают, что «количество служебных слов в его произведениях оказалось настолько неодинаковым, что появляется необходимость представить Шолохова в виде двух авторов, которых мы условно назвали: Шолохов I и подозреваемый Шолохов II» [2]. В соответствии с полученными данными, авторы статьи делают вывод о том, что «Статистические результаты, полученные в результате анализа авторского инварианта, подтверждают гипотезу, что части 1, 2, 3, 4, 5 и в значительной мере часть 6 романа “Тихий Дон”» написаны не М.А. Шолоховым» [2].

Схожая методика была апробирована Г. Хетсо в работе «Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах «Время» и «Эпоха» [3], ставшей одним из наиболее известных в аспекте атрибуции псевдонимных и анонимных статей, которые были опубликованы в журналах «Время» и «Эпоха», с опорой на математические методы. Исследование включало в себя 15 лингвостатистических параметров, 10 грамматических параметров, а также дополнительные критерии (например, лексический спектр текста на уровне словаря и на уровне текста, а также индекс разнообразия лексики). Подробное критическое исследование методики Г. Хетсо представлено во второй главе монографии «Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин»» [4].

В целом, идея авторского инварианта ставилась под сомнение более поздними исследователями. Из числа наиболее значимых работ отметим статью Т.В. Батуры «Формальные методы определения авторства текстов», где указывается, в частности, что значимым ограничением данного метода выступает «очень низкая разделительная способность оценки в случае большого числа авторов (потенциально метод может разделять лишь 10 авторских стилей)» [5]. Критические замечания, связанные с категорией инварианта, отмечаются и в статье Е.Б. Трофимовой «Инвариант: реальность или фикция» [6] (неясно, насколько обязательно присутствие инварианта во всей совокупности его вариантов; отсутствует строго формальная процедуры идентификации отношений такого рода), а также в работе А.А. Боронина «Инвариант в текстовой проекции (интерпретация персонажных субтекстов в условиях эксперимента)», где отмечается, что «инвариант как лингвистическая абстракция условен и не претендует на абсолютную точность» [7]. Особо этот вопрос рассматривается не только в аспекте анализа уже имеющихся произведений, но и с точки зрения генерации и моделирования новых текстов. В частности, в статье А.В. Мордвинова «Системный подход в моделировании текста» [8] отмечается, что при воплощении текста в конкретной языковой форме «происходит потеря информации, которая выражается, в том числе, и в потере части авторских инвариантов», и, как следствие, происходит частичное искажение интегрального понятия «авторский инвариант».

В соответствии с проведенными исследованиями, мы можем предположить, что само понятие авторского инварианта как некоей постоянной величины, сохраняющейся во всех произведениях автора, может быть подвергнуто сомнению. Стиль автора может меняться со временем; более того, даже в пределах одного временного промежутка на отдельные аспекты построения структуры текста могут быть наложены определенные ограничения. Как отмечает О.В. Карелова, «своеобразие индивидуального слога обусловлено как индивидуальными особенностями автора, так и историческими событиями, происходящими в тот или иной период времени и влияющими на творчество писателя» [9]. К таким глобальным историческим событиям, к примеру, может быть отнесена цензура, характерная для определенных временных промежутков (в частности, И. С. Чирскова указывает на то, что «в борьбе с цензурой русская литература оттачивала язык общения с читателем, научившись выражаться емко, иносказательно и многозначительно» [10], что не могло не отразиться на отдельных особенностях авторского идиостиля, реализуемого в тексте произведения). Таким образом, говорить об авторском инварианте в аспекте индивидуально-авторского стиля нужно с большой аккуратностью, учитывая стилистические особенности подвергаемых анализу произведений.

Методика исследования

Однако проводимые в последние десятилетия исследования указывают на отсутствие подобных инвариантов. Используемые характеристики авторских инвариантов противоречат второму или третьему пункту основных свойств, заявленных в изначальном определении. Рассмотрим несколько примеров:

1) Проанализируем в качестве возможного авторского инварианта **частоту употребления служебных слов**. Заметим, что описание проведенных экспериментов в [2] не позволяет сделать вывод о том, что найденный параметр «количество служебных слов» действительно является авторским инвариантом. Как отмечают сами авторы, он ведет себя по-разному на больших и малых текстах. Для его стабилизации требуется анализ фрагментов объемом 16000 слов. При этом, как указано в [2], «близкие значения инварианта отнюдь не означают, что исследуемые произведения написаны одним автором. Как мы отмечали, встречаются разные писатели с близкими значениями инварианта. Например, Леонов и Фадеев, у которых эти числа равны соответственно 23,08 и 23,40». Это утверждение явно противоречит третьему пункту основных свойств авторского инварианта.

2) **Методика Г. Хетсо**, заявленная в работе «Принадлежность Достоевскому: к вопросу об атрибуции Ф.М. Достоевскому анонимных статей в журналах «Время» и «Эпоха» [3]. Проведенные исследования, описанные в монографии [4] показывают, что ни один из 15 признаков и методик их оценки нельзя использовать как авторский инвариант самостоятельно или в комбинации с другими. Главное отличие описанного эксперимента от оригинального исследования, проведенного Г. Хетсо, заключается в том, что была исследована вторая группа текстов, отсутствующая в исследованиях Г. Хетсо. Это было связано с выявлением наличия ошибки второго рода: необходимо было сравнить тексты Достоевского со статьями, автором которых точно не является писатель. Поэтому в исследование вошли тексты М.М. Достоевского, Н.Н. Страхова и А.А. Григорьева, где редакторское вмешательство Ф.М. Достоевского исключено. По всем параметрам тексты Григорьева, Страхова и М.М. Достоевского попали в группу статей, автором которых является Ф.М. Достоевский. Наиболее эффективными оказались параметры, связанные с лексикой, то есть признаки № 15 – индекс разнообразия лексики (точность составила всего 61%, не отвергнутыми оказались 7 текстов из 18 проанализированных) и № 14 – лексический спектр текста на уровне текста (точность 56%). У остальных признаков точность оказалась еще ниже. Такие результаты явно противоречат третьему пункту основных свойств авторского инварианта.

3) **Методика оценки парной связи грамматических классов**. Она достаточно подробно описана в монографии [11]. Методика основана на гипотезе, в соответствии с которой «...стиль автора проявляется в «пристрастии» к определенным грамматическим связям, частота появления которых в тексте высока. Остальные связи соответствует слабым, несущественным статистическим связям на уровне грамматических классов; частоты этих связей малы ..., а их появление в сильной мере случайно ...». В указанной монографии описан эксперимент успешного анализа с помощью данной методики разных списков «Повести временных лет». В монографии [4] описан эксперимент о возможности использования данной методики для разделения произведений следующих писателей: Ф.М. Достоевского, М.М. Достоевского, В.П. Мещерского, К.П. Победоносцева, Я.П. Полонского, А.А. Григорьева, Н.Н. Страхова, П.К. Щебальского. Оказалось, что даже полный перебор настроечных параметров методики не позволяет последовательно разделить тексты этих авторов. Вновь наблюдается явное противоречие третьему пункту основных свойств авторского инварианта. Заметим, что проведенные эксперименты (после нахождения соответствующих настроечных параметров) позволяли корректно разграничить произведения двух и трех авторов.

4) **Методика «деревья решений»**. Данная методика и ее обобщение под названием «Случайный лес» успешно используется в машинном обучении. В частности, в статье [12] деревья решений анализируются как один из подходов к автоматической классификации текстов с опорой на понятие авторского стиля и с высокой точностью различения таких индивидуальных стилей. Эксперименты, описанные в монографии [4], позволяют сделать следующее заключение: точность данной методики достигала 95%. Точность отделения Ф.М. Достоевского от других авторов оказывалась ниже, чем при попытке разделить двух конкретных авторов. Для решения задачи по различению произведений двух разных авторов использовались разные деревья, что вновь противоречит третьему пункту основных свойств авторского инварианта.

