

Министерство науки и высшего образования  
Российской Федерации

УНИВЕРСИТЕТ ИТМО

Некоммерческое партнерство ПРИОР Северо-Запад

## **КОМПЬЮТЕРНАЯ ЛИНГВИСТИКА И ВЫЧИСЛИТЕЛЬНЫЕ ОНТОЛОГИИ**

**Выпуск 9**

**Труды XXVIII Международной  
объединённой научной конференции  
«Интернет и современное общество»,  
IMS-2025, Санкт-Петербург,  
23–25 июня 2025 г.**

**Сборник научных трудов**

# **ИТМО**

Санкт-Петербург

2025

УДК 81'33  
ББК 81.1  
К63

Рецензенты:

*д-р филол. наук, проф. А. В. Колмогорова, канд. филол. наук Т. Ю. Шерстинова*

Редколлегия:

*Л. Н. Беляева, О. А. Митрофанова, А. В. Чижик (председатель редколлегии)*

Ответственный редактор издания:

*канд. культурологии А. В. Чижик*

К63

**Компьютерная лингвистика и вычислительные онтологии.** Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г.) Сборник научных трудов. — СПб.: Университет ИТМО, 2025. — 133 с.

ISSN 3033-5582

В сборник включены тексты научных статей, представленные на XXVIII Международной объединенной научной конференции «Интернет и современное общество» (Internet and Modern Society – IMS). Работы прошли рецензирование и отобраны в результате конкурсной процедуры. Сборник снабжен авторским указателем.

Издание адресовано научным работникам, преподавателям, аспирантам и магистрантам, изучающих междисциплинарные проблемы влияния информационно-коммуникационных технологий на трансформацию социальных и политических отношений в современном обществе.

Информация о конференции «Интернет и современное общество» представлена на сайте объединенной конференции (<http://ims.itmo.ru>).

Все статьи и тезисы докладов конференции IMS публикуются в открытом доступе (лицензия Creative Commons — CC-BY 3.0 Unported). Сборники научных статей, издаваемые в рамках конференции IMS с 2011 года, размещаются в Научной электронной библиотеке (<http://elibrary.ru/>) и Российском индексе научного цитирования (РИНЦ).

Подготовка конференции осуществлялась при поддержке Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, Комитета информатизации и связи и Комитета по науке и высшей школе Санкт-Петербурга.

УДК 81'33  
ББК 81.1

**ИТМО**

**ИТМО (Санкт-Петербург)** — национальный исследовательский университет, научно-образовательная корпорация. Альма-матер победителей международных соревнований по программированию, один из ведущих вузов России по подготовке кадров для цифровой экономики. Приоритетные направления: IT и искусственный интеллект, фотоника, робототехника, квантовые коммуникации, трансляционная медицина, Life Sciences, Art&Science, Science Communication.

Лидер федеральных программ «Приоритет-2030» и «Передовые инженерные школы». С 2022 года ИТМО работает в рамках новой модели развития — научно-образовательной корпорации. В её основе академическая свобода, поддержка начинаний студентов и сотрудников, распределенная система управления, приверженность открытому коду, бизнес-подходы к организации работы. Образование в университете основано на выборе индивидуальной траектории для каждого студента.

По версии SuperJob, ИТМО занимает первое место в Санкт-Петербурге и второе в России по уровню зарплат выпускников в сфере IT. Университет в топе международных рейтингов среди российских вузов. Входит в топ-5 российских университетов по качеству приема на бюджетные места. Рекордсмен по поступлению олимпиадников в Санкт-Петербурге. С 2019 года ИТМО самостоятельно присуждает ученые степени кандидата и доктора наук.

© Университет ИТМО, 2025  
© Авторы, 2025

## XXVIII Международная объединённая научная конференция «Интернет и современное общество» (IMS 2025)

Санкт-Петербург, 23–25 июня 2025 г.

<http://ims.itmo.ru>

Конференция «Интернет и современное общество» (Internet and Modern Society – IMS) проводится в Санкт-Петербурге ежегодно с 1998 года. С 2014 года конференция проводится в международном формате.

Объединённая конференция «Интернет и современное общество» в 2025 году была проведена при поддержке Министерства цифрового развития, связи и массовых коммуникаций Российской Федерации, Комитета по науке и высшей школе и Комитета по информатизации и связи Санкт-Петербурга. Отдельные специализированные мероприятия проводились в сотрудничестве с проектами, реализуемыми при поддержке Российского научного фонда и Санкт-Петербургского научного фонда.

Конференция названа объединённой, так как научная программа конференции консолидирует серию специализированных международных и российских научных конференций, симпозиумов, семинаров, круглых столов и других мероприятий, посвящённых специальным вопросам развития технологий информационного общества. Отдельные специализированные и проблемно-ориентированные мероприятия проводятся в сотрудничестве с партнёрскими организациями.

Основу научной программы конференции 2025 года составили международные компоненты, включающие сессии на русском и английском языках:

- **VIII Международная конференция по электронному управлению** (Digital Transformation in Governance and Society – DTGS-2025);
- **международный семинар «Компьютерная лингвистика»** (Computational Linguistics – CompLing-2025);
- **международный семинар «Искусство и инновации в музеях»** (International Art and Innovation in Museums Seminar – AIMS 2025);
- **международный семинар «Киберпсихология и цифровая педагогика»** (Cyberpsychology and Post-AI Education – PsyAI-2025);
- **научно-практический симпозиум «Этико-правовые аспекты цифровой трансформации».**

Традиционно в программу конференции были включены сессии научных докладов:

- **Информационные системы для науки и образования;**
- **Культурология киберпространства;**
- **Цифровая урбанистика;**
- **Электронное обучение и дистанционные образовательные технологии.**

Программу объединённой конференции расширили специализированные мероприятия, ориентированные не только на исследователей, но и на экспертное сообщество и молодых ученых:

- **международный симпозиум «Interactive Systems & Information Society Technologies» (InterSys-2025)**, организованный четырьмя университетами: Университетом ИТМО (Санкт-Петербург, Россия), Новосибирским государственным техническим университетом (Новосибирск, Россия), Институтом технологий и науки Бирла (Birla Institute of Technology & Science; кампус в Дубае, ОАЭ), Федеральным университетом Параны (Federal University of Paraná; Куритиба, Бразилия);

- **научно-практический семинар «Цифровые городские сервисы: потенциал и барьеры развития»** (при поддержке проекта РНФ и СПбНФ № 23-18-20079 «Исследование социальной результативности электронного взаимодействия граждан и власти в Санкт-Петербурге на примере городских цифровых сервисов», в сотрудничестве с Комитетом по информатизации и связи Санкт-Петербурга, СПб ГКУ «Центр информационного сопровождения» и Комитетом цифрового развития Ленинградской области);
- **специализированная сессия «IT-Новации в цифровизации госсектора»** с участием проектов-призёров V Национального конкурса «ПРОФ-IT. Инновация» (при поддержке Экспертного центра электронного государства и Оргкомитета Всероссийского форума «ПРОФ-IT»);
- **научный симпозиум «Цифровое государство в глобальной перспективе»** в сотрудничестве с журналом «Россия в глобальном мире», входящим в перечень ВАК;
- **экспертная дискуссия «Многолетние рассуждения об этике в сфере ИИ: где практика и результат?»**;
- **Young Scholars' Poster Session «Digital Transformation in Governance and Society» (Young DTGS-2025).**

На конференцию IMS-2025 было подано 260 заявок авторами из России, Беларуси, Германии, Индии, Иордании, Испании, Италии, Казахстана, Китая, Объединённых Арабских Эмиратов, Соединённых Штатов Америки, Узбекистана, Южной Африки и других стран. В научную программу конференции вошло 156 докладов.

Отбор докладов на конференцию и текстов для публикации производится по результатам двойного слепого рецензирования членами программного комитета с использованием международной системы сопровождения научных конференций Microsoft's Conference Management Toolkit. В 2025 году в рецензировании научных текстов приняли участие более 100 членов программного комитета и приглашённых рецензентов со всего мира, сформировавших около 420 рецензий.

Общее количество участников конференции составило более 400 человек.

Благодаря информационной и организационной поддержке, которую оказали органы власти Санкт-Петербурга и Ленинградской области, в 2025 году в научно-практических мероприятиях и круглых столах конференции IMS-2025 приняли участие около 50 сотрудников исполнительных органов государственной власти, органов местного самоуправления и подведомственных учреждений.

В 2025 году международный симпозиум «Interactive Systems & Information Society Technologies» прошёл в формате двух сессий. Первая сессия предваряла основные треки конференции IMS и состоялась 5–7 мая в Дубае в Институте технологий и науки Бирла (Birla Institute of Technology & Science). Научная программа первой сессии симпозиума включила в себя 15 докладов, подготовленных авторскими коллективами из России, Индии, Объединённых Арабских Эмиратов и других стран.

По результатам объединённой конференции IMS-2025 издаются два сборника научных трудов (серийные издания) и сборник тезисов на русском языке:

- **Информационное общество: образование, наука, культура и технологии будущего (ISSN 3033-5574)**, вып. 9;
- **Компьютерная лингвистика и вычислительные онтологии (ISSN 3033-5582)**, вып. 9;
- **Интернет и современное общество: сборник тезисов докладов IMS-2025.**

Статьи, представленные для докладов на английском языке и прошедшие рецензирование, включены в сборник, подготовленный совместно с зарубежными партнерами конференции.

Сборник публикуется в издательстве Springer (индексация в базе Scopus). Также в сборник включены научные статьи, отобранные на конкурсной основе за авторством молодых учёных — участников Young DTGS-2025.

Оргкомитет конференции сотрудничает с профильными научными журналами и использует возможность рекомендации лучших докладов, заслушанных и обсужденных на конференции, для публикации в журналах в доработанном виде с представлением более подробной информации о проведенных исследованиях.

- С 2017 года конференция сотрудничает с научным журналом "**International Journal of Open Information Technologies**" (<http://injoit.org>, ВАК, РИНЦ), издаваемым в МГУ им. М. В. Ломоносова, по формированию специального номера. В 2025 г. такой номер планируется к изданию.
- С 2022 года началось партнерство с научным журналом "**Journal on Interactive Systems**" (<https://sol.sbc.org.br/journals/index.php/jis>, Scopus (Q3)), Бразилия. В 2025 г. ряд докладов, представленных на английском языке, рекомендован для публикации в доработанном виде в этом журнале.
- Международный научный электронный журнал «**Культура и технологии**» (<http://cat.ifmo.ru/>) регулярно публикует лучшие статьи авторов IMS по своей тематике.

Авторам конференции IMS-2025 оргкомитет предложил также направить заявки на публикацию статей в следующих научных журналах, входящих в перечень ВАК и соответствующих профилю конференции:

- «**Россия в глобальном мире**» (<https://russiaglobal.spbstu.ru/>);
- «**Экономика. Право. Инновации**» (<https://ecinn.itmo.ru/>, K3);
- "**PolitBook**" (<https://www.politbook.online/>, K2);
- «**Вопросы государственного и муниципального управления**» / "**Public Administration Issues**" (<https://vgmu.hse.ru/>, K1, Scopus (Q3), RSCI, «Белый список»).

Электронные версии сборников конференции размещаются в свободном доступе (лицензия Creative Commons – CC-BY 3.0 Unported) на сайте материалов конференции «Интернет и современное общество» (<http://ojs.itmo.ru>). С 2017 года всем статьям присваивается международный идентификатор DOI, а информация на уровне метаданных размещается в информационной системе CrossRef (<https://search.crossref.org>). Метаданные сборников размещаются в Научной электронной библиотеке (<https://elibrary.ru>), а все статьи и тезисы индексируются в Российском индексе научного цитирования (РИНЦ).

Информация обо всех сборниках и специальных номерах журналов, опубликованных с 2011 года, представлена на сайте конференции со ссылками на первоисточники – <https://ims.itmo.ru/proceedings/>.

## **ПРОГРАММНЫЙ КОМИТЕТ КОНФЕРЕНЦИИ**

### **Председатель Программного комитета:**

Васильев В. Н., докт. техн. наук, чл.-корр. РАН, ректор Университета ИТМО

### **Заместители председателя Программного комитета:**

Борисов Н. В., докт. физ.-мат. наук, заведующий кафедрой информационных систем в искусстве и гуманитарных науках СПбГУ, председатель оргкомитета конференции

Чугунов А. В., канд. полит. наук, директор Центра технологий электронного правительства ИДУ Университета ИТМО, генеральный директор НП ПРИОР Северо-Запад, ученый секретарь конференции

### **Члены Программного комитета:**

Бабина О. И., канд. филол. наук, Южно-Уральский государственный университет

Бакаев М. А., канд. техн. наук, Новосибирский государственный технический университет

Балаян А. А., канд. полит. наук, НИУ «Высшая школа экономики» — Санкт-Петербург

Барандова Т. Л., канд. социол. наук, Северо-Западный институт управления РАНХиГС

Блинова О. В., канд. филол. наук, Санкт-Петербургский государственный университет

Богданова-Бегларян Н. В., д-р филол. наук, Санкт-Петербургский государственный университет

Богомяжкова Е. С., канд. социол. наук, Санкт-Петербургский государственный университет

Болгов Р. В., канд. полит. наук, Санкт-Петербургский государственный университет, Университет ИТМО

Бундин М. В., канд. юрид. наук, Нижегородский государственный университет им. Н. И. Лобачевского

Видясова Л. А., канд. социол. наук, Университет ИТМО

Вяхирева В. В., Нижегородский государственный университет им. Н. И. Лобачевского

Глазкова А. В., канд. техн. наук, Тюменский государственный университет

Голубинская А. В., Нижегородский государственный университет им. Н. И. Лобачевского

Демарева В. А., канд. психол. наук, Нижегородский государственный университет им. Н. И. Лобачевского

Денисов М. В., канд. экон. наук, Северо-Западный институт управления РАНХиГС

Джанелидзе М. Г., канд. экон. наук, Институт проблем региональной экономики РАН

Игнатъев А. В., д-р техн. наук, Волгоградский государственный технический университет

Игнатъева О. А., канд. социол. наук, Санкт-Петербургский государственный университет

Измалкова А. И., канд. психол. наук, НИУ «Высшая школа экономики»

Ильин И. В., д-р экон. наук, Санкт-Петербургский политехнический университет Петра Великого

Кабанов Ю. А., НИУ «Высшая школа экономики»

Камшилова О. Н., канд. филол. наук, Российский государственный педагогический университет им. А. И. Герцена

Кладько С. С., канд. филос. наук, Университет «НЕЙМАРК»

Ковальчук С. В., канд. техн. наук, Университет ИТМО

Коган М. В., канд. техн. наук, Санкт-Петербургский политехнический университет Петра Великого

Колмогорова А. В., д-р филол. наук, НИУ «Высшая школа экономики»

Конюховский П. В., д-р экон. наук, Российский государственный педагогический университет им. А. И. Герцена

Котельников Е. В., д-р техн. наук, Европейский университет в Санкт-Петербурге

Крижановский А. А., канд. техн. наук, Карельский научный центр РАН

Кунникова К. И., канд. психол. наук, Уральский федеральный университет

Куприенко И. В., Университет ИТМО

Ларионов И. Ю., канд. филос. наук, Санкт-Петербургский государственный университет

- Лукашевич Н. В., канд. физ.-мат. наук, Московский государственный университет им. М. В. Ломоносова
- Мамаев И. Д., канд. филол. наук, Санкт-Петербургский государственный университет, Балтийский государственный технический университет «ВОЕНМЕХ» им. Д. Ф. Устинова
- Мамонова И. Г., канд. искусствоведения, Санкт-Петербургский государственный университет
- Митрофанова О. А., канд. филол. наук, Санкт-Петербургский государственный университет
- Момотова Т. А., канд. полит. наук, НИУ «Высшая школа экономики», Санкт-Петербургский политехнический университет Петра Великого
- Носиков А. А., канд. полит. наук, Санкт-Петербургский государственный университет
- Перов В. Ю., канд. филос. наук, Санкт-Петербургский государственный университет
- Прокудин Д. Е., д-р филос. наук, Санкт-Петербургский государственный университет
- Разумникова О. М., д-р биол. наук, Новосибирский государственный технический университет
- Рашевский Н. М., канд. техн. наук, Волгоградский государственный технический университет
- Рюмин Д. А., канд. техн. наук, Санкт-Петербургский Федеральный исследовательский центр РАН, НИУ «Высшая школа экономики»
- Рябушко А. Н., канд. полит. наук, независимый исследователь
- Садовникова Н. П., д-р техн. наук, Волгоградский государственный технический университет
- Сидорова Е. А., канд. физ.-мат. наук, Институт систем информатики имени А. П. Ершова СО РАН
- Слав Ю. Э., Совет муниципальных образований Санкт-Петербурга
- Сморгунов Л. В., д-р филос. наук, Санкт-Петербургский государственный университет
- Соколов А. В., д-р полит. наук, Ярославский государственный университет им. П. Г. Демидова
- Стецко Е. В., канд. филос. наук, Санкт-Петербургский государственный университет
- Стырин Е. М., канд. социол. наук, НИУ «Высшая школа экономики»
- Сытник А. Н., канд. полит. наук, Санкт-Петербургский государственный университет
- Тимофеева М. К., д-р филол. наук, Новосибирский государственный университет, Институт математики им. С. Л. Соболева Сибирского отделения РАН
- Трутнев Д. Р., Университет ИТМО
- Толстикова И. И., канд. филос. наук, Университет ИТМО
- Федосов А. Ю., д-р пед. наук, Российский государственный социальный университет
- Филатова О. Г., д-р полит. наук, Университет ИТМО
- Фомина Н. В., канд. психол. наук, Приволжский исследовательский медицинский университет
- Хилов П. Е., Экспертный центр электронного государства
- Ходачек И. А., PhD, Европейский университет в Санкт-Петербурге
- Чижик А. В., канд. культурологии, Санкт-Петербургский государственный университет, Университет ИТМО
- Чугунов А. В., канд. полит. наук, Университет ИТМО
- Шмелёва И. А., канд. психол. наук, Университет ИТМО
- Шульгинов В. А., канд. филол. наук, НИУ «Высшая школа экономики»
- Hussain Ahmed CHOWDHURY, PhD, Birla Institute of Technology & Science, Pilani, Dubai, UAE
- Katarina COKRLIC, University of Bologna, Italy
- Wei DAI, PhD, Huazhong University of Science & Technology, China
- Ruben ELAMIRYAN, PhD, Russian-Armenian University, Armenia

Ashish GUPTA, PhD, Birla Institute of Technology & Science, Pilani, Dubai, UAE  
 Grigera JULIAN, PhD, LIFIA – Universidad Nacional de La Plata, Argentina  
 Tojo MATHEW, PhD, Birla Institute of Technology and Science, Pilani, Dubai, UAE  
 Radka NACHEVA, PhD, University of Economics, Bulgaria  
 Prabir PANDA, PhD, Motilal Nehru National Institute of Technology, India  
 Pranav PAWAR, PhD, Birla Institute of Technology and Science, Pilani, Dubai, UAE  
 Tamizharasan PERIYASAMY, PhD, Birla Institute of Technology and Science, Pilani, Dubai, UAE  
 Elakkiya R, PhD, Birla Institute of Technology and Science, Pilani, Dubai, UAE  
 Ravikumar S, PhD, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India  
 Thylashri S, PhD, Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, India  
 Suyash SHUKLA, PhD, Bennett University, India  
 Anna SMOLIAROVA, PhD, The Hebrew University of Jerusalem, Israel  
 Subramaniaswamy VAIRAVASUNDARAM, PhD, Vellore Institute of Technology, India  
 Can YANG, PhD, Chongqing University, China

**В рассмотрении заявок на доклад и публикацию также участвовали рецензенты:**

Атугодаге М. М., НИУ «Высшая школа экономики»  
 Блохина В. Д., НИУ «Высшая школа экономики»  
 Быкова А. П., НИУ «Высшая школа экономики»  
 Гуминская А. В., Национальный исследовательский ядерный университет «МИФИ»  
 Гусева Д. Д., Санкт-Петербургский государственный университет  
 Жеребцова Ю. А., Университет ИТМО  
 Кирина М. А., НИУ «Высшая школа экономики»  
 Морозов Д. А., Новосибирский государственный технический университет  
 Москвина А. Д., Санкт-Петербургский государственный университет  
 Низомутдинов Б. А., Университет ИТМО  
 Панфилов Г. О., Университет ИТМО  
 Пашков А. А., Новосибирский государственный технический университет  
 Селетков И. П., NeuroEstimator  
 Студеникина К. А., Московский государственный университет им. М. В. Ломоносова

Shivang AGARWAL, Indian Institute of Technology Jodhpur, India  
 Chiranjeevi BURA, University of Colorado, USA  
 Thiago CAMPOS, Universidade Tecnológica Federal do Paraná, Brazil  
 Rao Deepak DASARATHA, Infosys Technologies Ltd., India  
 Saravanan DEVADOSS, Sri Manakula Vinayagar Group of Institutions, India  
 Aditya GUPTA, Amazon Inc., USA  
 Balaji INGOLE, Gainwell Technologies LLC, USA  
 Deógenes JUNIOR, Federal University of Paraná, Brazil  
 Mohan Krishna MANNAVA, Independent Researcher, India  
 Praveen Kumar MYAKALA, Independent Researcher, India  
 Lada PETRUSHENKO, Hyperskill, Czech Republic  
 Bogdan ROMANOV, University of Tartu, Estonia  
 Isan SAHOO, Oracle, India  
 Ulka SHIROLE, A.C. Patil College of Engineering, India  
 Amit SINGH, Cisco Systems, USA  
 Daniil VOLKOVSKII, Paris 1 Panthéon-Sorbonne University, France  
 Zicheng WANG, Hunan University of Science and Technology, China  
 Vibhas Mohan ZANPURE, Amazon Inc., USA

**ОРГАНИЗАЦИОННЫЙ КОМИТЕТ****Председатель оргкомитета:**

Прокудин Д. Е., д-р филос. наук, доцент Санкт-Петербургского государственного университета, аналитик Центра юзабилити и смешанной реальности Университета ИТМО

**Члены оргкомитета:**

Бакаев М. А., канд. техн. наук, Новосибирский государственный технический университет (заместитель председателя)

Демарева В. А., канд. психол. наук, Нижегородский государственный университет им. Н. И. Лобачевского (заместитель председателя)

Кабанов Ю. А., НИУ «Высшая школа экономики», Университет ИТМО (заместитель председателя)

Метелева А. С., Университет ИТМО (информационный менеджер конференции)

Низомутдинов Б. А., Университет ИТМО, НП ПРИОР Северо-Запад (финансовый директор конференции)

Перов В. Ю., канд. филос. наук, Санкт-Петербургский государственный университет

Чижик А. В., канд. культурологии, Санкт-Петербургский государственный университет, Университет ИТМО (заместитель председателя)

Чугунов А. В., канд. полит. наук, Университет ИТМО, НП ПРИОР Северо-Запад (заместитель председателя)

Elakkiya R, PhD, Birla Institute of Technology and Science, Pilani, Dubai, UAE

## От редколлегии

Компьютерная лингвистика стремительно развивается и предлагает новые решения для фундаментальных вопросов языкознания и прикладных задач в самых разных областях. Статьи, вошедшие в этот сборник, отражают многообразие тем, интересующих в настоящее время лингвистов: лексикология и автоматизированные методы работы с терминологией, изучение и применение больших языковых моделей, развитие методов машинного перевода, современные методы синтаксического анализа, исследование художественных текстов, корпусные исследования с применением тематического моделирования.

В статье Е.П. Лаврентьевой и М.С. Коган «Использование алгоритмов машинного перевода в задаче субтитрования образовательного видеоконтента» рассматривается возможность применения моделей машинного перевода для автоматического создания субтитров к лекциям. Авторы описывают проведенный эксперимент и отмечают потенциал улучшения качества субтитров при использовании специализированных данных.

К.М. Черников и И.А. Суров в работе «Геометрия падежей в векторных моделях русского языка» исследуют, как падежная система русского языка отражается в векторном пространстве. Авторы описывают выявленные структурные особенности, которые могут быть использованы при автоматической обработке текстов.

В статье Е.Д. Шамаевой «Сравнение нейросетевых синтаксических анализаторов для русского языка» описываются результаты анализа качества работы пяти популярных парсеров (UDPipe, Stanza, Natasha, DeepPavlov, spaCy) на материалах проекта Universal Dependencies. Результат исследования — систематизация особенностей работы синтаксических анализаторов в зависимости от корпуса и характеристик данных.

Е.В. Васильева в статье «Проблемы исследования словообразовательного потенциала с использованием современных поисковых систем: автоматизированный отбор дериватов через Яндекс» предлагает методику выявления дериватов с помощью поискового API. Подход позволяет фиксировать новые и некодифицированные единицы, расширяя возможности анализа словообразовательных процессов.

В статье Е.В. Исаевой и Б.З. Сафарбекова «Оптимизация обработки терминологии в беспилотной авиации: новый подход к извлечению терминов с использованием промпт-инжиниринга» представлен программный комплекс DroneTerms AI, использующий большие языковые модели для автоматического извлечения и классификации терминов.

О.А. Митрофанова в работе «Динамика тем научных статей в корпусе текстов по компьютерной и корпусной лингвистике» анализирует изменения исследовательских интересов с помощью алгоритма BERTopic. Исследование показывает, как трансформируется тематический ландшафт в связи с развитием цифровых технологий и распространением больших языковых моделей.

Исследование К.А. Студеникиной, Е.А. Лютиковой и А.А. Герасимовой, лежащее в основе статьи «Оценка лингвистической компетенции больших языковых моделей на материале корпуса согласовательной вариативности», посвящено оценке лингвистической компетенции больших языковых моделей. Авторы анализируют, какие конструкции оказываются наиболее трудными для моделей и насколько выводы моделей совпадают с суждениями носителей языка.

Статья «Разработка системы анализа разноплановых характеристик поэтического текста» представляет исследование А.И. Панковой и Е.В. Ягуновой. Авторы описывают разработанную ими систему автоматизированного анализа стихов. Приложение реализует комплекс методов метрического, лексического, синтаксического и семантического анализа.

Таким образом, сборник объединяет статьи, которые отражают современный взгляд на фундаментальные вопросы компьютерной лингвистики и показывают перспективы практического применения результатов исследований. Мы надеемся, что материалы сборника будут полезны как исследователям, так и практикам, работающим на стыке лингвистики, информатики и прикладных областей.

Редактор сборника  
А. В. Чижик

# Использование алгоритмов машинного перевода в задаче субтитрирования образовательного видеоконтента

Е. П. Лаврентьева, М. С. Коган

Санкт-Петербургский политехнический университет Петра Великого

klp1782@yandex.ru, m\_kogan@inbox.ru

## Аннотация

В статье исследуется возможность интеграции алгоритмов машинного перевода (МП) в задачу субтитрирования образовательного видеоконтента и, в частности, в процесс создания высококачественных английских субтитров для русскоязычных видеолекций по лингвистике YouTube-канала «Постнаука». Метод улучшения качества перевода субтитров основан на дообучении модели машинного перевода на специализированном корпусе видеолекций. Для лучшего понимания проблемы автоматического субтитрирования образовательных видео был изучен ряд теоретических вопросов: природа образовательного видео как жанра аудиовизуального текста, специфика аудиовизуального перевода (АВП) и субтитрирования, в частности, а также эволюция и современное состояние алгоритмов машинного перевода (МП), и применение МП к АВП. В результате проведения экспериментального исследования было установлено, что перевод субтитров, осуществленный моделью, когерентен при изменении их форматирования на формат одна строка — одно предложение. Перевод является адекватным, за исключением терминологии, которую модель с трудом распознает и неправильно переводит во многих случаях. После дообучения модель начала распознавать больше терминов. В статье обсуждаются возможные причины, по которым не удалось правильно перевести русскую лингвистическую терминологию из видеолекций на английский язык. Результаты исследования доказали возможность применения алгоритмов МП для создания субтитров к образовательным видеороликам при условии предварительного редактирования, например, изменения формата текста. Обозначены будущие направления исследований.

**Ключевые слова:** субтитрирование, образовательный видеоконтент, машинный перевод, аудиовизуальный перевод

**Библиографическая ссылка:** Лаврентьева Е. П., Коган М. С. Использование алгоритмов машинного перевода в задаче субтитрирования образовательного видеоконтента // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 12–25. DOI: 10.17586/3033-5582-2025-9-12-25.

## 1. Введение

В нынешних условиях глобализации все менее редким явлением становится просмотр видеоконтента как на родном языке, так и на иностранном. Для понимания видео на иностранном языке многие используют субтитры. Субтитры также часто используются изучающими определенный язык, чтобы иметь возможность воспринимать аутентичный контент и, при необходимости, обращаться к своему родному языку для перевода. Как

и в случае с другими задачами перевода, субтитрирование все чаще рассматривается как задача, пригодная для автоматизации. Автоматизация процесса создания субтитров может значительно сократить время и усилия, необходимые для выполнения задачи. Кроме того, это помогает снизить затраты, связанные с наймом профессиональных переводчиков. Нередко необходимо обработать большое количество видеоконтента за короткий промежуток времени, что часто приходится делать создателям субтитров; в данном случае автоматическая генерация субтитров является большим подспорьем.

Английский язык — основное средство международного общения в различных сферах, что обуславливает актуальность автоматизации перевода субтитров на английский язык для видеоконтента, созданного на других языках. Зрители, которые изучают или не знают язык, на котором создано видео, смогут обратиться к английским субтитрам, чтобы понять содержание.

Вопрос автоматизации субтитрирования также актуален для видеоконтента на русском языке. Предоставление высококачественных субтитров к видеолекциям повысит их доступность для студентов, не владеющих русским языком или изучающих его, в том числе для иностранных студентов, обучающихся в российских вузах.

## 2. Обзор литературы

### 2.1. Достижения в области разработки алгоритмов машинного перевода

Разработка алгоритмов машинного перевода (МП) имеет долгую историю, насчитывающую почти столетие. Однако крупный прорыв в этой области был совершен в предыдущем десятилетии с появлением нейронного машинного перевода (НМП). Данный термин вошел в обиход, когда исследователи [1; 2] представили модели МП, основанные на принципе end-to-end. С тех пор были предложены различные архитектуры НМП, а также методы усовершенствования исходной архитектуры моделей.

Базовый тип модели НМП построен на рекуррентной нейронной сети (англ. Recurrent Neural Network, RNN). Основным недостатком стандартной модели RNN является ограниченная способность запоминать долгосрочные зависимости в данных, поскольку информация быстро распадается во время передачи по сети. Из-за этого качество перевода длинных предложений резко падает. Чтобы решить эту проблему, исследователи [2] ввели три модификации, которые стали широко использоваться в моделях НМП: механизм внимания, двунаправленное кодирование и управляемый рекуррентный блок (англ. Gated Recurrent Unit, GRU).

Помимо RNN, были предложены и другие архитектуры моделей. Часто используемой архитектурой является Transformer, предложенная в 2017 г. [3]. Она полагается исключительно на механизмы внимания без рекуррентности и сверток. На основе архитектуры Transformer было разработано множество моделей, наиболее известные примерами которых являются BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformer) и T5 (Text-to-Text Transfer Transformer). К моделям Transformer, применимым для задачи перевода текста, можно отнести T5, BART, FairSeq Machine Translation (FSMT) и многоязычные модели, такие как mT5, mBART, M2M100 и SeamlessM4T (также более позднюю версию SeamlessM4Tv2).

Популярным фреймворком для обучения и использования моделей МП (включая модели, основанные на архитектуре Transformer) является платформа MarianMT, разработанная исследователями из Эдинбургского университета. Платформа упрощает процесс обучения моделей Transformer, предоставляя предварительно реализованные компоненты для обработки данных. Доступно более 1000 моделей, поддерживающих широкий спектр языковых пар. Существуют также модели, поддерживающие несколько исходных и целевых языков.

Область, в которой проводится много исследований, — многоязычный перевод. Данная область привлекает внимание исследователей по разным причинам: одна из них — различие между структурами языков (например, перевод с изолирующего языка с плохой морфологией на агглютинативный язык с богатой морфологией). Другой причиной является тот факт, что большинству языков мира не хватает параллельных данных, из-за чего их называют «языками с ограниченными ресурсами» (англ. resource-poor languages, low-resource languages). Исследователи предлагают различные способы преодоления вышеупомянутых проблем. Существует два типа методов многоязычного перевода: первый основан на применении дополнительных данных, а второй отталкивается от улучшения качества модели [4].

Проблема перевода языков с ограниченными ресурсами — не хватает обучающих данных. Легче собрать большое количество одноязычных данных на языке с ограниченными ресурсами, нежели собрать параллельный корпус языковых пар, содержащих этот язык. В НМП обучающие данные обычно дополняются за счет использования одноязычного корпуса. Распространенным методом увеличения количества данных является обратный перевод [5; 6], при котором стандартная модель НМП обучается на небольшом параллельном корпусе и затем используется для перевода незначительного объема одноязычного текста (предложений на целевом языке) на другой язык, создавая таким образом искусственный двуязычный (параллельный) корпус. Этот корпус может быть использован для повторного обучения модели.

Новым и многообещающим методом, используемым в МП, является применение больших языковых моделей (англ. Large Language Models, LLM) для решения задачи перевода текстов. LLM стали большим прорывом в области обработки естественного языка за последние несколько лет [7; 8]. Они устанавливают новые эталонные показатели производительности для различных задач, включая МП [9; 10; 11; 12]. LLM показывают впечатляющие результаты перевода для основных языков благодаря их обширному предобучению на больших объемах непарных (одноязычных) данных. Основываясь на экспериментальных исследованиях, направленных на тестирование LLM в задаче МП, авторы [13] приходят к выводу, что модели позволили меньше зависеть от двуязычных данных при переводе на языки, хорошо представленные в данных. Также результат работы моделей свидетельствует о возможности перевода длинных предложений и целых документов. Однако необходимо проделать дополнительную работу, чтобы модели могли лучше адаптироваться к выполнению комплексных задач и предсказанию редких слов, а также к переводу текстов на языки с ограниченными ресурсами.

## 2.2. Современные сферы применения МП и возможности развития

Алгоритмы МП широко используются во многих областях благодаря низкой стоимости, эффективности и высококачественным результатам. Наиболее распространенной задачей остается перевод текстов, среди которых могут быть выделены [14; 4]: перевод веб-страниц, научной литературы, а также переводы в сферах бизнеса, обучения языку и межязыкового поиска информации. Хотя перевод текста остается наиболее распространенным применением МП, другие задачи, которые все чаще реализуются в реальных приложениях, — перевод изображений и устной речи, аудиовизуальный перевод (АВП) — перевод аудиовизуального контента.

Перевод изображений служит связующим звеном между компьютерным зрением и МП, поскольку он принимает изображения в качестве входных данных и переводит их на целевые языки. Перевод изображений включает в себя подзадачи, двумя важными из которых являются многоязычные подписи к изображениям [15; 16; 17] и оптическое распознавание символов [18]. Другим многообещающим направлением является синхронный перевод (англ. Simultaneous Translation, ST), при котором речь переводится в режиме реального времени. Перевод речи в текст (англ. Speech-to-Text, S2T) сочетает в себе функцию автоматического распознавания речи (англ. Automatic Speech Recognition,

ASR) и текста перевода на одном экране для удобства пользователя. Перевод по технологии S2S (Speech-to-Speech, преобразование речи в речь) позволяет пользователям прослушивать аудио на мобильных устройствах на родном языке или на любом предпочитаемом языке.

Традиционный АВП, выполняемый людьми, включает в себя все формы мультимодальной коммуникации, включая внутриязыковое (субтитры на том же языке) и межъязыковое субтитрирование, а также озвучивание, дубляж и аудиодескрипцию (устное изложение визуальных элементов в различных средствах массовой информации, цель которого — сделать визуальный контент доступным для слепых или слабовидящих людей) [19]. Технологии МП все чаще применяются для решения задач АВП, автоматизируя процесс субтитрирования и дубляжа. Возможность автоматизации процесса создания субтитров изучается с 2000-х гг. с разной степенью успеха [20; 21; 22; 23; 24].

Значительный прогресс в качестве, достигнутый НМП, вновь пробудил интерес к МП в области субтитрирования. Исследователи AppTek [25] разработали общую систему НМП специально для создания субтитров, интегрировав модуль для прогнозирования разрывов строк наряду с переводом. Проект MeMAD [26] принял аналогичную стратегию, применяя МП на уровне предложения или документа для повышения эффективности перевода за счёт расширения контекста. Однако согласование переведенного текста с исходными временными метками стало проблемой из-за несоответствий в расширении текста и сегментации, в результате чего некоторые субтитры рассинхронизировались с саундтреком. Авторы [27] рассматривают перевод субтитров как перевод разметки, в рамках которого границы субтитров и свойства источника переносятся непосредственно на выходные данные с использованием алгоритма проекции, основанного на выравнивании слов. Примечательно, что все эти методы зависят от наличия высококачественных подписей на языке оригинала.

Хотя субтитрирование является наиболее изученной задачей АВП, было проведено несколько исследований, в которых изучалась возможность автоматизировать дублирование. Исследователи [28] изучили применение нескольких онлайн-инструментов МП с бесплатным доступом для перевода на занятиях по дубляжу. Авторы обнаружили, что все ещё существует значительная трудность в использовании МП в дубляже. Однако, данную проблему можно преодолеть, включив постредактирование в этап подготовки сценария дубляжа.

Форматом аудиовизуального контента, где востребован АВП, является образовательное видео. Согласно авторам [29], использование видеороликов широко распространено в различных дисциплинах и в учебных контекстах, но, если эти видео не будут сделаны доступными для всех учащихся (включая лиц с нарушениями слуха и зрения, лиц с другим родным языком), они создадут новые проблемы [30; 29]. Другая проблема доступности образовательных видеороликов [31] заключается в том, что во многих странах образовательные технологии либо малодоступны, либо функционируют плохо, особенно в сельской местности. Исследователи [31] предложили пайплайн перевода видео с английского на урду, который состоит из четырёх компонентов: (1) транскрипция английского аудио в текст; (2) грамматика в транскрипции проверяется и совершенствуется; (3) английский текст переводится на урду; (4) для исходного видео создаётся закадровый голос на урду. Форматом образования, который становится все более популярным в последнее десятилетие, особенно во время и после пандемии COVID-19, являются MOOK, или массовые открытые онлайн-курсы, которые позволяют учащимся слушать курсы независимо от их географического положения. Лекции обычно читаются преподавателями из разных университетов, материал доступен с помощью видеозаписей. Лекции обычно читаются на одном языке и сопровождаются расшифровкой. Текстовая расшифровка создаётся волонтерами, либо автоматизированной системой распознавания речи. Чтобы поделиться информацией с более широкой аудиторией, говорящей на разных языках, расшифровки впоследствии переводятся на другие языки. Тем не менее, ввиду недоступности параллельных корпусов, к которым может быть открыт доступ широкой

публики, остро стоит задача разработки высококачественной системы машинного перевода лекций. Авторы [32] исследуют фреймворк для извлечения информации из параллельных корпусов для создания системы машинного перевода лекций высокого качества.

### 3. Данные и методология

С учётом обсуждаемых выше теоретических подходов и проблем автоматизации процесса создания субтитров, было принято решение провести экспериментальную оценку того, насколько применима модель МП для перевода субтитров к видеолекциям. Также предстояло установить, может ли дообучение модели на корпусе, состоящем из субтитров к видеолекциям, способствовать улучшению качества переведённых субтитров.

#### 3.1. Этап сбора корпуса

Для проведения эксперимента были выбраны пять видеолекций YouTube-канала «Постнаука» по дисциплине «Лингвистика», а именно: «Архитектура грамматики» (Екатерина Лютикова), «Корпусная лингвистика» (Владимир Плунгян), «Закон Гримма в германских языках» (Александр Пиперски), «Падежи в языках мира» (Петр Аркадьев) и «Языковые универсалии» (Яков Тестелец). Выбор тематики видеолекций обусловлен компетенциями авторов в области лингвистики. Это позволило обеспечить качественную экспертную оценку перевода терминов. Понимание терминологии необходимо в процессе оценки качества перевода, так как нередко бывают случаи, когда модель более-менее связно переводит текст, но узкоспецифичные термины переводит неправильно. Жанр видеолекции был выбран потому, что является актуальным жанром для студентов, в том числе иностранных, обучающихся в российских (в частности, петербургских) вузах.

Выбранные лекции имеют языковые особенности, характерные для образовательных видеороликов, в которых используется широкий спектр лингвистических терминов, обилие слов и структур, характерных для разговорной речи (большое количество слов-заменителей, запяток и фальстартов), а также большое количество длинных предложений с очень сложной синтаксической структурой. В некоторых из предложений порядок слов немного отличается от письменного языка. Все упомянутые выше лексико-стилистические характеристики были учтены на этапе создания параллельного тестового корпуса, который состоит из субтитров на языке оригинала (русском) и эталонного перевода на английском.

Для составления тестового корпуса, который использовался для тестирования модели МП до и после дообучения, необходимо было сначала загрузить оригинальные русские субтитры с YouTube в формате .txt, используя конвертер субтитров *Downsub*. Далее нужно было отредактировать субтитры в соответствии с правилами русской грамматики и *Russian Timed Text Style Guide* — руководством по созданию субтитров к видео на русском языке, предоставленным компанией Netflix. Ввиду того, что к данным лекциям не были предусмотрены английские субтитры, оригинальные субтитры были переведены с помощью МП, а затем отредактированы снова уже на английском языке. Движком, выбранным для работы, был Яндекс.Переводчик, поскольку он способен быстро переводить целые документы. Русские субтитры к пяти лекциям были собраны в единый документ .txt. Аналогичная процедура была проведена для английской версии субтитров. Для эффективной визуализации тестового корпуса русские и английские субтитры были выровнены с помощью скрипта Python. Корпус был преобразован в формат .csv и сохранен как *subtitles\_test.csv*. Всего корпус содержит 1 758 строк и 22 158 токенов (11 697 английских токенов и 10 461 русский) (см. табл. 1).

Таблица 1. Первые 15 строк тестового корпуса

ru	en
Говорить о падежах языков мира,	Talking about cases in world languages,
на самом деле, очень сложно,	is, in fact, very difficult
потому что для того,	because in order to
чтобы о них говорить, нужно сначала понять,	talk about them, you must first understand
а что о чем, собственно, мы говорим –	what exactly are we talking about –
что такое эти самые падежи	what are these cases
и каким образом мы можем, например,	and how can we, for example,
называть одним и тем же термином,	call them by the same term
использовать понятие падеж	and use the concept of case
по отношению к явлениям,	in relation to phenomena
которые в разных языках могут,	which can, in different languages,
на первый взгляд,	at first glance,
а иногда и не только на первый взгляд,	and sometimes not only at first glance,
быть весьма непохожи друг на друга.	Be very different from each other.

### 3.2. Этап экспериментального тестирования модели

Для проведения эксперимента была выбрана модель Helsinki-NLP/opus-mt-ru-en. Данная модель имеется в открытом доступе (в репозитории на github) и не требует большой мощности компьютера при работе. Помимо этого, модель была использована авторами ранее в других экспериментах, поэтому процесс её использования для данной задачи не вызвал затруднений. Помимо этого, как говорят разработчики, она свободна от коммерческих интересов и ограничений [33]. Модель, как и все модели OPUS-MT, обучается на открытых параллельных корпусах, доступных в репозитории больших параллельных текстов OPUS. Существует пять версий модели ru-en, выпущенных с 05.12.2019 по 26.02.2020. Было решено работать с последней версией, поскольку она демонстрирует самые высокие показатели метрик BLEU и chrF для используемых тестовых корпусов ([newstest2012.ru.en](#), [newstest2013.ru.en](#), [newstest2014-ruen.ru.en](#), [newstest2015-enru.ru.en](#), [newstest2016-enru.ru.en](#), [newstest2017-enru.ru.en](#), [newstest2018-enru.ru.en](#), [newstest2019-ruen.ru.en](#) и [Tatoeba.ru.en](#)).

При выборе подходящего обучающего корпуса необходимо было учесть следующие факторы: корпус соответствует тестовой выборке с точки зрения жанра и, по возможности, предметной области (лингвистика). К сожалению, общедоступный русско-английский корпус субтитров найти было невозможно, особенно при условии, что это должны быть субтитры к образовательному видеоконтенту. Стоит отметить, что направление языковой пары в данном случае не имеет решающего значения, поскольку модель обучается на данных обоих языков. В качестве обучающей выборки выбран англо-русский корпус TED2020. Корпус содержит около 4000 стенограмм TED и TED-x, сделанных в июле 2020 г. Эти стенограммы были переведены глобальным сообществом волонтеров более чем на 100 языков. В корпусе содержится около 390 015 предложений, 5 410 846 русских токенов и 6 586 927 английских. Формат стенограмм отличается от формата субтитров — вместо разделения строк в соответствии с максимальным количеством символов в строке, стенограммы разделяются по предложениям, что означает, что одна строка стенограммы равна одному предложению. В случае возникновения проблем с разным форматированием тестового и обучающего корпусов, было решено создать ещё одну копию тестового корпуса, в которой субтитры были разделены по предложениям, а по не максимальному количеству символов в строке. Хотя набор данных содержит доклады, посвящённые различным научным областям, таким как медицина, экология, экономика, нейробиология, математика, ботаника, в нем также имеются лингвистические термины. При выборе обучающего корпуса учитывались стилистические особенности говорящего и формат

выступления: в случае TED2020 это выступление в режиме реального времени перед зрителями в зале, что может объяснить чуть меньшую формальность дискурса выступающих, чем в условиях записи лекции в студии, как происходит в лекциях на канале «Постнаука». Тем не менее, в обоих корпусах речь выступающих устная, поэтому в них присутствуют элементы устной речи, перечисленные выше. Помимо этого, целевая аудитория как TED Talks, так и подписчиков канала «Постнаука» широка: она включает себя как ученых, так и обывателей, интересующихся определённой областью. Целевая аудитория вместе с целью лекций также определяет лексико-стилистическое оформление выступлений, которые содержат терминологию, свойственную данной области знаний, но не используют её в таком количестве, как в сугубо научных материалах. Обучающий корпус также был преобразован в формат .csv, как и тестовый.

Последним этапом подготовки набора данных был импорт файла .csv в таблицу Excel, чтобы сделать кириллицу (русские субтитры) читаемыми, поскольку в исходном файле .csv они не были расшифрованы. Результирующий файл обучающего корпуса называется subtitles\_train.xlsx (см. табл. 2).

Таблица 2. Фрагмент лекции по лингвистике в наборе данных TED2020

en	ru
But <b>comparative linguistics</b> can help us by focusing on <b>grammatical structure, patterns of sound changes, and certain core vocabulary.</b>	Но <b>сравнительная лингвистика</b> помогает нам, фокусируясь на <b>грамматической структуре, моделях фонетических изменений</b> и определённой <b>базовой лексике.</b>

Для тестирования модели перед дообучением было решено работать с библиотекой trnslate2 для эффективного вывода с моделями Transformer, разработанными OpenNMT, для оптимизации и ускорения процесса перевода. Оценка качества проводилась как с помощью метрик, так и вручную (самими авторами). Были выбраны метрики BLEU и chrF, поскольку именно эти метрики были использованы при первоначальной оценке качества перевода модели, когда она была разработана. Заявленные разработчиками баллы метрик, которые получила модель при её первоначальной оценке, сравнивались с полученными баллами после перевода моделью субтитров. Помимо данных метрик, было решено использовать COMET, поскольку она обеспечивает анализ перевода, напоминающий оценку человеком. Баллы BLEU для исходных тестовых выборок были в диапазоне 27,9 — 34,8, тогда как балл тестовой выборки «Постнаука» составил 25,6. Для chrF он составил 54,5 — 60,3 для исходных датасетов и 46,6 для «Постнауки». Как видно, оценки BLEU и chrF, полученные для датасета «Постнаука», несколько ниже показателей, полученных исходными датасетами. Это означает, что перевод довольно приличный, но нуждается в некоторой доработке. Оценка COMET составила 0,0782. Перевод МП также оценивался авторами данного исследования. Были сделаны выводы: модель плохо справляется с задачей перевода терминов, поскольку она не была обучена на терминологии, относящейся к данной предметной области. Ещё одна проблема, которую авторы заметили при анализе перевода — неспособность модели выявить кореференцию, которая существует между разными строками. Она переводит текст построчно, не рассматривая его как единое целое, что приводит к бессвязному текстовому фрагменту.

Следующим шагом было дообучение модели на обучающем корпусе. Было решено обучать модель на различном количестве эпох: на 4 и на 12, чтобы увидеть, при каком количестве эпох будут достигнуты оптимальные результаты. В примере с четырьмя эпохами качество перевода все ещё растёт (баллы BLEU не начали падать, а функции training loss и validation loss непрерывно падают). В случае с двенадцатью эпохами произошло небольшое переобучение модели (функция validation loss начала расти после четвёртой эпохи). Это означает, что лучше использовать при проверке на тестовой выборке модель, обученную на четырёх эпохах.



Было также решено проверить, улучшится ли качество перевода, если отформатировать тестовый корпус так, как отформатированы субтитры TED2020: один сегмент равен одному предложению. Метрики снизились как для модели до дообучения, так и для модели после дообучения, но они стали намного ближе друг к другу. Это означает, что тексты (эталон и МП перед дообучением по сравнению с эталоном и МП после дообучения) стали намного более похожи друг на друга. Переводы были оценены авторами. Было выявлено, что тексты стали значительно более связными, чем когда длина строки определялась максимальным количеством символов в строке (от 37 до 42), предписываемых **требованиями** субтитрования. Это тактика предварительного редактирования, используемая при работе с системами МП. В табл. 5 показана разница в степени связности предложения (генерируемого моделью, обученной на четырёх эпохах) до и после изменения формата субтитров.

**Таблица 5.** Иллюстрация улучшения согласованности текста

До изменения форматирования	После изменения форматирования
It was 1957. It's got a central syntax. It's an American scientist. Noam Homsky. And the last 50 years are a period. It's the syntax. It's the most active. It's one of the most linguistic areas.	In 1957, this syntax model was proposed by American scientist Noam Homsky, and the last 50 years have been the period where syntax is most active in all areas of linguistics.

## 5. Выводы и дальнейшее исследование

В данном исследовании был изучен вопрос, возможно ли применение модели машинного перевода для задачи перевода субтитров видеолекций с русского на английский язык, а также улучшит ли качество перевода дообучение модели на жанрово-специфическом параллельном корпусе. В результате работы был составлен русско-английский параллельный корпус субтитров пяти видеолекций по лингвистике YouTube-канала «Постнаука», который содержит 1 758 строк и 22 158 токенов.

Выбрана Helsinki-NLP/opus-mt/ru-en для определения качества перевода, полученного с помощью этой модели на тестовом корпусе «Постнаука» без дообучения. Было определено, что, хотя перевод получился до известной степени связным, неспособность распознавать специфические термины и форматирование тестового набора привели к тому, что оценки BLEU и chrF оказались ниже, чем у модели на её первоначальном тестовом наборе. Это побудило авторов провести дообучение модели на данных, специфичных для жанра видеолекции.

В качестве обучающего корпуса был использован большой англо-русский корпус TED2020, в котором, помимо лекций из других областей, содержится некоторое количество лекций по лингвистике и, соответственно, лингвистических терминов. В результате дообучения были получены две модели: одна, обученная на четырёх эпохах, и другая — на 12. Модель, обученная на четырёх эпохах, распознала больше терминов, чем модель, не проходившая дообучение. После изменения формата субтитров в тестовом датасете связность и когерентность результатов перевода значительно улучшились, что повысило воспринимаемую связность и облегчило экспертную оценку результатов перевода.

Было установлено, что, несмотря на наличие различных факторов, негативно влияющих на качество перевода, нельзя сказать, что после дообучения оно ухудшилось. На самом деле некоторые аспекты перевода улучшились и стали более понятными. Однако улучшение качества перевода все ещё необходимо. Будущей задачей для улучшения качества перевода моделей МП может стать включение в пайплайн алгоритма извлечения терминов, чтобы

определить, сколько узкоспециальных (лингвистических) терминов действительно есть в обучающем корпусе. Другое решение — дополнить обучающий корпус узкоспециальными данными, что, при нехватке подобного рода параллельных данных, можно осуществить путём добавления одноязычных данных с обратным переводом.

Интеграция алгоритмов МП в процесс создания субтитров — действительно перспективный метод повышения скорости их создания. При определённых доработках этот метод, будучи внедрённым в образовательный процесс, может помочь иностранным студентам в изучении русского языка, а также других предметов, преподаваемых на русском языке.

## Литература

- [1] Sutskever I., Vinyals O., Le Q. V. Sequence to sequence learning with neural networks // *Advances in neural information processing systems*. 2014. Vol. 27. DOI: 10.48550/arXiv.1409.3215.
- [2] Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate // *arXiv preprint*. 2014. DOI: 10.48550/arXiv.1409.0473.
- [3] Vaswani A. et al. Attention is all you need // *Advances in neural information processing systems*. 2017. Vol. 30. DOI: 10.48550/arXiv.1706.03762.
- [4] Wang H. et al. Progress in machine translation // *Engineering*. 2022. Vol. 18. P. 143–153.
- [5] Sennrich R., Haddow B., Birch A. Improving neural machine translation models with monolingual data // *arXiv preprint*. 2015. DOI: 10.48550/arXiv.1511.06709.
- [6] Poncelas A. et al. Investigating Backtranslation in Neural Machine Translation // *arXiv preprint*. 2018. DOI: 10.48550/arXiv.1804.06189.
- [7] Touvron H. et al. Llama: Open and efficient foundation language models // *arXiv preprint*. 2023. DOI: 10.48550/arXiv.2302.13971.
- [8] Touvron H. et al. Llama 2: Open foundation and fine-tuned chat models // *arXiv preprint*. 2023. DOI: 10.48550/arXiv.2307.09288.
- [9] Lyu C., Xu J., Wang L. New trends in machine translation using large language models: Case examples with chatgpt // *arXiv preprint*. 2023. DOI: 10.48550/arXiv.2305.01181.
- [10] Zhu W. et al. Multilingual machine translation with large language models: Empirical results and analysis // *arXiv preprint*. 2023. DOI: 10.48550/arXiv.2304.04675.
- [11] Zhang B., Haddow B., Birch A. Prompting large language model for machine translation: A case study // *International Conference on Machine Learning*. PMLR, 2023. P. 41092–41110.
- [12] Wang L. et al. Document-level machine translation with large language models // *arXiv preprint*. 2023. DOI: 10.48550/arXiv.2304.02210.
- [13] Pang J. et al. Salute the classic: Revisiting challenges of machine translation in the age of large language models // *Transactions of the Association for Computational Linguistics*. 2025. Vol. 13. P. 73–95.
- [14] Poibeau T. *Machine translation*. MIT Press, 2017. 296 p.
- [15] Vinyals O. et al. Show and tell: A neural image caption generator // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015. P. 3156–3164.
- [16] Lu J. et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017. P. 375–383.
- [17] Anderson P. et al. Bottom-up and top-down attention for image captioning and visual question answering // *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018. P. 6077–6086.
- [18] Xu Y., Li M., Cui L., Wei F., Zhou M. Layoutlm: Pre-training of text and layout for document image understanding // *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*. 2020. P. 1192–1200.

- [19] Doherty S., Kruger J. L. Assessing quality in human-and machine-generated subtitles and captions // Translation quality assessment: From principles to practice. 2018. P. 179–197.
- [20] Piperidis S., Demiros I., Prokopidis P. Infrastructure for a multilingual subtitle generation system // 9th International Symposium on Social Communication. Cuba: Santiago de Cuba, 2005. P. 24–28.
- [21] Melero M., Oliver A., Badia T. Automatic Multilingual Subtitling in the eTITLE Project // Proceedings of Translating and the Computer 28. 2006.
- [22] Volk M. The automatic translation of film subtitles. A machine translation success story? // Journal for Language Technology and Computational Linguistics. 2009. Vol. 24. No. 3. P. 113–125.
- [23] De Sousa S.C.M., Aziz W., Specia L. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles // Proceedings of the International Conference Recent Advances in Natural Language Processing. 2011. P. 97–103.
- [24] Aziz W., de Sousa S.C.M., Specia L. Cross-lingual sentence compression for subtitles // Proceedings of the 16th Annual conference of the European Association for Machine Translation. 2012. P. 103–110.
- [25] Matusov E., Wilken P., Georgakopoulou Y. Customizing neural machine translation for subtitling // Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). 2019. P. 82–93.
- [26] Koponen M. et al. MT for subtitling: User evaluation of post-editing productivity // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. 2020. P. 115–124.
- [27] Cherry C. et al. Subtitle Translation as Markup Translation // Interspeech. 2021. P. 2237–2241.
- [28] Mejías-Climent L., de los Reyes Lozano J. Traducción automática y posesición en el aula de doblaje: resultados de una experiencia docente. 2021.
- [29] Wilkens L., Heitplatz V. N., Bühler C. Designing accessible videos for people with disabilities // International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2021. P. 328–344.
- [30] Thompson T., Burgstahler S. E. Video for all: accessibility of video content and universal design of a media player // Universal Design in Higher Education. From Principles to Practice. 2015. P. 259–273.
- [31] Shaghghi N. et al. An English to Urdu Educational Video Translation Pipeline to Reinforce Mother-Tongue Based Learning // World Congress on Services. Cham: Springer International Publishing, 2021. P. 61 – 74.
- [32] Song H. et al. Bilingual Corpus Mining and Multistage Fine-Tuning for Improving Machine Translation of Lecture Transcripts // arXiv preprint, 2023. DOI: 10.48550/arXiv.2311.03696.
- [33] Tiedemann J., Thottingal S. OPUS-MT–building open translation services for the world // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. 2020. P. 479–480.

## Integrating Machine Translation Algorithms into the Process of Subtitling Educational Videos

E. P. Lavrentyeva, M. S. Kogan

Peter the Great Saint Petersburg Polytechnic University

The article investigates the possibility of integrating machine translation (MT) algorithms into the task of subtitling educational video content and, in particular, into the process of creating high-quality English subtitles for Russian video lectures on linguistics from the YouTube channel Postnauka. The method chosen to improve the quality of automatically generated subtitles for the lectures is fine-tuning a machine translation model on a large genre-specific corpus of video lecture subtitles. A number of theoretical issues was studied to better understand the problem of automatically subtitling educational videos, such as the nature of the educational video as a genre of audiovisual text, the specifics of audiovisual translation (AVT) and subtitling, and the evolution of and the state-of-the-art in MT algorithms with focus on the application of MT to audiovisual translation. As a result of conducting the experimental study, it was determined that machine translation models translate subtitles coherently when their formatting is changed to the format one line = one sentence. The translation is adequate and fluent except for terminology, which the models struggle to recognize and translate correctly in many cases. After the model was fine-tuned, it started recognizing more terms. Possible reasons for the failure to translate correctly Russian linguistic terminology from the video lectures into English are discussed. The research proved the possibility of applying machine translation algorithms to subtitling educational videos if some pre-editing is done, such as changing the format of the text. Future directions of research are outlined.

**Keywords:** subtitling, educational videos, machine translation, audiovisual translation

**Reference for citation:** Lavrentyeva E. P., Kogan M. S. Integrating Machine Translation Algorithms into the Process of Subtitling Educational Videos // Computational Linguistics and Computational Ontologies. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 12-25. DOI: 10.17586/3033-5582-2025-9-12-25.

### Reference

- [1] Sutskever I., Vinyals O., Le Q. V. Sequence to sequence learning with neural networks // Advances in neural information processing systems. 2014. Vol. 27. DOI: 10.48550/arXiv.1409.3215.
- [2] Bahdanau D., Cho K., Bengio Y. Neural machine translation by jointly learning to align and translate // arXiv preprint. 2014. DOI: 10.48550/arXiv.1409.0473.
- [3] Vaswani A. et al. Attention is all you need // Advances in neural information processing systems. 2017. Vol. 30. DOI: 10.48550/arXiv.1706.03762.
- [4] Wang H. et al. Progress in machine translation // Engineering. 2022. Vol. 18. P. 143–153.
- [5] Sennrich R., Haddow B., Birch A. Improving neural machine translation models with monolingual data // arXiv preprint. 2015. DOI: 10.48550/arXiv.1511.06709.
- [6] Poncelas A. et al. Investigating Backtranslation in Neural Machine Translation // arXiv preprint. 2018. DOI: 10.48550/arXiv.1804.06189.
- [7] Touvron H. et al. Llama: Open and efficient foundation language models // arXiv preprint. 2023. DOI: 10.48550/arXiv.2302.13971.
- [8] Touvron H. et al. Llama 2: Open foundation and fine-tuned chat models // arXiv preprint. 2023. DOI: 10.48550/arXiv.2307.09288.
- [9] Lyu C., Xu J., Wang L. New trends in machine translation using large language models: Case examples with chatgpt // arXiv preprint. 2023. DOI: 10.48550/arXiv.2305.01181.

- [10] Zhu W. et al. Multilingual machine translation with large language models: Empirical results and analysis // arXiv preprint. 2023. DOI: 10.48550/arXiv.2304.04675.
- [11] Zhang B., Haddow B., Birch A. Prompting large language model for machine translation: A case study // International Conference on Machine Learning. PMLR, 2023. P. 41092–41110.
- [12] Wang L. et al. Document-level machine translation with large language models // arXiv preprint. 2023. DOI: 10.48550/arXiv.2304.02210.
- [13] Pang J. et al. Salute the classic: Revisiting challenges of machine translation in the age of large language models // Transactions of the Association for Computational Linguistics. 2025. Vol. 13. P. 73–95.
- [14] Poibeau T. Machine translation. MIT Press, 2017. 296 p.
- [15] Vinyals O. et al. Show and tell: A neural image caption generator // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 3156–3164.
- [16] Lu J. et al. Knowing when to look: Adaptive attention via a visual sentinel for image captioning // Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. P. 375–383.
- [17] Anderson P. et al. Bottom-up and top-down attention for image captioning and visual question answering // Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. P. 6077–6086.
- [18] Xu Y., Li M., Cui L., Wei F., Zhou M. Layoutlm: Pre-training of text and layout for document image understanding // Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining. 2020. P. 1192–1200.
- [19] Doherty S., Kruger J. L. Assessing quality in human-and machine-generated subtitles and captions // Translation quality assessment: From principles to practice. 2018. P. 179–197.
- [20] Piperidis S., Demiros I., Prokopidis P. Infrastructure for a multilingual subtitle generation system // 9th International Symposium on Social Communication. Cuba: Santiago de Cuba, 2005. P. 24–28.
- [21] Mero M., Oliver A., Badia T. Automatic Multilingual Subtitling in the eTITLE Project // Proceedings of Translating and the Computer 28. 2006.
- [22] Volk M. The automatic translation of film subtitles. A machine translation success story? // Journal for Language Technology and Computational Linguistics. 2009. Vol. 24. No. 3. P. 113–125.
- [23] De Sousa S.C.M., Aziz W., Specia L. Assessing the post-editing effort for automatic and semi-automatic translations of DVD subtitles // Proceedings of the International Conference Recent Advances in Natural Language Processing. 2011. P. 97–103.
- [24] Aziz W., de Sousa S.C.M., Specia L. Cross-lingual sentence compression for subtitles // Proceedings of the 16th Annual conference of the European Association for Machine Translation. 2012. P. 103–110.
- [25] Matusov E., Wilken P., Georgakopoulou Y. Customizing neural machine translation for subtitling // Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers). 2019. P. 82–93.
- [26] Koponen M. et al. MT for subtitling: User evaluation of post-editing productivity // Proceedings of the 22nd Annual Conference of the European Association for Machine Translation. 2020. P. 115–124.
- [27] Cherry C. et al. Subtitle Translation as Markup Translation // Interspeech. 2021. P. 2237–2241.
- [28] Mejías-Climent L., de los Reyes Lozano J. Traducción automática y posesición en el aula de doblaje: resultados de una experiencia docente. 2021.
- [29] Wilkens L., Heitplatz V. N., Bühler C. Designing accessible videos for people with disabilities // International Conference on Human-Computer Interaction. Cham: Springer International Publishing, 2021. P. 328–344.

- [30] Thompson T., Burgstahler S. E. Video for all: accessibility of video content and universal design of a media player // *Universal Design in Higher Education. From Principles to Practice*. 2015. P. 259–273.
- [31] Shaghghi N. et al. An English to Urdu Educational Video Translation Pipeline to Reinforce Mother-Tongue Based Learning // *World Congress on Services*. Cham: Springer International Publishing, 2021. P. 61 – 74.
- [32] Song H. et al. Bilingual Corpus Mining and Multistage Fine-Tuning for Improving Machine Translation of Lecture Transcripts // *arXiv preprint*, 2023. DOI: 10.48550/arXiv.2311.03696.
- [33] Tiedemann J., Thottingal S. OPUS-MT–building open translation services for the world // *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*. 2020. P. 479–480.

## Сравнение нейросетевых синтаксических анализаторов для русского языка

Е. Д. Шамаева

МГУ имени М. В. Ломоносова

derinhelm@yandex.ru

### Аннотация

Статья посвящена сравнению качества работы нейросетевых синтаксических анализаторов русского языка UDPipe, Stanza, Natasha, DeepPavlov, spacy. Оценка производилась на тестовых выборках датасетов синтаксически размеченных предложений GSD, PUD, SynTagRus, Poetry, Taiga, входящих в проект Universal Dependencies. Наиболее высокую скорость работы показали анализаторы Natasha, UDPipe и spacy, наилучшее качество работы — анализаторы DeepPavlov, Stanza и spacy. На большинстве анализаторов и датасетов метрика UAS равна 1.0 не более чем для 40 % предложений, метрика LAS — не более чем для 25 % предложений. Помимо стандартной оценки с помощью измерения средних значений метрик UAS и LAS на всем тестовом наборе предложений и на наборах предложений с определенной длиной, в статье исследуются распределения метрик на тестовых предложениях. Кроме того, приведена информация о качестве работы синтаксических анализаторов на наборах токенов с определенными характеристиками: эталонным типом связи, глубиной в эталонном дереве зависимостей, расстоянием до эталонного главного токена. Представленные в статье статистические данные результатов сравнения могут быть использованы для выбора синтаксического анализатора, наиболее подходящего под прикладную задачу. Реализация исследования представлена по адресу [https://github.com/Derinhelm/parser\\_stat](https://github.com/Derinhelm/parser_stat).

**Ключевые слова:** синтаксис, синтаксические анализаторы, дерево зависимостей, проект Universal Dependencies

**Библиографическая ссылка:** Шамаева Е. Д. Сравнение нейросетевых синтаксических анализаторов для русского языка // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 26–47. DOI: 10.17586/3033-5582-2025-9-26-47.

### 1. Введение

Синтаксический анализ — этап обработки текста, на котором происходит выделение синтаксической структуры предложения. Результаты синтаксического анализа используются в таких задачах автоматической обработки текстов, как распознавание именованных сущностей [1, 2, 3, 4], выявление плагиата [5], перефразирование [6], лингвистический анализ текста [7]. Многие синтаксические анализаторы основаны на машинном обучении, в частности, на нейронных сетях [8, 9, 10, 11]. Обучение многих нейросетевых анализаторов происходит на синтаксически размеченных корпусах из проекта Universal Dependencies [12].

В прикладных задачах часто используются синтаксические анализаторы, входящие в системы автоматической обработки текста: анализаторы UDPipe [13], Stanza [14], Natasha,

DeepPavlov, spacy. Данная статья посвящена сравнению качества этих анализаторов на тестовых выборках пяти корпусов из проекта Universal Dependencies: GSD, PUD, SynTagRus, Poetry и Taiga. Для оценки использованы, как стандартные метрики UAS и LAS, так и способы оценки, основанные на измерении качества на наборе токенов.

## 2. Сравнение качества работы синтаксических анализаторов

### 2.1. Аналогичные работы

Для сравнения синтаксических анализаторов проводились соревнования CoNLL 2017 Shared Task [13], CoNLL 2018 Shared Task [15], GramEval 2020 Shared Task [16]. В этих соревнованиях для оценки качества работы анализатора применялась F1-мера метрики LAS с выравниванием<sup>1</sup> и дополнительно F1-мера метрики UAS с выравниванием. Анализаторы Stanza, Natasha, DeepPavlov, spacy не принимали участия в этих соревнованиях. В соревнованиях CoNLL 2017 Shared Task и CoNLL 2018 Shared Task использовались данные из проекта Universal Dependencies, соревнование GramEval 2020 Shared Task — на 6 наборах предложений различных категорий: новости, поэтические тексты, сообщения из социальных сетей и тексты электронной коммуникации, энциклопедические статьи, художественная литература, тексты 17 века.

Сравнение русскоязычных анализаторов, предоставляющих удобный интерфейс к предобученным моделям, проведено в проекте Naeval [17]. В сравнении участвовали анализаторы UDPipe, Stanza, Natasha, DeepPavlov, spacy. В проекте Naeval использовались метрики Unlabeled Attachment Score (UAS) и Labeled Attachment Score (LAS) без выравнивания. Эти метрики отличаются от стандартных метрик UAS и LAS с выравниванием и некорректно работают для предложений, у которых разбиение на токены, полученное с помощью анализатора, отличается от эталонного разбиения на токены из датасета. Поэтому в текущем исследовании метрики UAS и LAS без выравнивания не могут быть применены, а результаты, полученные в проекте Naeval, не могут быть корректно сопоставлены с результатами текущего исследования. В проекте Naeval сравнение проводилось на датасетах из соревнования GramEval 2020 Shared Task.

Таблица 1. Сравнение с другими исследованиями

Исследование	Данные	Метрики	Анализаторы
Соревнование CoNLL 2017 Shared Task	Данные из Universal Dependencies	F1-мера метрик UAS и LAS с выравниванием	UDPipe и другие
Соревнование CoNLL 2018 Shared Task	Данные из Universal Dependencies	F1-мера метрик UAS и LAS с выравниванием	UDPipe и другие
Соревнование GramEval 2020 Shared Task	Данные для GramEval 2020 Shared Task	F1-мера метрик UAS и LAS с выравниванием	UDPipe и другие
Проект Naeval	Данные для GramEval 2020 Shared Task	Метрики UAS и LAS без выравнивания	UDPipe, Stanza, Natasha, DeepPavlov, spacy
Данное исследование	GSD, PUD, SynTagRus, Poetry, Taiga (из Universal Dependencies)	F1-мера метрик UAS и LAS с выравниванием, оценка на наборе токенов	UDPipe, Stanza, Natasha, DeepPavlov, spacy

<sup>1</sup>Использовалась также метрика MLAS, учитывающая метку части речи и морфологические характеристики токенов. В данной работе рассматриваются только метрики, связанные с синтаксическими характеристиками токенов, поэтому метрика MLAS не учитывается.

В табл. 1 представлено сопоставление текущего исследования с другими проектами сравнения качества работы анализаторов.

## 2.2. Данные

В датасетах проекта Universal Dependencies синтаксическая информация о предложении хранится с помощью деревьев зависимостей (dependency tree). Пример дерева зависимостей приведен на рис. 1. В качестве вершин дерево зависимостей содержит элементы предложения, называемые токенами (слова, знаки препинания, числа). Ребра дерева зависимостей соответствуют отношениям между токенами, каждое ребро имеет метку типа связи и направлено от родительского (главного) токена к дочернему. Дерево зависимостей также содержит в качестве вершины вспомогательный токен root, с которым связан корневой токен данного дерева зависимостей (используется специальный тип связи root).

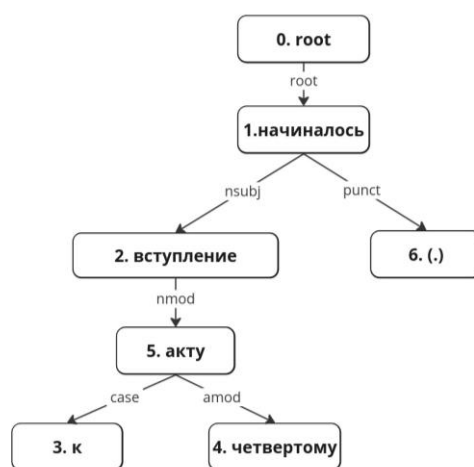


Рис. 1. Пример дерева зависимостей

На момент исследования в проект Universal Dependencies включены пять русскоязычных датасетов деревьев зависимостей: GSD, PUD, SynTagRus, Poetry, Taiga. Эти датасеты различаются с точки зрения входящих в них текстов.

Согласно [18], датасет GSD (Google Stanford Dependencies) основан на русскоязычных текстах энциклопедии Wikipedia. Датасет PUD (Parallel Universal Dependencies) содержит тексты из энциклопедии Wikipedia и новостные тексты. Все тексты, входящие в датасет PUD, изначально написаны на иностранных языках (английском, немецком, французском, итальянском, испанском) и переведены на русский язык вручную.

Датасет SynTagRus содержит тексты русской художественной прозы (начиная со второй половины XX в.) и научно-популярные, общественно-политические и информационные статьи из журналов и интернет-изданий (с 1980 г. по настоящее время). Датасет Poetry содержит русскоязычные поэтические тексты XIX — начала XXI. В датасете Taiga содержится большее количество текстов, относящихся к электронной коммуникации (блоги, социальные медиа).

Для исследования использовались тестовые выборки датасетов. Тестовая выборка датасета GSD содержит 601 предложение, PUD — 1000, SynTagRus — 8800, Poetry — 728, Taiga — 881.

### 2.3. Исследуемые синтаксические анализаторы

В данном исследовании для сравнения выбраны синтаксические анализаторы, входящие в системы автоматической обработки текста: Stanza, Natasha, DeepPavlov, spacy и UDPipe 1.0 (на момент исследования анализатор UDPipe 2.0 не предоставлял доступ к моделям для русского языка). Нейросетевые модели UDPipe, Stanza, DeepPavlov обучены на корпусе SynTagRus, Natasha и spacy — на новостных текстах из соревнования GramEval 2020 Shared Task<sup>2</sup>.

Анализаторы spacy и UDPipe 1.0 относятся к категории анализаторов на основе переходов. В этих анализаторах разбор предложения происходит поэтапно, за один проход по предложению слева направо. С помощью машинного обучения предсказывается преобразование (переход), которое необходимо совершить над предложением и частично построенным деревом зависимостей на данном этапе анализа.

Анализаторы DeepPavlov, Stanza и Natasha относятся к категории анализаторов на основе графов. В этих анализаторах с помощью машинного обучения предсказываются веса ребер между каждой парой токенов. Далее с помощью алгоритма Чу-Лью / Эдмондса происходит построение минимального остовного дерева. В анализаторах Stanza и DeepPavlov для предсказания весов ребер используется бифинная нейронная сеть [19] на основе двунаправленной архитектуры LSTM. В Stanza модель также дополнена двумя лингвистически мотивированными функциями: одна из них предсказывает порядок линейаризации двух слов в данном языке, а другая — расстояние между ними в линейном порядке. Анализатор Natasha получен с помощью дистилляции из анализатора DeepPavlov.

### 2.4. Оценка качества работы анализатора

Оценка качества работы анализатора проводится на наборе предложений с синтаксической разметкой. Существует два вида оценки [9]: на множестве предложений (с последующим усреднением метрик) и на множестве эталонных токенов из всех тестовых предложений. При первом способе оценки могут рассматриваться группы предложений (например, с определенной длиной). При втором способе оценки часто рассматривается множество эталонных токенов с определенными характеристиками (например, эталонным типом связи с главным токеном, расстоянием до эталонного главного токена, глубиной в эталонном дереве зависимостей).

Разбиение на токены в эталонной разметке из датасета и разбиение на токены, созданное анализатором, могут отличаться. Например, в предложении «*Он давным-давно знал всю партитуру наизусть.*» в датасете SynTagRus фрагмент «*давным-давно*» рассматривается как 1 токен, а синтаксический анализатор Stanza выделяет 3 токена: «*давным*», «*-*», «*давно*». Поэтому во втором способе оценки, на множестве токенов, используются только эталонные токены.

В обоих способах оценки может учитываться только совпадение главных токенов или совпадение и главных токенов, и типов связи.

### 2.5. Метрики UAS и LAS

Популярными метриками для оценки качества работы анализатора на предложении являются метрики UAS (Unlabeled Attachment Score) и LAS (Labeled Attachment Score) [20]. Метрика UAS вычисляется как процент токенов, для которых анализатор верно предсказал родительский токен, метрика LAS — как процент токенов, для которых анализатор верно предсказал и родительский токен и тип связи.

---

<sup>2</sup>Для UDPipe использовалась модель russian-syntagrus-ud-2.5-191206.udpipe, для Stanza — syntagrus\_charlm, для Natasha — slovnet\_syntax\_news\_v1, для DeepPavlov — syntax\_ru\_syntagrus\_bert, для spacy — ru\_core\_news\_lg.

Для метрик UAS и LAS вычисляются точность и полнота. В сравнении родительских токенов (и типа связи) участвуют только токены, присутствующие и в эталонном наборе токенов, и в наборе токенов, созданном анализатором. На рис. 2 приведены формулы точности, полноты и F1-меры для метрик UAS и LAS со следующими обозначениями:  $G$  — набор эталонных токенов предложения (gold token set),  $P$  — набор токенов, созданный анализатором (parser token set),  $gt$  — токен из эталонного набора,  $pt$  — токен из набора токенов, созданного анализатором,  $\mathbf{p}(t)$  — функция, возвращающая индексы начала и конца родительского токена для токена  $t$ ,  $\mathbf{d}(t)$  — функция, возвращающая тип связи, которым токен  $t$  связан с родительским токеном.

$$\begin{aligned}
 UAS\_precision &= \frac{\|gt|gt = pt, gt \in G, pt \in P, \mathbf{p}(gt) = \mathbf{p}(pt)\|}{\|pt|pt \in P\|} \\
 UAS\_recall &= \frac{\|gt|gt = pt, gt \in G, pt \in P, \mathbf{p}(gt) = \mathbf{p}(pt)\|}{\|gt|gt \in G\|} \\
 LAS\_precision &= \frac{\|gt|gt = pt, gt \in G, pt \in P, \mathbf{p}(gt) = \mathbf{p}(pt), \mathbf{d}(gt) = \mathbf{d}(pt)\|}{\|pt|pt \in P\|} \\
 LAS\_recall &= \frac{\|gt|gt = pt, gt \in G, pt \in P, \mathbf{p}(gt) = \mathbf{p}(pt), \mathbf{d}(gt) = \mathbf{d}(pt)\|}{\|gt|gt \in G\|} \\
 UAS\_F1 &= \frac{2 * UAS\_precision * UAS\_recall}{UAS\_precision + UAS\_recall} \\
 LAS\_F1 &= \frac{2 * LAS\_precision * LAS\_recall}{LAS\_precision + LAS\_recall}
 \end{aligned}$$

Рис. 2. Формулы вычисления F1-меры для метрик UAS и LAS

### 3. Результаты исследования

#### 3.1. Скорость работы анализаторов

Таблица 2. Усредненная скорость обработки одного предложения

	Taiga	Poetry	GSD	PUD	SynTagRus	Среднее время
Natasha	0.0008	0.0008	0.0011	0.0009	0.0009	0.0009
UDPipe	0.0023	0.0026	0.0044	0.0045	0.0038	0.0035
spacy	0.0027	0.0029	0.0036	0.0037	0.0034	0.0033
DeepPavlov	0.0256	0.0263	0.0286	0.0276	0.0273	0.0271
Stanza	0.0764	0.0818	0.1239	0.1240	0.1122	0.1037

В табл. 2 представлена усредненная скорость работы анализаторов. Усреднение выполнялось сначала внутри одного датасета, затем — между датасетами. Анализаторы продемонстрировали разную среднюю скорость работы на разных датасетах, что может быть связано с длиной тестовых предложений или другими особенностями тестовых данных. С точки зрения убывания скорости работы анализаторы упорядочены следующим образом: Natasha, UDPipe и spacy, DeepPavlov, Stanza.

#### 3.2. Сравнение анализаторов с помощью метрик UAS и LAS

В табл. 3 и 4 представлены усредненные значения метрик UAS и LAS по всем тестовым предложениям. Наилучшие результаты по метрике UAS показывают анализаторы Stanza и DeepPavlov, по метрике LAS — DeepPavlov, Stanza и spacy.

Таблица 3. Средние значения метрики UAS

	Taiga	Poetry	GSD	PUD	SynTagRus
Natasha	0.70	0.64	0.79	0.88	0.83
UDPipe	0.73	0.72	0.79	0.86	0.88
spacy	0.77	0.75	0.84	0.91	0.87
DeepPavlov	0.79	0.84	0.83	0.94	0.92
Stanza	0.79	0.82	0.85	0.93	0.94

Таблица 4. Средние значения метрики LAS

	Taiga	Poetry	GSD	PUD	SynTagRus
Natasha	0.64	0.58	0.75	0.84	0.78
UDPipe	0.66	0.65	0.71	0.79	0.84
spacy	0.70	0.69	0.80	0.87	0.82
DeepPavlov	0.72	0.78	0.75	0.86	0.89
Stanza	0.72	0.76	0.79	0.87	0.91

На большинстве датасетов средние значения метрик UAS и LAS невысокие. Среднее значение LAS превышает 0.91 только для анализатора Stanza на датасете SynTagRus. Только на датасетах PUD и SynTagRus средние значения метрики UAS превышают 0.9. Однако не для всех анализаторов разница между средними значениями метрик «наилучшего» и «наихудшего» анализаторов существенна. На датасете Poetry разница между этими значениями и для UAS, и для LAS составляет 0.2. Для остальных анализаторов разница между «наилучшим» и «наихудшим» составляет от 0.06 до 0.11 по UAS и от 0.08 до 0.13 по LAS.

Средние значения метрик существенно различаются в зависимости от датасета. И по метрике UAS, и по метрике LAS анализаторы демонстрируют наиболее высокое среднее значение метрик на датасетах SynTagRus и PUD, наиболее низкое — на датасетах Taiga и Poetry. Синтаксические анализаторы Natasha и spacy были обучены на новостных текстах, а остальные синтаксические анализаторы — на корпусе SynTagRus. Возможно, поэтому анализаторы Natasha и spacy показали более высокие результаты на корпусе PUD, чем на других корпусах, а на корпусе SynTagRus продемонстрировали наиболее низкие результаты среди всех анализаторов.

На рис. 3 показаны распределения значений метрик UAS и LAS, более подробные диаграммы распределений приведены в Приложении А. Соотношение количества предложений с высокими или низкими значениями метрик UAS и LAS достаточно сильно зависит от датасета. Для датасета SynTagRus большее количество предложений имеет более высокие значения метрик UAS и LAS. Для датасетов PUD и GSD возрастает доля предложений со средними значениями UAS и LAS, для датасетов Taiga и Poetry — доля предложений с низкими значениями UAS и LAS.

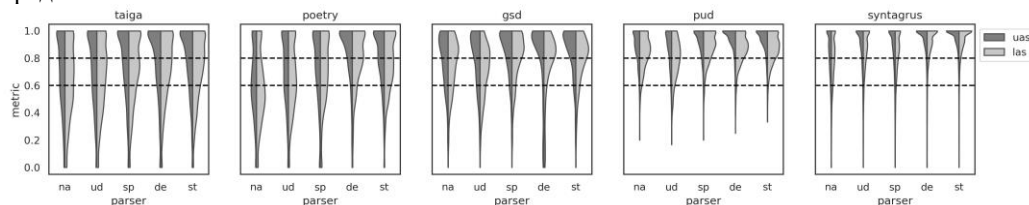


Рис. 3. Распределения средних значений метрик UAS и LAS

Почти для всех анализаторов и датасетов доля предложений с оценкой 1.0 для метрики UAS не превышает 40 %, для метрики LAS — 25 %<sup>3</sup>.

Статистика по среднему значению метрик UAS и LAS на предложениях с определенной длиной предложений представлена в Приложении В. Только в датасете SynTagRus группы предложений с длиной 30-40 токенов, 40-50 токенов и более 50 токенов содержат более 100 токенов. На этом датасете среднее значение метрик на предложениях с большей длиной меньше или равно среднему значению метрик на предложениях с меньшей длиной. На датасете SynTagRus анализаторы Stanza и DeepPavlov на большинстве предложений длины 1-10 токенов показали результат 1.0 по метрике UAS.

На предложениях длины 10-20 токенов на всех датасетах по обоим метрикам все анализаторы показывают результаты, близкие к результатам на полном множестве тестовых предложений. На датасетах Poetry, GSD и PUD для многих анализаторов на предложениях длиной 1-10 токенов средние значения метрик UAS и LAS превышают средние значения на всем наборе токенов, а на предложениях длиной 20-30 средние значения метрик ниже, чем средние значения метрик на всем наборе токенов.

### 3.3. Оценка на множествах эталонных токенов

В табл. 5 и 6 в процентах представлена доля эталонных токенов для которых верно определен главный токен или главный токен и тип связи соответственно. Для всех датасетов и всех анализаторов для более 50 % токенов верно предсказаны и главный токен, и тип связи.