5) **Дельта-метод**, основывающийся на расстоянии Дельта-Бёрроуза. Метод был предложен в 2001 году австралийским филологом Джоном Бёрроузом, и используется во многих исследованиях, большая часть которых посвящена установлению авторства различных произведений.

Метод работает следующим образом: для каждого текста строятся лексические спектры самых частотных слов, а затем рассматривается расстояние между этими спектрами. Эмпирическая закономерность заключается в том, что расстояние для текстов, написанных одним автором, оказывается меньше, чем для любой пары текстов, написанных разными людьми. Для сравнения расстояний между текстами и разбиения их на кластеры используются различные вариации

иерархического кластерного анализа, чаще всего – метод Варда. Визуализации результатов кластеризации представляется в виде дендрограммы.

Исследование, опубликованное в работе [13], показывает, что расстояние между произведениями многих авторов (например, между М. Булгаковым и Н. Островским, между Л. Леоновым и А. Фадеевым и т.д.) значительно меньше, чем между произведениями М. Шолохова «Тихий Дон» и «Поднятая целина». Из этого можно сделать вывод: авторский инвариант, определяемый с помощью данного метода, явно противоречит второму пункту основных свойств (его устойчивость на протяжении жизни автора может быть поставлена под сомнение).

В работе [14] описано применение различных модификаций данного метода, разного набора частотного словаря и преобразования исходных данных для определения авторства романов «Двенадцать стульев» и «Золотого теленка». Несмотря на в целом интересные полученные результаты в данной работе, можно заметить неустойчивость данной методики в области определения авторского стиля. Аналогичные результаты получены в работе [14] при сравнении стилей публицистических статей Ф.М. Достоевского, М.М. Достоевского, В.П. Мещерского, К.П. Победоносцева, Я.П. Полонского, А.А. Григорьева, Н.Н. Страхова, П.К. Щебальского. Из работы видно, что атрибутировать тексты В.П. Мещерского и К.П. Победоносцева от других авторов удалось успешно – их кластер собрался в результате разных экспериментов. Для А.А. Григорьева и М.М. Достоевского тексты группировались в кластер за исключением нескольких текстов – выбросов. Произведения Ф.М. Достоевского и П.К. Щебальского разбивались на несколько однородных относительно расстояния дельта-Берроуза кластеров.

Опять замечаем явное противоречие третьему пункту основных свойств авторского инварианта.

6) **Нейросетевые методы**. В последние десятилетия при анализе литературных текстов активно используют методы машинного обучения. Особое внимание в этом аспекте уделяется нейронным сетям, которые показали свою эффективность в машинном переводе, в решении задачи построения ответа на задаваемые вопросы, а также в рассмотрении некоторых других лингвистических проблем. Как следствие, возникает идея попробовать отыскать с их помощью авторский инвариант. В работе [4] описывается сравнение следующих методов для атрибуции текстов: дерево решений, рекуррентные сети без модификаций и с LSTM ячейками и трансформер. Лучшие результаты показал трансформер. Заметим, что в нейросетевых методах, как правило, отсутствует визуализация обоснования принятия решения. Это существенно затрудняет их использование в научных целях. Однако, проведенные эксперименты показали, что они тоже не могут достигнуть 100% точности на тестовых данных. Значит, тот набор признаков, который используют данные методы, не удовлетворяет третьему пункту основных свойств авторского инварианта.

Данный перечень может быть продолжен с опорой на другие признаки и методики.

Заметим, что существует целый ряд публикаций, где успешно решаются отдельные задачи по кластеризации текстов, например, [15, 16, 17]. Однако примененные там признаки и методы их оценки нельзя рассматривать как авторский инвариант, поскольку в данных работах не проводится исследование на наличие ошибок второго рода, устойчивость признаков за пределами рассматриваемых текстов и т. д. Следует заметить, что применение ансамблевых методов при атрибуции текстов заранее предполагает неиспользование авторского инварианта в описанном ранее смысле.

Заключение

Проведенный анализ показывает, что термин «авторский инвариант» следует разбить на две категории – «глобальный авторский инвариант» и «локальный авторский инвариант» – которые могут последовательно изучаться независимо друг от друга, а выбор между ними осуществляется с учетом тех конкретных задач, которые ставит перед собой исследователь.

Под глобальным авторским инвариантом понимается количественный показатель (группа количественных показателей), который удовлетворяет первому и второму свойству исходного определения. К его недостаткам следует отнести возможную существенную ошибку второго рода (принять гипотезу о принадлежности текста автору, если она не верна). Однако, за счет выбора критического параметра, его можно настроить на минимальную ошибку первого рода (отвергнуть верную гипотезу). Поэтому его надо использовать для опровержения гипотезы об авторстве, а не для ее принятия. Для нахождения глобального авторского инварианта нужно четко определить условия проведения эксперимента. Прежде всего, должна быть возможность его повторения, ясно определены признаки, размеры фрагментов, на которых они рассчитываются, круг текстов и т. д. Заметим, что размеры текстов в этом случае играют существенную роль, что было отмечено целым рядом исследователей (после проведения многочисленных экспериментов возникла своеобразная поговорка – «если что-то идет не так – смотри размер текста»). К глобальному авторскому инварианту можно отнести количественные показатели, рассмотренные в работах Н.А. Морозова, В.П. Фоменко, Т.Г. Фоменко, Г. Хетсо и пр.

Под локальным авторским инвариантом следует понимать количественный показатель, с помощью которого можно различить произведения разных авторов, который удовлетворяет следующим свойствам:

1) первое свойство, связанное со слабой контролируемостью такого анализируемого параметра автором, остается неизменным. В свою очередь, два других свойства авторского инварианта следует изменить:

2) искомый параметр, входящий в авторский инвариант, должен сохранять «постоянное значение» (слабо колебаться) для стилистически однородных произведений данного автора ***на протяжении определенного периода времени***. Данный временной промежуток не обязательно совпадает со всем периодом творчества писателя.

3) в решении задачи по разграничению разных авторов мы должны использовать различные характеристики авторских инвариантов.

Такое изменение связано с отсутствием универсального критерия, однозначно отличающего одного конкретного писателя от бесчисленного множества других авторов, пишущих на том же языке. Однако при попарном или групповом сравнении грамматических и лексических характеристик, присущих текстам определенных писателей, мы можем выявить существенные различия на разных уровнях языка, которые могут сыграть важную роль в решении вопросов, связанных с определением авторства текста. К локальному авторскому инварианту можно отнести количественные показатели, рассмотренные в большинстве работ, связанных с анализом текстов, например, в работах А.А. Рогова, М.А. Марусенко, А.В. Седова и пр.

В чем состоит преимущество разделения авторского инварианта на глобальный и локальный? Такой подход позволяет подбирать инструменты (как математические, так и собственно филологические) сообразно имеющейся проблеме, минимизируя вероятность появления ошибок и формирования ложных выводов. К примеру, если мы ставим перед собой задачу анализа специфики построения текстов определенного автора в отдельно взятом стихотворном сборнике, произведения для которого были написаны в пределах небольшого временного диапазона, нет никакого смысла обращаться к глобальному авторскому инварианту – достаточно его локальной версии, позволяющей противопоставить анализируемые тексты сходным (и по стилю, и по времени написания) текстам других авторов. В то же время, если филолог задается глобальным, комплексным, многоаспектным вопросом выделения неких индивидуально-авторских особенностей, проявляющихся на протяжении абсолютно всего творчества автора и выделяющих его на фоне не только отдельной эпохи, но и всего литературного наследия, включающего в себя многообразие текстов, написанных на определенном языке – то в этом случае должна быть задействована категория «глобальный авторский инвариант». Разумеется, делать это следует с учетом ранее

упомянутых оговорок и требований, предъявляемых к подобного рода исследованиям, поскольку без них достоверность выполненной работы может быть поставлена под сомнение.