**Таблица 5.** Доля эталонных токенов с верно предсказанным родительским токеном и типом связи (в процентах)

	Taiga	Poetry	GSD	PUD	SynTagRus
Natasha	62	53	71	82	75
UDPipe	65	62	69	77	82
spacy	69	67	78	87	82
DeepPavlov	72	75	70	86	88
Stanza	73	73	77	86	91

**Таблица 6.** Доля эталонных токенов с верно предсказанным родительским токеном и неверно предсказанным типом связи (в процентах)

	Taiga	Poetry	GSD	PUD	SynTagRus
Natasha	6	6	5	5	4
UDPipe	7	8	8	7	5
spacy	6	6	4	4	5
DeepPavlov	7	7	8	7	3
Stanza	7	6	6	6	3

Глубиной токена называется его глубина в дереве зависимостей — длина пути от вспомогательного корневого токена root до данного токена. В тестовых выборках рассматриваемых датасетов большинство токенов находятся на глубине менее 5. Качество работы анализатора на токенах определенной глубины зависит и от датасета, и от анализатора. Например, на датасете Taiga синтаксический анализатор Stanza показывает более высокие результаты на токенах глубины 4-6, а на датасете GSD —

<sup>3</sup>Исключения для UAS: анализатор Stanza на датасете SynTagRus, анализатор DeepPavlov на датасетах PUD и SynTagRus. Исключения для LAS: анализатор DeepPavlov на датасетах Poetry и SynTagRus, анализаторы Stanza и UDPipe — на датасете SynTagRus.

на токенах глубины 2-3. Распределения глубин эталонных токенов в тестовых выборках рассматриваемых датасетов приведено в Приложении С.

Длиной зависимости называется разница между номером токена и номером главного токена. В большинстве случаев на токенах с меньшей по модулю длиной зависимости анализаторы показывают более высокие результаты. Более подробная статистика по токенам с определенной длиной зависимости представлена в Приложении D.

Таблица со статистикой качества анализа по токенам с определенным типом связей представлена в Приложении E. Показано, что для некоторых эталонных типов связи качество работы анализатора ниже, чем на всем множестве токенов. Например, анализатор *Natasha* успешно определяет главный токен и тип связи для 75 % токенов из *SynTagRus* и только для 52 % токенов с эталонным типом связи *conj*. А анализатор *UDPipe* на датасете *Poetry* корректно определяет главный токен и тип связи для 62 %, а на множестве токенов из этого датасета с типом связи *case* — на 87 % токенов.

На датасете *Taiga* для большой доли токенов в корне дерева неверно определен главный токен — то есть во всем дереве неверно определен главный токен. Во всех остальных датасетах также есть предложения, в которых неверно определен корневым токен.

#### 4. Заключение

В данном исследовании проведено сравнение качества работы нейросетевых синтаксических анализаторов *Natasha*, *UDPipe*, *spacy*, *Stanza* и *DeepPavlov*. Тестирование анализаторов проведено на датасетах деревьев зависимостей из проекта *Universal Dependencies: SynTagRus, GSD, PUD, Taiga, Poetry*. Все анализаторы показывают более низкие результаты на датасетах *Taiga, Poetry, GSD* и более высокие — на датасетах *PUD, SynTagRus*.

С точки зрения скорости анализаторы разделились на 3 группы: высокую скорость работы показали анализаторы *Natasha, UDPipe, spacy*, среднюю — *DeepPavlov*, низкую — *Stanza*. Однако более высокое качество работы показали анализаторы *DeepPavlov, Stanza* и *spacy*. Практически на всех датасетах результат работы анализатора полностью совпадает с эталонным не более чем для 25 % тестовых предложений.

Кроме того, установлено, что качество работы синтаксического анализатора на предложении зависит от длины предложения, а корректность определения родительского токена и типа связи для токена зависит от таких свойств токена, как эталонный тип связи с главным токеном, глубина в эталонном дереве зависимостей, расстояние до эталонного родительского токена. Однако характер этой взаимосвязи зависит как от набора тестовых предложений, так и от используемого синтаксического анализатора. В дальнейших исследованиях планируется более глубоко изучить взаимосвязь между правильностью синтаксического анализа для предложения (токена) и характеристиками предложения (токена).

Автор благодарит Лукашевич Наталью Валентиновну и Волкову Ирину Анатольевну за советы при подготовке статьи.

#### Литература

- [1] Lin L., Ziyang C., Shuxing L., Yibin D., Hongxiao W., Zhihao L. Event extraction in complex sentences based on dependency parsing and longformer // *Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing*. Wuhan, China, 2024. P. 1-7. URL: <https://proceedings.mlr.press/v245/lin24a.html> (дата обращения: 29.03.2025).
- [2] Vasiliev S., Korobkin D., Fomenkov S. Extracting the component composition data of inventions from russian patents using dependency tree analysis // *2023 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM)*. Sochi,

- Russian Federation, 2023. P. 1030-1034. URL: <https://ieeexplore.ieee.org/document/10139170> (дата обращения: 29.03.2025).
- [3] Alonso M.A., Gómez-Rodríguez C., Vilares J. On the use of parsing for named entity recognition // *Applied Sciences*. 2021. No. 11(3). DOI: 10.3390/app11031090.
- [4] Nikolaev I. Knowledge and skills extraction from the job requirements texts // *Ontology of Designing*. 2023. No. 13(2). P. 282-293. URL: <https://journals.ssau.ru/ontology/article/view/27001> (дата обращения: 29.03.2025).
- [5] Taufiq U., Pulungan R., Suyanto Y. Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection // *Expert Systems with Applications*. 2023. Vol. 217. Article 119579. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0957417423000805> (дата обращения: 29.03.2025).
- [6] Liu T., Sun Y., Wu J., Xu X., Han Y., Li C., Gong B. Unsupervised paraphrasing under syntax knowledge // *Proceedings of the AAAI Conference on Artificial Intelligence*. 2023. No. 37(11). P. 13273-13281. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/26558> (дата обращения: 29.03.2025).
- [7] Corbetta C., Passarotti M., Moretti G. The rise and fall of dependency parsing in dante alighieri's divine comedy // *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)*. Torino, Italia, 2024. P. 50-56. URL: <https://aclanthology.org/2024.lt4hala-1.7/> (дата обращения: 29.03.2025).
- [8] Altıntaş M., Cüneyd Tantuğ A. Improving the performance of graph based dependency parsing by guiding bi-affine layer with augmented global and local features // *Intelligent Systems with Applications*. 2023. No. 18. Article 200190. URL: <https://www.sciencedirect.com/science/article/pii/S2667305323000157> (дата обращения: 29.03.2025).
- [9] Kulmizev A., Lhoneux M., Gontrum J., Fano E., Nivre J. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 2019. P. 2755-2768. URL: <https://aclanthology.org/D19-1277/> (дата обращения: 29.03.2025).
- [10] Mohammadshahi A., Henderson J. Graph-to-graph transformer for transition-based dependency parsing // *Findings of the Association for Computational Linguistics: EMNLP*. 2020. P. 3278-3289. URL: <https://aclanthology.org/2020.findings-emnlp.294/> (дата обращения: 29.03.2025).
- [11] Zuhra F.T., Saleem K., Naz S. An accurate transformer-based model for transition-based dependency parsing of free word order languages // *Journal of King Saud University - Computer and Information Sciences*. 2024. No. 36(6). P. 102-107. URL: <https://www.sciencedirect.com/science/article/pii/S1319157824001964> (дата обращения: 29.03.2025).
- [12] Marneffe M., Manning C.D., Nivre J., Zeman D. Universal dependencies // *Computational Linguistics*. 2021. No. 47(2). P. 255-308. URL: <https://aclanthology.org/2021.cl-2.11/> (дата обращения: 29.03.2025).
- [13] Straka M., Strakova J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipes // *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, 2017. P. 88-99. URL: <https://aclanthology.org/K17-3009/> (дата обращения: 29.03.2025).
- [14] Qi. P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A python natural language processing toolkit for many human languages // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. P. 101-108. URL: <https://aclanthology.org/2020.acl-demos.14/> (дата обращения: 29.03.2025).

- [15] Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. Brussels, Belgium 2018. P. 1-21. URL: <https://aclanthology.org/K18-2001/> (дата обращения: 29.03.2025).
- [16] Ляшевская О.Н., Шаврина Т.О., Трофимов И.В., Власова Н.А. Grameval 2020: дорожка по автоматическому морфологическому и синтаксическому анализу русских текстов // Annual International Conference Dialogue. 2020. P. 553-569. URL: <https://publications.hse.ru/pubs/share/direct/395318448.pdf> (дата обращения: 29.03.2025).
- [17] Проект Naeval. URL: <https://github.com/natasha/naeval> (дата обращения: 29.03.2025).
- [18] Droганova K., Lyashevskaya O., Zeman D. Data conversion and consistency of monolingual corpora: Russian ud treebanks // Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018). Linköping, Sweden. 2018. P. 53-66. URL: <https://ufal.mff.cuni.cz/biblio/attachments/2018-droganova-p7646871821297838363.pdf> (дата обращения: 29.03.2025).
- [19] Dozat T., Manning C.D. Deep biaffine attention for neural dependency parsing // International Conference on Learning Representations (ICLR). 2017. URL: <https://arxiv.org/abs/1611.01734> (дата обращения: 29.03.2025).
- [20] Zhang M.A. survey of syntactic-semantic parsing based on constituent and dependency structures // Science China Technological Sciences. 2020. No. 63(10). P. 1898-1920. DOI: 10.1007/s11431-020-1666-4.

## Приложение А. Распределение метрик UAS и LAS

На рис. 4 представлена схема визуализации данных *boxplot*, используемая в данном исследовании. Левая граница прямоугольника соответствует первому квартилю (Q1), правая — третьему (Q3). Второй квартиль (Q2) показан линией желтого цвета. Кроме того, отрезком показана граница выбросов. Статистически значимыми считаются данные находящиеся в диапазоне  $[Q1 - 1.5 IQR; Q3 + 1.5 IQR]$ , где IQR — межквартильный размах (разница между третьим и первым квартилями). Этим значениям соответствуют левый и правый концы отрезка, считающиеся статистическими выбросами.

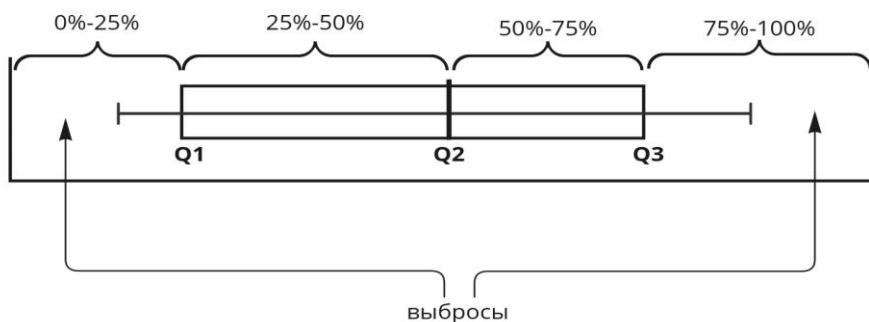


Рис. 4. Схема диаграммы *boxplot*

Соответствующие распределениям диаграммы вида *boxplot* показаны на рис. 5 и 6: линией серого цвета обозначено среднее значение метрики (соответствующая числовая величина указана над диаграммой *boxplot*), линией черного цвета — медианное значение метрики. Значения квартилей, отличающихся от 1.0, указаны под каждой диаграммой *boxplot*. Кроме того, под диаграммой *boxplot* указаны отличные от 1.0 границы статической значимости метрик.

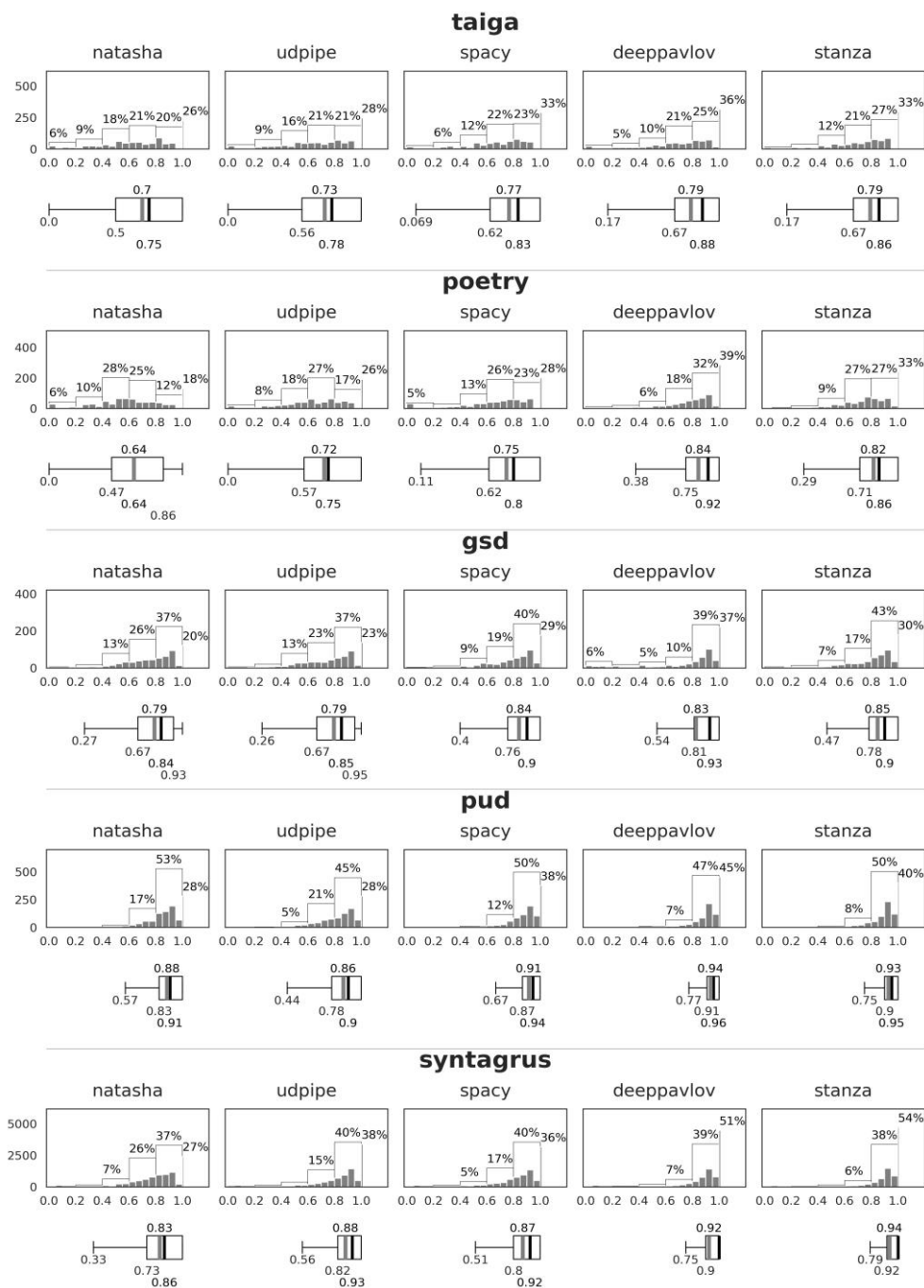


Рис. 5. Распределение метрики UAS

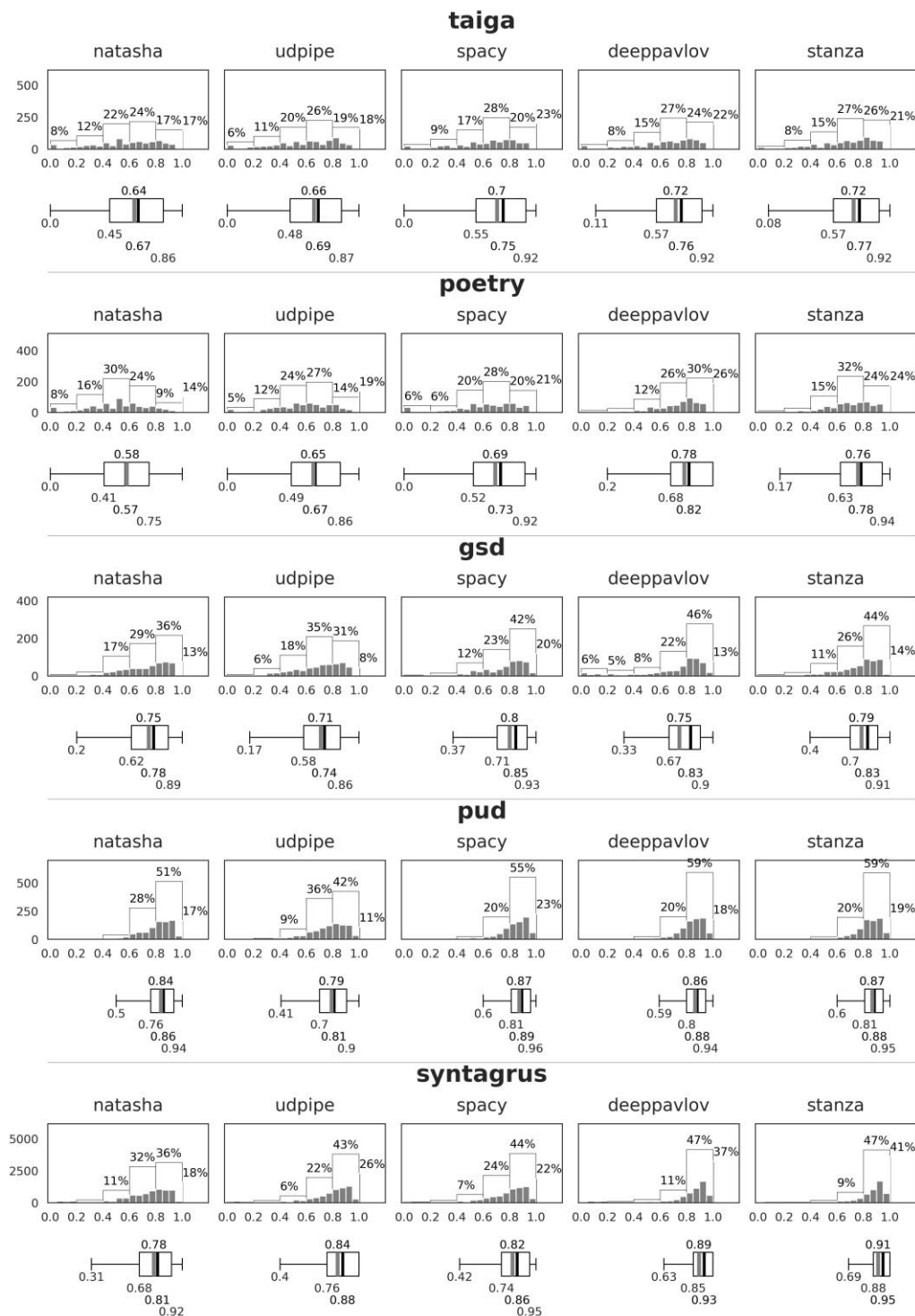


Рис. 6. Распределение метрики LAS

## Приложение В. Среднее значение метрик UAS и LAS на группах предложений с определенной длиной

В табл. 7 показано среднее значение метрик UAS и LAS на предложениях определенной длины. На рис. 7 и рис. 8 показаны распределения метрик UAS и LAS соответственно на группах предложений разной длины.

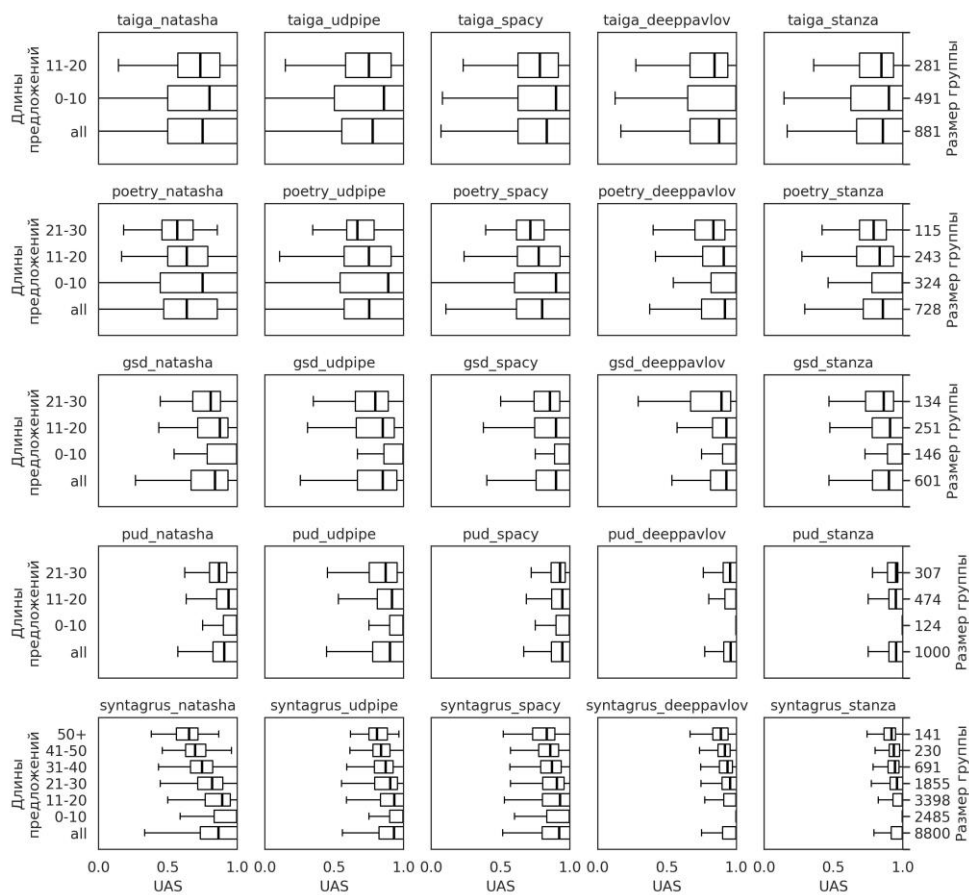


Рис. 7. Распределение метрик UAS в зависимости от длины предложения

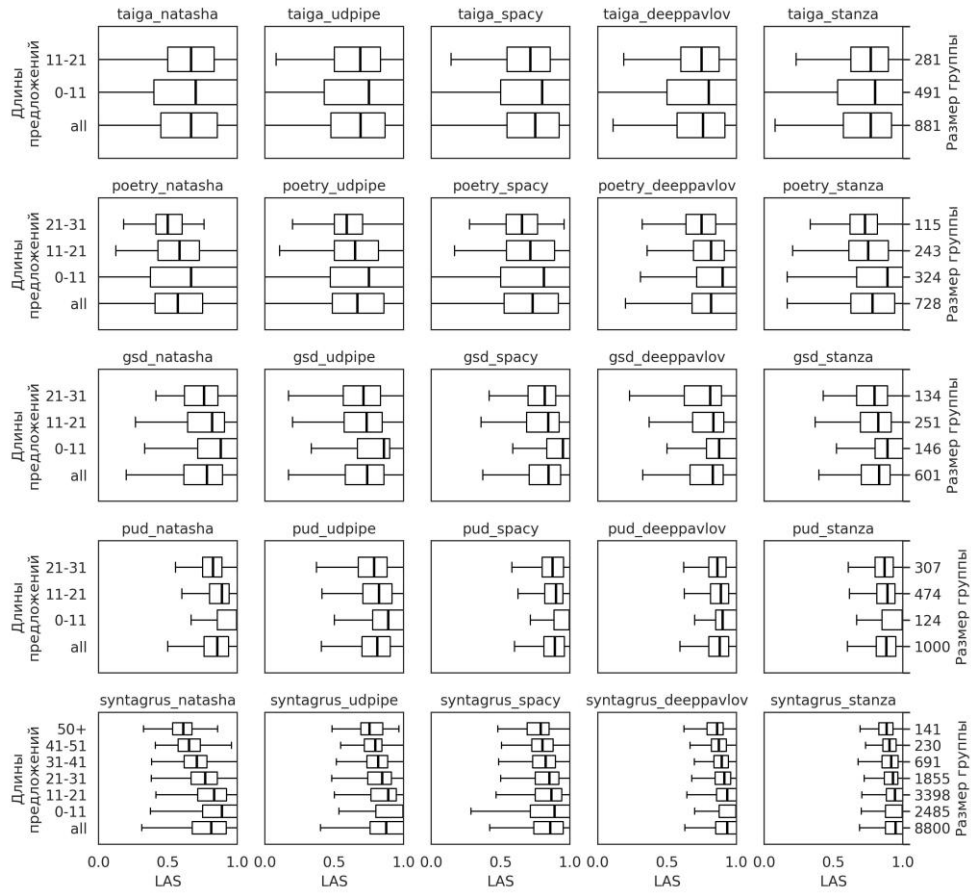


Рис. 8. Распределение метрик LAS в зависимости от длины предложения

Таблица 7. Среднее значение метрик UAS и LAS на группах предложений с определенной длиной

Датасет	Метрика	UAS							LAS						
		Длина	все	0-10	10-20	20-30	30-40	40-50	50+	все	0-10	10-20	20-30	30-40	40-50
Taiga	Natasha	0.70	0.71	0.71	0.63	0.62	0.61	0.48	0.64	0.64	0.65	0.58	0.58	0.56	0.45
	UDPipe	0.73	0.73	0.73	0.7	0.73	0.76	0.64	0.66	0.66	0.65	0.63	0.66	0.69	0.57
	spacy	0.77	0.78	0.75	0.74	0.78	0.7	0.68	0.70	0.71	0.69	0.69	0.72	0.65	0.63
	DeepPavlov	0.79	0.79	0.79	0.78	0.81	0.77	0.73	0.72	0.72	0.72	0.71	0.75	0.72	0.64
	Stanza	0.79	0.79	0.8	0.78	0.84	0.82	0.68	0.72	0.72	0.73	0.71	0.78	0.75	0.62
	Размер	881	491	281	81	15	4	9	881	491	281	81	15	4	9
Poetry	Natasha	0.64	0.68	0.64	0.56	0.53	0.5	0.46	0.58	0.63	0.57	0.5	0.47	0.45	0.41
	UDPipe	0.72	0.75	0.72	0.67	0.65	0.57	0.63	0.65	0.69	0.65	0.59	0.57	0.49	0.54
	spacy	0.75	0.76	0.76	0.71	0.71	0.68	0.67	0.69	0.71	0.7	0.65	0.66	0.62	0.59
	DeepPavlov	0.84	0.86	0.85	0.8	0.81	0.79	0.74	0.78	0.81	0.78	0.73	0.74	0.72	0.66
	Stanza	0.82	0.86	0.79	0.78	0.77	0.72	0.69	0.76	0.81	0.73	0.71	0.7	0.65	0.63
	Размер	728	324	243	115	30	9	7	728	324	243	115	30	9	7
GSD	Natasha	0.79	0.86	0.81	0.76	0.66	0.62	0.52	0.75	0.81	0.76	0.72	0.61	0.57	0.49
	UDPipe	0.79	0.87	0.79	0.76	0.74	0.7	0.65	0.71	0.77	0.7	0.68	0.66	0.61	0.58
	spacy	0.84	0.91	0.84	0.81	0.78	0.76	0.75	0.80	0.87	0.8	0.78	0.72	0.7	0.7
	DeepPavlov	0.83	0.91	0.85	0.77	0.71	0.64	0.57	0.75	0.82	0.76	0.71	0.64	0.56	0.52
	Stanza	0.85	0.91	0.85	0.81	0.8	0.77	0.72	0.79	0.84	0.79	0.76	0.74	0.7	0.65
	Размер	601	146	251	134	46	13	11	601	146	251	134	46	13	11
PUD	Natasha	0.88	0.94	0.91	0.85	0.81	0.75	-	0.84	0.9	0.86	0.81	0.76	0.71	-
	UDPipe	0.86	0.93	0.88	0.84	0.79	0.82	-	0.79	0.86	0.8	0.77	0.72	0.76	-
	spacy	0.91	0.94	0.92	0.91	0.89	0.86	-	0.87	0.92	0.88	0.86	0.83	0.81	-
	DeepPavlov	0.94	0.96	0.94	0.93	0.92	0.88	-	0.86	0.9	0.87	0.85	0.84	0.82	-
	Stanza	0.93	0.95	0.93	0.92	0.9	0.88	-	0.87	0.9	0.87	0.86	0.84	0.82	-
	Размер	1000	124	474	307	79	16	0	1000	124	474	307	79	16	0
SynTa gRus	Natasha	0.83	0.87	0.85	0.8	0.74	0.7	0.64	0.78	0.83	0.8	0.75	0.69	0.66	0.6
	UDPipe	0.88	0.91	0.88	0.86	0.85	0.83	0.81	0.84	0.87	0.84	0.82	0.8	0.78	0.76
	spacy	0.87	0.87	0.87	0.87	0.85	0.83	0.82	0.82	0.82	0.83	0.82	0.8	0.79	0.77
	DeepPavlov	0.92	0.92	0.93	0.92	0.91	0.9	0.88	0.89	0.9	0.9	0.89	0.88	0.87	0.84
	Stanza	0.94	0.94	0.94	0.93	0.92	0.93	0.9	0.91	0.91	0.92	0.91	0.9	0.9	0.87
	Размер	8800	2485	3398	1855	691	230	141	8800	2485	3398	1855	691	230	141

## Приложение С. Распределение доли верно разобранных токенов определенной глубины

На рис. 9 показана доля эталонных токенов определенной глубины, для которых верно определен родительский токен. Белый цвет соответствует эталонным токенам фиксированной глубины, для которых верно определен главный токен, серый цвет — эталонным токенам с неверно определенным родительским токеном. В правом верхнем углу каждой диаграммы указана процентная доля токенов датасета, для которых верно определен главный токен.

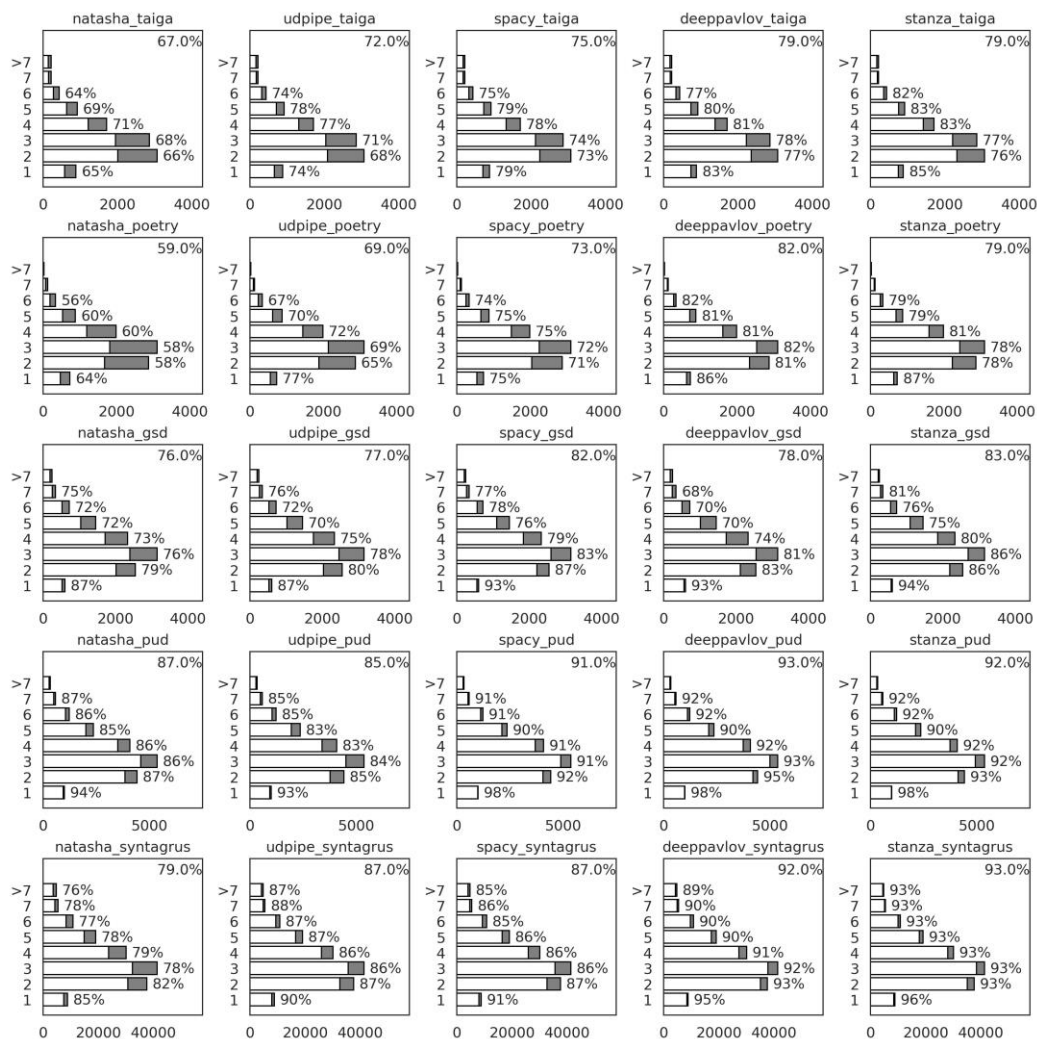


Рис. 9. Распределение доли верно разобранных токенов определенной глубины

## Приложение D. Распределение доли верно разобранных токенов с определенным расстоянием до главного токена

На рис. 10 отражена корректность работы анализатора на множестве эталонных токенов с определенным расстоянием до главного токена. Белый цвет соответствует токенам, для которых верно определен главный токен, серый цвет — токенам с неверно определенным главным токеном. В правом верхнем углу каждой диаграммы указана процентная доля токенов датасета, для которых верно определен главный токен.

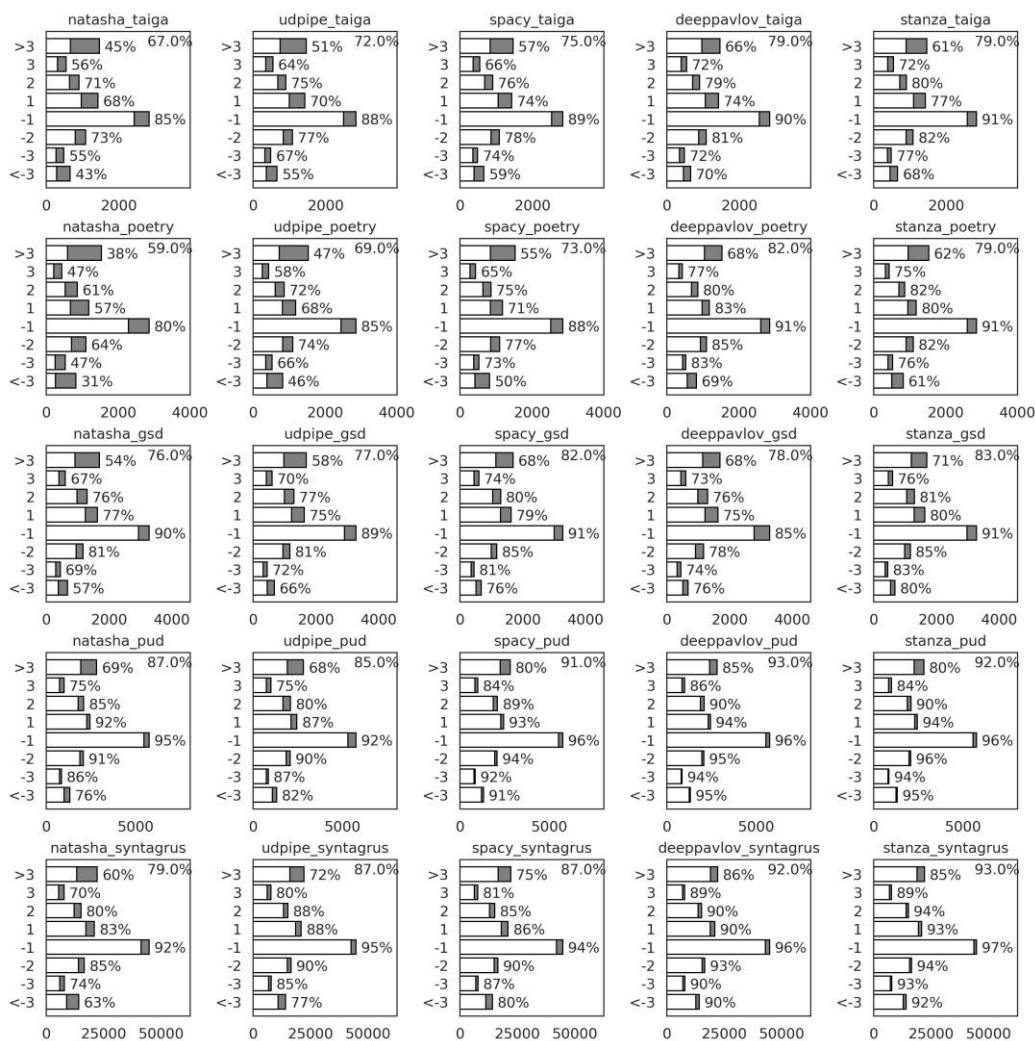


Рис. 10. Распределение доли верно разобранных токенов с определенным расстоянием до главного токена

### Приложение Е. Группировка по эталонному типу связи с главным

В табл. 8 представлена статистика по количеству эталонных токенов с верно определенным главным токеном (главным токеном и типом связи). Рассматриваются типы связи, которые входят в тройку наиболее частотных хотя бы в одном датасете. Обозначение *pnt* соответствует типу связи *punct*, обозначение *nm* – типу *nmod*.

Таблица 8. Доля корректных разобранных токенов с определенным типом связи

Датасет	Метрика	Главный токен						Главный токен + тип связи					
	Тип связи	Все	<i>pnt</i>	<i>case</i>	<i>conj</i>	<i>nm</i>	<i>obl</i>	Все	<i>pnt</i>	<i>case</i>	<i>conj</i>	<i>nm</i>	<i>obl</i>
Taiga	Natasha	67	60	92	46	73	69	62	60	91	43	66	61
	UDPipe	72	63	93	55	78	74	65	63	91	51	69	65
	spacy	75	69	93	61	78	78	69	67	93	55	74	72
	DeepPavlov	79	71	94	71	82	82	72	70	92	68	75	74
	Stanza	79	70	94	70	82	80	73	70	93	65	77	75
	Размер	10274	2099	887	625	613	564	10274	2099	887	625	613	564
Poetry	Natasha	59	55	86	31	58	55	53	55	84	27	49	46
	UDPipe	69	54	94	50	71	70	62	54	87	44	64	60
	spacy	73	68	92	59	71	71	67	68	90	54	65	64
	DeepPavlov	82	73	95	71	76	83	75	73	88	69	70	76
	Stanza	79	66	97	68	75	79	73	66	90	64	67	73
	Размер	10038	2130	841	850	396	648	10038	2130	841	850	396	648
GSD	Natasha	76	58	91	56	76	75	71	58	91	51	73	70
	UDPipe	77	58	93	59	78	77	69	58	92	55	74	69
	spacy	82	66	93	70	82	85	78	66	92	67	79	80
	DeepPavlov	78	59	87	75	82	79	70	59	85	74	77	69
	Stanza	83	63	94	75	85	84	77	63	93	74	81	78
	Размер	11385	2093	1262	531	1250	923	11385	2093	1262	531	1250	923
PUD	Natasha	87	85	96	72	84	78	82	85	95	70	78	74
	UDPipe	85	76	95	73	80	79	77	76	92	71	75	75
	spacy	91	89	96	84	87	87	87	89	95	83	81	83
	DeepPavlov	93	90	97	89	92	88	86	90	93	87	84	85
	Stanza	92	89	97	87	89	87	86	89	94	85	82	84
	Размер	19355	2977	2121	695	1934	1465	19355	2977	2121	695	1934	1465
SynTag Rus	Natasha	79	73	95	55	80	79	75	73	95	52	77	61
	UDPipe	87	81	98	70	85	86	82	81	97	68	81	80
	spacy	87	79	97	76	86	89	82	76	97	74	83	70
	DeepPavlov	92	85	98	87	91	93	88	85	98	86	88	89
	Stanza	93	90	99	87	91	92	91	90	98	86	89	88
	Размер	157718	29463	14943	76403	12179	132729	157718	29463	14943	76403	12179	132729

## Russian parser comparison

E. Shamaeva

Lomonosov Moscow State University

The article compares the quality of the Russian neural network syntax parsers UDPipe, Stanza, Natasha, DeepPavlov, spacy. The assessment was carried out dependency tree datasets GSD, PUD, SynTagRus, Poetry, Taiga from the Universal Dependencies project. The results obtained in the article can be used to select the parser that is most suitable for a task. The highest speed was demonstrated by Natasha, UDPipe and spacy, while the best quality was shown by DeepPavlov, Stanza and spacy. On most parsers and datasets, the sentences with a UAS of 1.0 does not exceed 40%, and with a LAS metric of 1.0 does not exceed 25%. In addition to the standard measuring the average values of UAS and LAS metrics on a test sentence set and on sets of sentences with a certain length, in the article the metric distributions are discussed. Moreover, such token sets, as the gold relation type, the depth in the gold dependency tree, the dependency length, are considered. The research implementation is published on the [https://github.com/Derinhelm/parser\\_stat](https://github.com/Derinhelm/parser_stat).

**Keywords:** syntax, parser, dependency tree, treebank, Universal Dependencies

**Reference for citation:** Shamaeva E. Russian parser comparison // Computational Linguistics and Computational Ontologies. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 26-47. DOI: 10.17586/3033-5582-2025-9-26-47.