При этом анализ любого из типов авторских инвариантов (как глобального, так и локального) возможен только с учетом комбинированных комплексных подходов, предусматривающих проведение атрибуции по нескольким параметрам сразу. Только многоаспектный разносторонний взгляд на проблему, комбинирующий в себе лингвостатистику, современные математические варианты обработки информации и традиционные методы филологии, позволит добиться необходимых результатов в контексте анализа индивидуально-авторского стиля и решения вопросов, связанных с атрибуцией текстов.

Материалы статьи были представлены на IV Международной конференции по инженерной и прикладной лингвистике «Пиотровские Чтения – 2022», посвященной 100-летию со дня рождения профессора Р.Г. Пиотровского в РГПУ им. А.И. Герцена 22 ноября 2022 г.

СПИСОК ИСТОЧНИКОВ

1. **Морозов Н.А.** Лингвистические спектры: средство для отличения плагиатов от истинных произведений того или иного известного автора. Стилеметрический этюд // Известия отделения русского языка и словестности Императорской Академии наук. 1915. Т. 20, № 4. С. 93–134.

2. **Фоменко В.П., Фоменко Т.Г.** Авторский инвариант русских литературных текстов. Предисловие А.Т. Фоменко // Фоменко А.Т. Новая хронология Греции: Античность в средневековье. М.: Изд-во МГУ, 1996. Т. 2. С. 768–820. URL: https://chronologia.org/seven2_2/add3.html (дата обращения: 25.12.2022)

3. **Kjetsaa G.** Attributed to Dostoevsky: the problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S., 1986. 82 p.

4. **Рогов А.А., Абрамов Р.В., Бучнева Д.Д., Захарова О.В., Кулаков К.А., Лебедев А.А., Москин А.А., Отливанчик А.В., Савинов Е.Д., Сидоров Ю.В.** Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин». Петрозаводск: Издательство «Острова», 2021. 400 с.

5. **Батура Т.В.** Формальные методы определения авторства текстов // Вестник НГУ. Серия: Информационные технологии. 2012. Т. 10, № 4. С. 81–94.

6. **Трофимова Е.Б.** Инвариант: реальность или фикция? // Языковое бытие человека и этноса. 2004. № 7. С. 167–171.

7. **Боронин А.А.** Инвариант в текстовой проекции (интерпретация персонажных субтекстов в условиях эксперимента) // Вопросы психолингвистики. 2011. № 14. С. 28–35.

8. **Мордвинов А.В.** Системный подход в моделировании текста // Вестник Нижегородского университета им. Н.И. Лобачевского. 2010. № 2-1. С. 185–190.

9. **Карелова О.В.** К вопросу изучения индивидуального стиля автора // Известия РГПУ им. А.И. Герцена. 2006. Т. 3, № 20. С. 24–29.

10. **Чирскова И.М.** Цензура как историко-культурный феномен в России XIX века // Вестник РГГУ. Серия: История. Филология. Культурология. Востоковедение. 2008. №10. С. 115–125.

11. **Милов Л.В., Бородкин Л.И., Иванова Т.В., Неберекутина Е.В., Полянская И.В., Романкова Н.В., Саркисова Г.И.** От Нестора до Фонвизина: Новые методы определения авторства. М.: Прогресс, 1994. 445 с.

12. **Шевелев О.Г., Петраков А.В.** Классификация текстов с помощью деревьев решений и нейронных сетей прямого распространения // Вестник Томского государственного университета. 2006. № 290. С. 300–307.

13. **Великанова Н.П., Орехов Б.В.** Цифровая текстология: атрибуция текста на примере романа М.А. Шолохова «Тихий Дон» // Мир Шолохова. 2019. Т. 1, № 11. С. 70–82.

14. **Масаева О.С.** Метод Дельта Бёрроуза // Научно-исследовательская работа обучающихся и молодых учёных: материалы 74-й Всероссийской (с международным участием) научной конференции обучающихся и молодых учёных. Петрозаводск, 2022. С. 262–266.

15. **Марусенко М.А.** Атрибуция анонимных и псевдонимных текстов как типичная задача распознавания образов // Историография и источниковедение отечественной истории. Санкт-Петербург, 2003. Вып. 3. С. 116–135.

16. **Седов А.В., Рогов А.А.** Анализ неоднородностей в тексте на основе последовательностей частей речи // Материалы VI Международной научно-практической конференции «Информационная среда вуза XXI века». 2012. С. 135–139.

17. **Щеголева Л.В., Лебедев А.А., Москин Н.Д.** Методы анализа данных в задаче разграничения фольклорных и авторских текстов // Вопросы языкознания. 2020. № 2. С. 61–74. DOI: 10.31857/S0373658X0008823-4

REFERENCES

[1] **N.A. Morozov,** Lingvisticheskiye spektry: sredstvo dlya otlicheniya plagiatov ot istinnykh proizvedeniy togo ili inogo izvestnogo avtora. Stilemetricheskiy etyud [Linguistic spectra: a tool for distinguishing plagiarisms from the true works of one or another famous author. Stylometric study], Izvestiya otdeleniya russkogo yazyka i slovestnosti Imperatorskoy Akademii nauk [Proceedings of the Department of the Russian Language and Literature of the Imperial Academy of Sciences]. 20 (4) (1995) 93–134.

[2] **V.P. Fomenko, T.G. Fomenko,** Avtorskiy invariant russkikh literaturnykh tekstov. Predisloviye A.T. Fomenko [Author's invariant of Russian literary texts. Foreword by A.T. Fomenko], Fomenko A.T. Novaya khronologiya Gretsii: Antichnost v srednevekovye [New chronology of Greece: Antiquity in the Middle Ages]. M.: Izd-vo MGU [Publishing House of Moscow State University], 2 (1996) 768–820. URL: https://chronologia.org/seven2_2/add3.html (data obrashcheniya: 25.12.2022)

[3] **G. Kjetsaa,** Attributed to Dostoevsky: the problem of attributing to Dostoevsky anonymous articles in Time and Epoch. Oslo: Solum Forlag A. S., 1986. 82 p.

[4] **A.A. Rogov, R.V. Abramov, D.D. Buchneva, O.V. Zakharova, K.A. Kulakov, A.A. Lebedev, A.A. Moskin, A.V. Otlivanchik, Ye.D. Savinov, Yu.V. Sidorov,** Problema atributsii v zhurnalakh «Vremya», «Epokha» i yezhenedelnike «Grazhdanin» [The problem of attribution in the magazines "Time", "Epoch" and the weekly "Citizen"]. Petrozavodsk: Izdatelstvo «Ostrova» [Islands Publishing House], 2021. 400 p.

[5] **T.V. Batura,** Formalnyye metody opredeleniya avtorstva tekstov [Formal methods for determining the authorship of texts], Vestnik NGU. Seriya: Informatsionnyye tekhnologii [Vestnik NSU. Series: Information Technologies]. 10 (4) (2012) 81–94.

[6] **Ye.B. Trofimova,** Invariant: realnost ili fiktsiya? [Invariant: reality or fiction?], Yazykovoye bytiye cheloveka i etnosa [The linguistic existence of a person and an ethnic group]. 7 (2004) 167–171.

[7] **A.A. Boronin,** Invariant v tekstovoy proyektsii (interpretatsiya personazhnykh subtekstov v usloviyakh eksperimenta) [Invariant in text projection (interpretation of character subtexts in experimental conditions)], Voprosy psikholingvistiki [Questions of psycholinguistics]. 14 (2011) 28–35.

[8] **A.V. Mordvinov,** Sistemnyy podkhod v modelirovanii teksta [System approach in text modeling], Vestnik Nizhegorodskogo universiteta im. N.I. Lobachevskogo [Vestnik of Lobachevsky University of Nizhni Novgorod]. 2-1 (2010) 185–190.

[9] **O.V. Karelova,** K voprosu izucheniya individualnogo stilya avtora [On the issue of studying the author's individual style], Izvestiya RGPU im. A.I. Gertsena [Izvestia: Herzen University journal of Humanities & Sciences]. 3 (20) (2006) 24–29.

[10] **I.M. Chirskova,** Tsenzura kak istoriko-kulturnyy fenomen v Rossii XIX veka [Censorship as a historical and cultural phenomenon in Russia in the 19th century], Vestnik RGGU. Seriya: Istoriya. Filologiya. Kulturologiya. Vostokovedeniye [RGGU Bulletin. Series: History. Philology. Cultural studies. Oriental studies]. 10 (2008) 115–125.

[11] **L.V. Milov, L.I. Borodkin, T.V. Ivanova, Ye.V. Neberekutina, I.V. Polyanskaya, N.V. Romankova, G.I. Sarkisova,** Ot Nestora do Fonvizina: Novyye metody opredeleniya avtorstva [From Nestor to Fonvizin: New methods authorship definitions]. M.: Progress, 1994. 445 p.

[12] **O.G. Shevelev, A.V. Petrakov,** Klassifikatsiya tekstov s pomoshchyu derevyev resheniy i neyronnykh setey pryamogo rasprostraneniya [Text classification with decision trees and feed-forward neural networks], Vestnik Tomskogo gosudarstvennogo universiteta [Tomsk State University Journal]. 290 (2006) 300–307.

[13] **N.P. Velikanova, B.V. Orekhov,** Tsifrovaya tekstologiya: atributsiya teksta na primere romana M.A. Sholokhova «Tikhiy Don» [Digital textology: text attribution on the example of the novel by M. A. Sholokhov "Quiet Flows the Don"], Mir Sholokhova [World of Sholokhov]. 1 (11) (2019) 70–82.

[14] **O.S. Masayeva,** Metod Delta Berrouza [The Burrow's Delta method], Nauchno-issledovatelskaya rabota obuchayushchikhsya i molodykh uchenykh: materialy 74-y Vserossiyskoy (s mezhdunarodnym uchastiyem) nauchnoy konferentsii obuchayushchikhsya i molodykh uchenykh [Research work of students and young scientists: materials of the 74th All-Russian (with international participation) scientific conference of students and young scientists]. Petrozavodsk, 2022. Pp. 262–266.

[15] **M.A. Marusenko,** Atributsiya anonimnykh i psevdonimnykh tekstov kak tipichnaya zadacha raspoznavaniya obrazov [Attribution of anonymous and pseudonymous texts as a typical task of pattern recognition], Istoriografiya i istochnikovedeniye otechestvennoy istorii [Historiography and source study of national history]. St. Petersburg, 3 (2003) 116–135.

[16] **A.V. Sedov, A.A. Rogov,** Analiz neodnorodnostey v tekste na osnove posledovatelnostey chastey rechi [Analysis of heterogeneities in the text based on sequences of parts of speech], Proceedings of the VI International Scientific and Practical Conference "Information Environment of the XXI Century University". 2012. Pp. 135–139.

[17] **L.V. Shchegoleva, A.A. Lebedev, N.D. Moskin,** Metody analiza dannykh v zadache razgranicheniya folklornykh i avtorskikh tekstov [Methods of data analysis in the task of distinguishing folklore and author's texts], Voprosy yazykoznaniya [Questions of Linguistics]. 2 (2020) 61–74. DOI: 10.31857/S0373658X0008823-4

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Рогов Александр Александрович
Alexander A. Rogov
E-mail: rogov@petrsu.ru
ORCID: https://orcid.org/0000-0002-8815-7920

Москин Николай Дмитриевич
Nikolai D. Moskin
E-mail: moskin@petrsu.ru
ORCID: https://orcid.org/0000-0001-5556-5349

Лебедев Александр Александрович
Alexander A. Lebedev
E-mail: perevodchik88@yandex.ru
ORCID: https://orcid.org/0000-0001-9939-9389

Поступила: 30.12.2022; Одобрена: 29.01.2023; Принята: 17.03.2023.
Submitted: 30.12.2022; Approved: 29.01.2023; Accepted: 17.03.2023.

Terra Linguistica. 2023, Том 14, № 1. С. 98–107.
Terra Linguistica. 2023, Vol. 14, No. 1. Pp. 98–107.

Материалы конференции «Пиотровские Чтения – 2022»

Conference materials "R. Piotrowski's readings – 2022"

Материалы конференции
УДК 8'33
DOI: https://doi.org/10.18721/JHSS.14109

ИНЖЕНЕРНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА СЕГОДНЯ: ХРОНИКА IV МЕЖДУНАРОДНОЙ КОНФЕРЕНЦИИ «ПИОТРОВСКИЕ ЧТЕНИЯ – 2022»

О.Н. Камшилова ✉ , **Л.Н. Беляева** , **К.Р. Пиотровская**

Российский государственный педагогический университет им. А.И. Герцена, Санкт-Петербург, Российская Федерация

✉ onkamshilova@gmail.com

Аннотация. Предлагается обзор работы конференции «Пиотровские Чтения – 2022», проходившей в РГПУ им. А.И. Герцена (Санкт-Петербург, Россия) 22 ноября 2022 г. Конференция проводилась в честь 100-летней годовщины со дня рождения Раймонда Генриховича Пиотровского (1922–2009) – российского ученого, профессора, доктора филологических наук, Заслуженного деятеля науки, основателя школы инженерной (компьютерной) лингвистики, одного из создателей первых систем машинного перевода в России, основоположника инженерно-лингвистической стратегии в научно-исследовательской и практической методической работе, методологии исследований в области реализации экспериментально-доказательной парадигмы в гуманитарных исследованиях. Дается краткий очерк научного наследия Р.Г. Пиотровского. Описываются различные методологические подходы, исследовательские и образовательные практики в области прикладной лингвистики, представленные участниками конференции.

Ключевые слова: Р.Г. Пиотровский, инженерная лингвистика, прикладная лингвистика, корпусная лингвистика, машинный перевод, обучающие компьютерные системы.

Для цитирования: Камшилова О.Н., Беляева Л.Н., Пиотровская К.Р. Инженерная и прикладная лингвистика сегодня: хроника IV Международной конференции «Пиотровские Чтения – 2022» // Terra Linguistica. 2023. Т. 14. № 1. С. 98–107. DOI: 10.18721/JHSS.14109

Conference materials
DOI: https://doi.org/10.18721/JHSS.14109

LANGUAGE ENGINEERING AND APPLIED LINGUISTICS TODAY: THE CHRONICLE OF THE IV INTERNATIONAL CONFERENCE "R. PIOTROWSKI'S READINGS – 2022"

O.N. Kamshilova ✉, **L.N. Beliaeva**, **X.R. Piotrowska**

Herzen State Pedagogical University of Russia,
St. Petersburg, Russian Federation

✉ onkamshilova@gmail.com

Abstract. This chronicle provides an overview of the IV International Conference on Language Engineering and Applied Linguistics "R. Piotrowski's Readings – 2022" held on November 22, 2022, in Herzen State University (St. Petersburg, Russia). The conference was organized to mark the 100th anniversary of Rajmund G. Piotrowski's birth (1922–2009), a Russian scientist, professor, Honored scientist of Russia. R.G. Piotrowski was the founder of Language Engineering School, pioneer of MT in Russia, initiator of engineering-linguistic strategy in research and practical methodological work, and evidence-based paradigm in methodology of humanitarian research. The article presents a brief outline of R.G. Piotrowski's scientific legacy. It focuses on various methodological approaches and research practices in the field of engineering and applied linguistics contributed by the conference participants.

Keywords: R.G. Piotrowski, engineering linguistics, applied linguistics, corpus linguistics, machine translation, language training computer systems.