## Reference

- [1] Lin L., Ziyang C., Shuxing L., Yibin D., Hongxiao W., Zhihao L. Event extraction in complex sentences based on dependency parsing and longformer // Proceedings of 2024 International Conference on Machine Learning and Intelligent Computing. Wuhan, China 2024. P. 1-7. URL: <https://proceedings.mlr.press/v245/lin24a.html> (accessed date: 29.03.2025).
- [2] Vasiliev S., Korobkin D., Fomenkov S. Extracting the component composition data of inventions from russian patents using dependency tree analysis // 2023 International Conference on Industrial Engineering, Applications and Manufacturing (ICIEAM). Sochi, Russian Federation, 2023. P. 1030-1034. URL: <https://ieeexplore.ieee.org/document/10139170> (accessed date: 29.03.2025).
- [3] Alonso M.A., Gómez-Rodríguez C., Vilares J. On the use of parsing for named entity recognition // Applied Sciences. 2021. No. 11(3). DOI: 10.3390/app11031090.
- [4] Nikolaev I. Knowledge and skills extraction from the job requirements texts // Ontology of Designing. 2023. No. 13(2). P. 282-293. URL: <https://journals.ssau.ru/ontology/article/view/27001> (accessed date: 29.03.2025).
- [5] Taufiq U., Pulungan R., Suyanto Y. Named entity recognition and dependency parsing for better concept extraction in summary obfuscation detection // Expert Systems with Applications. 2023. Vol. 217. Article 119579. URL: <https://www.sciencedirect.com/science/article/abs/pii/S0957417423000805> (accessed date: 29.03.2025).
- [6] Liu T., Sun Y., Wu J., Xu X., Han Y., Li C., Gong B. Unsupervised paraphrasing under syntax knowledge // Proceedings of the AAAI Conference on Artificial Intelligence. 2023. No. 37(11). P. 13273-13281. URL: <https://ojs.aaai.org/index.php/AAAI/article/view/26558> (accessed date: 29.03.2025).
- [7] Corbetta C., Passarotti M., Moretti G. The rise and fall of dependency parsing in dante alighieri's divine comedy // Proceedings of the Third Workshop on Language Technologies

- for Historical and Ancient Languages (LT4HALA). Torino, Italia, 2024. P. 50-56. URL: <https://aclanthology.org/2024.lt4hala-1.7/> (accessed date: 29.03.2025).
- [8] Altıntaş M., Cüneyd Tantuğ A. Improving the performance of graph based dependency parsing by guiding bi-affine layer with augmented global and local features // *Intelligent Systems with Applications*. 2023. No. 18. Article 200190. URL: <https://www.sciencedirect.com/science/article/pii/S2667305323000157> (accessed date: 29.03.2025).
- [9] Kulmizev A., Lhoneux M., Gontrum J., Fano E., Nivre J. Deep contextualized word embeddings in transition-based and graph-based dependency parsing - a tale of two parsers revisited // *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, 2019. P. 2755-2768. URL: <https://aclanthology.org/D19-1277/> (accessed date: 29.03.2025).
- [10] Mohammadshahi A., Henderson J. Graph-to-graph transformer for transition-based dependency parsing // *Findings of the Association for Computational Linguistics: EMNLP*. 2020. P. 3278-3289. URL: <https://aclanthology.org/2020.findings-emnlp.294/> (accessed date: 29.03.2025).
- [11] Zuhra F.T., Saleem K., Naz S. An accurate transformer-based model for transition-based dependency parsing of free word order languages // *Journal of King Saud University - Computer and Information Sciences*. 2024. No. 36(6). P. 102-107. URL: <https://www.sciencedirect.com/science/article/pii/S1319157824001964> (accessed date: 29.03.2025).
- [12] Marneffe M., Manning C.D., Nivre J., Zeman D. Universal dependencies // *Computational Linguistics*. 2021. No. 47(2). P. 255-308. URL: <https://aclanthology.org/2021.cl-2.11/> (accessed date: 29.03.2025).
- [13] Straka M., Strakova J. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe // *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Vancouver, Canada, 2017. P. 88-99. URL: <https://aclanthology.org/K17-3009/> (accessed date: 29.03.2025).
- [14] Qi. P., Zhang Y., Zhang Y., Bolton J., Manning C.D. Stanza: A python natural language processing toolkit for many human languages // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. 2020. P. 101-108. URL: <https://aclanthology.org/2020.acl-demos.14/> (accessed date: 29.03.2025).
- [15] Zeman D., Hajič J., Popel M., Potthast M., Straka M., Ginter F., Nivre J., Petrov S. Conll 2018 shared task: Multilingual parsing from raw text to universal dependencies // *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*. Brussels, Belgium, 2018. P. 1-21. URL: <https://aclanthology.org/K18-2001/> (accessed date: 29.03.2025).
- [16] Lyashevskaya O.N., Shavrina T.O., Trofimov I.V., Vlasova N.A. Grameval 2020: Russian full morphology and universal dependencies parsing // *Annual International Conference Dialogue*. 2020. P. 553-569. URL: <https://publications.hse.ru/pubs/share/direct/395318448.pdf> (accessed date: 29.03.2025).
- [17] Naeval. URL: <https://github.com/natasha/naeval> (дата обращения: 29.03.2025).
- [18] Droганова K., Lyashevskaya O., Zeman D. Data conversion and consistency of monolingual corpora: Russian ud treebanks // *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*. Linköping, Sweden, 2018. P. 53-66. URL: <https://ufal.mff.cuni.cz/biblio/attachments/2018-droganova-p7646871821297838363.pdf> (accessed date: 29.03.2025).
- [19] Dozat T., Manning C.D. Deep biaffine attention for neural dependency parsing // *International Conference on Learning Representations (ICLR)*. 2017. URL: <https://arxiv.org/abs/1611.01734> (accessed date: 29.03.2025).

- [20]Zhang M.A. survey of syntactic-semantic parsing based on constituent and dependency structures // Science China Technological Sciences. 2020. No. 63(10). P. 1898-1920. DOI: 10.1007/s11431-020-1666-4.

## Геометрия падежей в векторных моделях русского языка

К. М. Черников, И. А. Суров

Университет ИТМО

`kirill242006@yandex.ru, ilya.a.surov@itmo.ru`

### Аннотация

Падежные системы естественных языков выражают их смысловую структуру, информативную для задач лингвистического анализа. Без связи с архитектурами современных языковых моделей эта информация практически не используется, что приводит к смысловой неопределённости и другим трудоёмким проблемам машинного обучения. В данной работе такая связь установлена для векторных моделей GloVe и FastText, чувствительных к морфологии. Для этого рассмотрены примерно 3 тыс. наиболее используемых существительных русского языка, падежные формы которых кодируются моделями в виде 300-мерных векторов. Структура падежей в этом массиве данных изучалась методом линейного дискриминантного анализа. В 300-мерных пространствах обеих моделей найдено четырёхмерное подпространство, оси которого разделяют падежные классы с точностью от 75 % до 90 %. В этом подпространстве падежные словоформы образуют четырёхмерный тетраэдр, в котором падежам соответствуют вершины, центр и соединяющие их лучи. Разделительные функции осей и симметрии падежных распределений совпадают для обеих моделей. Найденная структура падежной семантики открывает возможности машинного анализа русскоязычных текстов, недоступные для английского и китайского. Полученный результат намечает пути природоподобного развития машинных моделей языков с сильной морфологией.

**Ключевые слова:** векторные модели, дистрибутивная семантика, семантическое пространство, падеж, геометрия, русский язык, морфология, дискриминантный анализ

**Библиографическая ссылка:** Черников К. М., Суров И. А. Геометрия падежей в векторных моделях русского языка // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 48–59. DOI: 10.17586/3033-5582-2025-9-48-59.

### 1. Введение

Естественные языки — особые форматы кодирования информации, в большинстве которых семантика и прагматика сообщения связаны с морфологией (формы слов) и синтаксисом (согласование слов в предложениях) сложным и неоднозначным образом. В современном машинном обучении правила такого кодирования практически не используются. Большие языковые модели (БЯМ) грамматикой не пользуются вовсе, рассматривая всевозможные формы слов как уникальные языковые единицы. Неудивительно, что ресурсоёмкость таких моделей [1, 2] ставит задачу разработки более экономных парадигм языкового моделирования.

В этой связи целесообразно исследовать возможности машинного кодирования смысло-грамматических закономерностей естественных языков. Важнейшая такая закономерность определяется падежной системой, являющейся грамматическим ядром языков флективной

группы (индоевропейские, семитские) [3, 4, 5]. При этом отмечается, что падеж — не поверхностный синтаксический феномен, а глубинная структура смысловой организации языка [6]. Гуманитарные исследования падежных систем обычно носят описательно-классификационный характер, не предполагающий алгоритмического воплощения.

В данной работе предпринят шаг по сокращению этого разрыва. Идея в том, что если падеж действительно несёт устойчивую смысловую функцию, то она должна быть отражена в структуре векторных языковых моделей, обладающих свойством алгебраического представления смысловых отношений [7, 8]. Такая связь позволила бы извлекать семантику и прагматику падежей из существующих языковых моделей, а также пользоваться ей для задач машинного анализа текстов.

## 2. Данные

В качестве источника падежных форм использовалась база данных [9], составленная на основе словаря М. Хагена [10]. В этой базе содержится 767 694 словоформы существительных русского языка с указанием падежа, числа (ед. / множ.), одушевлённости (одуш. / неодуш.) и рода (муж. / жен. / сред. / общ.). Из этого числа к именительному, винительному, родительному, дательному и предложному падежам отнесено примерно по 125 тыс. словоформ, к творительному примерно 140 тыс. К местному, звательному, частичному и счётному падежам отнесено в сумме около 350 слов, которые в последующем не использовались.

Эти словоформы переводились в численный вид посредством двух предобученных моделей русского языка, в которых слова представлены в виде 300-мерных вещественных векторов:

- модель семейства GloVe [11], тренированная на массиве русской художественной литературы [12];
- модель семейства FastText [13], тренированная на текстах Википедии и базы Common Crawl [14].

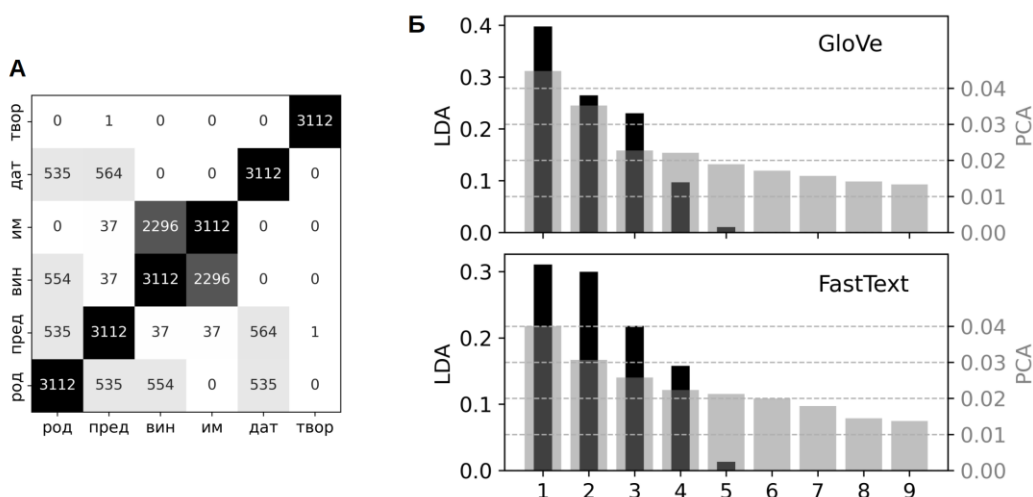
Принципы обучения этих моделей таковы, что точность векторного представления слова тем выше, чем чаще оно встречается в массиве тренировочных текстов. В этой связи экзотические слова и термины вроде «абазия», «аптеригот» и т. п., составляющие большую часть исходной базы [9], в представлении этих моделей дают малоинформативный, шумовой результат. Для минимизации возникающей отсюда ошибки из базы выбирались только словоформы 10 тыс. наиболее употребимых существительных русского языка в единственном числе [15].

Кроме того, оставались только существительные, для которых в исходной базе и языковой модели имеются все шесть основных падежей. Для моделей GloVe [12] и FastText [14] таких существительных оказалось 2 891 и 3 112, так что полное число словоформ составило  $2891 \times 6 = 17\,346$  и  $3112 \times 6 = 18\,672$  единиц соответственно. Как отмечено выше, каждая из этих словоформ кодируется 300-мерным вектором вещественных чисел.

## 3. Методы

Выявление предполагаемой структуры падежей проводилось с помощью снижения исходной размерности векторного представления словоформ с исходных 300 до малого количества измерений. Для этого использовался метод линейного дискриминантного анализа [16], который позволяет найти линейные комбинации 300 векторных компонент, которые наилучшим образом разделяют элементы выборки на падежные классы. Число таких комбинаций не превышает число классов минус один, поэтому для разделения шести падежей их может быть не более 5. Каждая такая комбинация определяет в 300-мерном

векторном пространстве линейное измерение — ось, в проекции на которую векторные представления падежей разделяются более или менее точно. Эта точность, то есть относительная разделяющая способность каждого найденного измерения, показана чёрным цветом на графиках рис. 1Б. Для обеих моделей видно, что информативность пятого измерения много меньше остальных. Это объясняется тем, что для большинства существительных формы именительного и винительного падежей совпадают, в результате чего два из шести классов оказываются во многом неразличимыми. Как видно из матрицы совпадений на графике рис. 1А, для остальных падежей совпадение форм отсутствует или значительно меньше.



**Рис. 1. А:** Матрица совпадений падежных словоформ в рассматриваемой выборке данных. **Б:** разделятельная информативность 5 осей в 300-мерных пространствах языковых моделей GloVe и FastText, найденных методом LDA (левые оси, чёрные столбцы). Серый: собственные значения для первых 9 главных компонент (PCA, правые оси, серые столбцы).

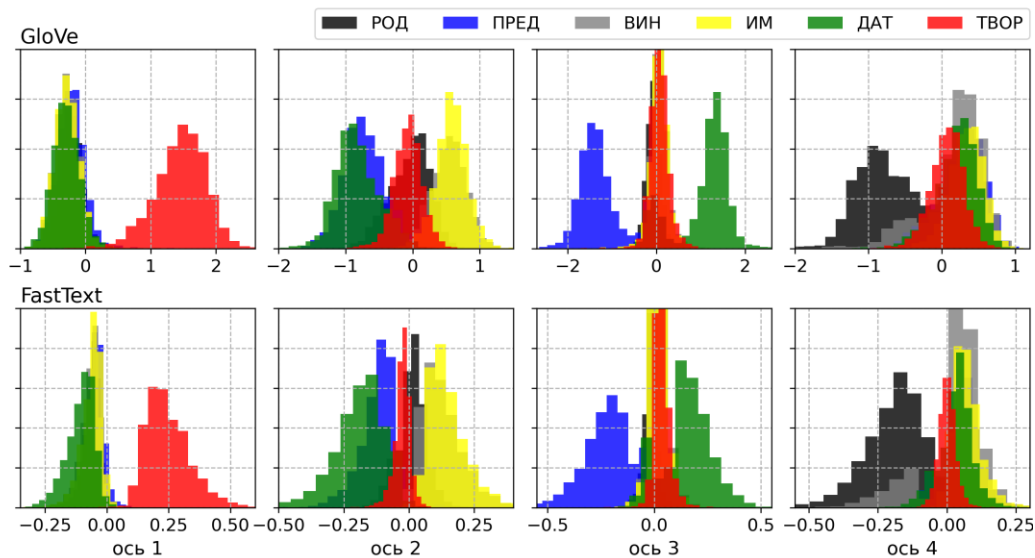
Абсолютная информативность найденных измерений определяется точностью классификации словоформ по падежам. Для моделей GloVe и FastText эта точность, вычисленная обычным методом перекрёстной проверки, составила 0.77 и 0.75 соответственно. При исключении из набора данных винительного падежа эти значения возрастают до 0.92 и 0.90. Эти величины свидетельствуют, что в 300-мерном пространстве векторных признаков существует малое число направлений, с высокой точностью разделяющих словоформы по их принадлежности к падежам.

Найденный набор осей много меньше числа главных компонент в массивах из 17 346 и 18 672 трёхсотмерных элементов, собственные значения первых, 9 из которых для сравнения показаны на графиках рис. 1Б серыми столбцами. Для обеих языковых моделей все 300 главных компонент являются информативными, тогда как для разделения падежей достаточно четырёх. Структура падежных словоформ выявляется в этом четырёхмерном пространстве путём проекции соответствующих векторных представлений на его базисные оси, найденные методом LDA.

## 4. Результаты

Проекция вектор-представлений на каждую из найденных осей порождает некоторое распределение словоформ каждого из падежей по этой оси. Гистограммы полученных распределений показаны на графике рис. 2, где моделям GloVe и FastText соответствует первый и второй ряд графиков. При этом использована следующая цветовая кодировка:

родительный — чёрный, предложный — синий, винительный — серый, именительный — жёлтый, дательный — зелёный, творительный — красный.



**Рис. 2.** Распределение падежных словоформ в проекциях на каждую из четырёх главных осей в порядке убывания информативности (рис. 1Б) для моделей GloVe и FastText

Четыре главные оси, пронумерованные по величине их собственных значений (рис. 1), в пространстве языковой модели GloVe (верхний ряд рис. 2) выполняют следующие функции:

- первая ось разделяет творительный падеж ( $X > 0$ ) и все остальные ( $X < 0$ );
- вторая ось отделяет предложный и дательный ( $Y < 0$ ) от родительного и творительного ( $Y = 0$ ) и винительного с именительным ( $Y > 0$ );
- третья ось отделяет предложный ( $Z < 0$ ) и дательный ( $Z > 0$ ) от всех остальных ( $Z = 0$ );
- четвёртая ось отделяет родительный ( $W < 0$ ) от всех остальных ( $W > 0$ );

То же самое верно в пространстве языковой модели FastText (нижний ряд рис. 2). При этом в силу подавляющего совпадения именительных и винительных словоформ (график рис. 1А), эти падежи практически неразличимы в обоих моделях.

Более полная картина взаимного расположения падежей получается путём проекции векторных представлений словоформ на плоскости, образуемые парами найденных осей. При этом исследуемая структура представляется в виде цветных карт как показано на рис. 3.

Три таких карты для пространства GloVe показаны в верхнем ряду рис. 3: каждому из падежей соответствует компактная область в четырёхмерном пространстве. При этом в плоскости осей 2 и 3 (средний график) именительный, предложный и дательный образуют треугольник, в центре которого расположен творительный (красный) и родительный (чёрный). В осях 1-3 (левый график) творительный располагается над плоскостью осей 2-3, в которой лежит пятёрка остальных падежей. В осях 1-4, напротив, из плоскости остальных падежей выделяется родительный. Здесь лучше других выделяется винительный падеж (серый), в остальных проекциях скрытый под именительным (жёлтым). Совместно, все 5 падежей образуют симметричную треугольную бипирамиду — тетраэдр в четырёхмерном пространстве.

Аналогичная геометрия имеет место в пространстве модели FastText. В этом пространстве, однако, место точек каждого из падежей представляется вытянутым намного сильнее; вместо «облаков» падежам соответствуют скорее направления — лучи векторного

пространства. В результате именительный (жёлтый), предложный (синий) и дательный (зелёный) на левом и правом графиках образуют не плоское основание тетраэдра, а тройку лучей к его вершинам — родительному и творительному.

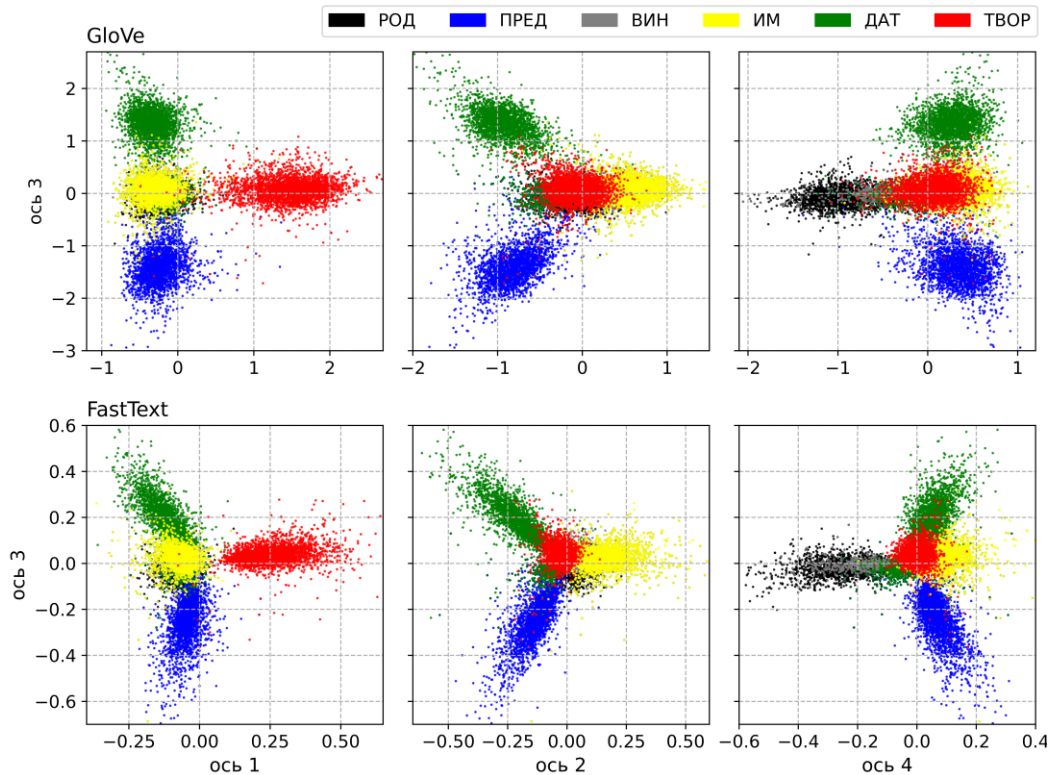


Рис. 3. Распределение падежных словоформ в проекциях на три пары из четырёх главных осей (рис. 2) для моделей GloVe и FastText

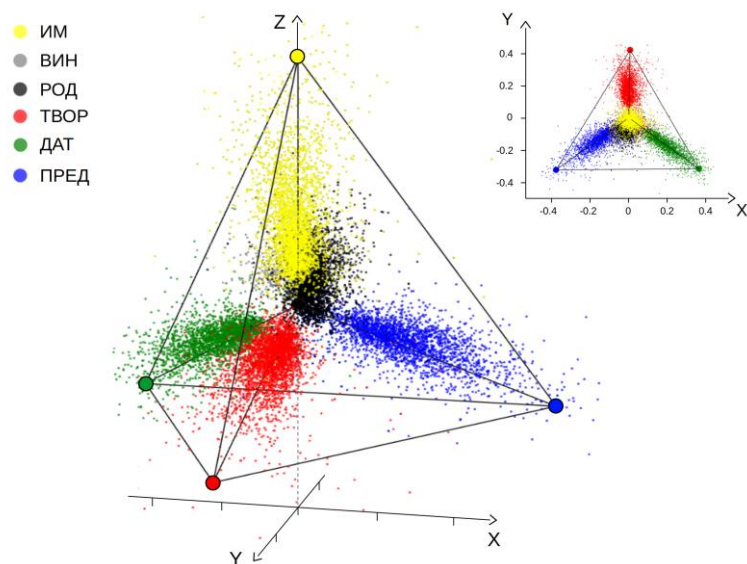


Рис. 4. Распределение падежных словоформ в трёхмерном подпространстве модели FastText

Пространственная карта падежных словоформ в осях 1, 2, 3 пространства FastText показана на графике рис. 4. Точками показаны центры масс словоформ именительного, предложного, творительного и дательного, вынесенные на тройное расстояние от начала координат для удобства наблюдения геометрии. Построенный на этих точках тетраэдр повернут так, что его вершиной является именительный падеж, в центре расположен родительный, а основание тетраэдра образовано предложным, творительным и дательным. Интерактивная версия графика доступна по адресу <https://github.com/newpotatato/Osgud>.

## 5. Обсуждение

Целевые функции использованных моделей (предсказание соседних слов текста в FastText и аппроксимация матрицы совпадений в GloVe) не используют для векторизации слов ни падежной, ни иной грамматической информации. Обе эти модели взяты в готовом виде из открытых источников; какой-либо последующей настройки или дообучения не проводилось. Совпадение геометрии падежей в различных моделях указывает на устойчивость и воспроизводимость наблюдаемой структурной организации. Отметим некоторые из её возможных применений. В области машинной обработки языка полученный результат подтверждает свойство векторных моделей к выявлению закономерностей, напрямую не связанных с их целевыми функциями [8, 17, 18]. В отличие от смысловых отношений типа «король-мужчина  $\approx$  королева-женщина» [19, 20], обнаруженная геометрия падежей специфична для русского языка. Эта специфика открывает возможности машинного анализа русскоязычных текстов, недоступные для английского, китайского и других языков со слабой морфологией.

Кодируя типы функциональных связей между понятиями, падежная система обозначает ситуативные роли описываемых сущностей, процессов и отношений между ними [3], классифицируя триплеты «субъект — отношение — объект» [21] по типам функциональных связей. В естественной речи эта информация разрешает полисемии, анафоры и другие неопределённости. В обычных методах машинного анализа, напротив, морфологическая информация сначала стирается лемматизацией текста, а затем восстанавливается по контексту сложными методами [22].

В отличие от двухпадежного английского и беспадёжного китайского, русскоязычный текст содержит морфологическую семантику в явном виде. В этой связи обнаруженная структура падежей востребована в задачах текстового поиска [23, 24] и извлечения информации из текстов [25, 26, 27, 28, 29], направленного sentiment-анализа [29], инженерии знаний, онтологий [30, 31, 32] и семантических сетей [33, 34, 35]. В этих целях возможно расширение разработанного подхода на склонения, спряжения, союзы, предлоги, другие «стоп-слова» [5, 36, 37] и грамматические конструкции.

Векторная геометрия падежей также открывает возможности совершенствования машинных моделей русского и других флективных языков. Каждая из двух рассмотренных моделей извлекла эту геометрию из обучающих текстов, не делая каких-либо предположений о структуре языка. Этот вычислительно-ёмкий процесс повторяется при обучении каждой новой модели, выявляющей неявные языковые закономерности с чистого листа. Эта работа отчасти обуславливает огромную энергоёмкость современных БЯМ, на порядки превышающую аналогичные затраты учащегося языку человека. В отличие от искусственных нейросетей, естественное мышление при этом учится не на пустом месте, а на основе врождённой когнитивной системы [38].

Если гипотеза о глубинной семантике падежей [3] верна, то обнаруженная геометрия может послужить машинным аналогом этой когнитивной предустановки. Аналогичные смысловые структуры обнаружены в арабском языке [39, 40]. Интерес представляет сопряжение этих результатов с инструментами грамматической и логической алгебры [41, 42], а также с процессной онтологией и семантическими факторами Ч. Огуда «оценка — сила — активность» в моделях семейства word2vec [43, 44].

Представленный результат показывает, что развитие машинной лингвистики в этом направлении возможно на основе существующих языковых моделей и методов машинного обучения.

## 6. Заключение

Полученные результаты подтвердили гипотезу о наличии падежной информации в машинных моделях русского языка. Структурное совпадение этой геометрии в машинных моделях разных семейств указывает на неслучайный характер явления. Осуществимость отмеченных практических применений зависит от его общности, судить о которой на основе двух рассмотренных моделей трудно. В этой связи целесообразен поиск падежных и других морфологических структур на большей выборке языков и языковых моделей. Наряду с линейным дискриминантным анализом для этого могут быть использованы другие методы снижения размерности.

Исследование выполнено при поддержке гранта Российского научного фонда № 23-71-01046 «Семантическое моделирование данных на основе комплекснозначных матричных разложений».

## Литература

- [1] Aljbour J., Wilson T., Patel P. Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption: EPRI White Paper no. 3002028905. Electric Power Research Institute, 2024. P. 35.
- [2] Chen S. How much energy will AI really consume? The good, the bad and the unknown // Nature. 2025. Vol. 639. No. 8053. P. 22-24. URL: <https://www.nature.com/articles/d41586-025-00616-z> (дата обращения: 19.03.2025).
- [3] Дрёмов А.Ф. Системная теория падежей и ее место в эволюции взглядов на падеж в лингвистике XX века // Русский язык: исторические судьбы и современность. Международный конгресс. М.: МГУ, 2001. С. 164-165.
- [4] Крючкова Л.С. Предложно-падежная система русского языка: взаимодействие формы, семантики, функции // Вестник РУДН, серия Русский и иностранные языки и методика их преподавания. 2014. № 1. С. 59-63.
- [5] Азарова И.В., Захаров В.П., Москвина А.Д. Семантическая структура русских предложно-падежных конструкций // Компьютерная лингвистика и вычислительные онтологии. 2018. Вып. 2. С. 9-16. DOI: 10.17586/2541-9781-2018-2-9-16
- [6] Мосина Н. М. История изучения падежной категории // Вестник Санкт-Петербургского Университета. Серия 9, 2009. № 3. С. 222-226.
- [7] Mikolov T. et al. Distributed representations of words and phrases and their compositionality // NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013.
- [8] Lenci A. Distributional Models of Word Meaning // Annu. Rev. Linguist. 2018. Vol. 4. No. 1. P. 151-171. DOI: 10.1146/annurev-linguistics-030514-125254
- [9] Korol Y. Russian morphology SQL dump. 2017. URL: <https://github.com/sshra/database-russian-morphology> (дата обращения: 19.03.2025).
- [10] Хаген М. Полная парадигма Русского языка. Морфология. Частотный словарь. 2014.
- [11] Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation // Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). 2014. P. 1532-1543.
- [12] Kukushkin A. navec\_hudlit\_v1\_12B\_500K\_300d\_100q.tar. URL: <https://github.com/natasha/navec> (дата обращения: 19.03.2025).
- [13] Bojanowski P. et al. Enriching Word Vectors with Subword Information. ArXiv, 2017. URL: <https://arxiv.org/abs/1607.04606> (дата обращения: 19.03.2025).

- [14]Meta AI. 2023. URL: <https://huggingface.co/facebook/fasttext-ru-vectors/tree/main/model.bin>; <https://fasttext.cc/blog/2016/08/18/blog-post.html> (дата обращения: 19.03.2025).
- [15]Hingston W. 10000-russian-words.txt. 2018. URL: <https://github.com/hingston/russian/> (дата обращения: 19.03.2025).
- [16]McLachlan G.J. Discriminant analysis and statistical pattern recognition. New York: Wiley, 2004. 526 p. URL: [https://scikit-learn.org/stable/modules/generated/sklearn.discriminant\\_analysis.LinearDiscriminantAnalysis.html](https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html) (дата обращения: 19.03.2025).
- [17]Erk K. Vector Space Models of Word Meaning and Phrase Meaning: A Survey // *Linguistics and Language Compass*. 2012. Vol. 6. No. 10. P. 635–653. DOI: 10.1002/Inco.362
- [18]Günther F., Rinaldi L., Marelli M. Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions // *Perspectives on Psychological Science*. 2019. Vol. 14. No. 6. P. 1006-1033. DOI: 10.1177/1745691619861372
- [19]McGregor S., Purver M., Wiggins G. Words, Concepts, and the Geometry of Analogy // *Electronic Proceedings in Theoretical Computer Science*. 2016. Vol. 221. P. 39-48.
- [20]Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations // *Proceedings of NAACL-HLT*. 2013. P. 746-751.
- [21]Nickel M. et al. A Review of Relational Machine Learning for Knowledge Graphs // *Proceedings of the IEEE*. 2016. No. 1 (104). P. 11-33. DOI: 10.1109/JPROC.2015.2483592
- [22]Yadav A., Patel A., Shah M. A comprehensive review on resolving ambiguities in natural language processing // *AI Open*. 2021. Vol. 2. P. 85-92.
- [23]Гаврилкина А. С., Максимов Н. В., Голицына О. Л. К идентификации ситуативных ролей сущностей в контексте задачи семантического информационного поиска // *Компьютерная лингвистика и вычислительные онтологии*. 2024. № 7. С. 21-31. DOI: 10.17586/2541-9781-2024-7-21-31
- [24]Максимов Н. В., Голицына О. Л., Монанков К. В. Модель механизма документального информационного поиска на графах знаний // *Компьютерная лингвистика и вычислительные онтологии*. 2021. № 5. С. 39-50. DOI: 10.17586/2541-9781-2021-5-39-50
- [25]Рубашкин В. Ш. Онтологическая семантика: Знания. Онтологии. Онтологически ориентированные методы информационного анализа текстов. М.: Физматлит, 2012. 346 с.
- [26]Артамонова Е. В., Лештаев С. В. Преобразование естественно-языковых текстов в rdf-граф // *Технические науки*, 2016. № 11. С. 214-218.
- [27]Сидорова Е.А. Подход к моделированию процесса извлечения информации из текста на основе онтологии // *Онтологии проектирования*. 2018. Том 8. № 1. С. 134-151. DOI: 10.18287/2223-9537-2018-8-1-134-151
- [28]Зулкарнеев Р. Х. [и др.]. Методы и модели извлечения знаний из медицинских документов // *Информатика и автоматизация*, 2022. № 6(21). С. 1169-1210. DOI: 10.15622/ia.21.6.4
- [29]Brauwers G., Frasincar F. A Survey on Aspect-Based Sentiment Classification // *ACM Computing Surveys*. 2021. Vol. 55. P. 1-37. DOI: 10.1145/3503044
- [30]Луценко Е. В. Инженерия знаний и интеллектуальные системы. Краснодар: ВЦСКИ «Эйдос», 2020. 642 с.
- [31]Дуга С.В., Труфанов А.И. Сеть знаний как концепция систем поддержки принятия решения в предварительном следствии // *Безопасность информационных технологий*. 2020. № 3(27). С. 54-65. DOI: 10.26583/bit.2020.3.05
- [32]Лебедев А.А. и др. Онто-графовые механизмы глубинного семантического поиска // *Научно-Техническая Информация Серия 2. Информационные процессы и системы*. 2022. № 7. С. 1-17. DOI: 10.36535/0548-0027-2022-07-1

- [33] Батура Т.В. Методы и системы семантического анализа текстов // Программные продукты и системы. 2016. № 12. С. 1-29.
- [34] Бесмертный И.А. Визуализация знаний на основе семантической сети // Программирование, 2010. № 4. С. 16-24.
- [35] Артюшина Л.А. Методы представления информации в простых семантических сетях // Научно-технический вестник информационных технологий, механики и оптики. 2020. № 3(20). С. 382–393. DOI: 10.17586/2226-1494-2020-20-3-382-393.
- [36] Захаров В.П., Михайлова В.Д. Контекстная грамматика предложных конструкций русского языка // Компьютерная лингвистика и вычислительные онтологии. 2018. № 1. С. 57-71. DOI: 10.17586/2541-9781-2017-1-57-71
- [37] Выборная В.В., Гончарова А.М., Родина А.А. Частотные характеристики предлогов и их значений в базе данных предложных конструкций // Компьютерная лингвистика и вычислительные онтологии, 2024. № 8. С. 61-69. DOI: 10.17586/2541-9781-2024-8-61-69
- [38] Пинкер С. Чистый лист: Природа человека. Кто и почему отказывается признавать ее сегодня. М: Альпина нон-фикшн, 2018. 608 с.
- [39] Adi T. et al. Muhkam Algorithmic Models of Real World Processes for Intelligent Technologies // International Journal of Robotics Applications and Technologies. 2014. Vol. 1. No. 2. P. 56-82. DOI: 10.4018/ijrat.2013070105
- [40] Adi T. A Framework of Cognition and Conceptual Structures Based on Deep Semantics // International Journal of Conceptual Structures and Smart Applications. 2016. Vol. 3. No. 1. P. 1-19. DOI: 10.4018/ijcssa.2015010101
- [41] Piedeleu R. et al. Open System Categorical Quantum Semantics in Natural Language Processing // Computing Research Repository. 2015. Vol. 1502.00831. URL: <https://arxiv.org/abs/1502.00831> (дата обращения: 19.03.2025).
- [42] Widdows D., Cohen T. Reasoning with vectors: A continuous model for fast robust inference // Logic Journal of IGPL. 2015. No. 2(23). P. 141-173. DOI: 10.1093/jigpal/jzu028
- [43] Surov I. Opening the Black Box: Finding Osgood's Semantic Factors in Word2vec Space // Informatics and Automation. 2022. Vol. 21. No. 5. P. 916-936. DOI: 10.15622/ia.21.5.3
- [44] Суров И.А. Процессная онтология и квантование информации // Знания-Онтологии-Теории. Новосибирск. 2023. С. 255–265.

### **Geometry of cases in vector models of Russian language**

K. M. Chernikov, I. A. Surov

ITMO University

Case systems of natural languages encode their semantic structure, central for natural language practice and informative for linguistic analysis. Without connection to the architectures of modern language models, however, this information is practically ignored, which results in semantic ambiguity and other computationally intensive problems of machine learning. This article establishes the lacking connection for the baseline machine models GloVe and FastText, which are sensitive to linguistic morphology. To this end we considered ~3000 of the most frequently used nouns of the Russian language, six major case forms of which are encoded in these models by 300-dimensional vectors. The structure of cases in this dataset was subjected to linear discriminant analysis. As a result, the 300-dimensional spaces of both models were found to contain a four-dimensional subspace, the axes of which discriminate case classes with an accuracy of 75 to 90 percent. In this subspace, case word forms constitute a four-dimensional tetrahedron, in which the cases correspond to the vertices, the center, and the rays between them. Functions of the axes and the discovered symmetries coincide for both models. The found structure of case semantics suggests ways for machine analysis of Russian-language texts, unavailable for English and

Chinese. The developed approach to vector representation of grammar opens up new possibilities for improving machine models of fusional languages.

**Keywords:** vector embedding, language models, distributional semantics, semantic space, geometry, Russian, case, morphology, discriminant analysis

**Reference for citation:** Chernikov K. M., Surov I. A. Geometry of cases in vector models of Russian language // *Computational Linguistics and Computational Ontologies*. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 48-59. DOI: 10.17586/3033-5582-2025-9-48-59.

## Reference

- [1] Aljbour J., Wilson T., Patel P. Powering Intelligence: Analyzing Artificial Intelligence and Data Center Energy Consumption: EPRI White Paper no. 3002028905. Electric Power Research Institute, 2024. P. 35.
- [2] Chen S. How much energy will AI really consume? The good, the bad and the unknown // *Nature*. 2025. Vol. 639. No. 8053. P. 22-24. URL: <https://www.nature.com/articles/d41586-025-00616-z> (accessed date: 19.03.2025).
- [3] Drjomov A.F. Sistemnaja teorija padezhej i ee mesto v jevoljucii vzgljadov na padezh v lingvistike XX veka // *Russkij jazyk: istoricheskie sud'by i sovremennost'*. Mezhdunarodnyj kongress. M.: MGU, 2001. P. 164-165. (In Russian)
- [4] Krjuchkova L.S. Predložno-padezhnaja sistema russkogo jazyka: vzaimodejstvie formy, semantiki, funkcii // *Vestnik RUDN, serija Russkij i inostrannye jazyki i metodika ih prepodavanija*. 2014. No. 1. P. 59-63. (In Russian)
- [5] Azarova I.V., Zaharov V.P., Moskvina A.D. Semanticheskaja struktura russkih predložno-padezhnyh konstrukcij // *Komp'juternaja lingvistika i vychislitel'nye ontologii*. 2018. Iss. 2. P. 9-16. DOI: 10.17586/2541-9781-2018-2-9-16 (In Russian)
- [6] Mosina N.M. Istorija izuchenija padezhnoj kategorii // *Vestnik Sankt-Peterburgskogo Universiteta*. Serija 9. 2009. No. 3. P. 222-226. (In Russian)
- [7] Mikolov T. et al. Distributed representations of words and phrases and their compositionality // *NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems*. 2013.
- [8] Lenci A. Distributional Models of Word Meaning // *Annu. Rev. Linguist.* 2018. Vol. 4. No. 1. P. 151-171. DOI: 10.1146/annurev-linguistics-030514-125254 (In Russian)
- [9] Korol Y. Russian morphology SQL dump. 2017. URL: <https://github.com/sshra/database-russian-morphology> (accessed date: 19.03.2025).
- [10] Hagen M. Polnaja paradigma Russkogo jazyka. Morfologija. Chastotnyj slovar'. 2014.
- [11] Pennington J., Socher R., Manning C.D. Glove: Global vectors for word representation // *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014. P. 1532-1543.
- [12] Kukushkin A. `navec_hudlit_v1_12B_500K_300d_100q.tar` URL: [https://github.com/natasha/navec\\_2023](https://github.com/natasha/navec_2023) (accessed date: 19.03.2025).
- [13] Bojanowski P. et al. Enriching Word Vectors with Subword Information. ArXiv. 2017. URL: <https://arxiv.org/abs/1607.04606> (accessed date: 19.03.2025).
- [14] AI at Meta, 2023. URL: <https://huggingface.co/facebook/fasttext-ru-vectors/tree/main/model.bin>; <https://fasttext.cc/blog/2016/08/18/blog-post.html> (accessed date: 19.03.2025).
- [15] Hingston W. 10000-russian-words.txt. 2018. URL: <https://github.com/hingston/russian/> (accessed date: 19.03.2025).