Citation: O.N. Kamshilova, L.N. Beliaeva, X.R. Piotrowska, Language engineering and applied linguistics today: The chronicle of the IV International conference "R. Piotrowski's Readings – 2022", Terra Linguistica, 14 (1) (2023) 98–107. DOI: 10.18721/JHSS.14109

Значимым событием в сфере инженерной и прикладной лингвистики явилась IV международная научная конференция «Пиотровские Чтения – 2022», посвященная 100-летней годовщине со дня рождения Раймонда Генриховича Пиотровского (1922–2009) – российского ученого, профессора, доктора филологических наук, Заслуженного деятеля науки, основателя школы инженерной (компьютерной) лингвистики, одного из создателей первых систем машинного перевода в России, основоположника инженерно-лингвистической стратегии в научно-исследовательской и практической методической работе, методологии исследований в области реализации экспериментально-доказательной парадигмы в гуманитарных исследованиях. Конференция проходила 22 ноября 2022 г. в Российском государственном педагогическом университете им. А.И. Герцена.

В эпоху научного прорыва во многих областях знаний, прорыва, опирающегося на точность и адекватность применяемых методов и достоверность получаемых результатов, на независимость применяемых методов от узких и зачастую спекулятивных целей исследования, на применение математических методов и моделей, на создание новых технологий, в частности, технологий компьютерного анализа и автоматической переработки различной информации, особенно важны люди, для которых эти новые подходы являются сутью их исследований.

Раймонд Генрихович Пиотровский в середине прошлого века был среди тех, для кого этот прорыв был не только красивыми словами, но делом, которому он посвятил жизнь, науке, которой оставался верен до последнего дня, продолжая работать и находить для себя все новые направления исследований в рамках как теоретического языкознания, так и той отрасли науки, которую принято называть прикладной лингвистикой. Под этим термином принято понимать лингвистические исследования, опирающиеся на объективные методы анализа, на оценку достоверности и репрезентативности получаемых результатов. Проводимые под руководством Р.Г. Пиотровского

исследования, основанные на применении методов вероятностного прогнозирования и лингвостатистики к огромному эмпирическому материалу [1–2], позволили показать, что в отличие от большинства искусственных систем переработки, хранения и передачи информации язык представляет собой открытую динамическую неравновесную метасистему. Он считал (задолго до того как термин *синергетика* стал модным словом), что метасистема постоянно балансирует между сохранностью и изменением, «порядком» и «хаосом», дискретностью и непрерывностью, чёткостью составляющих элементов и множеств, с одной стороны, и их нечёткостью, с другой. Однако, несмотря на присутствие этих противостоящих свойств, язык и индивидуальная речемыслительная деятельность человека сохраняют внутреннюю устойчивость благодаря заложенным в них синергетическим механизмам самоорганизации и саморегулирования. С его точки зрения обнаружение, изучение, а затем и моделирование этих механизмов является одной из важнейших и пока ещё не решенных задач языкознания на рубеже XX и XXI веков [3].

Основным направлением исследований, которые проводились под руководством Р.Г. Пиотровского, было создание воспроизводящих инженерно-лингвистических моделей речемыслительной деятельности человека. Такое моделирование предполагало использование результатов лингво-статистических и инженерно-лингвистических исследований различных предметных областей. Лингвистическое моделирование и его принципы являются базой исследований в области прикладной и инженерной лингвистики. Именно установление принципов создания модели, подтверждение ее онтологической и эпистемологической состоятельности являются важными компонентами любого исследования. Понятие воспроизводящей инженерно-лингвистической модели (ВИЛМ) было введено в работах Р.Г. Пиотровского еще в середине 80-х годов прошлого века. ВИЛМ – это искусственно созданная формальная система, построение и поведение которой, с одной стороны, имитирует структуру и поведение реальных лингвистических объектов, а с другой – позволяет воспроизводить эти объекты с помощью компьютера и, следовательно, оценивать адекватность модели [4].

Следует отметить, что применение информационных технологий и, как следствие, создание и использование различных информационных систем является объективной реальностью и осознанной необходимостью, поскольку именно на основе их использования сегодня осуществляется научное и культурное взаимодействие, являющееся условием развития общества в целом и каждого отдельного человека в частности.

На основе проведенных исследований были сформулированы основные подходы и требования к системам автоматической переработки текстов – лингвистическим автоматам (ЛА) [5–9]. С точки зрения архитектуры ЛА представляет собой сбалансированный комплекс аппаратных, программных, лингвистических, а иногда и лингводидактических средств, взаимодействующих с мощной базой лингвистических данных и знаний – лингвистической информационной базой (ЛИБ). Реализация различных вариантов лингвистических автоматов в разной комплектации и с разными функциями позволила в рамках создания инновационной образовательной среды перейти к решению задач создания автоматизированных рабочих мест (АРМ). Такое АРМ представляет собой комплекс баз данных и знаний, а также средств обучения и контроля. Разработка среды предполагает выработку навыков и умений работы с текстами в одно- и многоязычной среде в аспекте как лингвистического, так и литературоведческого анализа. Для решения этой задачи АРМ должен быть реализован как комплекс лингвистических, лингвометодических и программных средств [10–12].

Вероятностно-статистические исследования языка и речи были первым этапом в создании тех семиотических моделей процесса коммуникации, которые сегодня составляют теоретическую основу исследований, научным руководителем и вдохновителем которых был и останется Раймонд Генрихович. В теоретическом плане сегодня этот подход входит в новую парадигму доказательной лингвистики, сущность которой состоит в поиске скрытых от прямого наблюдения

системно-семиотических синергетических механизмов, обеспечивающих самоорганизацию и саморазвитие систем языка и речи.

В последние годы Р.Г. Пиотровский активно участвовал в инновационной образовательной программе РГПУ им. А.И. Герцена «Создание инновационной системы подготовки специалистов в области гуманитарных технологий в социальной сфере». Для Р.Г. Пиотровского было особенно важно, что подготовка «гуманитарных технологов» предусматривает преодоление разрыва, возникшего в прошлом между предметно-центричностью в подготовке педагога и потребностью гуманизировать и гармонизировать отношения людей друг к другу, а также к общественным обязанностям и труду, к природе и национальным ресурсам, к культурно историческим ценностям и традициям. Он считал, что для успешного развития общества требуется формирование человека, владеющего современными информационно-коммуникативными технологиями. Это становится возможным при условии достижения нового качества образования, целью которого является воспитание педагога, способного видеть человека, и в первую очередь своего учащегося, как уникальную целостность и развивать его на основе комплексных законов динамики общественной жизни, науки, техники и искусства.

Благодаря трудам Р.Г. Пиотровского были выработаны и реализованы на практике не только технологии организации автоматических словарей и практических систем машинного перевода [13–15], но и методология исследований в области реализации экспериментально-доказательной парадигмы в гуманитарных исследованиях и инженерно-лингвистической и синергетической стратегии в научных исследованиях в области различных направлений филологии и в практической методической работе [16–17].

В память о нашем выдающемся коллеге и учителе, основателе отечественной школы инженерной лингвистики, одном из создателей первых систем машинного перевода в России Раймонде Генриховиче Пиотровском неоднократно проводились научные мероприятия, подтверждающие и развивающие идеи, заложенные Р.Г. Пиотровским и его научной школой: в 2010 г.: Санкт-Петербургским государственным университетом был выпущен сборник научных статей памяти Р.Г. Пиотровского [18], в 2010 г. Минским государственным лингвистическим университетом была проведена международная конференция и выпущен сборник материалов в двух томах [19], 35-летию выхода в свет книги Р.Г. Пиотровского «Инженерная лингвистика и теория языка» (1979) была посвящена VII международная научная конференция «Прикладная лингвистика в науке и образовании» в РГПУ им. А.И. Герцена [20], в 2009 г. и в 2013 г. в Герценовском университете прошли вечера памяти, а с 2017 г. регулярно проводятся Пиотровские Чтения – международная конференция по инженерной и прикладной лингвистике [21–22]. Учредители и организаторы Пиотровских Чтений – научная школа «Прикладные исследования языка и речи, школа Р.Г. Пиотровского» (входит в реестр ведущих научных и научно-педагогических школ Санкт-Петербурга), Центр теоретических и прикладных компьютерных исследований в филологии, филологический факультет РГПУ им. А.И. Герцена совместно с кафедрой математической лингвистики филологического факультета СПбГУ и Санкт-Петербургским Федеральным исследовательским центром Российской академии наук.

Круг тем, входивших в программу Чтениий, касался развития, применения и оценки современных систем машинного перевода, создания и ведения автоматических словарей и терминологических баз, статистических измерений в лингвистике, корпусных методов исследования текста и речи, разработки компьютерных лингводидактических ресурсов и, в целом, современных подходов в области доказательной лингвистики.

Научные и образовательные проблемы инженерной лингвистики и прикладной лингвистики обсуждались как с точки зрения математиков, лингвистов и профессиональных переводчиков, так и с позиций инженеров и разработчиков технической и математической инфраструктуры, что позволило конструктивно обсудить теоретические вопросы и инновации в практической деятельности и новые образовательные практики.

В работе конференции приняло участие более 70 исследователей-лингвистов, филологов, математиков, IT-специалистов и преподавателей из университетов Гродно и Минска (Республика Беларусь), Парижа (Франция), Батуми (Грузия). География Российской Федерации была представлена учеными из университетов Москвы, Санкт-Петербурга, Воронежа, Смоленска, Петрозаводска, Тольятти, а также компаний «PROMT» и «Глобус».

С приветственным словом на открытии юбилейной конференции выступили Председатель Программного комитета, Л.Н. Беляева и Председатель Международного Программного комитета Чтений, директор ФИЦ РАН А.Л. Ронжин. Деятельность Р.Г. Пиотровского в середине XX века привела к созданию национальных центров инженерной лингвистики в республике Беларусь, Молдове, Узбекистане, Казахстане, Киргизии, Грузии, Азербайджане, Армении, Дагестане. К участникам конференции обратились ученики и последователи Раймонда Генриховича – И.В. Совпель, доктор технических наук, заведующий лабораторией интеллектуальных информационных систем Белорусского государственного университета (Минск, республика Беларусь), Р. Д. Кожамбердина, кандидат филологических наук, доцент Педагогического университета (Шимкент, Казахстан), С. В. Соколова, кандидат технических наук, генеральный директор ООО «ПРОМТ».

Программу Пленарного заседания составили доклады, отражающие развитие основных научных направлений, которыми занимался профессор Пиотровский. Об успехах в развитии отечественных систем машинного перевода, о начальных этапах и пионерских решениях в лаборатории машинного перевода и группе «Статистика речи», возглавляемых Р.Г. Пиотровским, рассказала генеральный директор ООО «ПРОМТ» С. В. Соколова (Санкт-Петербург). Вероятностно-статистические исследования языка продолжаются сегодня в трудах Смоленской школы стилеметрии (доклад С.Н. Андреева «Распределение единиц описания в цикле Осипа Мандельштама “Армения”», СмолГУ, Смоленск). Лингвистическое моделирование, семиотика школы Пиотровского получили продолжение в исследовании Е.А. Шингаревой-Славин (Париж) «Семиотика инженерно-лингвистической школы Р.Г. Пиотровского как триггер открытия 3D-измерения в модели знака Ф. де Соссюра: от эллипса к эллипсоиду», исследования в сфере психиатрической лингвистики нашли отражение в докладе В.Э. Пашковского (СПбГУ, Санкт-Петербург) «Речь при детском аутизме. Семиотические аспекты». Современные технологии автоматической обработки текста и речи были продемонстрированы научными коллективами ФИЦ РАН (Санкт-Петербург) в докладах «Метод автоматического тегирования документов для научно-просветительского ресурса “Пушкин Цифровой”» (А.Л. Тесля, А.Л. Ронжин, Г.Н. Беляк, В.В. Головин, С.Г. Николова) и «Влияние машинного перевода на распознавание эмоций и сентимента в русскоязычных текстах» (А.А. Двойникова, И.А. Кагиров, А.А. Карпов).

Обсуждение тем и направлений научных исследований, представленных на Пленарном заседании, продолжилось во время работы секций. В рамках конференции были проведены два секционных заседания, посвященные проблемам применения инженерно-лингвистических технологий в исследованиях текста и речи и разработке и использованию инженерно-лингвистических ресурсов для исследовательских и лингводидактических задач.

На секционном заседании «Инженерно-лингвистические технологии в исследованиях текста и речи» участники Чтений продолжили обсуждение современных технологий автоматической обработки текста и речи и процедур сентимент-анализа. Коллективами ФИЦ РАН (Санкт-Петербург) были представлены доклады «Выявление репрезентативных акустических признаков в задаче автоматического распознавания вовлеченности собеседников» (А.А. Двойникова, А.А. Карпов) и «Влияние предобработки текстовых данных на распознавание эмоций» (А.А. Двойникова, К.О. Кондратенко), этой же теме был посвящен доклад «Запись и апробация набора речевых данных для распознавания негативных эмоций в речи» (А.А. Поволоцкая, В.В. Евдокимова, П.А. Скрелин). Особое внимание на этой секции было уделено развитию современных методов анализа при идентификации авторства, эта тема нашла отражение в докладах «Дистрибутив-

но-семантические модели в автороведении: необходимость или дань моде?» (Т.А. Литвинова, Воронеж), «О смене парадигмы авторского инварианта» (Н.Д. Москин, А.А. Лебедев, А.А. Рогов, Петрозаводск), «Отношение правдоподобия: к разрешению избыточного многообразия при решении задач авторской идентификации» (М.А. Марусенко, С. Петербург). Активно развивается направление статистических исследований текста в работах Смоленской школы стилеметрии [23], которая была представлена еще одним докладом «Динамика текста и динамика стиля» (В.С. Андреев). Идеи Р.Г. Пиотровского о необходимости создания и ведения лингвистических информационных баз (ЛИБ) не только продуктивны в системах машинного перевода и ведении автоматических словарей, но созвучны активно развивающимся сегодня исследованиям, опирающимся на лингвистические базы данных и корпусы текстов. Вопросами соотношения качественного и количественного анализа для решения задач анализа текста традиционно занимаются ученые из Санкт-Петербургского государственного университета (см, например [24]). На секции коллегами из СПбГУ были представлены доклады «Европейская католическая гимнография в сопоставительном аспекте: качественно-количественный анализ текстового материала» (М.В. Корышев) и «Диалогические тексты в мультимодальных тематических моделях» (О.А. Митрофанова). Один из аспектов языкового моделирования рассматривался в докладе «Метод лексико-семантических полей в анализе политического дискурса о международном вооруженном конфликте в Сирии (на материале публикаций в российской и французской прессе) (Е.Д. Власова)

На втором секционном заседании «Разработка и использование инженерно-лингвистических ресурсов для исследовательских и лингводидактических задач» обсуждались проблемы исследования и результаты применения технологий автоматического анализа текстов. В докладе Л.Н. Беляевой и О.Н. Камшиловой (С. Петербург) «Машинный перевод в научном и учебном пространстве: за и против» рассматривались новые условия, возможности и проблемы использования результатов машинного перевода. Развитию методов современной прикладной лексикографии посвящены доклады «Тезаурусная лексикография в прикладной лингвистике и компьютерной лингводидактике» (Ю.И. Горбунов, О.Ю. Горбунова, Тольятти), «Языковые ресурсы для задач лингвистического мониторинга» (Л.В. Рычкова, Гродно, республика Беларусь). Результаты корпусных исследований текстов рассматривались в докладах «Значимая лексика на примере публицистического дискурса немецкой научно-культурной интеллигенции» (М.В. Корышев, М.В. Хохлова, С. Петербург), «Частотный словарь художественной прозы в контексте социополитики (на материале «Корпуса русского рассказа 1900-1930 гг.») (О.А. Гребенников, Т. Г. Скребцова, М.В. Хохлова, С. Петербург). Детальный анализ обучающих систем, продолжающих направление разработки и применения инженерно-лингвистических технологий для поддержки обучения был представлен в докладах «Методы компьютерной лингводидактики при изучении финского языка» (И.Н. Ларченков, Л.А. Ларченкова, С. Петербург), «CALL-технологии в средней школе: разработка и внедрение» (У.В. Матвеева, В.Р. Нымм, К.Р. Пиотровская, С. Петербург).