- [16]McLachlan G. J. Discriminant analysis and statistical pattern recognition. New York: Wiley, 2004. 526 p. URL: [https://scikit-learn.org/stable/modules/generated/sklearn.discriminant\\_analysis.LinearDiscriminantAnalysis.html](https://scikit-learn.org/stable/modules/generated/sklearn.discriminant_analysis.LinearDiscriminantAnalysis.html) (accessed date: 19.03.2025).
- [17]Erk K. Vector Space Models of Word Meaning and Phrase Meaning: A Survey // Linguistics and Language Compass, 2012. Vol. 6. No. 10. P. 635-653. DOI: 10.1002/Inco.362
- [18]Günther F., Rinaldi L., Marelli M. Vector-Space Models of Semantic Representation from a Cognitive Perspective: A Discussion of Common Misconceptions // Perspectives on Psychological Science, 2019. Vol. 14. No. 6. P. 1006-1033. DOI: 10.1177/1745691619861372
- [19]McGregor S., Purver M., Wiggins G. Words, Concepts, and the Geometry of Analogy // Electronic Proceedings in Theoretical Computer Science. 2016. Vol. 221. P. 39-48.
- [20]Mikolov T., Yih W., Zweig G. Linguistic Regularities in Continuous Space Word Representations // Proceedings of NAACL-HLT. 2013. P. 746-751.
- [21]Nickel M. [et al.]. A Review of Relational Machine Learning for Knowledge Graphs // Proceedings of the IEEE. 2016. No. 1(104). P. 11-33. DOI: 10.1109/JPROC.2015.2483592
- [22]Yadav A., Patel A., Shah M. A comprehensive review on resolving ambiguities in natural language processing // AI Open, 2021. Vol. 2. P. 85-92.
- [23]Gavrilkina A.S., Maksimov N.V., Golicyna O.L. K identifikacii situativnyh rolej sushhnostej v kontekste zadachi semanticheskogo informacionnogo poiska // Komp'juternaja lingvistika i vychislitel'nye ontologii, 2024. Vol. 1. No. 7. P. 21-31. DOI: 10.17586/2541-9781-2024-7-21-31 (In Russian)
- [24]Maksimov N.V., Golicyna O.L., Monankov K.V. Model' mehanizma dokumental'nogo informacionnogo poiska na grafah znaniy // Komp'juternaja lingvistika i vychislitel'nye ontologii. 2021. No. 5. P. 39-50. DOI: 0.17586/2541-9781-2021-5-39-50 (In Russian)
- [25]Rubashkin V.Sh. Ontologicheskaja semantika: Znaniya. Ontologii. Ontologicheski orientirovannye metody informacionnogo analiza tekstov. M.: Fizmatlit, 2012. 346 p. (In Russian)
- [26]Artamonova E.V., Leshtaev S.V. Preobrazovanie estestvenno-jazykovyh tekstov v rdf-graf // Tehnicheskie nauki. 2016. No. 11. P. 214-218. (In Russian)
- [27]Sidorova E.A. Podhod k modelirovaniju processa izvlechenija informacii iz teksta na osnove ontologii // Ontologii proektirovanija. 2018. Vol. 8. No. 1. P. 134-151. DOI: 10.18287/2223-9537-2018-8-1-134-151 (In Russian)
- [28]Zulkarneev R. H. [i dr.]. Metody i modeli izvlechenija znaniy iz medicinskih dokumentov // Informatika i avtomatizacija. 2022. No. 6(21). P. 1169-1210. DOI: 10.15622/ia.21.6.4 (In Russian)
- [29]Brauwers G., Frasincar F. A Survey on Aspect-Based Sentiment Classification // ACM Computing Surveys. 2021. Vol. 55. P. 1-37. DOI: 10.1145/3503044
- [30]Lucenko E. V. Inzhenerija znaniy i intellektual'nye sistemy. Krasnodar: VCSKI «Jejdos». 2020. 642 p. (In Russian)
- [31]Duga S.V., Trufanov A.I. Set' znaniy kak koncepcija sistem podderzhki prinjatija reshenija v predvaritel'nom sledstvii // Bezopasnost' informacionnyh tehnologij. 2020. No. 3(27). P. 54-65. DOI: 10.26583/bit.2020.3.05 (In Russian)
- [32]Lebedev A.A. et al. Onto-grafovyje mehanizmy glubinnogo semanticheskogo poiska // Nauchno-Tehnicheskaja Informacija Serija 2. Informacionnye Processy I Sistemy. 2022. No. 7. P. 1-17. DOI: 10.36535/0548-0027-2022-07-1 (In Russian)
- [33]Batura T.V. Metody i sistemy semanticheskogo analiza tekstov // Programmnye produkty i sistemy. 2016. P. 1-29. (In Russian)
- [34]Bessmertnyj I.A. Vizualizacija znaniy na osnove semanticheskoi seti // Programmirovanie, 2010. No. 4. P. 16-24. (In Russian)
- [35]Artjushina L.A. Metody predstavlenija informacii v prostyh semanticheskikh setjah // Nauchno-tehnicheskij vestnik informacionnyh tehnologij, mehaniki i optiki. 2020. No. 3(20). P. 382-393. DOI: 10.17586/2226-1494-2020-20-3-382-393 (In Russian)

- [36]Zaharov V.P., Mihajlova V.D. Kontekstnaja grammatika predlozhnyh konstrukcij russkogo jazyka // Komp'juternaja lingvistika i vychislitel'nye ontologii. 2018. No. 1. P. 57-71 DOI: 10.17586/2541-9781-2017-1-57-71 (In Russian)
- [37]Vybornaja V. V., Goncharova A. M., Rodina A. A. Chastotnye karakteristiki predlogov i ih znachenij v baze dannyh predlozhnyh konstrukcij // Komp'juternaja lingvistika i vychislitel'nye ontologii. 2024. No. 8. P. 61-69. DOI: 10.17586/2541-9781-2024-8-61-69 (In Russian)
- [38]Pinker S. Chistyj list: Priroda cheloveka. Kto i pochemu otkazyvaetsja priznavat' ee segodnja. M: Al'pina non-fikshn, 2018. 608 p. (In Russian)
- [39]Adi T. et al. Muhkam Algorithmic Models of Real World Processes for Intelligent Technologies // International Journal of Robotics Applications and Technologies. 2014. Vol. 1. No. 2. P. 56-82. DOI: 10.4018/ijrat.2013070105
- [40]Adi T. A Framework of Cognition and Conceptual Structures Based on Deep Semantics // International Journal of Conceptual Structures and Smart Applications, 2016. Vol. 3. No. 1. P. 1-19. DOI: 10.4018/ijcssa.2015010101
- [41]Piedeleu R. et al. Open System Categorical Quantum Semantics in Natural Language Processing // Computing Research Repository. 2015. Vol. 1502.00831. URL: <https://arxiv.org/abs/1502.00831> (accessed date: 19.03.2025).
- [42]Widdows D., Cohen T. Reasoning with vectors: A continuous model for fast robust inference // Logic Journal of IGPL. 2015. № 2(23). P. 141-173. DOI: 10.1093/jigpal/jzu028
- [43]Surov I. Opening the Black Box: Finding Osgood's Semantic Factors in Word2vec Space // Informatics and Automation. 2022. Vol. 21. No. 5. P. 916-936. DOI: 10.15622/ia.21.5.3
- [44]Surov I.A. Processnaja ontologija i kvantovanie informacii // Znanija-Ontologii-Teorii, Novosibirsk, 2023. P. 255-265.

## Оптимизация обработки терминологии в беспилотной авиации: новый подход к извлечению терминов с использованием промпт-инжиниринга

Е. В. Исаева<sup>1</sup>, Б. З. Сафарбеков<sup>2</sup>

<sup>1</sup>Пермский государственный национальный исследовательский университет

<sup>2</sup>Национальный исследовательский технологический университет «МИСИС»

ekaterinaisae@gmail.com, behruzsafarbekov3@gmail.com

### Аннотация

Современные методы автоматизированной обработки текстов играют ключевую роль в анализе и систематизации специализированной терминологии. В области беспилотных авиационных систем (БАС) точное извлечение и классификация терминов крайне важны для разработки стандартов, обмена знаниями и совершенствования технологий искусственного интеллекта в аэронавтике. Однако, несмотря на растущий интерес к обработке естественного языка (Natural Language Processing — NLP), автоматизированные методы обработки терминологии в БАС остаются недостаточно разработанными. Большинство подходов либо сосредоточены на статистическом, контекстно-независимом выделении терминов, либо требуют предварительно размеченных корпусных данных, что ограничивает их применимость в динамично развивающихся технических областях. В данной статье предлагается новый метод автоматического извлечения и классификации терминологии БАС, основанный на использовании больших языковых моделей и промпт-инжиниринга. Метод включает в себя многоэтапную обработку документов, включающую предобработку текста, анализ терминов с помощью NLP-моделей и сохранение результатов в базе данных. Подход обеспечивает работу с многословными терминами без предварительной ручной разметки. Метод демонстрирует более высокую точность выделения терминов по сравнению с традиционными методами. Он позволяет выявить ключевые термины БАС и классифицировать их по функциональности в технической документации. Подход вносит вклад в область автоматизированной терминологической обработки, открывая новые возможности для стандартизации данных в области БАС, интеграции с онтологиями и создания интеллектуальных систем управления терминологией.

**Ключевые слова:** автоматизированное распознавание терминов, автоматическая идентификация терминов, промпт-инжиниринг, беспилотные авиационные системы, обработка естественного языка, искусственный интеллект, терминологическая обработка, большие языковые модели, машинное обучение

**Библиографическая ссылка:** Исаева Е. В., Сафарбеков Б. З. Оптимизация обработки терминологии в беспилотной авиации: новый подход к извлечению терминов с использованием промпт-инжиниринга // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 60–77. DOI: 10.17586/3033-5582-2025-9-60-77.

## 1. Введение

В последние годы наблюдается стремительный рост объема технической документации в высокотехнологичных отраслях, особенно в области беспилотных авиационных систем (БАС). Эти документы насыщены специализированной терминологией, требующей точного выделения, классификации и обработки для последующего использования в системах стандартизации, управления знаниями, перевода и автоматизированного анализа. В связи с растущим интересом к интеграции искусственного интеллекта в инженерные процессы особую актуальность приобретают инструменты, способные автоматически извлекать и структурировать специфические для данной области термины из разнородных источников.

Цель данного исследования — разработать и протестировать программный комплекс DroneTerms AI, предназначенный для автоматического распознавания, извлечения и классификации терминов, встречающихся в технических текстах, связанных с БАС. Проект опирается на возможности больших языковых моделей (Large Language Models — LLM), промпт-инжиниринга и модульной архитектуры для эффективной обработки неструктурированных документов .pdf и .doc/.docx.

Предложенная система актуальна как для исследовательской, так и для прикладной деятельности. Она позволяет значительно снизить трудозатраты на терминологическую обработку, повысить точность и согласованность терминов, а также создать основу для дальнейшей интеграции с онтологиями и интеллектуальными системами управления знаниями в области беспилотных технологий.

## 2. Обзор литературы

Проблема автоматического распознавания терминов (Automatic Text Recognition — ATR) и создания специализированных словарей привлекает значительное внимание с 1970-х годов. Подходы к ATR можно разделить на статистические, лингвистические и гибридные методы.

В ранних работах основное внимание уделялось статистическому подходу к автоматической индексации терминов путем присвоения весов, основанных на характеристиках отдельных массивов документов [1]. Этот подход до сих пор используется для решения смежной задачи извлечения ключевых слов. При данном подходе вычисляется частотность слов после фильтрации стоп-слов, и нормализуется путем увеличения весов слов, которые часто встречаются в данном документе, но не встречаются в других документах, например, в Википедии. В настоящее время для ATR и извлечения ключевых слов используются следующие статистические и вероятностные алгоритмы: частота термина (Term Frequency — TF), т.е. количество раз, когда термин встречается в документе или на веб-странице, обратная частота встречаемости в документе (Inverse document frequency — IDF), т.е. количество документов, содержащих данное слово, относительно всех документов, TF-IDF, т.е. произведение указанных выше показателей ( $TF \cdot IDF$ ), и первое вхождение слова, т.е. место первого появления данного слова в тексте или на веб-странице [1].

В конце 1980-х годов подходы сместились от чисто статистических методов к лингвистическим, например, морфологическому анализу. Авторы выделили такие особенности терминов, как монореферентность, т.е. обозначение термином конкретного понятия в определенной предметной области, междометная полисемия и омонимия, а также использовали структуру терминологических систем, в которых термины связаны родовидовыми отношениями, отношениями «часть-целое» и причинно-следственными связями для настройки алгоритмов ATR [2]. Лингвистические методы также включают составление

списка слов без синонимов в Wordnet, определение семантических значений, синтаксических позиций слов, разметку частей речи (POS), что позволяет выявить специфику ключевых слов и терминов [1]. В работе [3] описан удобный и эффективный метод машинного обучения без учителя, использующий шаблоны предложений и частеречные паттерны для извлечения терминологии из научного текста. Авторы начинают с нескольких начальных обучаемых шаблонов для определения терминологических лексем и их частеречных паттернов, затем строят новые шаблоны, которые могут соответствовать большему количеству предложений, чтобы найти больше паттернов, включающих термины, и, наконец, используют полученные частеречные паттерны и шаблоны предложений для извлечения терминологических терминов из нового научного текста.

Для извлечения многословных терминов независимо от предметной области был предложен метод ATR, сочетающий статистический анализ и лингвистическую фильтрацию. Метод включает два основных компонента: C-value, который повышает точность извлечения вложенных терминов за счет учета их частоты и степени включенности в более длинные конструкции, и NC-value, который дополнительно учитывает связанные с терминами слова (существительные, прилагательные, глаголы), сопровождающие их в тексте. Авторы продемонстрировали, что их подход повышает точность по сравнению с обычным частотным анализом, особенно для вложенных и многословных терминов [5].

В статье [1] представлен метод ACI-rank, общий алгоритм ранжирования терминов, сочетающий частотный, лингвистический и структурный анализ для автоматического извлечения ключевых слов. ACI-rank учитывает не только частоту встречаемости термина и его позицию в документе, но и информационную насыщенность контекста, в котором он встречается. Алгоритм был протестирован на новостных статьях и научных публикациях и показал конкурентоспособную точность по сравнению с другими методами извлечения [2]. Тем не менее, хотя метод и эффективен в извлечении ключевых фраз, он не ориентирован на технические многословные термины и не использует современные модели глубокого обучения, что ограничивает его применение в узкоспециализированных областях, таких как БАС.

Интеграция подходов, опирающихся на корпусные данные и NLP, показала перспективность в извлечении терминологии и информации, специфичной для конкретной предметной области, что было продемонстрировано в исследовании, проведенном на базе корпуса военных данных США [4]. Процедура включала следующие этапы:

- подготовка корпуса (US Army Field Manual 8-10-6) в текстовых форматах ANSI и UTF-8;
- создание списка ключевых слов с помощью программы AntConc 3.5.8;
- создание списка наиболее родственных терминологий с использованием гибридного подхода, предложенного А. Тонгпун-Патанасорн;
- проверка наличия терминологических словосочетаний путем анализа кластеров ключевых слов;
- создание текстовой базы данных путем синтаксического анализа корпуса с помощью системы обработки естественного языка CRS;
- извлечение знаний о предметной области из текстовой базы данных с использованием выявленных терминов [4].

Метод, предложенный в [4], имеет ряд недостатков, таких как неспособность дифференцировать стоп-слова, общеязыковые и специализированные слова, а также низкая эффективность при работе со сложными терминами. Данный подход, основанный на правилах, не позволяет выявить неявные семантические отношения между терминами, что может снизить качество извлекаемой информации, и не может быть легко масштабирован на другие языки, требуя ручного обновления правил. Кроме того, несмотря

на автоматизацию ряда процессов, для проверки результатов по-прежнему требуются терминологи и эксперты.

Пошаговый алгоритм выявления многокомпонентных терминов в русских научно-технических текстах предложен Ю. И. Бутенко.

Метод успешно преодолевает отмеченные выше ограничения. Алгоритм работает следующим образом:

1. Выявление грамматических характеристик каждого слова в тексте с помощью морфолого-синтаксического анализа;
2. Исключение частей речи, которые не могут входить в состав русских многокомпонентных терминов;
3. Исключение стоп-слов, образующих свободные словосочетания с терминами;
4. Сопоставление оставшихся цепочек слов с типовыми терминологическими словосочетаниями в базе данных;
5. Проверка терминов по терминологическому словарю для выявления терминов-кандидатов;
6. Разрешение неоднозначных ситуаций экспертами-лингвистами [5].

В своих более поздних исследованиях, чтобы исключить ручную лингвистическую фильтрацию на заключительном этапе ATR, автор заменяет шаг 6 на следующий:

6. Окончательный отбор терминов с учетом грамматических зависимостей (левое определение наследует грамматические признаки ядерного элемента; правое определение остается неизменным) [6].

Описанный выше метод улучшает обработку многокомпонентных терминов, но ограничен в работе с различными предметными областями, допускает ошибки при морфологическом анализе и по-прежнему требует экспертной проверки.

Гибридные методы сочетают в себе вышеупомянутые методики, включающие машинное обучение, нейронные сети и модели на основе трансформаторов. Традиционные языковые модели, такие как BERT, обученные предсказывать скрытые слова в тексте, относительно недавно появились в ATR и нейромашинном переводе (Neural Machine Translation — NMT) [7]. Интересным решением является дообучение нейросети-трансформера для выполнения NMT, как это реализовано в модели BART (Bidirectional and Autoregressive Transformer) [8]. Этот подход сочетает двунаправленный кодер, подобный BERT, и авторегрессивный декодер, как в GPT [9]. Таким образом, модель обучается созданию контекстуальных языковых репрезентаций и становится способной к генерации и переводу текста. В работе [10] архитектура BART была адаптирована для обработки больших монологических корпусов на 25 языках, в результате чего была создана мультязычная версия mBART, которая может быть дообучена для решения задач машинного перевода (MT).

Последние достижения в области ATR включают модели на основе трансформеров, работающие на уровне предложений, демонстрирующие лучшие результаты по сравнению с предыдущими базовыми моделями, особенно в многоязычной идентификации терминов из разных предметных областей [11]. Эти модели включают классификацию лексем и последовательностей. В работе [11] авторы предлагают новый метод идентификации и категоризации терминов, основанный на моделях-трансформерах на уровне предложения. Они сравнивают три модели.

Модель 1: токен-классификатор на основе языковой модели XLM-RoBERTa, определяющий, является ли каждое слово в предложении частью термина.

Модель 2: классификатор последовательностей на основе языковой модели XLM-RoBERTa, который анализирует n-граммы (до 6 слов) и классифицирует их как термины или общеупотребимые слова.

Модель 3: NMT модель на основе предобученной модели mBART, которая учится преобразовывать входные предложения в последовательности терминов, используя запятые в качестве разделителей.

Авторы отмечают, что и первая, и вторая модели одинаково эффективны и точны, при этом вторая требует больше вычислительных ресурсов. Результаты показывают, что трансформерные модели, особенно языковая модель на основе классификатора лексем, более эффективны, чем ранее использовавшиеся базовые модели на основе BERT для автоматического извлечения терминов. Однако третья модель превосходит эти подходы, обеспечивая точность F1 достигающую 69,8 % (по сравнению с 48,1 % для метода на основе BERT). Авторы делают вывод, что модель NMT и классификатор лексем обладают потенциалом для работы с неконтиуальными терминами (многословными терминами, последовательность элементов которых может нарушаться вкраплением других слов), которых нет в существующих наборах данных. Они также демонстрируют высокую устойчивость к переходу между языками и предметными областями, т.е. модели, обученные на одном языке, могут хорошо работать на другом языке, что позволяет извлекать термины без сложной предварительной обработки текста [11].

Однако, несмотря на то, что трансферное обучение показало свою перспективность для ATR, особенно для языков с низким уровнем ресурсов, учитывая достигнутую на данный момент точность F1, их производительность все еще не является превосходной, требует вмешательства эксперта и совместных усилий по созданию аннотированных наборов данных и общих терминологических баз данных [12].

### 3. Методы, использованные при разработке системы ATR для БАС

#### 3.1. Архитектура системы

В число требований к разрабатываемой системе входили извлечение и очистка текста из документов *.pdf* и *.doc/.docx*, автоматическая обработка для идентификации и классификации технических терминов БАС, экспертная постобработка и сохранение результатов в структурированном виде с поддержкой различных форматов выходных данных. Учитывая специфику предметной области, разрабатываемая система должна обеспечивать конфиденциальность обрабатываемой информации, особенно при взаимодействии с внешними API искусственного интеллекта. Таким образом, требуется реализация механизмов безопасного хранения ключей доступа и защиты передаваемых данных. С точки зрения масштабируемости и сопровождения, архитектура разрабатываемой программной системы должна обеспечивать возможность легкого расширения функциональности и интеграции новых компонентов. В связи с этим требованием система была спроектирована как модульная структура, позволяющая независимо модифицировать и обновлять отдельные компоненты системы, не влияя на работу других модулей.

Система имеет трехслойную архитектуру.

1. Презентационный слой. Этот уровень реализован в виде консольного интерфейса, который предоставляет пользователю возможность управлять обработкой документов и получать информацию о результатах работы системы. Данный уровень представлен модулями *main.py* и *interface.py*, которые обеспечивают взаимодействие с пользователем и координируют работу базовых компонентов системы.
2. Слой бизнес-логики. Этот уровень обеспечивает базовые алгоритмы обработки текста, а именно: извлечение, очистку текста, извлечение терминов из очищенного текста с помощью нейросетевой модели, обработку ответа модели и сохранение терминов из обработанного ответа в базе данных глоссария.

3. Слой доступа к данным. Этот слой обеспечивает взаимодействие с файловой системой. В нем определяется логика извлечения текстового содержимого из документов различных форматов и сохранения обработанной информации. В состав данного слоя входят модули: *text\_extractor.py* и *data\_saver.py*, которые взаимодействуют с хранилищами данных, включая директории: *data/*, *results/* и файл базы данных — *db/db\_terms.json*.

Взаимодействие между представленными слоями реализовано через отдельные интерфейсы, что обеспечивает гибкость при масштабировании. Общая архитектура системы, развернутой в контейнере Docker, представлена на рис. 1.

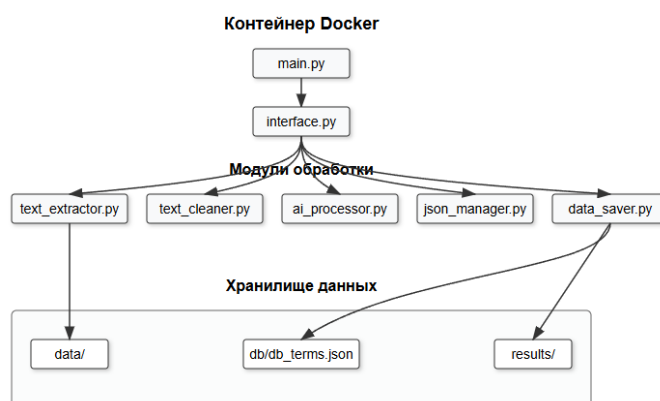


Рис. 1. Общая архитектура системы в контейнере Docker

Модульная структура позволяет легко адаптировать систему к новым требованиям без необходимости существенного пересмотра действующего кода.

### 3.2. Структура базы данных

В рамках разработки системы была спроектирована и реализована легковесная база данных на основе формата JSON. Выбор JSON в качестве формата хранения данных обусловлен его читабельностью и простотой использования, не требующих профессиональных навыков развертывания отдельной СУБД. Дополнительным преимуществом такого подхода является возможность версионного контроля данных, записанных в JSON с помощью Git, что особенно важно при коллективной работе над терминологической базой данных.

База данных реализована в виде структурированного JSON-документа, где каждый термин представляет собой отдельную запись с набором атрибутов. Основными характеристиками термина являются его название, определение, английский эквивалент, оценка релевантности и временная метка последнего обновления. Такая структура обеспечивает оптимальный баланс между простотой доступа к данным и информативностью хранимой информации.

Программное обеспечение автоматически создает необходимые каталоги и файлы при первом запуске, что значительно упрощает процесс первоначального развертывания и настройки, не требуя «ручного» внешнего вмешательства.

Важным аспектом реализации является обеспечение целостности данных. Система автоматически проверяет целостность структуры JSON при каждой операции и обеспечивает корректную обработку исключительных ситуаций при операциях ввода-вывода. Структура базы данных, выбранная при проектировании, показана на рис. 2.



Рис. 2. Структура базы данных

Управление базой данных и выполнения основных операций (добавления новых терминов, получения информации о существующих записях, проверки наличия терминов в базе данных и обновления базы данных) осуществляется через интерфейс JsonTermManager.

### 3.3. Пользовательский интерфейс

В текущей версии системы спроектирован консольный пользовательский интерфейс, который в последующих версиях планируется реализовать в виде веб-интерфейса. На рис. 3 показана схема всех возможных переходов между экранами пользовательского интерфейса.



Рис. 3. Логика пользовательского интерфейса

Пользовательский интерфейс также обеспечивает визуализацию результатов обработки текста. Найденные термины отображаются в структурированном виде с их определениями, переводами и оценкой релевантности для предметной области БАС. При сохранении результатов пользователь получает информацию о созданных файлах и количестве сохраненных и измененных терминов.

### 3.4. Алгоритмы обработки текстовых данных

Основной процесс обработки данных разделен на четыре ключевых алгоритмических блока: извлечение текста, очистка данных, извлечение терминов с помощью ИИ и сохранение результатов.

На первом этапе происходит извлечение текстовой информации из документов различных форматов. Система поддерживает работу с файлами: *.pdf* и *.doc/.docx*. В процессе извлечения учитывается структура документа и обеспечивается правильное сохранение последовательности текстов, что очень важно для дальнейшего анализа.

Как известно, предварительная обработка текста, такая как удаление стоп-слов, цифр и знаков препинания, а также стемминг или лемматизация, существенно влияют на качество результатов ATR [2], второй этап обработки данных в нашей программе включает

алгоритмы очистки и нормализации текста, такие как удаление специальных символов и лишних пробелов, нормализация переносов строк и абзацев, унификация кодировки текста, исправление типографских артефактов, удаление технических метаданных.

Алгоритм очистки реализован в виде последовательности регулярных выражений и преобразований текста, что обеспечивает высокую производительность и надежность обработки. После очистки текста следует этап обработки данных с помощью нейросетевых моделей для извлечения релевантных терминов из текста, их перевода и подбора определений (толкований) идентифицированных терминов. Блок-схема алгоритма обработки данных представлена на рис. 4.



Рис. 4. Логика алгоритма обработки данных

Третий этап — взаимодействие с моделями искусственного интеллекта для ATR. Алгоритм использует технологию промпт-инжиниринга для создания эффективных запросов к большим языковым моделям.

Процесс анализа включает в себя следующее:

- разбиение текста на фрагменты оптимального размера;
- генерация контекстно-зависимых запросов;
- обработка ответов модели и извлечение структурированной информации;
- оценка релевантности найденных терминов;
- объединение результатов, полученных из различных фрагментов текста.

Особое внимание уделяется обработке ответов нейросетевых моделей и их преобразованию в структурированный формат. Система анализирует полученные данные, выделяя термины, их определения и переводы, а также рассчитывает индекс релевантности для каждого термина, исходя из контекста его использования и частоты встречаемости в тексте.

Заключительный этап включает в себя алгоритмы сохранения обработанных данных.

В системе реализовано сохранение результатов в нескольких форматах:

- запись в CSV-файл для обеспечения универсальной совместимости;
- создание XLSX-файла с форматированием для удобства лингвистической работы;
- обновление базы данных JSON на основе релевантности терминов.

При сохранении данных применяются алгоритмы проверки на дублирование терминов. Термины с релевантностью выше 80 % автоматически добавляются в основную базу данных, при этом проверяется наличие похожих записей и при необходимости обновляются существующие определения.

### 3.5. Модуль для извлечения текста из файлов (*text\_extractor.py*)

Одна из задач модуля — извлечение текстовой информации из файлов различных форматов, таких как *.pdf* и Microsoft Word. Класс *TextExtractor* позволяет последовательно обработать входной файл и решить эту задачу (рис. 5).



Рис. 5. Обработка документов *.pdf* и *.doc/.docx*

Для работы с указанными форматами файлов используются специализированные библиотеки, такие как *PyPDF2* и *python-docx*. Конструктор класса `__init__` принимает в качестве параметра объект класса *TextCleaner*, что обеспечивает последовательную обработку данных. Следующий метод, *extract\_raw\_text*, отвечает за извлечение текста из файла. Он определяет тип файла по его расширению и вызывает соответствующую приватную функцию обработки (*\_extract\_from\_pdf* для файлов *.pdf* или *\_extract\_from\_word* для документов *.doc/.docx*). Эти методы считывают данные постранично и объединяют их в одну строку. Если поступает файл иного формата, пользователю выдается сообщение об ошибке.

В будущем возможно расширение функциональности модуля, например, поддержка дополнительных форматов файлов (например, TXT или парсинг какого-либо сайта БАС для поиска терминов с этого сайта) или использование более продвинутых библиотек для работы с документами PDF и Word.

### 3.6. Модуль очистки текста (*text\_cleaner.py*)

Модуль *text\_cleaner.py* отвечает за очистку извлеченных на предыдущем этапе текстов. Модуль получает извлеченный текст и удаляет из него нерелевантную информацию, такую как гиперссылки, рисунки, имена авторов, специальные символы и другие элементы, которые могут помешать дальнейшему анализу текста, классификации лексических единиц и извлечению знаний.

Основной класс, *TextCleaner*, представляет собой совокупность базовых методов очистки и более сложных алгоритмы для последовательной обработки текста. Реализация основана на использовании регулярных выражений из библиотеки *re*, что позволяет эффективно обрабатывать текстовые данные (рис. 6).



Рис. 6. Логика работы модуля *TextCleaner*

Базовыми компонентами класса *TextCleaner* являются атрибуты класса *self.patterns*, которые составляют словарь регулярных выражений, используемых для очистки текста. Каждое регулярное выражение отвечает за определенный тип данных, подлежащих удалению или преобразованию.

Следует отметить, основными источниками терминологических глоссариев являются тексты ГОСТов, научных статей и нормативных документов, которые, как правило, содержат библиографические разделы. В связи с этим было принято решение разработать функцию для удаления раздела «Литература» или «Библиография». Для этого используется регулярное выражение для поиска ключевых слов, ассоциируемых с данными разделами.

Аналогичным образом осуществляется поиск и удаление рисунков и других графических элементов, имен авторов и др.

Метод *clean\_text* является основным методом комплексной очистки текста. Он последовательно выполняет все шаги очистки: удаляет раздел литературы, ссылки на рисунки, имена авторов, тире, подчеркивания и специальные символы, заменяет множественные знаки препинания и пробелы, удаляет пустые строки и пробелы в начале и конце текста. В конце метод возвращает очищенный текст.

### 3.7. Модуль взаимодействия с нейросетевыми моделями (*ai\_processor.py*)

Модуль *ai\_processor.py* отвечает за взаимодействие с внешними большими языковыми моделями через *OpenAI API* с использованием прокси *tonica.im*. Модуль реализован в виде класса *AIProcessor* для интеграции системы обработки текста с внешними сервисами искусственного интеллекта. *AIProcessor* обеспечивает выбор нейросетевой модели, отправку текста на обработку и получение результатов. Основными функциями модуля являются: загрузка конфигурации API из переменных окружения, выбор модели для обработки текста, отправка текста на анализ и извлечение терминов, обработка ошибок, возникающих при взаимодействии с внешними сервисами. Модуль поддерживает работу с несколькими моделями искусственного интеллекта, предоставляя пользователю возможность выбрать подходящую модель (рис. 7).

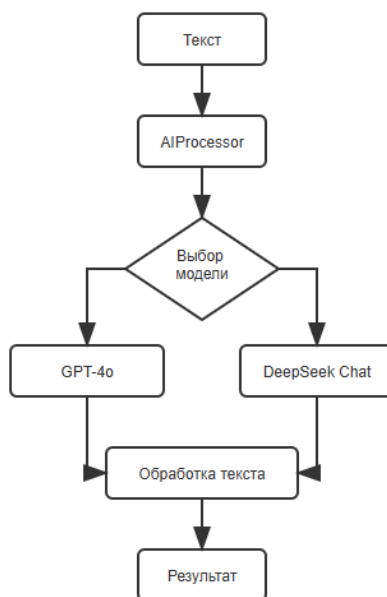


Рис. 7. Работа модуля взаимодействия с нейросетевыми моделями

На рис. 8 показан API-запрос к *OpenAI*, а именно к модели *gpt-4o*. Этот класс состоит из следующих атрибутов класса: *self.client* (экземпляра клиента *OpenAI*, инициализированного с помощью базового URL и ключа API, загруженного из переменных окружения) и *Self.available\_models* (словарь доступных моделей). Текущая реализация поддерживает две модели: *gpt-4o*, последнюю новейшую модель в семействе GPT и *Deepseek*. Переменная *self.model\_name* используется для хранения имени выбранной модели ИИ.

```

completion = self.client.chat.completions.create(
    model=self.model_name,
    messages=[
        {
            "role": "user",
            "content": [
                {
                    "type": "text",
                    "text": f"""Извлеки из текста термины, связанные с
                    беспилотными авиационными системами (БАС) и
                    беспилотными летательными аппаратами (БПЛА).
                    Верни результат строго в следующем формате:

                    Термин: [термин]
                    Определение: [определение]
                    Перевод: [перевод]
                    Релевантность: [процент]%

                    Текст для анализа:
                    {cleaned_text}"""
                }
            ]
        }
    ]
)

```

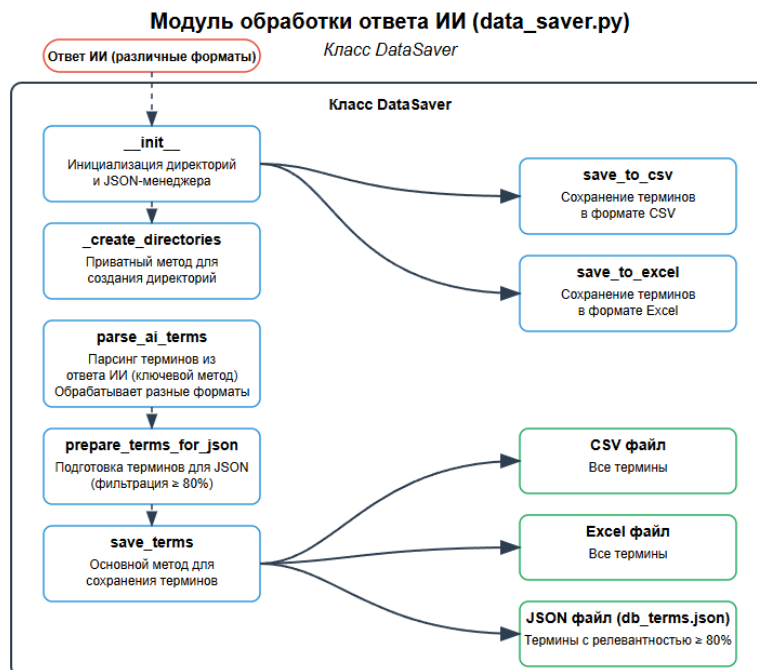
Рис. 8. Реализация API-запроса (промпта) к OpenAI

Основными элементами класса являются конструктор `__init__` для инициализации клиента OpenAI, загрузки переменных окружения с помощью библиотеки `dotenv` и определения доступных моделей, функция `Select_model` для выбора модели пользователем и сохранения в атрибут `self.model_name`, метод `Process_text` для обработки через Monica API в GPT-4o или Deepseek в зависимости от выбора пользователя. Затем ответ нейросети обрабатывается с помощью модуля `data_saver.py` (а этот модуль, в свою очередь, вызывает модуль `json_manager.py`).

### 3.8. Модуль обработки ответов нейронной сети (`data_saver.py`)

Модуль `data_saver.py` отвечает за обработку, хранение и управление данными, извлеченными из ответа нейросетевой модели. Основная задача модуля – парсинг терминов, их структурирование и последующее сохранение в различных форматах (CSV, Excel, JSON). Модуль разработан с учетом того, что может быть получено несколько вариантов ответов и, соответственно, может потребоваться обработка всех вариантов. Схема работы данного модуля представлена на рис. 9.

Отметим, что `_create_directories` — приватный метод для создания директорий, указанных в `self.directories`. Если каталог уже существует, он не создается. Метод `parse_ai_terms` необходим для парсинга терминов из текстового ответа нейросетевой модели. Он разбирает ответ, извлекая термины, их определения, переводы и релевантность. Методы `save_to_csv` и `save_to_excel` позволяют сохранить данные в форматах CSV и Excel соответственно, а `Prepare_terms_for_json` подготавливает термины для сохранения в JSON. Термины фильтруются по релевантности ( $\geq 80\%$ ). Все термины, найденные с помощью модели искусственного интеллекта, сохраняются в файлы excel и csv независимо от их релевантности, а также в JSON, если релевантность больше или равна 80 %.

Рис. 9. Логика модуля *DataSaver*

### 3.9. Модуль управления локальной базой данных терминов (*json\_manager.py*)

Файл *json\_manager.py* управляет локальной базой данных терминов, хранящихся в формате JSON. Основной класс *JsonTermManager* включает в себя всю логику работы с базой данных, в том числе создание файла базы данных, загрузку и сохранение данных. Атрибут класса *self.db\_folder* указывает путь к директории, в которой хранится база данных (папка *db* по умолчанию). Атрибут *self.db\_file* — имя JSON-файла базы данных (*db\_terms.json* по умолчанию). Атрибут *self.db\_path* определяет полный путь к файлу базы данных, объединяя *self.db\_folder* и *self.db\_file*. При создании экземпляра класса автоматически проверяется наличие папки и файла базы данных. Если они отсутствуют, создаются пустые структуры.

Рассмотрим методы класса. Начнем с *\_ensure\_db\_exists*. Этот метод является приватным и проверяет существование папки и файла базы данных. Если они отсутствуют, создается пустая директория и JSON-файл с пустым словарем. *\_load\_terms* — приватный метод для загрузки терминов из JSON-файла. Он возвращает данные в виде словаря. Если файл поврежден или пуст, возвращается пустой словарь. *\_save\_terms* — также приватный метод для сохранения терминов в JSON-файл. На вход он получает словарь с терминами и записывает их в файл. Метод *add\_terms* добавляет новые термины в базу данных. Новые термины передаются в виде словаря, который объединяется с существующими данными. После этого обновленные данные сохраняются обратно в файл. Термины в базе данных хранятся в виде словаря, где ключ — термин, а значение — информация о термине. Отметим, что поддержка обновления данных (метод *add\_terms*) позволяет не только добавлять новые термины, но и обновлять существующие, если их ключи совпадают.

#### 4. Результаты: тестирование и качественный анализ моделей искусственного интеллекта

Был проведен качественный анализ использования моделей DeepSeek и GPT-4o. Результаты этих анализов приведены в Таблицах 1 и 2 соответственно, где «+» — точное нахождение термина, «+-» — найден неточный, альтернативный или видоизменённый термин, а «-» значит не термин не найден.

Таблица 1. Результат качественного анализа модели DeepSeek

Ожидаемые термины (DeepSeek)	Предсказанные термины	Соответствие
беспилотные авиационные системы (БАС)	беспилотные авиационные системы (БАС)	+
БПЛА (беспилотные летательные аппараты)	БПЛА (беспилотные летательные аппараты)	+
дроны	дроны	+
мультикоптеры	мультикоптеры	+
беспилотники	беспилотники	+
авиационные системы с дистанционным управлением	авиационные системы с дистанционным управлением	+
автономный полет		-
системы навигации		-
полезная нагрузка	полезная нагрузка БПЛА	+-
УТМ-системы	УТМ-системы	+
Операторы БПЛА	Операторы БПЛА	+
коптеры	коптеры	+
БПЛА вертолетного типа	БПЛА вертолетного типа	+
беспилотники самолетного типа	беспилотники самолетного типа	+
системы управления		-
алгоритмы компьютерного зрения		-
беспилотные авиационные комплексы	беспилотные авиационные комплексы	+

Таким образом, получены следующие метрики модели DeepSeek: Precision: 92.3 %, Recall: 70.6 % и F1 мера: 80 %. Это достаточно хороший результат. Далее рассмотрим результаты GPT-4o.

Результаты приведены в таблице 2. Метрики модели GPT-4o равны: Precision: 70 %, Recall: 41.2 % и F1 мера: 51.9 %. Сравнительный анализ результатов показывает, что модель DeepSeek извлекает термины из технического текста более точно и полно по сравнению с GPT-4o. Таким образом, для задач извлечения терминов из специализированных документов в данной предметной области использование DeepSeek является более предпочтительным выбором, поскольку она обеспечивает более надёжные и качественные результаты.

Таблица 2. Результат качественного анализа модели GPT-4o

Ожидаемые термины (GPT-4o)	Предсказанные термины	Соответствие
беспилотные авиационные системы (БАС)	беспилотные авиационные системы (БАС)	+
БПЛА (беспилотные летательные аппараты)	БПЛА (беспилотные летательные аппараты)	+
дроны		-
мультикоптеры	мультикоптер	+-

Продолжение таблицы 2

Ожидаемые термины (GPT-4o)	Предсказанные термины	Соответствие
беспилотники		-
авиационные системы с дистанционным управлением	авиационные системы с дистанционным управлением	+
автономный полет	автономный полет	+
системы навигации		-
полезная нагрузка		-
UTM-системы	UTM-системы	+
Операторы БПЛА		-
коптеры	коптер вертолетного типа	+-
БПЛА вертолетного типа		-
беспилотники самолетного типа	беспилотники самолетного типа	+
системы управления		-
алгоритмы компьютерного зрения	компьютерное зрение	+-
беспилотные авиационные комплексы	беспилотные авиационные комплексы	+

## 5. Обсуждение результатов

Данное исследование является первым, в котором модели GPT-4o и Deepseek применяются для решения задачи АТР в области БАС. Разработанный программный комплекс DroneTerms AI продемонстрировал высокую эффективность при извлечении русскоязычных многокомпонентных терминов из технических текстов, связанных с БАС. Основными преимуществами являются модульная архитектура, применение промпт-инжиниринга с большими языковыми моделями, а также поддержка семантической классификации терминов без предварительной разметки корпуса. Инструмент успешно справляется со сложными синтаксическими структурами и способен адаптироваться к различным предметным областям, обеспечивая масштабируемость и интеграцию с терминологическими базами данных.

Нейросетевые модели на основе трансформеров, описанные в [11], в частности XLM-R и mBART, показывают схожие результаты. Эти модели применяются на уровне предложений и обеспечивают извлечение терминов без ручной предварительной обработки. Методология, предложенная в работе [13], которая сочетает статистический анализ с контекстуализацией для повышения точности извлечения многословных терминов, также имеет фундаментальные сходства. Весьма актуальны и гибридные методы, объединяющие корпусные и NLP-подходы, как в [4], где терминологическая информация используется в качестве логического звена для поиска семантических единиц в военном корпусе на английском языке.

Исследования, основанные на традиционных статистических и структурных лингвистических методах, демонстрируют разные подходы. Например, С. Ананиаду предлагает использовать формальные морфологические модели, включающие латинские и греческие компоненты, особенно в английской медицинской терминологии [14], без учета контекста или нейронных сетей.

Ю. И. Бутенко фокусируется на терминах с правосторонними определениями, разрабатывая формальные модели на основе морфологических особенностей русского языка и анализируя ошибки морфологических анализаторов [8]. В отличие от DroneTerms AI, ее метод частично зависит от лемматизации и грамматических паттернов.

Еще один подход, рассмотренный в Обзоре литературы — ACI-rank, который использует комбинацию статистических, структурных и лингвистических признаков эффективен при извлечении ключевых слов из HTML-страниц, он ограничен при работе со сложной технической лексикой и не использует современные большие языковые модели [1].

Несмотря на удовлетворительные результаты, предложенный в данной работе метод имеет ряд ограничений. Прежде всего, метод обеспечивает классификацию, как метод машинного обучения без учителя, то есть классификация терминов (например, по функциям) осуществляется с помощью predetermined подсказок, без обучения на размеченных данных. В связи с этим возникает еще одна проблема — отсутствие встроенной оценки точности, т. е. текущая реализация не позволяет автоматически проверять термины по эталонному корпусу или терминологическому словарю, а предложенные метрики построены на субъективной «ручной» разметке небольших текстовых файлов. Наконец, система характеризуется ограниченной лингвистической универсальностью, то есть она все еще настроена преимущественно на русскоязычные тексты в области БАС.