Тем самым в докладах и обсуждениях были затронуты все аспекты исследований, входивших в сферу научных интересов Р.Г. Пиотровского.

В рамках Чтений состоялось открытие выставки «Р.Г. Пиотровский: Каким мы его помним» в фундаментальной библиотеке им. Императрицы Марии Федоровны. Это совместный проект, подготовленный отделом редкой книги библиотеки (С.Е. Колоскова) и кафедрой методики обучения математике и информатике (К.Р. Пиотровская). Выставка была посвящена научному творчеству, общественной деятельности Р.Г. Пиотровского, а также истории развития созданных им направлений: инженерная лингвистика, лингвистическая синергетика и психиатрическая лингвистика[1].

Цель Пиотровских Чтений – 2022 заключалась не только в том, чтобы почтить память большого ученого, но и в том, чтобы обсудить современное состояние и достижения тех отраслей

[1] https://lib.herzen.spb.ru/news/show/1072

прикладной лингвистики, начало которым во многом было положено трудами Р.Г. Пиотровского и его учеников. Различные методологические подходы и исследовательские практики в своей совокупности позволили всесторонне рассмотреть феномен развития прикладной лингвистики, реализуемой на базе достижений математики, лингвистики, филологии и новых технических средств поддержки исследовательской и образовательной деятельности. Проведение форумов такого рода поддерживает преемственность в отечественной науке, просвещение и воспитание молодого поколения ученых.

СПИСОК ИСТОЧНИКОВ

1. **Пиотровский Р.Г.** Информационные измерения языка. Л.: Наука, Ленинградское отделение, 1968. 116 с.

2. **Piotrovsky R.G.** Mathematische Linguistik / R.G. Piotrovsky, K.B. Bektaev, A.A. Piotrowskaja. Vol. 27. Bochum: N. Brockmeyer, 1985. 514 p.

3. **Пиотровский Р.Г.** Лингвистическая синергетика: исходные положения, первые результаты, перспективы. СПб.: Филологический факультет Санкт-Петербургского государственного университета, 2006. 159 с.

4. **Пиотровский Р.Г.** Лингвистический автомат и его речемыслительное обоснование. Минск: МГЛУ, 1999. 196 с.

5. **Пиотровский Р.Г.** Лингвистический автомат и Машинный фонд русского языка // Вопросы языкознания. 1987. № 4. С. 69–73.

6. **Беляева Л.Н., Пиотровский Р.Г.** Введение. На пути к лингвистическому автомату // Статистика речи и формализованный анализ текстовых единиц / Л.Н. Беляева (ответственный редактор). Л.: ЛГПИ, 1990. С. 3–8.

7. **Беляева Л.Н., Пиотровский Р.Г.** Как строить лингвистический автомат // Русский язык: прошлое, настоящее, будущее. Материалы Всероссийской научной конференции. Часть II, D. Компьютеризация лингвистических исследований. Сыктывкар: Альманах «Говор»,1996. С. 45–66.

8. **Беляева Л.Н.** Лингвистические автоматы в современных информационных технологиях. СПб.: Изд-во РГПУ им. А.И. Герцена, 2001. 130 с.

9. **Пиотровская К.Р.** Обучающий лингвистический автомат. Санкт-Петербург: Интерлайн, 2002. 38 с.

10. **Пиотровская К.Р.** Модели, программные и информационные средства учебного АРМ переводчика : специальность 05.25.05 «Информационные системы и процессы»: автореф. дисс. канд. техн. наук / Пиотровская Ксения Раймондовна. Киев, 1993. 19 с.

11. **Беляева Л.Н., Зайцева Н.Ю., Пиотровский Р.Г., Романов Ю.В.** Работа лингвистического автомата с языками разной типологии // Структурная и прикладная лингвистика. СПб: филол.ф-т СПбГУ, 2004. С. 260–277.

12. **Беляева Л.Н., Джепа Т.Л.** Автоматизированное рабочее место переводчика: лингвистические ресурсы и технологии // Структурная и прикладная лингвистика. 2012. № 9. С. 109–128.

13. **Пиотровский Р.Г.** Лингвистические уроки машинного перевода // Вопросы языкознания. 1985. № 4. С. 18–27.

14. **Beliaeva L., Piotrovskij R.** The Structural Approach to Machine-Aided Translation and Modelling of Unitary Linguistic Data Base // Sixth Internat. Conference on Computer and the Humanities. June 6-8, 1983.Raleigh: North Carolina State Univ., 1983. Pp. 48–49.

15. **Beliaeva L, Piotrovskij R., van Nunen P.** Man-Machine Inter-Action in Linguistic Automation // 2nd Internation. Maaststricht-Lodz Duo Colloquium on Translation and Meaning. Abstracts. Maastricht, 1995. P. 25.

16. **Пиотровский Р.Г.** О лингвистической синергетике // Научно-техническая информация. Серия 2: Информационные процессы и системы. 1996. № 12. С. 1–12.

17. **Пиотровский Р.Г.** Статистические модели текста и опыт их лингво-синергетического анализа // Научно-техническая информация. Серия 2: Информационные процессы и системы. 2007. № 8. С. 1–12.

18. Памяти профессора Р.Г. Пиотровского: межвузовский сборник. Древняя и Новая Романия. Выпуск 9. СПб.: Санкт-Петербургский государственный университет, 2010. 292 с.

19. Актуальные проблемы теоретической и прикладной лингвистики. Материалы международной конференции, посвященной памяти профессора Р.Г. Пиотровского. В 2 частях. Минск: МГЛУ, 2010.

20. Прикладная лингвистика в науке и образовании. Материалы VII международной конференции. СПб., «Книжный дом», 2014. 232 с.

21. **Ronzhin A., Bogdanov S., Laptev V., Belyaeva L., Piotrovskaya X., Kamshilova O.** Preface to R. Piotrowski's readings in language engineering and applied linguistics (R. Piotrowski's readings LE & AL'2017). Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Saint Petersburg. Russia. November 27, 2017. CEUR Workshop Proceedings 2233. Pp. 1–9. URL: http://ceur-ws.org/Vol-2233/

22. **Ronzhin A., Bogdanov S., Laptev V., Beliaeva L., Piotrowska X., Kamshilova O.** Preface to R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019). Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019). Russia. February 13, 2020. CEUR Workshop Proceedings 2552. Pp. 1–6. URL: http://ceur-ws.org/Vol-2552/

23. **Андреев В.С.** Экспоненциальное распределение частей речи в стихотворном тексте: опыт стилеметрического анализа // Общество. Коммуникация. Образование. 2021. Т. 12. № 4. С. 94–104. DOI: 10.18721/JHSS.12407

24. **Митрофанова О.А., Гаврилик Д.А.** Эксперименты по автоматическому выделению ключевых выражений в стилистически разнородных корпусах русскоязычных текстов // Terra Linguistica. 2022. Т. 13. № 4. С. 22–40. DOI: 10.18721/JHSS.13402

REFERENCES

[1] **R.G. Piotrovskiy,** Informatsionnyye izmereniya yazyka [Information measurements of language]. L. Nauka, Leningradskoye otdeleniye, 1968.

[2] **R.G. Piotrovsky,** Mathematische Linguistik / R.G. Piotrovsky, K.B. Bektaev, A.A. Piotrowskaja. Vol. 27. Bochum: N. Brockmeyer, 1985.

[3] **R.G. Piotrovskiy,** Lingvisticheskaya sinergetika : iskhodnyye polozheniya, pervyye rezultaty, perspektivy [Linguistic synergetics: starting points, first results, prospects]. SPb. SPbSU, 2006.

[4] **R.G. Piotrovskiy,** Lingvisticheskiy avtomat i yego rechemyslitelnoye obosnovaniye [The language automaton and its speech-cogitative justification]. Minsk. MGLU, 1999.

[5] **R.G. Piotrovskiy,** Lingvisticheskiy avtomat i Mashinnyy fond russkogo yazyka [Linguistic automaton and Machine fund of the Russian language], Voprosy yazykoznaniya. 4 (1987) 69–73.

[6] **L.N. Belyayeva, R.G. Piotrovskiy,** Vvedeniye. Na puti k lingvisticheskomu avtomatu [Introduction. Towards a linguistic automaton], Statistika rechi i formalizovannyy analiz tekstovykh yedinits [Speech statistics and formalized analysis of text units] / L. N. Belyayeva ed. L. LGPI, 1990. Pp. 3–8.

[7] **L.N. Belyayeva, R.G. Piotrovskiy,** Kak stroit lingvisticheskiy avtomat [How to build a linguistic automaton], Russkiy yazyk: proshloye, nastoyashcheye, budushcheye. Materialy Vserossiyskoy nauchnoy konferentsii.[Russian language: past, present, future. Proc. of the conference] Part II, D. Kompyuterizatsiya lingvisticheskikh issledovaniy [Computerization of linguistic research]. Syktyvkar: Almanakh "Govor", 1996. Pp. 45–66.

[8] **L.N. Belyayeva,** Lingvisticheskiye avtomaty v sovremennykh informatsionnykh tekhnologiyakh [Linguistic automata in modern information technologies]. SPb.: Izd-vo RGPU im. A.I.Gertsena, 2001.

[9] **K.R. Piotrovskaya,** Obuchayushchiy lingvisticheskiy Avtomat [Educational linguistic automation]. SPb. Interlayn, 2002.

[10] **K.R. Piotrovskaya,** Modeli, programmnyye i informatsionnyye sredstva uchebnogo ARM perevodchika: spetsialnost 05.25.05 "Informatsionnyye sistemy i protsessy": avtoref. diss. kand. tekhn. nauk [Models, software and information tools of the educational workstation of a translator: specialty 05.25.05 "Information systems and processes": diss. abstract of cand. techn. sci.]. Kiyev, 1993.

[11] **L.N. Belyayeva, N.Yu. Zaytseva, R.G. Piotrovskiy, Yu.V. Romanov,** Rabota lingvisticheskogo avtomata s yazykami raznoy tipologii [The work of a linguistic automaton with languages of different typologies], Strukturnaya i prikladnaya lingvistika [Structural and Applied Linguistics]. SPb. SPbSU, 2004. Pp. 260–277.

[12] **L.N. Belyayeva, T.L. Dzhepa,** Avtomatizirovannoye rabocheye mesto perevodchika: lingvisticheskiye resursy i tekhnologii [Automated Translator Workplace: Linguistic Resources and Technologies], Strukturnaya i prikladnaya lingvistika [Structural and Applied Linguistics]. 9 (2012) 109–128.

[13] **R.G. Piotrovskiy,** Lingvisticheskiye uroki mashinnogo perevoda [Linguistic lessons of machine translation], Voprosy yazykoznaniya. 4 (1985) 18–27.

[14] L. Beliaeva, R. Piotrovskij, The Structural Approach to Machine-Aided Translation and Modelling of Unitary Linguistic Data Base. Proc. of Sixth Int. Conference on Computer and the Humanities. Raleigh: North Carolina State Univ., 1983, pp. 48–49.

[15] **L. Beliaeva, R. Piotrovskij, P. van Nunen,** Man-Machine Inter-Action in Linguistic Automation/ Proc/ of 2nd Internation. Maaststricht-Lodz Duo Colloquium on Translation and Meaning. Maastricht, 1995, p. 25.

[16] **R.G. Piotrovskiy,** O lingvisticheskoy sinergetike [On Linguistic Synergetics], Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnyye protsessy i sistemy [Scientific and technical information. Series 2: Information Processes and Systems]. 12 (1996) 1–12.

[17] **R.G. Piotrovskiy,** Statisticheskiye modeli teksta i opyt ikh lingvo-sinergeticheskogo analiza [Statistical text models and the experience of their linguo-synergetic analysis], Nauchno-tekhnicheskaya informatsiya. Seriya 2: Informatsionnyye protsessy i sistemy [Scientific and technical information. Series 2: Information Processes and Systems]. 8 (2007) 1–12.

[18] Pamyati professora R.G. Piotrovskogo. Drevnyaya i Novaya Romaniya. Vypusk 9 [In memory of Professor R. G. Piotrovsky. Ancient and New Romania. V. 9]. SPb.: SPbSU, 2010.

[19] Aktualnyye problemy teoreticheskoy i prikladnoy lingvistiki [Actual problems of theoretical and applied linguistics], Materialy mezhdunarodnoy konferentsii, posvyashchennoy pamyati professora R.G. Piotrovskogo. V 2 chastyakh [Proceedings of the international conference dedicated to the memory of Professor R.G. Piotrovsky. In 2 parts]. Minsk: MGLU, 2010.

[20] Prikladnaya lingvistika v nauke i obrazovanii [Applied linguistics in science and education], Materialy VII mezhdunarodnoy konferentsii [Proc. of the VII International Conference]. SPb., "Knizhnyy dom", 2014.

[21] **A. Ronzhin, S. Bogdanov, V. Laptev, L. Belyaeva, X. Piotrovskaya, O. Kamshilova,** Preface to R. Piotrowski's readings in language engineering and applied linguistics (R. Piotrowski's readings LE & AL'2017). Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics. Saint Petersburg. Russia. November 27, 2017. CEUR Workshop Proceedings 2233. RR. 1–9. Available at: http://ceur-ws.org/Vol-2233/

[22] **A. Ronzhin, S. Bogdanov, V. Laptev, L. Beliaeva, X. Piotrowska, O. Kamshilova,** Preface to R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019). Proceedings of the R. Piotrowski's Readings in Language Engineering and Applied Linguistics (PRLEAL-2019). Russia. February 13, 2020. CEUR Workshop Proceedings 2552. Pp. 1–6. Available at: http://ceur-ws.org/Vol-2552/

[23] **V.S. Andreev,** Exponential distribution of parts of speech in verse text: experience in stylometric analysis, Society. Communication. Education, 12 (4) (2021) 94–104. DOI: 10.18721/JHSS.12407

[24] **O.A. Mitrofanova, D.A. Gavrilic,** Experiments on automatic keyphrase extraction in stylistically heterogeneous corpus of Russian texts, Terra Linguistica, 13 (4) (2022) 22–40. DOI: 10.18721/JHSS.13402

СВЕДЕНИЯ ОБ АВТОРАХ / INFORMATION ABOUT AUTHORS

Камшилова Ольга Николаевна
Olga N. Kamshilova
E-mail: onkamshilova@gmail.com
ORCID: https://orcid.org/0000-0002-1488-2206

Беляева Лариса Николаевна
Larisa N. Beliaeva
E-mail: lauranbel@gmail.com
ORCID: https://orcid.org/0000-0002-8622-4595

Пиотровская Ксения Раймондовна
Xenia R. Piotrowska
E-mail: krp62@mail.ru
ORCID: https://orcid.org/0000-0003-2557-9461

Поступила: 14.02.2023; Одобрена: 10.03.2023; Принята: 17.03.2023.
Submitted: 14.02.2023; Approved: 10.03.2023; Accepted: 17.03.2023.