## 6. Выводы

В данной статье представлен программный комплекс DroneTerms AI, реализующий автоматическое распознавание и классификацию терминов в области беспилотных авиационных систем. Система сочетает в себе обработку технической документации, промпт-инжиниринг и использование больших языковых моделей для идентификации многословных терминов, включая конструкции с правыми и левыми определениями.

Работа вносит вклад в область обработки терминологии, предлагая модульную и масштабируемую архитектуру, адаптируемую к различным предметным областям. В отличие от традиционных лингвистических подходов, основанных на частотном подходе, DroneTerms AI извлекает термины с учетом контекста и семантики. Поскольку программа задумывалась как инструмент для автоматического построения глоссариев технических переводчиков, помимо ATR, были решены задачи перевода терминов на английский язык и генерации толкования термина. Такой глоссарий может быть загружен в системы поддержки переводческой деятельности — CAT-системы.

Тем не менее, предложенный метод имеет ограничения. Он не содержит объективного встроенного механизма автоматической валидации терминов и ориентирован в основном на русскоязычные тексты.

С практической точки зрения DroneTerms AI может применяться в системах технической документации, стандартизации, машинного перевода, а также при создании терминологических словарей и баз знаний. Особенно перспективно его использование в организациях, работающих с большими массивами технической документации в высокотехнологичных отраслях.

В дальнейших исследованиях планируется расширить языковую и отраслевую палитру, интегрировать механизмы обучения с учителем и объективной оценки, а также протестировать платформу на многоязычных корпусах с включением английских, французских, испанских и китайских технических текстов, с которыми работает наш исследовательский коллектив. Это позволит сделать DroneTerms AI более универсальным инструментом для решения глобальных задач управления терминологией.

## Литература

- [1] Shah H., Fränti P. Combining statistical, structural, and linguistic features for keyword extraction from web pages // *Applied Computing and Intelligence*. 2022. Vol. 2. No. 2. P. 115-132.
- [2] Ananiadou S. *Towards a Methodology for Automatic Term Recognition* / PhD thesis. Manchester: University of Manchester Institute of Science and Technology, 1988.
- [3] Shao W., Hua B., Song L. A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Text // *Data Inf Manag*. 2021. Vol. 5. No. 3. P. 329-335.

- [4] Chen L.-C., Chang K.-H., Yang S.-C. Integrating corpus-based and NLP approach to extract terminology and domain-oriented information: an example of US military corpus // *Acta Scientiarum. Technology*. 2022. Vol. 44. e60486.
- [5] Butenko I. I., Sapozhkov A. M., Stroganov Y. V. Method for the extraction of Russian-language multicomponent terms from scientific and technical texts // *Journal of Applied Informatics*. 2021. Vol. 16. No. 96. P. 21-27.
- [6] Бутенко Ю.И. Метод извлечения многокомпонентных терминологических единиц с правыми определениями из научно-технических текстов // *Вестник НГУ. Серия: Информационные технологии*. 2024. № 22(3). С. 5-14. DOI: 10.25205/1818-7900-2024-22-3-5-14
- [7] Zhu J., Xia Y., Wu L., He D., Qin T., Zhou W., Li H., Liu T.Y. Incorporating BERT into Neural Machine Translation. URL: <https://github.com/bert-nmt/bert-nmt> (дата обращения: 25.07.2025).
- [8] Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension // *Proceedings of the Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics (ACL). 2020. P. 7871-7880. DOI: 10.18653/V1/2020.ACL-MAIN.703
- [9] Radford A., Narasimhan K. Improving language understanding by generative pre-training // [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (дата обращения: 25.07.2025).
- [10] Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L. Multilingual denoising pre-training for neural machine translation // *Trans Assoc Comput Linguist. MIT Press Journals*. 2020. Vol. 8. P. 726-742. DOI: 10.1162/TACL\_A\_00343/96484/MULTILINGUAL-DENOISING-PRE-TRAINING-FOR-NEURAL
- [11] Lang C., Wachowiak L., Heinisch B., Gromann D. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. P. 3607-3620. DOI: 10.18653/v1/2021.findings-acl.316
- [12] Di Nunzio G. M., Marchesin S., Silvello G. A systematic review of Automatic Term Extraction: What happened in 2022? // *Digital Scholarship in the Humanities*. 2023. Vol. 38. No. Supplement\_1. P. i41-i47. DOI: 10.1093/lc/fqad030
- [13] Frantzi K. T., Ananiadou S. The C-value/NC-value domain-independent method for multiword term extraction // *Journal of Natural Language Processing*. 1999. Vol. 6. No. 3. P. 145-179.
- [14] Ananiadou, S. A Methodology for Automatic Term Recognition // *COLING 1994*. Volume 2: The 15th International Conference on Computational Linguistics. 1994. URL: <https://aclanthology.org/C94-2167/> (дата обращения 25.07.2025).

## Optimising Terminology Processing in Unmanned Aviation: A New Approach to Term Extraction Using Prompt Engineering

Ekaterina Isaeva<sup>1</sup>, Behruz Safarbekov<sup>2</sup>

<sup>1</sup> Perm State University, <sup>2</sup> National University of Science and Technology MISIS

Modern automated text processing techniques play a key role in the analysis and systematisation of specialised terminology. In the field of Unmanned Aerial Systems (UAS), accurate extraction and classification of terms is critical to standards development, knowledge sharing, and the improvement of artificial intelligence technologies in aeronautics. However, despite the growing interest in natural language processing (NLP), automated methods for terminology processing in UAS remain underdeveloped. Most approaches either focus on statistical, context-insensitive term extraction or require premarked corpora, which limits their applicability in dynamic technical domains. This paper proposes a new method for automated extraction and classification of UAS terminology based on large language models and prompt engineering. The method involves multistage document processing including text preprocessing, term analysis using NLP models, and storing the results in a database. The approach provides handling multi-word terms without preliminary manual markup. The method shows high accuracy of term extraction compared to traditional methods. It allows us to identify key UAS terms and classify them by their functionality in technical documentation. The approach contributes to the field of automated terminological processing, opening new opportunities for standardisation of data in the field of UAS, integration with ontologies, and creation of intelligent terminology management systems.

**Keywords:** Automated Term Recognition, Automated Term Identification, Unmanned Aircraft Systems, NLP, Artificial Intelligence, Terminological Processing, Large Language Models, Machine Learning, Prompt Engineering

**Reference for citation:** Isaeva E., Safarbekov B. Optimising Terminology Processing in Unmanned Aviation: A New Approach to Term Extraction Using Prompt Engineering // *Computational Linguistics and Computational Ontologies*. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 61-77. DOI: 10.17586/3033-5582-2025-9-61-77.

### References

- [1] Shah H., Fränti P. Combining statistical, structural, and linguistic features for keyword extraction from web pages // *Applied Computing and Intelligence*. 2022. Vol. 2. No. 2. P. 115-132.
- [2] Ananiadou S. Towards a Methodology for Automatic Term Recognition: PhD thesis. Manchester: University of Manchester Institute of Science and Technology, 1988.
- [3] Shao W., Hua B., Song L. A Pattern and POS Auto-Learning Method for Terminology Extraction from Scientific Text // *Data Inf Manag*. 2021. Vol. 5. No. 3. P. 329-335.
- [4] Chen L.-C., Chang K.-H., Yang S.-C. Integrating corpus-based and NLP approach to extract terminology and domain-oriented information: an example of US military corpus // *Acta Scientiarum. Technology*. 2022. Vol. 44. e60486.
- [5] Butenko I.I., Sapozhkov A.M., Stroganov Y.V. Method for the extraction of Russian-language multicomponent terms from scientific and technical texts // *Journal of Applied Informatics*. 2021. Vol. 16. No. 96. P. 21-27.
- [6] Butenko Yu.I. The Method of Extracting Multi-component Terminological Units with Right-Hand Definitions from Scientific and Technical Texts // *Vestnik NSU. Series: Information Technologies*. 2024. Vol. 22. No. 3. P. 5-14. DOI: 10.25205/1818-7900-2024-22-3-5-14

- [7] Zhu J., Xia Y., Wu L., He D., Qin T., Zhou W., Li H., Liu T.Y. Incorporating BERT into Neural Machine Translation. URL: <https://github.com/bert-nmt/bert-nmt> (accessed date: 25.07.2025).
- [8] Lewis M., Liu Y., Goyal N., Ghazvininejad M., Mohamed A., Levy O., Stoyanov V., Zettlemoyer L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension // Proceedings of the Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics (ACL). 2020. P. 7871-7880. DOI: 10.18653/V1/2020.ACL-MAIN.703
- [9] Radford A., Narasimhan K. Improving language understanding by generative pre-training // [https://cdn.openai.com/research-covers/language-unsupervised/language\\_understanding\\_paper.pdf](https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf) (accessed date: 25.07.2025).
- [10] Liu Y., Gu J., Goyal N., Li X., Edunov S., Ghazvininejad M., Lewis M., Zettlemoyer L. Multilingual denoising pre-training for neural machine translation // Trans Assoc Comput Linguist. MIT Press Journals. 2020. Vol. 8. P. 726-742. DOI: 10.1162/TACL\_A\_00343/96484/MULTILINGUAL-DENOISING-PRE-TRAINING-FOR-NEURAL
- [11] Lang C., Wachowiak L., Heinisch B., Gromann D. Transforming Term Extraction: Transformer-Based Approaches to Multilingual Term Extraction Across Domains // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Stroudsburg, PA, USA: Association for Computational Linguistics, 2021. P. 3607-3620. DOI: 10.18653/v1/2021.findings-acl.316
- [12] Di Nunzio G.M., Marchesin S., Silvello G. A systematic review of Automatic Term Extraction: What happened in 2022? // Digital Scholarship in the Humanities. 2023. Vol. 38. No. Supplement\_1. P. i41-i47. DOI: 10.1093/llc/fqad030
- [13] Frantzi K.T., Ananiadou S. The C-value/NC-value domain-independent method for multi-word term extraction // Journal of Natural Language Processing. 1999. Vol. 6. No. 3. P. 145-179.
- [14] Ananiadou S. A Methodology for Automatic Term Recognition // COLING 1994. Volume 2: The 15th International Conference on Computational Linguistics. 1994. URL: <https://aclanthology.org/C94-2167/> (accessed date: 25.07.2025).

## **Проблемы исследования словообразовательного потенциала с использованием современных поисковых систем: автоматизированный отбор дериватов через Яндекс**

Е. В. Васильева

Иркутский государственный университет

elvavi2301@yandex.ru

### **Аннотация**

Один из ключевых этапов исследования деривационного потенциала тех или иных слов — отбор производных единиц, позволяющий выявить в составе словообразовательных парадигм имеющиеся реализации и «пустые места» — лакуны. Однако словообразовательные словари не отражают весь перечень дериватов, используемых в языке, а толковые словари не всегда фиксируют новые лексические единицы. В то же время интернет представляет собой ценный источник данных о реальном функционировании производных слов, хотя в настоящее время отсутствует общепринятый метод их автоматизированного извлечения.

В данной работе предлагается один из подходов к отбору дериватов с использованием поисковой системы Яндекс. Метод включает в себя конструирование гипотетических производных на основе производящих основ и аффиксов, участвующих в словообразовании, их автоматизированную верификацию через Яндекс Search API, а также создание датасета, содержащего найденные производные, ссылки на соответствующие интернет-страницы и сниппеты, в которых встречаются дериваты. Дополнительно предлагается классификация собранных контекстов по их стилистической принадлежности на основе заданного алгоритма, анализирующего фрагменты интернет-ссылок. Это позволит в дальнейшем определять, насколько сбалансированно используются некодифицированные производные слова в разных сферах. Представленный метод позволяет более объективно оценивать словообразовательный потенциал лексем, в частности — прилагательных, обозначающих психические характеристики человека, а также исследовать причины деривационной лакунарности в сфере производства отадекватных имен лиц. Исследование имеет прикладное значение для автоматизированного составления динамических словообразовательных ресурсов и расширения возможностей компьютерной лексикографии.

**Ключевые слова:** словообразовательный потенциал, дериват, Яндекс Поиск, парсинг, поисковые системы, обработка естественного языка

**Библиографическая ссылка:** Васильева Е. В. Проблемы исследования словообразовательного потенциала с использованием современных поисковых систем: автоматизированный отбор дериватов через Яндекс // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 78–92. DOI: 10.17586/3033-5582-2025-9-78-92.

## 1. Введение

На первый взгляд, отбор дериватов для исследования не должен сопровождаться какими-либо трудностями, если лингвист имеет доступ к словообразовательным словарям: в лексикографическом источнике он находит нужное производящее, анализирует его деривационные парадигмы и делает вывод о наличии или отсутствии в них того или иного производного.

Однако информация, сосредоточенная в таких словарях, ограничивается объёмом лексикографических источников, на которые опирались составители. Если в них не была представлена та или иная лексическая единица, она не попадала в словообразовательный словарь. В этом случае данные соответствующего деривационного словаря оказываются недостаточными. Это характерно, например, для «Словообразовательного словаря русского языка» А. Н. Тихонова [1]. Он составлялся на базе ряда лексикографических источников, список которых к моменту его создания был весьма представительным [2, с. 14-15], однако все равно был неполным, поскольку не учитывал значительную часть словарного состава национального языка.

Во-первых, в него не попадала жаргонная, диалектная, просторечная лексика, так как она не была представлена в исходных словарях. В результате такие, например, слова, как *амбициозник*, *агрессивник*, *щедряга* и т. п. в нем отсутствуют. Это создаёт впечатление лакунарности в соответствующей зоне деривационной парадигмы: *амбициозный* → \*, *агрессивный* → \*, *щедрый* → \*.

Во-вторых, словарная база русского языка существенно расширилась и сегодня включает не только академические источники [3; 4], но и доступные онлайн-ресурсы, в числе которых как краудсорсинговые словари (например, викисловарь), так и любительские или специализированные проекты (многослов.рф, sinonim.org). В этих источниках зафиксированы единицы, отсутствующие в академических словарях: *шкодливец*, *взбалмошник*, *жизнерадостник*, *смешливец* и др.

В-третьих, стали доступны инструменты, позволяющие составить более полное представление о динамике языкового развития, — поисковые системы Google и Яндекс. С их помощью обнаруживаются некодифицированные дериваты *безынициативник*, *добротряпачник*, *жалостливец* и многие другие.

Игнорирование этих возможностей может приводить к недостоверным выводам: если исследователь будет основываться только на данных словообразовательного словаря, он рискует ошибочно заключить, что в конкретной деривационной парадигме отсутствует производное с тем или иным значением.

Кроме того, важно отметить, что материал [1] не только не отражает новые слова, но и фиксирует производные, устаревшие и малоупотребительные в современной речевой практике, например, существительное *брюзгливец*. При поиске в интернете оно обнаруживается только в четырех контекстах, в то время как ряд других дериватов, не отмеченных в словообразовательном словаре, имеет значительно большее количество употреблений: *болтливец* — 16, *злопамятник* — 33, *амбициозник* — 109 и др.

Складывается ситуация, когда исследователь может заключить, что в словообразовательной парадигме той или иной лексемы нет лакуны, так как соответствующее семантическое место занято (при том, что занято оно неупотребительным словом), или, наоборот, констатировать наличие лакуны — при том, что в современном узусе слово употребляется достаточно частотно.

Как кажется, эта проблема осознаётся многими учёными. В работах, связанных с данной тематикой, исследователи все чаще привлекают материалы более современных словарей, не ограничиваясь словником [1].

Использование современных лексикографических источников — это существенный шаг на пути к получению актуальных данных об участии слов в деривационных процессах. Однако, чтобы анализ словообразовательного потенциала лексем был более

полон, необходимо использовать не только словари, но и современные интернет-ресурсы. Последние позволяют оперативно проверить утверждение о том, что «всегда имеется вероятность, что данное морфемное образование произведено «где-то кем-то» в речи» [5, с. 390].

Например, в [6] указывается, что существительное *дольмен* по-прежнему не имеет дериватов. Обращение же к Яндекс Поиск позволяет обнаружить производное *дольменный*: *дольменный город* (<https://dzen.ru/a/XVukn42hzgCtYffG?ysclid=m8d8onneg0329414055>), *дольменный камень* (<https://www.b17.ru/article/11430/?ysclid=m8d8q50dou398608916>) и др. Если игнорировать такие факты при обсуждении проблем лакуарности, можно сделать некорректные выводы и начать подводить теоретическую базу под «отсутствие» единиц, которые на самом деле существуют в речевой практике.

Таким образом, интернет становится важным инструментом для выявления актуальных производных, не всегда фиксируемых в словарях. Однако для нахождения некодифицированных производных, используемых в речи, необходим метод, обеспечивающий осмысленный подход к их поиску.

Конструирование гипотетических дериватов позволяет систематизировать процесс отбора материала, а также точно определить круг производных, подлежащих исследованию. Этот метод подробно рассматривается в работах Н. Б. Лебедевой и М. Г. Шкуропацкой [5; 7]. Авторы предлагают образовывать дериваты на основании знаний о сочетаемости мотивирующих основ и аффиксов, участвующих в словопроизводстве, и оценивать эти «лабораторные слова» [5, с. 391] с точки зрения их возможности функционировать в речи, опираясь на мнение информантов и анализ употребимости дериватов в языке [7, с. 345].

В рамках данного подхода моделирование производных осуществляется с использованием каждого из 28 суффиксов, участвующих в образовании отадъективных имен лиц, включая их алломорфы [8, с. 166-177]. Всего в процесс конструирования дериватов вовлекается 901 прилагательное, обозначающее психические характеристики человека. Данные лексические единицы отобраны для исследования на основе анализа психологической и лингвистической литературы (см. подробнее [9]). Далее все возможные сочетания основ и формантов проверяются на фонологическую допустимость. Для этой цели привлекаются данные Грамматического словаря русского языка А. А. Зализняка [10]. В результате список полученных «лабораторных слов» составляют 41915 единиц: *агрессивняга, агрессивец, агрессивнюга, агрессивнюк, агрессивный* и др.

На следующем этапе определяется статус каждого из отфильтрованных производных: фиксируется ли дериват в реальном употреблении или остается потенциальным. Наиболее эффективным инструментом для такой проверки являются поисковые системы интернета. При этом в поле наблюдения попадают не только нарицательные существительные, но и имена собственные, которые рассматриваются как равноправные проявления деривационного потенциала, поскольку отражают мотивировочный признак так же, как и нарицательные.

## 2. Разработка алгоритма сбора данных

Для сбора контекстов, в которых встречаются дериваты, и создания датасета с необходимой для исследования информацией был разработан алгоритм, реализованный на языке программирования Python. В качестве источника данных используется поисковая система Яндекс, а взаимодействие с ней осуществляется через Яндекс Search API. Алгоритм позволяет формировать запросы, отправлять их в поисковую систему, получать и обрабатывать результаты для последующего лингвистического анализа.

Принцип его работы заключается в последовательном выполнении нескольких этапов. Сначала задаётся список поисковых запросов, представляющих собой возможные

дериваты исходного слова, например: *занудняга, занудник, зануднец, занудный* и др. Далее каждый запрос передается API в виде JSON-файла, содержащего параметры поиска.

### 2.1. Поисковые запросы

Образец JSON-файла:

```
{
  "query": {
    "search_type": "SEARCH_TYPE_RU",
    "query_text": "агрессивец",
    "family_mode": "FAMILY_MODE_MODERATE",
    "page": 0,
    "fix_typos_mode": "FIX_TYPO_MODE_OFF"
  },
  "sort_spec": {
    "sort_mode": "SORT_MODE_BY_RELEVANCE",
    "sort_order": "SORT_ORDER_DESC"
  },
  "group_spec": {
    "group_mode": "GROUP_MODE_DEEP",
    "groups_on_page": 100,
    "docs_in_group": 1
  },
  "max_passages": 4,
  "region": "213",
  "l10n": "LOCALIZATION_RU",
  "folder_id": "***",
  "response_format": "FORMAT_XML",
  "user_agent": "***"
}
```

В приведенном JSON-объекте каждая из четырех частей выполняет свою функцию.

Блок `query` задает параметры поискового запроса, включая тип поиска, текст запроса, фильтрацию результатов, номер запрашиваемой страницы и настройки исправления опечаток. В качестве типа поиска выбрано значение `'SEARCH_TYPE_RU'`, так как исследуемый материал представлен на русском языке. Текст запроса — изменяемый параметр, который на каждом этапе подставляется из списка сгенерированных дериватов. `Family_mode` отвечает за фильтрацию результатов, и его значение `'FAMILY_MODE_MODERATE'` установлено по умолчанию. В этом режиме из выдачи исключаются материалы, относящиеся к категории «для взрослых». Параметр `page` определяет номер страницы поисковой выдачи (нумерация начинается с нуля). Для получения максимального количества результатов мы используем значения 0 и 1, что позволяет извлечь около 200 документов. При значении 2 поисковая платформа перестает отвечать на запросы, что ограничивает дальнейший сбор данных. Последний параметр в этом блоке — `fix_typos_mode`, который управляет исправлением опечаток. Эта опция отключена намеренно (`'FIX_TYPO_MODE_OFF'`), поскольку многие из сконструированных производных могут быть автоматически исправлены системой из-за своей редкости или отсутствия в интернет-пространстве.

Раздел `sort_spec` задает параметры упорядочивания результатов. Значение `'SORT_MODE_BY_RELEVANCE'` определяет сортировку по релевантности в отличие от упорядочивания по дате изменения документа. Параметр `'SORT_ORDER_DESC'` устанавливает порядок отображения — от более новых результатов к более старым.

В свою очередь, `group_spec` отвечает за группировку найденных документов, позволяя задать количество групп и документов в каждой из них. Так, параметр `group_mode`

указывает на метод группировки результатов. Здесь по умолчанию выбрано значение 'GROUP\_MODE\_DEEP' — группировка по доменам. Это значит, что все документы из одного и того же веб-сайта (домена) собираются в одну группу. Например, если искомое слово встречается на нескольких страницах одного сайта, то все эти страницы будут находиться в одной группе и в выдаче будет указано, что все они принадлежат одному и тому же домену. В нашем случае это полезно, поскольку позволяет собрать статистику, как часто искомый дериват встречается на одном ресурсе или в пределах одной категории сайтов. `Groups_on_page` задает максимальное количество групп, которые могут быть возвращены на одной странице результатов поиска (`page`). Диапазон от 1 до 100. Мы определяем значение 100, поскольку нам важно собрать большее число контекстов с минимальным количеством обращений к поисковой системе. Параметр `docs_in_group` регулирует количество разных документов (URL-адресов) из одного домена, которые могут попасть в одну группу. В нашем случае он равен 1. Это значит, что в выдаче от одного домена будет только один документ, даже если там есть другие релевантные страницы. Такой выбор значения позволяет получить более широкий охват результатов выдачи, так как из одного домена выбирается только один наиболее релевантный документ, что помогает избежать засилья контекстов из одного источника. Это в свою очередь уменьшает вероятность дублирования, что делает выборку более разнообразной и репрезентативной.

Параметр `max_passages` указывает максимальное число текстовых фрагментов (предложений), которые могут быть извлечены из найденных документов при формировании их сниппета. Чем больше объем сниппета, тем ниже вероятность, что исследователю придется обращаться к полному документу (интернет-странице). Более широкий контекст употребления деривата позволяет лучше определить его семантику.

Код региона (`region`) определяет географическую область поиска. В нашем случае указываем код Москвы — 213. Такой выбор обусловлен тем, что Москва — крупнейший город России с самым высоким интернет-трафиком, что позволяет получить больше релевантных результатов.

Параметр `l10n` задает языковую локализацию (в данном случае — русский язык). Поле `folder_id` представляет собой идентификатор сервисного аккаунта, от имени которого выполняются запросы. Он предоставляется пользователю при регистрации на Yandex Cloud. Формат ответа указывается в параметре `response_format` (например, XML), а `user_agent` передает строку, имитирующую информацию о браузере и операционной системе пользователя. Более подробную информацию о параметрах и всех их значениях можно найти в документации Яндекс Search API [11].

Выполнение запроса для каждого варианта искомого слова отдельно может быть трудоемким, поэтому мы автоматизировали этот процесс, написав Python-скрипт, позволяющий отправлять большее количество запросов сразу.

Автоматически для каждого поискового запроса попеременно создается отдельный JSON-файл (`body_X.json`, где `X` — номер файла и порядковый номер деривата из общего перечня производных), содержащий все необходимые параметры поиска. Для отправки запросов к Яндекс Search API используется инструмент `grpcurl`, который работает аналогично `curl`, но предназначен для gRPC-запросов. Каждый запрос передается на сервер с использованием Bearer-токена, предоставленного при регистрации в Yandex Cloud. Полученные результаты попеременно сохраняются в файлы (`result_X.json`), чтобы их можно было проанализировать позже.

## 2.2. Декодирование и сохранение результатов поиска

После выполнения множества поисковых запросов и получения ответов в JSON-формате следующий этап алгоритма — обработка и декодирование результатов.

Имеющийся перечень файлов с результатами проверяется вторым Python-скриптом на наличие пустых файлов. Они могут возникать, если определенный дериват не имеет контекстов употребления. Скрипт пропускает такие файлы, что предотвращает ошибки при обработке всего массива данных.

Закодированная информация выгружается из JSON-файла, после чего извлекается ключ `rawData`, содержащий результаты поиска в кодировке Base64. С использованием `base64.b64decode` строка декодируется в XML-формат.

Результаты сохраняются в отдельную папку, а каждый файл получает название `result_X.xml`, где X — все тот же порядковый номер деривата. Это позволяет структурировать данные для дальнейшего анализа.

Образец содержимого XML-файла:

```
<group>
<categ attr="d" name="litres.ru"/>
<doccount>4</doccount>
<relevance priority="all"/>
<doc id="Z629ACA5FB63964FA">
  <relevance priority="all"/>
  <saved-copy-
url>https://yandexwebcache.net/yandbtm?fmode=inject&tm=1742381050&tld=ru&lang=
ru&la=1741456640&text=%D0%B7%D0%B0%D0%BD%D1%83%D0%B4%D0%BD%
D0%B8%D0%BA&url=https%3A//www.litres.ru/book/mett-heyg/devochka-kotoraya-
spasla-rozhdestvo-26552836/chitat-
onlayn/&l10n=ru&mime=html&sign=a361baf99c0ee50ac843580b34cd520c&keyno=0</s
aved-copy-url>
  <url>https://www.litres.ru/book/mett-heyg/devochka-kotoraya-spasla-rozhdestvo-
26552836/chitat-onlayn/</url>
  <domain>www.litres.ru</domain>
  <title>Читать онлайн «Девочка, которая спасла Рождество», Мэтт Хейг...</title>
  <modtime>20171027T210158</modtime>
  <size>6837</size>
  <charset>utf-8</charset>
  <mime-type>text/html</mime-type>
  <passages>
    <passage>
      <hlword>Занудник</hlword>
      служил Помощником заместителя главы Цеха игрушек, которые прыгают и
      вращаются, считался незаменимым сотрудником и никогда не жаловался на то,
      что приходится задерживаться после работы.
    </passage>
  </passages>
  <properties>
    <_PassagesType>0</_PassagesType>
    <lang>ru</lang>
    <extended-text> Занудник был одним из лучших работников Мастерской игрушек.
    Многим этот нервный маленький эльф казался странным, но Отцу Рождество он
    нравился. Занудник служил Помощником заместителя главы Цеха игрушек,
    которые прыгают и вращаются, считался незаменимым сотрудником и никогда не
    жаловался на то, что приходится задерживаться после работы. </extended-text>
  </properties>
</doc>
</group>
```

Данный образец содержит информацию, структурированную в виде группы (<group>), включающей несколько важных элементов, каждый из которых представляет определенные сведения о документе. XML-файл как правило содержит от 1 до 100 таких групп.

Рассмотрим только данные, необходимые для дальнейшей работы. Элемент <categ> с атрибутом name указывает на домен источника — litres.ru. <doccount> выводит количество веб-страниц данного домена, в которых встречается искомое слово. <url> содержит ссылку на оригинальный текст. Также указывается дата последнего обновления документа, в данном случае — 27 октября 2017 г.

Важную часть данных представляет текстовый раздел. В нем содержатся фрагменты текста, в которых встречается интересующий нас дериват. Блок <passage> включает сокращенный вариант сниппета, а элемент <hlword> выделяет слово из текста, наиболее точно соответствующее запросу. В идеале оно должно совпадать с введенным в поиск дериватом.

Элемент <extended-text> представляет собой расширенный сниппет, который дает более широкий контекст употребления производного. Это позволяет оценить значение искомого слова и его релевантность без необходимости перехода на веб-страницу.

### 2.3. Создание датасета

Далее производится структурированная обработка XML-файлов с помощью третьего Python-скрипта, который на выходе позволяет получить датасет в формате CSV, включающий в себя текстовые запросы, метаданные найденных документов и их содержимое (см. табл. 1).

Каждый XML-файл парсится с помощью библиотеки xml.etree.ElementTree [12] путем вычленения объектов дерева XML. Извлекаются такие элементы, как запрос, количество документов в домене, URL, дата последнего обновления документа, а также фрагменты текста, в которых встречается искомое слово.

Однако в процессе работы мы сталкиваемся с рядом проблем, требующих решения: дублирование данных при формировании датасета и наличие контекстов без целевого деривата.

#### 2.3.1. Дублирование данных при формировании датасета

Один и тот же текст может быть в разных доменах, что не учитывается системой, она распознает его каждый раз как уникальный, так как URL-адреса страниц различаются. В результате в датасете обнаруживаются дубликаты (см. табл. 1). Это впоследствии может исказить статистический анализ данных, поскольку создает ложное впечатление о количестве употреблений того или иного производного. Так, система может посчитать, что существительное *занудник* встречается в 5 разных контекстах, что не является верным.

Для решения этой проблемы можно применять метод группировки и консолидации повторяющихся контекстов. Его суть заключается в объединении одинаковых текстов, что позволяет сформировать перечень содержащих их URL, а также определить общее количество документов, в которых они встречаются.

Процесс начинается с проверки наличия одинаковых текстовых фрагментов. Для этого осуществляется группировка по столбцу «Сниппет». Поля «Дериват» и «Вид. слово» при этом остаются неизменными и принимают первое значение внутри группы. Для столбца «Кол-во док.» выполняется суммирование, что позволяет определить общее количество вхождений сниппета во всех доменах. «URL» и «Год» объединяются, и все обнаруженные значения записываются через пробел.

Таблица 1. Фрагмент датасета отобранных дериватов с дубликатами

Дериват	URL	Год	Кол-во док.	Сниппет	Выд. слово
занудник	<a href="https://www.litres.ru/">https://www.litres.ru/&lt;...&gt;</a>	2017	4	Занудник был одним из лучших работников Мастерской игрушек. <...>	занудник
занудник	<a href="https://librebook.me/">https://librebook.me/&lt;...&gt;</a>	2017	2	Занудник был одним из лучших работников Мастерской игрушек. <...>	занудник
занудник	<a href="https://royallib.com/">https://royallib.com/&lt;...&gt;</a>	2017	1	Занудник был одним из лучших работников Мастерской игрушек. <...>	занудник
занудник	<a href="https://klumba.guru/">https://klumba.guru/&lt;...&gt;</a>	2019	1	В народе заманиху именуют волчьей ягодой. Это растение также имеет такие названия, как эхинопанакс и оплопанакс высокий, шипник, киргун, занудник <...>	занудник
занудник	<a href="https://landshaftnik.com/">https://landshaftnik.com/&lt;...&gt;</a>	2020	1	В народе заманиху именуют волчьей ягодой. Это растение также имеет такие названия, как эхинопанакс и оплопанакс высокий, шипник, киргун, занудник <...>	занудник

В результате такой работы мы получаем датафрейм, в котором каждая строка представляет уникальный сниппет (см. табл. 2). Вместо 5 контекстов мы видим только 2.

Таблица 2. Фрагмент датасета отобранных дериватов без дубликатов

Дериват	URL	Год	Кол-во док.	Сниппет	Выд. слово
занудник	<a href="https://www.litres.ru/">https://www.litres.ru/&lt;...&gt;</a> <a href="https://librebook.me/">https://librebook.me/&lt;...&gt;</a> <a href="https://royallib.com/">https://royallib.com/&lt;...&gt;</a>	2017	7	Занудник был одним из лучших работников Мастерской игрушек. <...>	занудник
занудник	<a href="https://klumba.guru/">https://klumba.guru/&lt;...&gt;</a> <a href="https://landshaftnik.com/">https://landshaftnik.com/&lt;...&gt;</a>	2019 2020	2	В народе заманиху именуют волчьей ягодой. Это растение также имеет такие названия, как эхинопанакс и оплопанакс высокий, шипник, киргун, занудник <...>	занудник

### 2.3.2. Шум в данных: контексты без целевого деривата

Вторая проблема — это наличие контекстов, в которых отсутствует искомый дериват. Появление таких сниппетов связано с особенностями работы поисковых систем.

При формировании выдачи они ориентируются не только на точное совпадение запроса, но и на семантическое соответствие, что может приводить к включению нерелевантных результатов. Например, если запрос содержит производное слово, алгоритм может учитывать страницы, где встречаются однокоренные лексические единицы или семантически связанные выражения, но сам искомый дериват отсутствует. Так, мы можем искать дериват *занудник*, но в представленных текстах максимально сходными с запросом словами будут такие лексемы, как *занудный*, *зануда*, *Ника*, а для производного *занудняшка* — *зануда*, *няшка* и т.п. (см. табл. 3).

Кроме того, проблема может возникать из-за ограничений в обработке сниппетов. Поисковая система формирует их автоматически, вырезая небольшие фрагменты текста, которые могут не содержать само производное, даже если оно присутствует в полной версии документа. В результате в итоговый датасет попадают контексты, не содержащие нужное слово, что увеличивает уровень шума в данных.

Для того чтобы исключить такие сниппеты, мы автоматически сравниваем два столбца «Дериват» и «Выд. слово». Если они совпадают, в тексте присутствует необходимое производное, если нет — то в данном контексте оно отсутствует.

Таблица 3. Фрагмент датасета с контекстами без целевого деривата

Дериват	URL	Год	Кол-во док.	Сниппет	Выд. слово
занудник	<a href="https://poiski.p&lt;br/&gt;ro/&lt;...&gt;">https://poiski.p ro/&lt;...&gt;</a>	2018	1	Ник Зануда.	ник, зануда
занудныхш	<a href="https://librebo&lt;br/&gt;ok.me/&lt;...&gt;">https://librebo ok.me/&lt;...&gt;</a>	2017	4	Мирное пребывание Эрика на каникулах по случаю окончания "проекта века" было омрачено порцией занудных Ш-вопросов <...>	занудных, ш
занудняшка	<a href="https://otvet.ya&lt;br/&gt;.guru/&lt;...&gt;">https://otvet.ya .guru/&lt;...&gt;</a>	2020	1	Я жду тебя, как свершения чуда! Ну что ж не звонишь, моя няшка зануда.	няшка, зануда
занудняш	<a href="https://landsha&lt;br/&gt;ftnik.com/&lt;...&gt;">https://landsha ftnik.com/&lt;...&gt;</a>	2021	1	Потому что я – няшка Зануда	зануда, няша

Таким образом, посредством написанного Python-скрипта, позволяющего не только извлечь данные из XML-файлов, но и решить указанные выше проблемы, формируется структурированный набор данных, содержащий необходимую информацию, очищенную от шума и дубликатов.

### 2.3.3. Обогащение данных: от извлечения к генерации новой информации

Однако ценность сформированного датасета заключается не только в упорядоченном представлении извлеченной информации из Яндекс Поиска, но и в возможности порождения новых сведений на основе уже имеющихся данных. Один из таких примеров — формирование столбца «Стиль», который определяется через анализ поля «URL».

Для автоматической классификации стиля текстов на основе URL-адресов используется следующий алгоритм.

Сначала определяется перечень ключевых слов, соответствующих трем стилям: разговорному (comments, forum, otvet и т. п.), публицистическому (news, article, media, blog и т. п.), художественному (proza, stih, book и т. п.).

Далее реализуется функция классификации, которая проверяет, содержится ли в URL хотя бы одно ключевое слово из заданных списков, и возвращает соответствующее значение. Если ни одного совпадения не выявлено, стиль маркируется как неизвестный.

Затем эта функция применяется ко всем URL в датасете, результат записывается в новый столбец «Стиль» (см. табл. 4).

Это позволяет автоматически классифицировать контексты по стилистической принадлежности, что расширяет возможности последующего исследования употребимости дериватов.

Следует отметить, что при отнесении текстов к тому или иному стилю в настоящем исследовании используется условная, упрощенная классификация, не предполагающая строгого следования критериям функциональной стилистики. Это обусловлено спецификой интернет-дискурса, который характеризуется высокой стилистической гибридностью.

На наш взгляд, стилистическая триада «разговорный — публицистический — художественный», используемая при автоматической классификации сниппетов, может рассматриваться как ориентировочная модель, позволяющая в компактной и практичной форме отразить спектр сфер употребления производных.

Таблица 4. Фрагмент датасета отобранных дериватов

Дериват	URL	Год	Кол-во док.	Сниппет	Стиль
занудник	<a href="https://4italka.site/&lt;...&gt;">https://4italka.site/&lt;...&gt;</a>	2023	2	Занудник смущённо покраснел и торопливо водрузил очки <...>	Худож.
занудник	<a href="https://forum.sources.ru/&lt;...&gt;">https://forum.sources.ru/&lt;...&gt;</a>	2018	1	А чем Фармер те не нравится. Занудник. У меня от его Дюн, пока дочитал, мозги уже из ушей лезли.	Разг.
занудныш	<a href="https://galartemenko.livejournal.com/&lt;...&gt;">https://galartemenko.livejournal.com/&lt;...&gt;</a>	2017	1	Я с удовольствием продолжаю шествовать по граням своего внутреннего занудныша!	Публ.
занудняк	<a href="https://politikus.info/&lt;...&gt;">https://politikus.info/&lt;...&gt;</a>	2014	1	Это диагноз — быть затюканным, недовольным всем и вся, занудняком.	Разг.

Однако, несмотря на очевидные преимущества такого подхода, он не охватывает всех нюансов, поэтому остается необходимость в ручной обработке данных, позволяющей уточнить результаты определения стиля.

Кроме того, особого внимания требуют сами контексты, включенные в датасет, так как не все из них соответствуют условиям исследования. Важным критерием отбора является наличие дериватов, обозначающих человека по его психическому качеству. Например, *занудник* должно именовать лицо, склонное к занудству, однако это же существительное используется и для номинации растения (см. табл. 1). Поиск и отбор таких контекстов осуществляется вручную.

### 3. Анализ фрагмента собранного датасета

Собранные данные позволяют сделать выводы о реализованном деривационном потенциале прилагательных в сфере образования имен лиц.

Так, например, с помощью Яндекса Поиска установлено, что из 45 гипотетических производных со значением ‘носитель признака (лицо)’ от прилагательного *занудный* в речи используются только 4: *занудняк*, *занудник*, *занудныш* и *зануднюка*. Из них наиболее часто встречается субстантив *занудник*, остальные используются значительно реже (см. рис. 1).

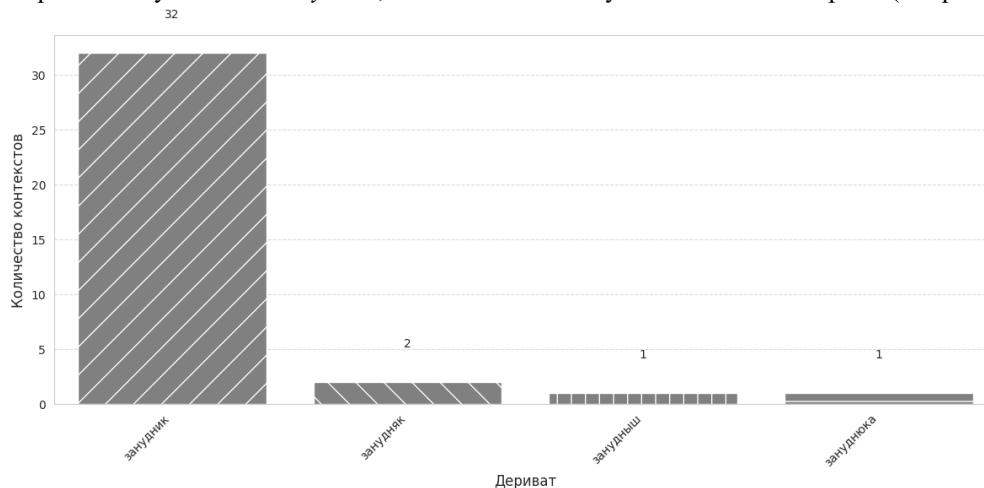


Рис. 1. Количество употреблений дериватов прилагательного *занудный*

Именно существительное *занудник* является подтверждением того, что адектив *занудный* реализует свой деривационный потенциал в сфере производства имен лиц. Остальные производные демонстрируют эпизодическое появление, что может указывать на зарождающуюся возможность, на данный момент активно не проявляемую в речи.

Так, субстантив *занудняк* встречается только в разговорных контекстах в период с 2014 по 2019 гг., лексема *занудныш* появляется лишь раз в блоге в 2017 г., а дериват *зануднюка* – в комментарии в социальной сети в 2022 г.:

(1) *Это диагноз-быть затюканным, недовольным всем и вся занудняком* (<https://metalarea.org/forum/index.php?showtopic=66870&st=2070>);

(2) *Я с удовольствием продолжаю шествовать по граням своего внутреннего занудныша!* (<https://galartemenko.livejournal.com/109712.html>);

(3) *Даниил, когда жирок там, где надо, это thicc, учи матчасть, зануднюка* ([https://vk.com/wall-125057438\\_47849](https://vk.com/wall-125057438_47849)).

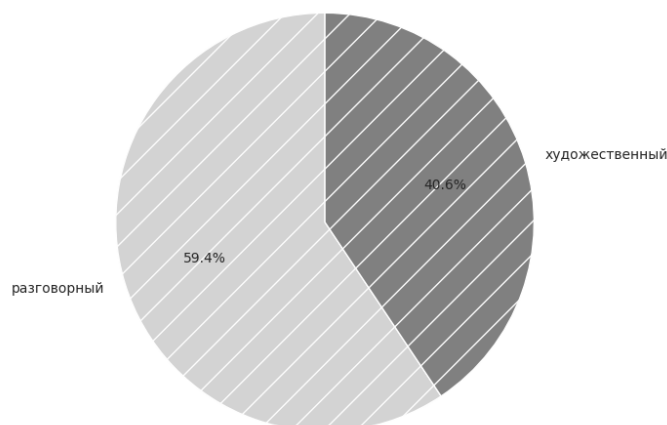


Рис. 2. Распределение примеров употребления деривата *занудник* по стилям

Результаты анализа распределения контекстов по стилям показывают, что субстантив *занудник*, в отличие от трех других, используется в речи носителями языка в разных сферах (см. рис. 2).

Дериват встречается в 40 % контекстов художественного стиля и в 60 % текстов разговорного характера:

(4) *Занудник* кивнул и тихо прошептал: — *Всё верно, всё верно* ([https://4italka.site/detskoe/zarubejnaya\\_literatura\\_dlya\\_detey/537161/fulltext.htm](https://4italka.site/detskoe/zarubejnaya_literatura_dlya_detey/537161/fulltext.htm)).

(5) *Денис, ты занудник*)), *но я исправлю* (<https://mosyagin.livejournal.com/101941.html>).

Производное не является устаревшим, примеры его употребления встречаются стабильно в интернет-пространстве как минимум последние 5 лет. (см. рис. 3).

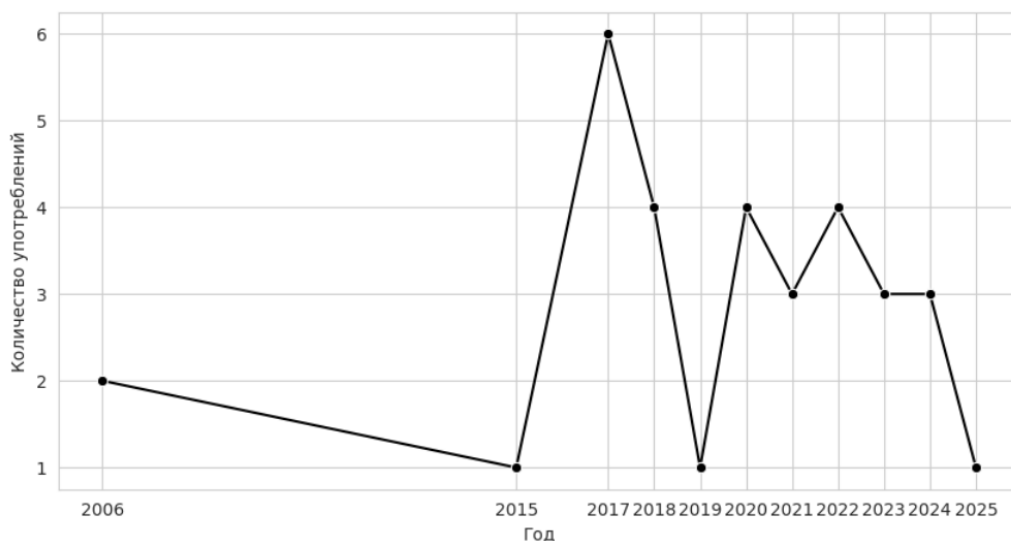


Рис. 3. Распределение контекстов с дериватом *занудник* по годам

Таким образом, информация, собранная в датасете, позволяет не только сделать вывод о том, какие прилагательные реализуют свой деривационный потенциал в сфере производства имен лиц, но и проанализировать статистику характера употреблений найденных существительных.

В дальнейшем датасет, содержащий информацию о возможных дериватах всех прилагательных, обозначающих психические характеристики человека, позволит создать полную картину того, какие адъективы образуют существительные со значением 'носитель признака (лицо)', какие суффиксы чаще, чем другие, участвуют в производстве субстантивов, какие дериваты, несмотря на отсутствие фиксации в словарях, активно используются в речи, а какие имеют единичное употребление и т.п.

#### 4. Заключение

Итогом данной работы стала разработка алгоритма, которая позволяет в сжатые сроки зафиксировать факты использования сконструированных дериватов в речи и решить тем самым ряд исследовательских проблем.

Во-первых, наш алгоритм позволяет оперативно фиксировать появление новых производных слов в больших текстовых массивах, тем самым решая проблему замедленного процесса выявления новообразований. Словари, как правило, обновляются с опозданием, а дериваты могут активно использоваться в речи задолго до их фиксации

в лексикографических источниках. Более того, даже при наличии больших корпусов сложно оперативно находить дериваты в потоках новых текстов. Алгоритм интегрируется с поисковой системой Яндекс, позволяя фиксировать изменения в языке в реальном времени.

Во-вторых, без автоматизированных инструментов трудно оценить количество употреблений и характер использования производных. Наш алгоритм справляется с этой задачей, собирая статистику о количестве снипшотов того или иного слова, распределении контекстов с ним по различным источникам и стилям. Автоматизированный сбор информации также позволяет отслеживать динамику появления и распространения новых слов. Если дериват встречается часто в различных текстах, это свидетельствует о его закреплении в языке. Напротив, если его употребление ограничено единичными примерами, это указывает на его эпизодический характер или экспериментальность.

Предложенный алгоритм обеспечивает более объективный и масштабный сбор данных по функционированию производных единиц, значительно повышая точность и воспроизводимость лингвистического анализа.

## Литература

- [1] Тихонов А.Н. Словообразовательный словарь русского языка. В 2 т. М.: АСТ; Астрель, 2008.
- [2] Тихонов А.Н. Основные понятия русского словообразования // Тихонов А.Н. Словообразовательный словарь русского языка. В 2 т. Т. 1. 3-е изд., испр. и доп. М.: АСТ; Астрель, 2008. С. 24-57.
- [3] Словарь современного русского литературного языка. В 17 т. / Академия наук СССР, Институт языкознания. М.; СПб.: Издательство АН СССР, 1949–1965.
- [4] Словарь русского языка. В 4 т. / гл. ред. А. П. Евгеньева. М.: Русский язык, 1981–1984.
- [5] Лебедева Н.Б. Экспериментальное исследование производного слова: соотношение выводимости как словообразовательного признака и узуальности как лексического признака // Новые явления в славянском словообразовании: система и функционирование. Доклады XI Международной научной конференции Комиссии по славянскому словообразованию при Международном комитете славистов / отв. ред. Е.В. Петрухина. М.: Макс Пресс, 2010. С. 390-399.
- [6] Лю С., Катышев П.А. Факторы реализации деривационного потенциала знаменательных одиночных слов // Международный аспирантский вестник. Русский язык за рубежом. 2024. № 3. С. 55-59. DOI: 10.37632/PI.2024.15.14.010
- [7] Шкуропацкая М.Г. Методика исследования системной и эмпирической реализации словообразовательного типа в узуальной и потенциальной лексике // Мир науки, культуры, образования. 2013. № 6(43). С. 343-345.
- [8] Русская грамматика: в 2 т. / гл. ред. Н. Ю. Шведова. М.: Наука, 1982. Т. 1. 783 с.
- [9] Васильева Е.В., Ташлыкова М.Б. Прилагательные лексико-семантической группы ‘психические характеристики человека’: проблема классификации // В мире научных открытий. 2014. № 11–10(59). С. 3922-3938.
- [10] Зализняк А.А. Грамматический словарь русского языка: Словоизменение. М.: Русский язык, 2008. 792 с. URL: <https://gramdict.ru/> (дата обращения: 27.03.2025).
- [11] Выполнение поисковых запросов с помощью API v2 в синхронном режиме // Документация Yandex Search API. Yandex Cloud. URL: [https://yandex.cloud/ru/docs/search-api/operations/web-search-sync?utm\\_referrer=about%3Ablank](https://yandex.cloud/ru/docs/search-api/operations/web-search-sync?utm_referrer=about%3Ablank) (дата обращения: 27.03.2025).
- [12] Python Software Foundation. XML processing modules – xml.etree.ElementTree. URL: <https://docs.python.org/3/library/xml.etree.elementtree.html> (дата обращения: 27.03.2025).

## Issues in Studying Derivational Potential Using Modern Search Engines: Automated Selection of Derivatives via Yandex

E. V. Vasileva

Irkutsk State University

One of the key stages in studying the derivational potential of certain words is the selection of derived units, which helps identify both existing realizations and lacuna within word-formation paradigms. However, word-formation dictionaries do not reflect the full range of derivatives used in the language, and explanatory dictionaries do not always record newly emerging lexical units. At the same time, the Internet serves as a valuable source of data on the actual usage of derived words, although there is currently no widely accepted method for their automated extraction. This study proposes an approach to selecting derivatives using the Yandex search engine. The method involves constructing hypothetical derivatives based on productive stems and affixes involved in word formation, their automated verification via the Yandex Search API, and the creation of a dataset containing the identified derivatives, links to corresponding web pages, and snippets in which these derivatives appear. Additionally, a classification of the collected contexts is proposed based on their stylistic affiliation, using a predefined algorithm that analyzes web link fragments. This will allow for a better understanding of how non-standard derivatives are distributed across different domains. The proposed method enables a more objective assessment of the derivational potential of lexemes, particularly adjectives describing psychological characteristics of a person, as well as an investigation of the reasons behind derivational lacunarity in the formation of denominal personal nouns. The research has applied significance for the automated compilation of dynamic word-formation resources and the expansion of computational lexicography capabilities.

**Keywords:** derivational potential, derivative, Yandex Search, parsing, search engines, Natural Language Processing

**Reference for citation:** Vasileva E. V. Issues in Studying Derivational Potential Using Modern Search Engines: Automated Selection of Derivatives via Yandex // Computational Linguistics and Computational Ontologies. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 78-92. DOI: 10.17586/3033-5582-2025-9-78-92.

### Reference

- [1] Tihonov A.N. Slovoobrazovatel'nyj slovar' russkogo yazyka. In 2 vols. M.: AST; Astrel, 2008. (In Russian)
- [2] Tihonov A.N. Osnovnye ponyatiya russkogo slovoobrazovaniya // Tihonov A.N. Slovoobrazovatel'nyj slovar' russkogo yazyka. In 2 vols. Vol. 1. M.: AST; Astrel', 2008. P. 24-57. (In Russian)
- [3] Slovar' sovremennogo russkogo literaturnogo yazyka. In 17 vols. / Akademiya nauk SSSR, Institut yazykoznaniiya. M.; L.: Izdatel'stvo AN SSSR, 1949–1965. (In Russian)
- [4] Slovar' russkogo yazyka. In 4 vols. / Ed. A. P. Evgen'eva. M.: Russkiĭ yazyk, 1981–1984. (In Russian)
- [5] Lebedeva N.B. Eksperimental'noe issledovanie proizvodnogo slova: sootnoshenie vyvodimosti kak slovoobrazovatel'nogo priznaka i uzual'nosti kak leksicheskogo priznaka // Novye yavleniya v slavyanskom slovoobrazovanii: sistema i funkcionirovanie. Proceedings of the 11th International Scientific Conference of the Commission on Slavic Word Formation under the International Committee of Slavists / ed. by E.V. Petruhina. M.: Maks Press, 2010. P. 390-399. (In Russian)

- [6] Lyu S., Katyshev P.A. Faktory realizacii derivacionnogo potenciala znamenatel'nyh odinochnyh slov // *Mezhdunarodnyj aspirantskij vestnik. Russkij yazyk za rubezhom*. 2024. No. 3. P. 55-59. DOI: 10.37632/PI.2024.15.14.010 (In Russian)
- [7] Shkuropackaya M.G. Metodika issledovaniya sistemnoj i empiricheskoj realizacii slovoobrazovatel'nogo tipa v uzual'noj i potencial'noj leksike // *Mir nauki, kul'tury, obrazovaniya*. 2013. No. 6(43). P. 343-345. (In Russian)
- [8] *Russkaya grammatika: In 2 vols.* / Ed. N.Yu. Shvedova. M.: Nauka, 1982. Vol. 1. 783 p. (In Russian)
- [9] Vasileva E.V., Tashlykova M.B. Prilagatel'nye leksiko-semanticheskoj grupy 'psikhicheskie kharakteristiki cheloveka': problema klassifikatsii // *V mire nauchnykh otkrytij*. 2014. No. 11–10(59). P. 3922-3938. (In Russian)
- [10] Zaliznyak A.A. *Grammaticheskiĭ slovar' russkogo yazyka: Slovoizmenenie*. M.: Russkiĭ yazyk, 2008. 792 p. URL: <https://gramdict.ru/> (accessed date: 27.03.2025). (In Russian)
- [11] Vypolnenie poiskovyh zaprosov s pomoshch'yu API v2 v sinhronnom rezhime // Dokumentaciya Yandex Search API. Yandex Cloud. URL: [https://yandex.cloud/ru/docs/search-api/operations/web-search-sync?utm\\_referrer=about%3Ablank](https://yandex.cloud/ru/docs/search-api/operations/web-search-sync?utm_referrer=about%3Ablank) (accessed date: 27.03.2025). (In Russian)
- [12] Python Software Foundation. XML processing modules — xml. etree. Element Tree. URL: <https://docs.python.org/3/library/xml.etree.elementtree.html> (accessed date: 27.03.2025).

## Динамика тем научных статей в корпусе текстов по компьютерной и корпусной лингвистике

О. А. Митрофанова

Санкт-Петербургский государственный университет

`o.mitrofanova@spbu.ru`

### Аннотация

Цель исследования заключается в анализе изменений тематики научных статей в корпусе текстов по компьютерной и корпусной лингвистике (ТКиКЛ) с применением современных методов тематического моделирования. Проведенные эксперименты основываются на применении комбинированного алгоритма BERTopic для анализа научных текстов, что позволило получить более специфичные темы по сравнению с традиционными методами. Значимость работы — в формировании методологии для семантического поиска научной информации и отслеживания тенденций в области компьютерной и корпусной лингвистики, что может быть использовано для совершенствования наукометрических инструментов. Результаты исследования демонстрируют динамику изменений научных интересов в области компьютерной и корпусной лингвистики под влиянием цифровизации общества и технологического прогресса.

**Ключевые слова:** компьютерная лингвистика, корпусная лингвистика, корпус текстов, динамическое тематическое моделирование, BERTopic

**Библиографическая ссылка:** Митрофанова О. А. Динамика тем научных статей в корпусе текстов по компьютерной и корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 93–99. DOI: 10.17586/3033-5582-2025-9-93-99.

### 1. Введение

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

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

## 2. Корпус текстов по компьютерной и корпусной лингвистике

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

Тексты корпуса распределены по годам и представлены в неразмеченном и лемматизированном виде. В корпусе есть следующие типы информации: при сохранении общей структуры статей (авторство, заголовок, набор ключевых выражений, аннотация, текст статьи) проведена автоматическая разметка ключевых выражений с применением библиотеки *RuTermExtract* [13], генерация аннотаций с помощью моделей экстрактивной и абстрактивной суммаризации в библиотеке *sumy* [14] и с помощью модели *ruT5* [15], систематизация и разметка терминологизированных именованных сущностей, мультимодальное тематическое моделирование корпуса с автоматическим назначением меток тем, а также экспертная разметка рубрик в корпусе. В частности, были выделены следующие рубрики: 1) *Общие вопросы корпусной лингвистики*, 2) *Создание, разработка и применения корпусов*, 3) *Статистические исследования на материале корпусов*, 4) *Корпусы и лексикография*, 5) *Морфология и синтаксис в корпусах*, 6) *Семантика в корпусах*, 7) *Параллельные корпусы и машинный перевод*, 8) *Обучающие корпусы*, 9) *Исторические корпусы*, 10) *Речевые и мультимедийные корпусы*, 11) *Корпусы художественных текстов*. Имеющаяся в корпусе хронологическая разметка позволяет получить данные об изменениях в тематике статей, алгоритмом выбора для анализа данных является динамическое тематическое моделирование (Dynamic Topic Modelling).

## 3. Динамическое тематическое моделирование

Процедуры тематического моделирования направлены на построения семантически интерпретируемой модели корпуса текстов, в которой устанавливаются связи между документами, их словарем и темами (скрытыми факторами). Каждый текст в корпусе с той или иной вероятностью соотносится с одной или несколькими темами, которые могут пересекаться [16]. Из множества алгоритмов тематического моделирования, допускающих расширение до мультимодальных версий (LSA, NMF, pLSA, LDA, LDA2Vec, Top2Vec и т. д.) мы выбрали BERTopic [17; 18]. BERTopic — комбинированная модель, основанная на трансформере BERT, к векторам которой применяется снижение размерности UMAP, кластеризация HDBSCAN, ранжирование слов-тематизаторов по метрикам c-TF-IDF и MMR.

На первом этапе работы с BERTopic проводится векторизация корпусных данных, фильтрация словаря по частоте лемм и распространенности в текстах корпуса. Параметры обучения тематической модели подбираются эмпирически (минимальный объем темы

— 10 слов-тематизаторов, выделение в корпусе  $n$ -грамм и т. д.). В модель входит 14 тем, соотносимых с группами статей из корпуса ТКиКЛ. Примеры тем: (0) *орган, услуга, правительство, портал, власть...* (1) *коллакат, граф, кластеризация, онтология...* (2) *жест, реплика, УРК, разговор, ЭДЕ...* (8) *стих, слоговой, метрический...* (9) *житие, цитата, агиографический, СКАТ, рукопись...* (12) *латышский, румынский, транслитерация...* Для каждой темы модель выбирает наиболее типичный документ, например, для темы (10) это статья {Рогозина Е. А. «Разметка содержательной структуры житийных текстов в корпусе агиографических текстов СКАТ» // Труды международной конференции «Корпусная лингвистика — 2008». СПб., 2008}. Внутри темы слова-тематизаторы ранжируются по значению меры ассоциации: (9) *житие* (0.129), *цитата* (0.082), *агиографический* (0.071), *СКАТ* (0.068), *рукопись* (0.052), *житийный* (0.049), *словоуказатель* (0.035), *склонение* (0.031), *Алексеев* (0.030), *Евангелие* (0.029) и т. д. Визуализация результатов тематического моделирования представлена на рис. 1, показывающем цепочечную организацию тем с их незначительным наложением друг на друга.

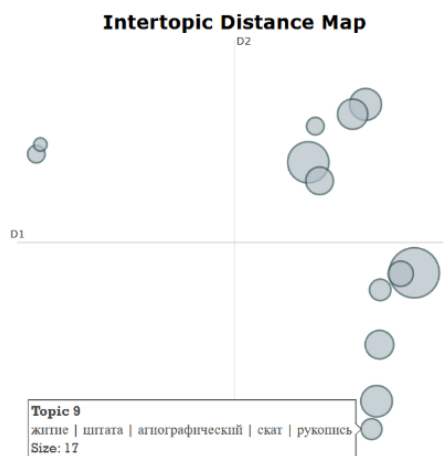


Рис. 1. Визуализация тематической модели ТКиКЛ

На втором этапе для построения динамической тематической модели в режиме «Topics over Time» определяются временные точки, относительно которых будут оцениваться изменения в наполнении тем. В корпусе ТКиКЛ данными точками являются годы проведения конференций. Результат представлен на рис. 2.

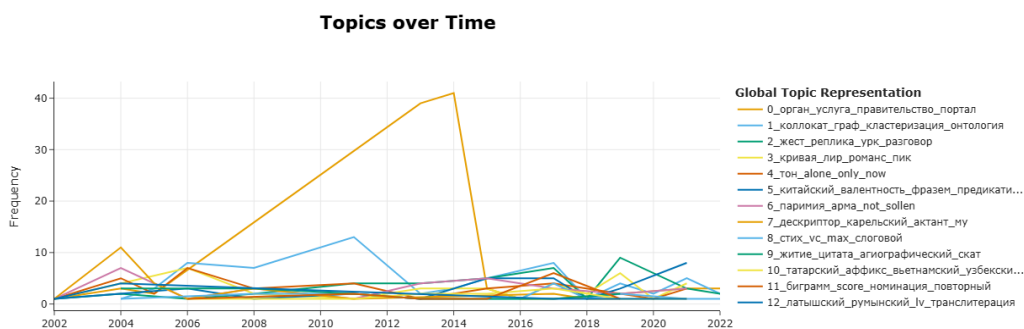


Рис. 2. Динамическое тематическое моделирование корпуса ТКиКЛ

Можно рассмотреть изменение во времени каждой из тем по отдельности, например, на рис. 3 отражена динамика темы (0) *орган, услуга, правительство,..*: в 2002 г. она была сосредоточена на проблемах электронных библиотек (*читальный, зал, навигация,..*), в 2013–2014 гг. доминировала проблематика электронного правительства (*услуга, орган, портал, власть, гражданин, правительство,..*), в 2015 г. фокус тематического наполнения сместился на формальные онтологии, в 2016 г. — на компьютерные тезаурусы (*YARN, WordNet, RussNet, синсет,..*), а к 2022 г. адаптировалась к проблематике чат-ботов и клиент-ориентированной коммуникации (*бот, пациент, настроение,..*).

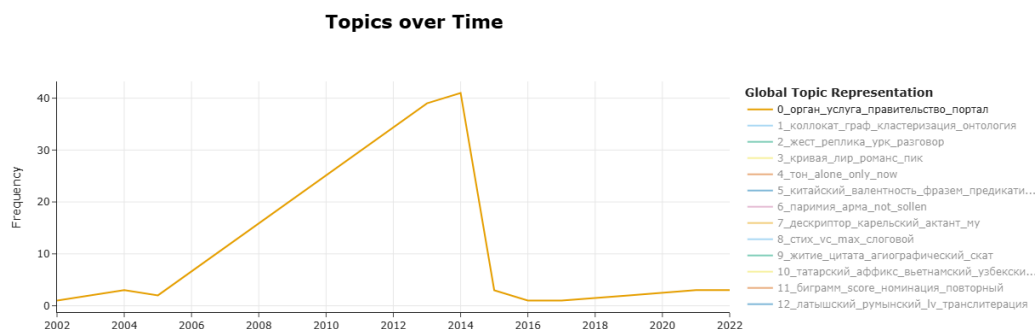


Рис. 3. Динамика темы (0) орган, услуга, правительство,..

На рис. 4 показаны изменения темы (2) *жест, реплика, УРК,..* в 2002 г. в ее составе были термины *фонетика, звукозапись*, в 2004 г. — *диалоговый, тестирование*, в 2013 г. — *коррекция, артикуляторный, самоисправление*, в 2017 г. — *УРК, метакоммуникация, реплика, пересказ* и т. д.

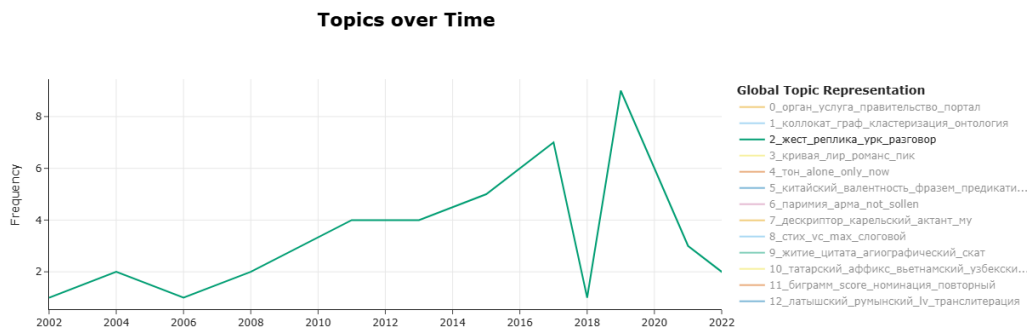


Рис. 4. Динамика темы (2) жест, реплика, УРК,..

Рис. 5 демонстрирует эволюцию темы (9) *жизние, цитата, агиографический,..*: в 2004 г. тема была связана с исследованием диахронических и диалектных корпусов текстов в целом (*ижгорский, песня, рукопись, старолатышский, житие,..*), в 2008 г. акцент сместился в сторону корпуса агиографических текстов и построения словоуказателей (*жизние, агиографический, СКАТ, словоуказатель,..*), к 2011 г. — в сторону исследования цитат и морфологической аннотации диахронических корпусов текстов (*жизние, цитата, СКАТ, агиографический, склонение,..*).

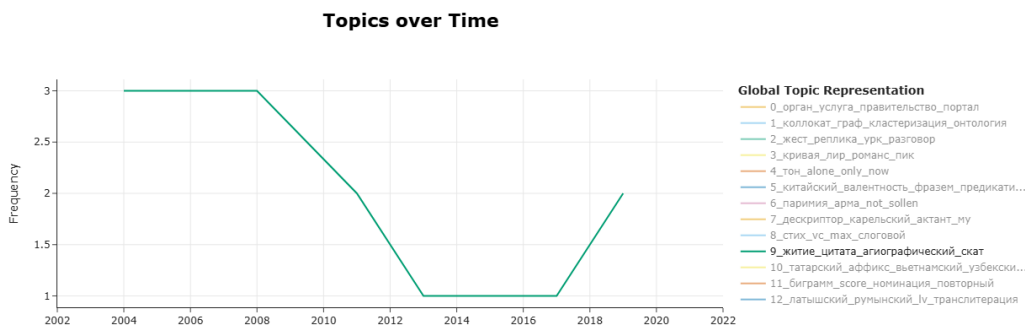


Рис. 5. Динамика темы (9) жите, цитата, агиографический,...

Тем самым, в ходе экспериментов были применены расширенные возможности библиотеки BERTopic, прежде всего, возможность настройки обучения тематических моделей с учетом хронологических меток документов, а также режим визуализации как всего набора тем, так и групп тем и отдельных тем по выбору пользователей.

#### 4. Заключение

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

- комбинированный алгоритм тематического моделирования BERTopic, примененный к данным корпуса, генерирует более конкретные по наполнению темы, чем алгоритмы LSA, LDA, NMF [10], при этом темы являются более специфичными по сравнению с рубриками конференций (см. раздел 2);
- динамическое моделирование, проведенное средствами BERTopic в режиме «Topics over Time» позволяет отследить изменение фокуса внимания исследователей, работающих над узкими темами, например, в области диахронической корпусной лингвистики, а также зарегистрировать резкие скачки в тематике статей, связанные с социально-политическими процессами и технологическим прогрессом, прежде всего, с цифровизацией общественной жизни.

#### Литература

- [1] Диалог-21. Компьютерная лингвистика и интеллектуальные технологии. URL: <https://dialogue-conf.org> (дата обращения: 08.07.2025).
- [2] AINL. Artificial Intelligence and Natural Language. URL: <https://ainlconf.ru/> (дата обращения: 08.07.2025).
- [3] AIST. International Conference on Analysis of Images, Social Networks and Texts. URL: <https://aistconf.org/> (дата обращения: 08.07.2025).
- [4] SPECOM. International Conference on Speech and Computer. URL: <https://specom.nw.ru/> (дата обращения: 08.07.2025).
- [5] FRUCT. URL: <https://fruct.org/> (дата обращения: 08.07.2025).
- [6] Dimensions. URL: <https://www.dimensions.ai/> (дата обращения: 08.07.2025).
- [7] OpenAlex. URL: <https://openalex.org/> (дата обращения: 08.07.2025).
- [8] ResearchRabbit. URL: <https://www.researchrabbit.ai/> (дата обращения: 08.07.2025).
- [9] Митрофанова О.А., Адамова М.А., Букреева Л.А., Зернова А.К., Литвинова А.А., Павликова В.С., Сологуб П.С. Корпус текстов по корпусной лингвистике: состав и этапы формирования // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции

- «Интернет и современное общество», IMS-2024, Санкт Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 13-29. DOI: 10.17586/2541-9781-2024-8-13-29
- [10] Митрофанова О.А., Голубев Р.В., Гусьяцкая П.А., Макеев К.В., Плюснина Е.А., Сухан Д.Д., Трошина А.В., Уткина А.А. Разработка тематических моделей корпуса по корпусной лингвистике с автоматическим назначением меток тем // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 30-44. DOI: 10.17586/2541-9781-2024-8-30-44
- [11] Сухан Д.Д., Плюснина Е.А. Метаразметка и визуализация данных в корпусе текстов по корпусной лингвистике // Компьютерная лингвистика и вычислительные онтологии. Выпуск 8 (Труды XXVII Международной объединенной научной конференции «Интернет и современное общество», IMS-2024, Санкт-Петербург, 24–26 июня 2024 г. Сборник научных статей). СПб.: Университет ИТМО, 2024. С. 45-60. DOI: 10.17586/2541-9781-2024-8-45-60
- [12] Митрофанова О.А., Адамова М.А., Букреева Л.А., Голубев Р.В., Гусьяцкая П.А., Зернова А.К., Литвинова А.А., Макеев К.В., Павликова В.С., Плюснина Е.П., Сологуб П.С., Сухан Д.Д., Трошина А.В., Уткина А.А. Интеллектуальный анализ данных в корпусе текстов по корпусной и компьютерной лингвистике // International Journal of Open Information Technologies. 2024. Т. 12. № 12. С. 11-26.
- [13] RuTermExtract. URL: <https://pypi.org/project/rutermextract/> (дата обращения: 08.07.2025).
- [14] sumy. URL: <https://github.com/miso-belica/sumy> (дата обращения: 08.07.2025).
- [15] ruT5. URL: <https://huggingface.co/ai-forever/ruT5-base> (дата обращения: 08.07.2025).
- [16] Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URL: <http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf> (дата обращения: 08.07.2025).
- [17] BERTopic. URL: <https://github.com/MaartenGr/BERTopic> (дата обращения: 08.07.2025).
- [18] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint. 2022. URL: <https://arxiv.org/abs/2203.05794/> (дата обращения: 08.07.2025).

## Topic Dynamics of Scientific Articles in the Text Corpus on Computational and Corpus Linguistics

O. A. Mitrofanova

Saint-Petersburg State University

The purpose of the study is to analyze changes in the subject matter of scientific articles in the text corpus on computational and corpus linguistics using modern methods of topic modeling. The experiments carried out are based on the complex BERTopic algorithm for scientific texts analysis, which allowed us to obtain more specific topics as compared with traditional methods. The significance of our work deals with the development of a methodology for semantic search of scientific information and tracking trends in the field of computational and corpus linguistics, which can be used to improve scientometric tools. The study demonstrates important changes in the focus of scientific interests of specialists in the field of computational and corpus linguistics related to society digitalization of and technological progress.

**Keywords:** computational linguistics, corpus linguistics, text corpus, dynamic topic modeling, BERTopic

**Reference for citation:** Mitrofanova O. A. Topic Dynamics of Scientific Articles in the Text Corpus on Computational and Corpus Linguistics // Computational Linguistics and Computational Ontologies. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 93-99. DOI: 10.17586/3033-5582-2025-9-93-99.

## Reference

- [1] Dialogue-21. Computational Linguistics and Intellectual Technologies. URL: <https://dialogue-conf.org> (accessed date: 08.07.2025).
- [2] AINL. Artificial Intelligence and Natural Language. URL: <https://ainlconf.ru/> (accessed date: 08.07.2025).
- [3] AIST. International Conference on Analysis of Images, Social Networks and Texts. URL: <https://aistconf.org/> (accessed date: 08.07.2025).
- [4] SPECOM. International Conference on Speech and Computer. URL: <https://specom.nw.ru> (accessed date: 08.07.2025).
- [5] FRUCT. URL: <https://fruct.org/> (accessed date: 08.07.2025).
- [6] Dimensions. URL: <https://www.dimensions.ai/> (accessed date: 08.07.2025).
- [7] OpenAlex. URL: <https://openalex.org/> (accessed date: 08.07.2025).
- [8] ResearchRabbit. URL: <https://www.researchrabbit.ai/> (accessed date: 08.07.2025).
- [9] Mitrofanova O.A., Adamova M.A., Bukreeva L.A., Zernova A.K., Litvinova A.A., Pavlikova V.S., Sologub P.S. Text Corpus on Corpus Linguistics: Composition and Stages of Formation // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). St. Petersburg: ITMO University, 2024. P. 13-29. DOI: 10.17586/2541-9781-2024-8-13-29 (In Russian)
- [10] Mitrofanova O.A., Golubev R.V., Gusyatskaya P.A., Makeev K.V., Pliusnina E.A., Sukhan D.D., Troshina A.V., Utkina A.A. Development of Topic Models of the Corpus on Corpus Linguistics with Automatic Topic Labels Assignment // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). St. Petersburg: ITMO University, 2024. P. 30-44. DOI: 10.17586/2541-9781-2024-830-44 (In Russian)
- [11] Sukhan D.D., Pliusnina E.A. Meta Tagging and Visualization for the Corpora Linguistics Texts Corpora // Computational Linguistics and Computational Ontologies. Vol. 8 (Proceedings of the XXVII International Joint Scientific Conference «Internet and Modern Society», IMS-2024, St. Petersburg, June 24–26, 2024). St. Petersburg: ITMO University, 2024. P. 45-60. DOI: 10.17586/2541-9781-2024-8-45-60 (In Russian)
- [12] Mitrofanova O.A., Adamova M.A., Bukreeva L.A., Golubev R.V., Gusyatskaya P.A., Zernova A.K., Makeev K.V., Litvinova A.A., Pavlikova V.S., Pliusnina E. P., Sologub P. S., Sukhan D. D., Troshina A. V., Utkina A. A. Data Mining in the Text Corpus on Corpus and Computational Linguistics// International Journal of Open Information Technologies. 2024. Vol. 12. No. 12. P. 11-26. (In Russian)
- [13] RuTermExtract. URL: <https://pypi.org/project/rutermextract/> (accessed date: 08.07.2025).
- [14] sumy. URL: <https://github.com/miso-belica/sumy> (accessed date: 08.07.2025).
- [15] ruT5. URL: <https://huggingface.co/ai-forever/ruT5-base> (accessed date: 08.07.2025).
- [16] Vorontsov K.V. Probabilistic topic modeling: ARTM regularization theory and BigARTM open source library. URL: <http://www.machinelearning.ru/wiki/images/d/d5/Voron17survey-artm.pdf> (accessed date: 08.07.2025).
- [17] BERTopic. URL: <https://github.com/MaartenGr/BERTopic> (accessed date: 08.07.2025).
- [18] Grootendorst M. BERTopic: Neural topic modeling with a class-based TF-IDF procedure // arXiv preprint. 2022. URL: <https://arxiv.org/abs/2203.05794> (accessed date: 08.07.2025).

## **Оценка лингвистической компетенции больших языковых моделей на материале корпуса согласовательной вариативности**

К. А. Студеникина, Е. А. Лютикова, А. А. Герасимова

МГУ им. М. В. Ломоносова

xeanst@gmail.com, lyutikova2008@gmail.com,  
anastasiagerasimova432@gmail.com

### **Аннотация**

Данное исследование нацелено на оценку сходства и различия лингвистической компетенции носителей языка и больших языковых моделей (БЯМ). Материалом для сравнения служит созданный нами корпус вариативного согласования (КВас). Он содержит 6803 предложения с оценками по шкале от 1 до 7, полученными при проведении 20 синтаксических экспериментов по изучению вариативного согласования. Корпус фиксирует средние оценки русских предложений с различными условиями согласования, полученные от носителей языка, и позволяет выяснить, как БЯМ справляются с градуальной оценкой приемлемости. Мы приводим результаты тестирования четырех моделей: преимущественно русскоязычных YandexGPT 5 Pro и GigaChat 2 Max, а также мультиязычных Llama 3.3 70B и Mistral Large. Для каждой модели было опробовано два режима тестирования: zero-shot, содержащий только инструкцию, и few-shot, где добавлены тренировочные предложения и их оценки. Поскольку данные в КВас демонстрируют различный уровень приемлемости в зависимости от экспериментальных условий, подсчет только средней ошибки для предсказанных моделями оценок будет недостаточно показателен. Нами разработана метрика, позволяющая оценить, какая доля контрастов между экспериментальными условиями, релевантными для людей, выявляется с помощью БЯМ. Результаты показывают, что среднее значение ошибки меньше для предложений без вариативного согласования. Примеры с согласовательной вариативностью оказываются сложнее для БЯМ. Качество моделей проседает для одного и того же типа конструкций — сочинения. Модели значительно лучше определяют контрасты для конструкций с постпозитивными относительными предложениями, количественными конструкциями и управляющими квантификаторами. Наиболее точное совпадение при выделении значимых контрастов по сравнению с носителями языка демонстрирует Mistral. Наименьшее количество контрастов выделяет модель Llama. Русскоязычные модели занимают промежуточную позицию, при этом YandexGPT превосходит GigaChat. Добавление примеров в режиме few-shot улучшает среднее качество, но различие незначительно. Результаты показывают, что качество решения задачи градуальной оценки приемлемости сильно отличается для разных классов лингвистических феноменов. Сравнение моделей демонстрирует, что для достижения лучшего качества наиболее важным оказывается количество параметров модели, которое, однако, может быть компенсировано объемом русскоязычных данных при обучении.

**Ключевые слова:** языковая способность, синтаксис, согласование, обработка естественного языка, большие языковые модели, бенчмарк

**Библиографическая ссылка:** Студеникина К. А., Лютикова Е. А., Герасимова А. А. Оценка лингвистической компетенции больших языковых моделей на материале корпуса согласовательной вариативности // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 100-119. DOI: 10.17586/3033-5582-2025-9-100-119.

## 1. Введение

Современные БЯМ демонстрируют уровень обработки естественного языка, близкий к человеческому. Они успешно применяются для понимания и генерации текстов в диалоге с пользователем. Сравнение БЯМ осуществляется с помощью бенчмарков — наборов для их оценки и тестирования. Большинство бенчмарков призвано оценить семантические и прагматические способности БЯМ: владение верной фактической информацией, выстраивание причинно-следственных связей. Например, бенчмарк SuperGLUE [1] включает такие задачи, как разрешение семантической неоднозначности, логический вывод и поиск ответа на вопрос, а бенчмарк MMLU [2] содержит вопросы с множественным выбором из 57 предметных областей, различных по уровню сложности.

Помимо адекватности с точки зрения содержания, языковое поведение БЯМ должно быть приближено к человеческому по формальным лингвистическим критериям. Носителям языка свойственна способность не только порождать правильные (приемлемые) высказывания, но и отличать их от предложений с ошибками (неприемлемых). Аналогичные компетенции в идеальном случае ожидаются и от БЯМ. Проблема состоит в том, что оценка языковых способностей моделей всегда происходит бинарно. Так, бенчмарк CoLA [3] подразумевает классификацию предложений на приемлемые и неприемлемые, а бенчмарк BLiMP [4] — выбор более приемлемого предложения из минимальной пары. В теоретической лингвистике приемлемость предложений считается градуальным понятием, которое складывается из двух аспектов: (i) грамматической корректности, т. е. соответствия правилам и ограничениям врожденной грамматики; (ii) факторов употребления: частотности слов, длины предложения, синтаксической сложности и т. д. [5; 6; 7; 8]. Существование промежуточных оценок приемлемости предполагает, что отношение к приемлемости как к шкале является ключевой характеристикой языковой компетенции.

Данное исследование направлено на то, чтобы ответить на вопрос, способны ли БЯМ оценивать приемлемость предложений по градуальной шкале. Для достижения этой цели используем хорошо изученный феномен русского морфосинтаксиса — вариативное согласование. Правила согласования, будучи довольно однозначными в стандартном случае, допускают множество альтернативных стратегий с неканоническими контролерами, такими как количественные и сочиненные группы. Вариативность согласования позволяет нам анализировать влияние различных факторов на выбор конкретной стратегии согласования как для языковых моделей, так и для людей-носителей.

Структура статьи: в разделе 2 — разработанный нами набор данных, в разделе 3 обсуждается эксперимент с тестированием БЯМ и его результаты, в разделе 4 подводятся итоги.

## 2. Корпус вариативного согласования

Нами был создан новый бенчмарк под названием КВаС (корпус вариативного согласования), предназначенный для тестирования языковых моделей на задаче градуальной оценки приемлемости в русском языке. Согласование является одним из видов синтаксической связи между словами в предложении. Одна единица, так называемая мишень согласования, копирует некоторые признаки другой единицы, контролера

согласования. В стандартном случае признаки контролера согласования, такие как род, лицо, число, определяются однозначно и копируются мишенью. Это позволяет точно разграничить грамматичные (1) vs. неграмматичные (2) предложения. Вариативное согласование возникает в том случае, если возможно неоднозначное вычисление признаков мишени. Это происходит при наличии нескольких потенциальных контролеров: в конструкциях с определительным предложением, где в качестве контролера может выступать именная вершина и союзное слово *кто* (3), а также в сочинительных конструкциях, где контролерами оказываются конъюнкты (4). Кроме того, вариативность наблюдается при неканоническом контролере: в конструкциях с количественными существительными (5) и управляющими квантификаторами (6). Тогда выбор стратегии согласования определяется тем, насколько данные конструкции демонстрируют свойства количественных групп с числительными, а насколько — свойства именных групп с генитивным зависимым.

- (1) В субботу **Марина** полила цветы и пропылесосила.
- (2) \*По вторникам **дедушки** забираем детей из школы.
- (3) **Те, кто** заплатит / заплатят за подписку, продлят доступ.
- (4) **Петя и я** идем / ?иду / ?идут / \*идет домой.
- (5) На льду **свалилась** / ?свалились / \*свалилось **уйма людей**.
- (6) **Двое из нас** придут / ?придет / \*придем в гости.

В качестве данных для корпуса были использованы результаты экспериментальных исследований вариативного согласования в русском языке, проведенных в ходе «Практикума по экспериментальному синтаксису»<sup>4</sup> в 2022–2024 гг. и опубликованных в Базе анкет феноменов с согласовательной вариативностью<sup>5</sup>. В каждом исследовании использовался факторный подход, при котором две или более независимых переменных (фактора) изменяются для изучения их влияния на зависимую переменную. При этом один фактор всегда соответствовал стратегии согласования, а зависимая переменная была оценкой приемлемости. Другие факторы в экспериментах описывали различные особенности предложений, которые могли повлиять на выбор согласования: например, порядок слов, семантику контролера и мишени. Комбинирование этих факторов позволяет выяснить, изменяется ли приемлемость стратегии согласования в зависимости от грамматического окружения.

Во всех исследованиях применялась методика извлечения суждений при помощи шкалы Ликерта от 1 до 7 [9] — стандартный инструмент в области экспериментального синтаксиса, позволяющий количественно оценить приемлемость различных структур предложений. Данный метод является наиболее мощным в сравнении с другими методами извлечения суждений, что было показано в экспериментах с симуляцией выборов [10]. В отличие от более простых методов, таких как использование звездочки (\*) или вопросительного знака (?), обычно применяемых в формальном синтаксисе для обозначения степени приемлемости, шкала Ликерта представляет собой более тонкий и систематический исследовательский инструмент. Нечетное количество делений на шкале допускает небинарные суждения, что особенно важно для оценки влияния синтаксических манипуляций со стимулами.

Чтобы уменьшить вклад лексического содержания на приемлемость, стимулы были распределены по листам таким образом, чтобы каждое предложение (лексикализация) появлялось в каждом листе при одном условии (определенной комбинации значений факторов). Пример лексикализации из эксперимента с относительным предложением

<sup>4</sup> Курс проводится к.ф.н. А. А. Герасимовой и д.ф.н. Е. А. Лютиковой на кафедре теоретической и прикладной лингвистики филологического факультета МГУ им. М. В. Ломоносова. Более подробная информация доступна на сайте Московской группы экспериментального синтаксиса, URL: <https://expsynt.com/>.

<sup>5</sup> URL: <https://expsynt.com/table.html/>.

с легкой или именной вершиной приведен в (3) и (7). В эксперименте изучалось вариативное согласование, возникающее при наличии конфликта признака числа между вершиной *те / те пользователи* и относительным местоимением *кто*. Рассматривалось два фактора: стратегия согласования и структура вершины.

(7) **Те (пользователи), кто заплатит / заплатят** за подписку, продлят доступ.

Наряду со стимулами, в каждый эксперимент включались предложения-филлеры без вариативности в согласовании. Грамматичные примеры представляли собой полностью правильные предложения, в то время как неграмматичные содержали ошибки согласования. Эти два типа филлеров используются в качестве порога (не)грамматичности при анализе оценок в экспериментах [11]. Филлеры также могут служить основой для проверки способности модели оценивать невариативное согласование.

Корпус вариативного согласования для русского языка доступен в гитхаб-репозитории<sup>6</sup>. Его объем составляет 6803 предложения. Корпус включает результаты 20 экспериментальных исследований, в каждом из них приняло участие от 66 до 137 человек (среднее количество участников — 96, стандартное отклонение — 20). Для каждого эксперимента был проведен отсев респондентов с отклоняющимися значениями ответов. Поиск осуществлялся с помощью фильтров, традиционно используемых в экспериментальном синтаксисе: (i) оценки тренировочных предложений и филлеров не должны сильно отклоняться от эталона (2 для неграмматичных, 6 для грамматичных), (ii) время ответа не должно быть меньше 300 миллисекунд, (iii) респондент не должен использовать преимущественно одну или две оценки на шкале, (iv) респондент должен правильно отвечать на контрольные вопросы по содержанию предложений [12]. После процедуры отсева выбросов в каждом эксперименте были удалены ответы от 5 до 15 участников.



Рис. 1. Распределение контекстов для бенчмарка KVaS

Распределение контекстов в корпусе представлено на рис. 1. Наиболее многочисленную группу составили примеры с несколькими потенциальными контролерами (3741 предложение): конструкции с постпозитивным определительным предложением (689 предложений) и сочинительные конструкции (3052 предложения). Они позволяют

<sup>6</sup> URL: <https://github.com/Xeanst/KVaS>.

проанализировать влияние большого количества факторов, поэтому представляют наибольший интерес. Также было собрано значительное количество примеров с неканоническим контролером (2410 предложений): конструкции с количественными существительными (884 предложения) и управляющими квантификаторами (1526 предложений). Контексты, не демонстрирующие вариативность, представляют наименьшую группу (652 предложения): грамматичные конструкции с корректным согласованием (326 предложений) и неграмматичные конструкции с ошибками в согласовании (326 предложений).

Корпус содержит следующие данные: код эксперимента (автора и год) в поле *experiment*, предложение в поле *sentence*, усредненную по всем участникам оценку в поле *response*, округленную оценку в поле *response\_round*, тип конструкции для предложений с вариативным согласованием и степень грамматичности для предложений с невариативным согласованием (филлеров) в поле *subtype*, наличие вариативного согласования в поле *type*, номер лексикализации в поле *lexicalization*. Фрагмент корпуса представлен в табл. 1.

Таблица 1. Фрагмент корпуса вариативного согласования для русского языка

experiment	sentence	response	response_round	subtype	type	lexicalization
Davidjuk_2023	Я и Максим прогуляет последний урок.	2.56	3	coordination	stimul	32
Belova_2023	Шестеро из нас справятся с заменой лампочек.	5.27	5	quantifier	stimul	11
Krainova_2023	Эти миллион марок хранились в альбомах.	3.88	4	quantitative_nouns	stimul	28
Dorofeeva_2023	Каждую мать, кто явилась на собрание родителей, спрашивает классный руководитель.	3.43	3	relative_clauses	stimul	34
Davidjuk_2023	И Максим, и я прогуляют последний урок.	2.70	3	coordination	stimul	32
Vrubel_2022	На солнечной опушке в лесу спала и принцесса, и рыцарь.	3.00	3	coordination	stimul	5
Pasko_2024	На аукционе треть скульптур покупаешь коллекционер.	1.81	2	bad	filler	505
Danilova_2023	Клиент попросил вегетарианское или постное меню.	6.03	6	good	filler	51

Помимо этого, каждому предложению соответствует грамматическая разметка, которая описывает конкретные факторы, потенциально влияющие на его степень приемлемости. Например, для сочиненные конструкций это тип конъюнктов (два местоимения / два существительных / одно местоимение и одно существительное), тип союза (конъюнктивный *и* / дизъюнктивный *или*, одинарный / двойной), линейное расположение подлежащего и сказуемого (прямой порядок SV/ обратный порядок VS), соотношение рода конъюнктов (оба мужского рода / оба женского рода / один мужского рода и один женского), симметричность предиката (симметричный / несимметричный), одушевленность конъюнктов (одушевленные / неодушевленные). В табл. 2 приведен пример разметки для данных одного из экспериментов [13]. К информации из основного корпуса добавляется тип конъюнктов в поле *conjuncts* (*я* + имя собственное / имя собственное + *я*), стратегия согласования в поле *agreement* (1 лицо единственное число / 3 лицо единственное число / 1 лицо множественное число / 3 лицо множественное число), тип союза в поле *conjunction* (*и*, *и..и*, *или*, *или..или*) и порядок слов в поле *order* (подлежащее + сказуемое / сказуемое + подлежащее).

Таблица 2. Фрагмент грамматической разметки для одного эксперимента [13]

sentence	response	response – round	subtype	type	conjuncts	agreement	conjunction	order
И я, и Гриша сыграем шахматную партию.	4.50	4	coordination	stimul	I_noun	1pl	and_and	SV
Влад и я покажу короткую дорогу.	2.44	2	coordination	stimul	noun_I	1sg	and	SV
Дима или я послушают это голосовое сообщение.	4.88	5	36	stimul	noun_I	3pl	or	SV
Или Митя, или я составит этот длинный список.	2.20	2	coordination	stimul	noun_I	3sg	or_or	SV
Лёва оформит этот деловой договор.	6.44	6	good	filler	NaN	NaN	NaN	NaN

Бенчмарк КВаС имеет свои преимущества и недостатки по сравнению с другими корпусами оценки приемлемости. С одной стороны, существующие корпуса, такие как CoLA и BLiMP, охватывают более широкий спектр грамматических явлений, чем КВаС, что позволяет определить, какие грамматические феномены представляют наибольшую сложность для БЯМ. С другой стороны, они осуществляют только бинарную оценку языковой компетенции, которая может показаться чрезмерно упрощенной и недостаточно детализированной. КВаС является первым набором данных для градуальной оценки приемлемости по шкале от 1 до 7. Он позволяет проверить, способны ли языковые модели,

обученные на неразмеченных текстовых данных, обнаруживать тонкие морфосинтаксические и семантические различия между предложениями. Поскольку каждый пример в бенчмарке КВаС был оценен большим количеством респондентов, имеющаяся разметка является надежной и отражает обобщенную лингвистическую компетенцию носителей русского языка. Следовательно, КВаС может быть использован для изучения того, насколько согласованность оценок БЯМ для различных лексикализаций соответствует результатам носителей русского языка.

Кроме того, наш корпус позволяет ранжировать БЯМ на основе уровня их языковой компетенции, особенно в области грамматической вариативности. В следующем разделе мы представим результаты эксперимента, в котором различные БЯМ решали задачу градуальной оценки приемлемости. Оценки по шкале Ликерта, полученные от БЯМ, будут сопоставлены с эталонными человеческими оценками, содержащимися в бенчмарке КВаС.

### 3. Градуальная оценка приемлемости с помощью БЯМ

#### 3.1. Метод

Ранее для оценки лингвистической компетенции БЯМ использовались либо тонкая настройка для задачи бинарной классификации по приемлемости, как в случае с CoLA и аналогами, либо прямое сравнение вероятностных метрик, как для корпусов семейства VLiMP. В данном исследовании мы изучаем грамматические предпочтения БЯМ, напрямую анализируя ответы на вербальную инструкцию (промпт). Аналогичная процедура была использована с датасетом MMLU, с помощью которого тестируются общие знания о мире [2]: БЯМ предлагается ответить на вопрос с несколькими вариантами ответов путем генерации номера правильного ответа. Поскольку большие языковые модели специально дообучаются на инструктивных наборах данных, подобная формулировка задачи является наиболее естественной. Наша методика не только проверяет способность модели предсказывать, что неприемлемое предложение менее вероятно, чем приемлемое, но и исследует представления моделей о понятии приемлемости. Мы обращаемся к модели так же, как к носителю естественного языка: фактически воспроизводим методику эксперимента на оценку приемлемости по шкале Ликерта от 1 до 7, которая использовалась для опроса людей-носителей и разметки данных. Поэтому, на наш взгляд, предложенная методика является наиболее валидной.

Мы поставили цель сравнить модели, обученные в основном на русскоязычных данных, с мультиязычными. Ожидалось, что первые модели будут лучше подходить для градуальной оценки приемлемости вариативных феноменов русского языка. Данное предположение основано на предыдущих наблюдениях при сравнении языковой способности БЯМ, не обучавшихся на русскоязычных данных, с моделями, дообученными на русских текстах [14]. В этом исследовании мы протестировали четыре модели, доступные по API: YandexGPT 5 Pro и GigaChat 2 Max как преимущественно русскоязычные, а также Llama 3.3 70B и Mistral Large как мультиязычные.

В примере (8) приводится итоговая формулировка инструкции, при которой был получен наилучший результат. Было опробовано две методики: (i) zero-shot — запрос содержит только инструкцию и целевое предложение, (ii) few-shot — помимо инструкции и целевого предложения, запрос содержит два тренировочных примера: предложения и предполагаемые оценки. Пример тренировочной пары приводится в (9). В эксперименте, проводимом с людьми-носителями, также присутствуют тренировочные предложения. Они помогают участникам лучше понять задание. Отличие состоит в том, что людям не предлагаются конкретные оценки для тренировочных предложений, они оцениваются наравне с дальнейшими примерами.

(8) Тебе нужно оценить предложение по шкале от 1 до 7. Если предложение звучит хорошо, так можно сказать, поставь ему высокую оценку (6 или 7). Если предложение

звучит плохо, «не по-русски», так сказать нельзя, поставь ему низкую оценку (1 или 2). Некоторые предложения могут казаться не очень хорошими, но в принципе допустимыми. Таким предложениям поставь среднюю оценку (от 3 до 5).

{Инструкция для режима few-shot}

Оцени предложение по шкале от 1 до 7: «{Целевое предложение}». Ответ одной цифрой, ничего не добавляя.

(9) Например, предложение «Перед парой Ярослав надели и пиджак, и жилетку.» содержит ошибку. Ему стоит поставить оценку 2. Предложение «В апреле Марина посеяла семена и помидора, и тыквы.» является вполне естественным. Ему можно поставить оценку 6.

Тестирование проводилось на данных бенчмарка KBaC, а именно на 6803 предложениях из 20 экспериментов. Кратко опишем исследовательскую цель каждого из них. Эксперименты 1–12 направлены на изучение вариативного согласования с сочинительными конструкциями. В экспериментах 1 и 2 рассматривается предикативное согласование по числу с сочиненным подлежащим, где конъюнктами являются существительные. Эксперименты 3–8 посвящены изучению предикативного согласования по лицу и числу при сочинении личного местоимения 1 лица и имени собственного в позиции подлежащего. Эксперименты 9 и 10 рассматривают стратегии предикативного согласования по лицу и числу с подлежащим, выраженным сочиненными местоимениями 1 и 2 лица. Эксперименты 11 и 12 исследуют числовую вариативность существительного в именных группах с сочиненными модификаторами. Эксперименты 13 и 14 изучают вариативность предикативного согласования в постпозитивном определительном предложении. Эксперименты 15 и 16 рассматривают предикативное согласование по лицу и числу с управляющими квантификаторами. Эксперименты 17–20 посвящены согласованию с количественными существительными и числительными. Список релевантных условий и примеры лексикализаций для данных 20 экспериментов приведены в конце статьи (табл.5).

### 3.2. Результаты

На рис. 2 представлено значение средней абсолютной ошибки (mean average error, MAE) для предсказанных моделями оценок и суждений людей. Результаты показывают, что Mistral демонстрирует самое низкое значение ошибки для филлеров и стимулов. Также низкая ошибка для филлеров наблюдается для YandexGPT, однако ошибка для стимулов сильно выше. Модель GigaChat демонстрирует такую же ошибку для стимулов, как YandexGPT, но более высокое значение ошибки для филлеров. Llama дает наименее точные предсказания для стимулов, однако для филлеров значение ошибки невелико. В целом можно заметить, что модели более точно предсказывают оценки для предложений без вариативного согласования (филлеров), чем при наличии вариативности (стимулы).

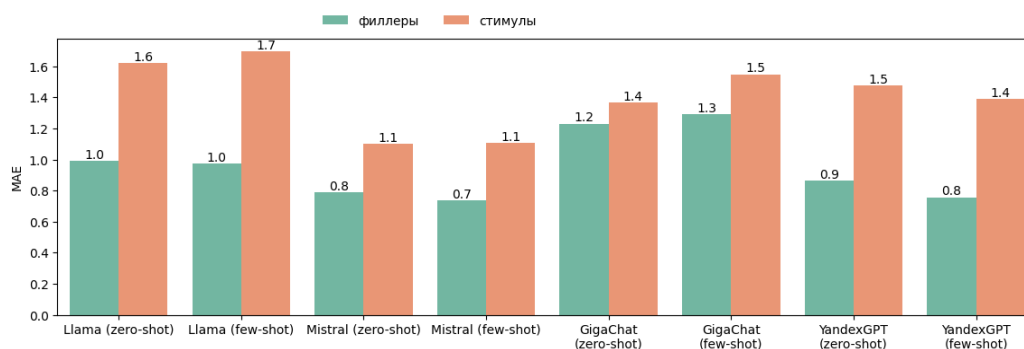
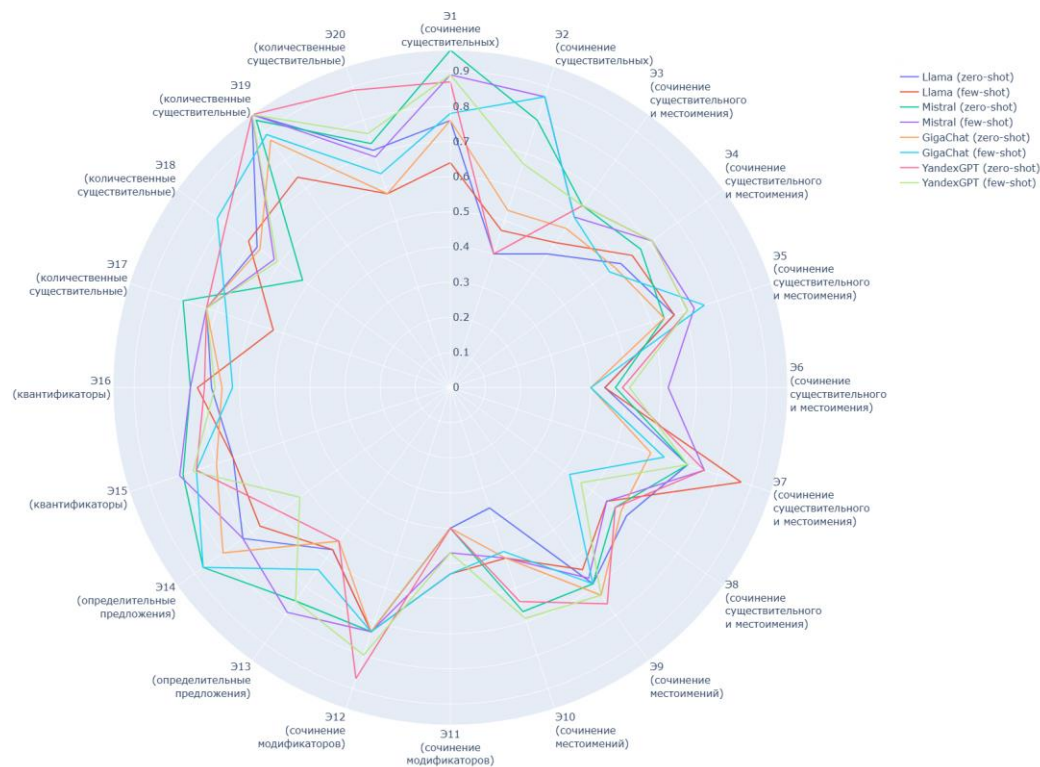


Рис. 2. Средняя абсолютная ошибка для предсказанных моделями оценок

Важно отметить, что стимульные предложения в бенчмарке КВаС демонстрируют различный уровень приемлемости в зависимости от конкретных экспериментальных условий, которые влияют на приемлемость той или иной стратегии согласования. Поэтому только подсчет MAE для предсказанных моделями оценок не будет достаточно показательным. Необходимо определить, могут ли БЯМ улавливать контрасты между экспериментальными условиями так же, как и люди. Для этого следует оценить, демонстрируют ли условия, которые показали статистически значимые различия в экспериментах на людях, также существенные различия в оценках БЯМ. И наоборот, остаются ли условия, незначимые для людей, незначимыми для моделей. Конечным показателем является процент совпадения результатов парных сравнений между оценками людей и ответами БЯМ. Для определения значимости мы использовали как параметрический критерий Стьюдента, так и непараметрический критерий Манна–Уитни, поскольку неясно, являются ли данные из БЯМ интервальными или порядковыми.

Результаты сравнения четырех моделей для 20 экспериментов по разработанной нами метрике представлены ниже. Табл. 3 и рис. 3 демонстрируют значение метрики для непараметрического критерия, табл. 4 и рис. 4 — для параметрического.



**Рис. 3.** Радарная диаграмма, отображающая долю совпадения значимых различий между оценками людей и моделей, посчитанных с помощью непараметрического критерия Манна–Уитни

**Таблица 3.** Доля совпадения значимых различий между оценками людей и моделей, посчитанная с помощью непараметрического критерия Манна–Уитни

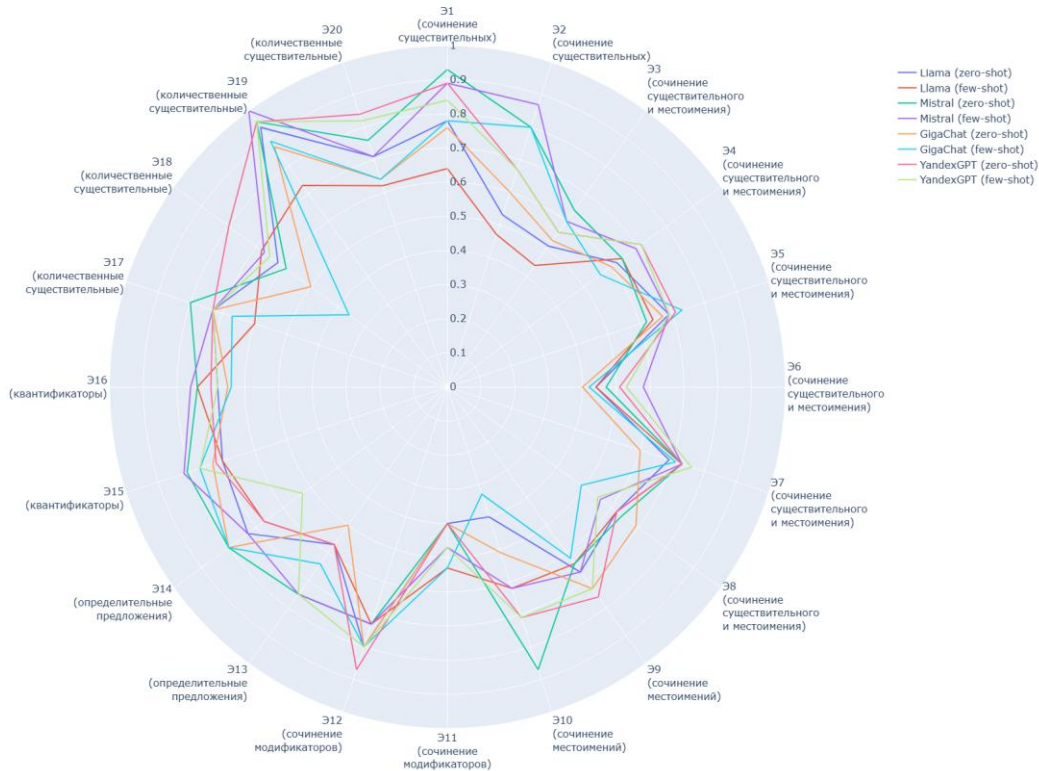
	Llama (zero-shot)	Llama (few-shot)	Mistral (zero-shot)	Mistral (few-shot)	Giga Chat (zero-shot)	Giga Chat (few-shot)	Yandex GPT (zero-shot)	Yandex GPT (few-shot)	$\mu$
Э1	0.76	0.64	0.96	0.89	0.76	0.78	0.87	0.89	0.82
Э2	0.40	0.47	0.80	0.87	0.53	0.87	0.40	0.67	0.63
Э3	0.47	0.51	0.64	0.60	0.56	0.60	0.64	0.64	0.58
Э4	0.60	0.64	0.67	0.71	0.58	0.56	0.71	0.71	0.65
Э5	0.67	0.67	0.64	0.73	0.64	0.76	0.71	0.71	0.69
Э6	0.44	0.44	0.47	0.62	0.40	0.40	0.49	0.51	0.47
Э7	0.71	0.87	0.71	0.76	0.60	0.64	0.76	0.71	0.72
Э8	0.62	0.55	0.58	0.55	0.60	0.42	0.58	0.46	0.54
Э9	0.69	0.64	0.69	0.67	0.73	0.69	0.76	0.73	0.70
Э10	0.36	0.51	0.67	0.51	0.51	0.49	0.64	0.69	0.55
Э11	0.40	0.53	0.40	0.47	0.40	0.53	0.40	0.47	0.45
Э12	0.73	0.73	0.73	0.73	0.73	0.73	0.87	0.80	0.76
Э13	0.57	0.57	0.75	0.79	0.54	0.64	0.54	0.75	0.64
Э14	0.73	0.67	0.87	0.73	0.80	0.87	0.60	0.53	0.72
Э15	0.65	0.65	0.80	0.81	0.70	0.76	0.76	0.77	0.74
Э16	0.68	0.72	0.74	0.74	0.65	0.62	0.70	0.67	0.69
Э17	0.73	0.53	0.80	0.73	0.73	0.67	0.73	0.73	0.71
Э18	0.68	0.71	0.52	0.62	0.67	0.82	0.79	0.61	0.68
Э19	0.96	0.74	0.94	0.96	0.87	0.89	0.96	0.96	0.91
Э20	0.71	0.58	0.73	0.69	0.58	0.64	0.89	0.76	0.70
$\mu$	0.63	0.62	0.71	0.71	0.63	0.67	0.69	0.69	–
$\sigma$	0.15	0.11	0.14	0.12	0.12	0.14	0.16	0.13	–

**Таблица 4.** Доля совпадения значимых различий между оценками людей и моделей, посчитанная с помощью параметрического критерия Стьюдента

	Llama (zero-shot)	Llama (few-shot)	Mistral (zero-shot)	Mistral (few-shot)	Giga Chat (zero-shot)	Giga Chat (few-shot)	Yandex GPT (zero-shot)	Yandex GPT (few-shot)	$\mu$
Э1	0.78	0.64	0.93	0.89	0.76	0.78	0.89	0.84	0.81
Э2	0.53	0.47	0.80	0.87	0.60	0.80	0.67	0.67	0.68
Э3	0.51	0.44	0.64	0.60	0.53	0.60	0.56	0.56	0.55
Э4	0.62	0.64	0.64	0.69	0.60	0.56	0.71	0.71	0.65
Э5	0.69	0.64	0.62	0.69	0.67	0.73	0.71	0.69	0.68
Э6	0.44	0.44	0.47	0.58	0.40	0.42	0.51	0.53	0.47
Э7	0.69	0.73	0.73	0.73	0.60	0.71	0.73	0.76	0.71
Э8	0.62	0.62	0.64	0.56	0.69	0.49	0.62	0.55	0.6
Э9	0.67	0.64	0.64	0.67	0.73	0.62	0.76	0.73	0.68
Э10	0.40	0.62	0.87	0.62	0.51	0.33	0.71	0.71	0.6
Э11	0.40	0.53	0.40	0.47	0.40	0.53	0.40	0.47	0.45
Э12	0.80	0.73	0.73	0.73	0.80	0.80	0.87	0.8	0.78
Э13	0.57	0.57	0.75	0.75	0.50	0.64	0.57	0.75	0.64
Э14	0.73	0.67	0.80	0.73	0.80	0.8	0.67	0.53	0.72
Э15	0.70	0.70	0.81	0.82	0.73	0.77	0.72	0.77	0.75
Э16	0.68	0.74	0.74	0.76	0.65	0.64	0.70	0.68	0.7
Э17	0.73	0.60	0.80	0.73	0.73	0.67	0.73	0.73	0.72
Э18	0.62	0.68	0.59	0.67	0.50	0.36	0.8	0.65	0.61
Э19	0.94	0.73	0.96	1.00	0.87	0.89	0.96	0.96	0.91

Продолжение таблицы 4

	Llama (zero-shot)	Llama (few-shot)	Mistral (zero-shot)	Mistral (few-shot)	Giga Chat (zero-shot)	Giga Chat (few-shot)	Yandex GPT (zero-shot)	Yandex GPT (few-shot)	$\mu$
Э20	0.71	0.62	0.76	0.71	0.64	0.64	0.84	0.82	0.72
$\mu$	0.64	0.62	0.72	0.71	0.64	0.64	0.71	0.70	—
$\sigma$	0.14	0.09	0.14	0.12	0.13	0.15	0.13	0.12	—



**Рис. 4.** Радарная диаграмма, отображающая долю совпадения значимых различий между оценками людей и моделей, посчитанных с помощью параметрического критерия Стьюдента

Результаты показывают, что наименьшее количество релевантных контрастов для людей и БЯМ наблюдается при оценивании предложений в экспериментах 3, 6, 8, 10 и 11. Все они исследуют различные типы сочинительных конструкций: именные группы с сочинительными модификаторами, сочинение имени собственного и личного местоимения или только личных местоимений в позиции подлежащего. Следовательно, именно этот текст оказывается наиболее сложным для БЯМ. Возможное объяснение состоит в том, что при согласовании мишени со всей сочинительной группой отсутствует эксплицитный контролер, обладающий необходимыми признаками, – они вычисляются по особым правилам из признаков каждого конъюкта. Наибольшее количество правильно предсказанных контрастов наблюдается для экспериментов 14, 15, 17, 19 и 20. Эксперимент 14 направлен на изучение предикативного согласования по роду в определительном предложении, эксперимент 15 посвящен предикативному согласованию при наличии управляющего квантификатора. Эксперименты 17, 19 и 20 исследуют конструкции с количественными числительными. Во всех случаях варианты согласования не равновероятны, но в высокой степени выводимы из предикторов. В целом вариативность, вызванная наличием

сочинительной конструкции, параметризуется в БЯМ значительно хуже, чем вариативность, возникающая при наличии определительного предложения, управляющих квантификаторов и количественных существительных.

Наилучшее качество демонстрирует мультиязычная модель Mistral Large (zero-shot & few-shot mean Mann-Whitney = 0.71, zero-shot mean Student = 0.72, few-shot mean Student = 0.71). Русскоязычная модель YandexGPT 5 Pro оказывается на втором месте (zero-shot & few-shot mean Mann-Whitney = 0.69, zero-shot mean Student = 0.71, few-shot mean Student = 0.70). Третье место занимает другая русскоязычная модель — GigaChat 2 Max (zero-shot mean Mann-Whitney = 0.63, few-shot mean Mann-Whitney = 0.67, zero-shot & few-shot mean Student = 0.64). Наконец, модель Llama 3 70B демонстрирует наименьшее значение метрики (zero-shot mean Mann-Whitney = 0.63, few-shot mean Mann-Whitney = 0.62, zero-shot mean Student = 0.64, few-shot mean Student = 0.62). Сопоставление режимов zero-shot и few-shot показывает, что их роль отличается для моделей. Добавление примеров в промпт может как повысить качество (GigaChat), так и понизить его (Llama). Для большинства моделей наблюдается меньшее значение стандартного отклонения в режиме few-shot по сравнению с режимом zero-shot.

Сравнение моделей демонстрирует, что наиболее важным фактором оказывается именно количество параметров модели: Mistral Large обладает 123 миллиардами параметров, протестированная нами версия Llama 3 имеет 70 миллиардов параметров. Объем русскоязычных данных при обучении хоть и не является решающим, но тоже играет некоторую роль. Так, точное количество параметров используемых русскоязычных моделей узнать не удалось. Известно, что облегченная версия YandexGPT 5 Lite имеет 8 миллиардов параметров, а GigaChat 2 Pro — 29 миллиардов параметров. Вероятнее всего, расширенные версии YandexGPT 5 Pro и GigaChat 2 Max, используемые в данном исследовании, обладают меньшим количеством параметров, чем Llama 3 70B и Mistral Large. Тем не менее, YandexGPT 5 Pro и GigaChat 2 Max показывают достаточно высокий результат, сравнимый или даже превышающий качество моделей с большим количеством весов. Можно предположить, что русскоязычность модели компенсирует меньшее количество параметров.

Следует отметить, что многие стратегии вариативного согласования и условия, определяющие выбор стратегии, скорее являются универсальными для различных языков. В то же время, набор и организация признаков, по которым происходит согласование, значительно меняется от языка к языку. Если многоязычные модели справляются не хуже русскоязычных, можно сделать вывод, что они способны усвоить стратегии без конкретного признакового наполнения.

#### 4. Заключение

В данной статье представлен первый бенчмарк для оценки БЯМ на задаче градуальной оценки приемлемости. В отличие от бинарной классификации по приемлемости, эта задача позволяет детально оценить лингвистическую компетенцию БЯМ. Представленный набор данных содержит результаты синтаксических экспериментов по оценке приемлемости на основе вариативного согласования в русском языке. Предложения-стимулы, характеризующиеся вариативностью согласования, занимают среднюю часть шкалы приемлемости, что делает их идеальным материалом для тестирования градуальной приемлемости. Филлеры, занимающие либо высокую (грамматичные), либо низкую (неграмматичные) позицию на шкале, показывают, насколько успешно модель оценивает стандартные случаи согласования.

Тестирование БЯМ на материале данного бенчмарка проводилось путем генерации ответа на заданную инструкцию. Четыре модели — YandexGPT 5 Pro, GigaChat 2 Max, Llama 3 70B и Mistral Large — были протестированы в двух режимах: zero-shot (только инструкции) и few-shot (инструкции плюс тренировочные предложения). Для сравнения

предсказаний моделей с ответами людей была разработана специальная метрика. Она показывает, насколько схожие грамматические контрасты выявляются моделью и человеком при вынесении суждений о приемлемости.

Для всех БЯМ наблюдается большее значение средней ошибки при предсказании оценок для стимулов, содержащих вариативное согласование, чем для филлеров с однозначным согласованием. Было проведено сравнение оценок для разных экспериментальных условий, а также для филлеров. Следует отметить, что контраст между грамматичными и неграмматичными филлерами выявили все тестируемые модели. Среди экспериментальных условий наиболее простыми феноменами для моделей оказались конструкции с определительными предложениями и управляющими квантификаторами, наиболее сложными – сочинительные конструкции определенных классов. Данный результат может быть мотивирован тем, насколько признаки мишени согласования выводимы из признаков контролера и других характеристик предложения.

Все модели продемонстрировали стабильный результат при переходе от режима zero-shot к режиму few-shot, качество увеличивалось лишь незначительно. Наилучший результат продемонстрировала Mistral Large, обладающая наибольшим количеством параметров. Модель YandexGPT 5 Pro оказалась на втором месте. GigaChat 2 Max продемонстрировала результат хуже, чем первые две модели, но лучше, чем Llama 3 70B. Следовательно, для задачи градуальной оценки приемлемости является значимым как объем русскоязычных данных при обучении, так и количество параметров модели.

Работа Е. А. Лютиковой и А. А. Герасимовой (сбор корпуса и разработка метрик) выполнена при поддержке программы развития МГУ, проект № 23-Ш02-10 «Языковая компетенция носителей естественного языка и нейросетевых моделей». Работа К. А. Студеникиной (оценка языковых моделей) выполнена при финансовой поддержке некоммерческого фонда развития науки и образования «Интеллект».

Таблица 5. Перечень релевантных условий для выборки экспериментов

Идентификатор эксперимента	Экспериментальные условия	Пример предложения
Эксперимент 1	симметр., SV, ед. ч	Теория и практика совмещается в новом курсе.
	симметр., SV, мн. ч	Теория и практика совмещаются в новом курсе
	несимметр., SV, ед. ч	Теория и практика освещается в новом курсе.
	несимметр., SV, мн. ч	Теория и практика освещаются в новом курсе
	симметр., VS, ед. ч	В новом курсе совмещается теория и практика.
	симметр., VS, мн. ч	В новом курсе совмещаются теория и практика.
	несимметр., VS, ед. ч	В новом курсе освещается теория и практика.
	несимметр., VS, мн. ч	В новом курсе освещаются теория и практика.
	грамм. филлер	Папа преподает биологию и химию в старшей школе
неграмм. филлер	Дети хотят конструктор и пазлом на Новый год.	
Эксперимент 2	совп., ед. ч.	В жаркое лето на огороде вырос и огурец, и кабачок.
	совп., мн. ч.	В жаркое лето на огороде выросли и огурец, и кабачок.
	несовп., ед. ч.	В жаркое лето на огороде выросла и тыква, и кабачок.
	несовп., мн. ч.	В жаркое лето на огороде выросли и тыква, и кабачок.
	грамм. филлер	В моём саду на даче растут и розы, и фиалки.
	неграмм. филлер	В моей комнате на кровати лежит подушки и одеяла.

Продолжение таблицы 5

Идентификатор эксперимента	Экспериментальные условия	Пример предложения
Эксперимент 3 (порядок VS, союз <i>и</i> )	<i>я и X</i> , 1 л. ед. ч.	На правом берегу остаюсь я и Слава.
	<i>я и X</i> , 1 л. мн. ч.	На правом берегу остаёмся я и Слава.
	<i>я и X</i> , 3 л. ед. ч.	На правом берегу остаётся я и Слава.
	<i>я и X</i> , 3 л. мн. ч.	На правом берегу остаются я и Слава.
	<i>X и я</i> , 1 л. ед. ч.	На правом берегу остаюсь Слава и я.
	<i>X и я</i> , 1 л. мн. ч.	На правом берегу остаёмся Слава и я.
	<i>X и я</i> , 3 л. ед. ч.	На правом берегу остаётся Слава и я.
	<i>X и я</i> , 3 л. мн. ч.	На правом берегу остаются Слава и я.
	грамм. филлер	В билетной кассе платят Лёша и Вася.
неграмм. филлер	В правом углу шепчу Кеша и Егор.	
Эксперимент 4 (порядок SV, <i>я + X</i> )	<i>и</i> , 1 л. ед. ч.	Я и Вова спою весёлую песню.
	<i>и</i> , 1 л. мн. ч.	Я и Вова споём весёлую песню.
	<i>и</i> , 3 л. ед. ч.	Я и Вова споёт весёлую песню.
	<i>и</i> , 3 л. мн. ч.	Я и Вова споют весёлую песню.
	<i>и..и</i> , 1 л. ед. ч.	И я, и Вова спою весёлую песню.
	<i>и..и</i> , 1 л. мн. ч.	И я, и Вова споём весёлую песню.
	<i>и..и</i> , 3 л. ед. ч.	И я, и Вова споёт весёлую песню.
	<i>и..и</i> , 3 л. мн. ч.	И я, и Вова споют весёлую песню.
	грамм. филлер	Гриша сварит овсяную кашу.
неграмм. филлер	Петя привлечёт молодые публику.	
Эксперимент 5 (порядок SV, <i>X + я</i> )	<i>и</i> , 1 л. ед. ч.	Вова и я спою весёлую песню.
	<i>и</i> , 1 л. мн. ч.	Вова и я споём весёлую песню.
	<i>и</i> , 3 л. ед. ч.	Вова и я споёт весёлую песню.
	<i>и</i> , 3 л. мн. ч.	Вова и я споют весёлую песню.
	<i>и..и</i> , 1 л. ед. ч.	И Вова, и я спою весёлую песню.
	<i>и..и</i> , 1 л. мн. ч.	И Вова, и я споём весёлую песню.
	<i>и..и</i> , 3 л. ед. ч.	И Вова, и я споёт весёлую песню.
	<i>и..и</i> , 3 л. мн. ч.	И Вова, и я споют весёлую песню.
	грамм. филлер	Витя почистит и диван, и кресло.
неграмм. филлер	Дима и Ваня уберём ванную комнату.	
Эксперимент 6 (порядок SV, <i>я + X</i> )	<i>или</i> , 1 л. ед. ч.	Я или Вова спою эту весёлую песню.
	<i>или</i> , 1 л. мн. ч.	Я или Вова споём эту весёлую песню.
	<i>или</i> , 3 л. ед. ч.	Я или Вова споёт эту весёлую песню.
	<i>или</i> , 3 л. мн. ч.	Я или Вова споют эту весёлую песню.
	<i>или..или</i> , 1 л. ед. ч.	Или я, или Вова спою эту весёлую песню.
	<i>или..или</i> , 1 л. мн. ч.	Или я, или Вова споём эту весёлую песню.
	<i>или..или</i> , 3 л. ед. ч.	Или я, или Вова споёт эту весёлую песню.
	<i>или..или</i> , 3 л. мн. ч.	Или я, или Вова споют эту весёлую песню.
	грамм. филлер	Артём почистит этот резиновый сапог.
неграмм. филлер	Яша запишет этот почтовую адрес.	
Эксперимент 7 (порядок SV, <i>X + я</i> )	<i>или</i> , 1 л. ед. ч.	Вова или я спою весёлую песню.
	<i>или</i> , 1 л. мн. ч.	Вова или я споём весёлую песню.
	<i>или</i> , 3 л. ед. ч.	Вова или я споёт весёлую песню.
	<i>или</i> , 3 л. мн. ч.	Вова или я споют весёлую песню.
	<i>или..или</i> , 1 л. ед. ч.	Или Вова, или я спою весёлую песню.
	<i>или..или</i> , 1 л. мн. ч.	Или Вова, или я споём весёлую песню.
	<i>или..или</i> , 3 л. ед. ч.	Или Вова, или я споёт весёлую песню.
	<i>или..или</i> , 3 л. мн. ч.	Или Вова, или я споют весёлую песню.
	грамм. филлер	Вася получит медаль или кубок.
неграмм. филлер	Дима или Ваня протрём эту грязную полку.	

Продолжение таблицы 5

Идентификатор эксперимента	Экспериментальные условия	Пример предложения
Эксперимент 8	<i>с</i> , 1 л. ед. ч.	Завтра я с Борей исполняю вариации.
	<i>с</i> , 2 л. ед. ч.	Завтра ты Борей исполняешь вариации.
	<i>с</i> , 3 л. ед. ч.	Завтра он с Борей исполняет вариации.
	<i>с</i> , 1 л. мн. ч.	Завтра мы с Борей исполняем вариации.
	<i>с</i> , 2 л. мн. ч.	Завтра вы Борей исполняете вариации.
	<i>с</i> , 3 л. мн. ч.	Завтра они с Борей исполняют вариации.
	<i>и</i> , 1 л. мн. ч.	Завтра я и Боря исполняем вариации.
	<i>и</i> , 2 л. мн. ч.	Завтра ты и Боря исполняете вариации.
	<i>и</i> , 3 л. мн. ч.	Завтра он и Боря исполняют вариации.
	грамм. филлер	Завтра вы с ней снимаете рекламный ролик.
неграмм. филлер	Весной Лада и Олег убирает кабинет отца.	
Эксперимент 9 (порядок SV)	<i>я и ты</i> , 1 л. мн.ч.	Я и ты строим крепость из снега.
	<i>я и ты</i> , 1 л. ед.ч.	Я и ты строю крепость из снега.
	<i>я и ты</i> , 2 л. ед.ч.	Я и ты строишь крепость из снега.
	<i>я и ты</i> , 3 л. мн. ч.	Я и ты строят крепость из снега.
	<i>ты и я</i> , 1 л. мн.ч.	Ты и я строим крепость из снега.
	<i>ты и я</i> , 1 л. ед.ч.	Ты и я строю крепость из снега.
	<i>ты и я</i> , 2 л. ед.ч.	Ты и я строишь крепость из снега.
	<i>ты и я</i> , 3 л. мн. ч.	Ты и я строят крепость из снега.
	грамм. филлер	Я погладила рубашки Жене и Захару.
неграмм. филлер	Лиза помыла лапы кошкой и собаками.	
Эксперимент 10 (порядок VS)	<i>я и ты</i> , 1 л. мн.ч.	Крепость из снега строим я и ты.
	<i>я и ты</i> , 1 л. ед.ч.	Крепость из снега строю я и ты.
	<i>я и ты</i> , 2 л. ед.ч.	Крепость из снега строишь я и ты.
	<i>я и ты</i> , 3 л. мн. ч.	Крепость из снега строят я и ты.
	<i>ты и я</i> , 1 л. мн.ч.	Крепость из снега строим ты и я.
	<i>ты и я</i> , 1 л. ед.ч.	Крепость из снега строю ты и я.
	<i>ты и я</i> , 2 л. ед.ч.	Крепость из снега строишь ты и я ы.
	<i>ты и я</i> , 3 л. мн. ч.	Крепость из снега строят ты и я.
	грамм. филлер	В зоопарке Женя увидел слона и жирафа.
неграмм. филлер	В понедельник Кира и Зина проспал.	
Эксперимент 11	ед. ч., <i>и</i>	Перед уходом Ира закрыла книжный и платяной шкаф.
	мн. ч., <i>и</i>	Перед уходом Ира закрыла книжный и платяной шкафы.
	ед. ч., <i>и..и</i>	Перед уходом Ира закрыла и книжный, и платяной шкаф.
	мн. ч., <i>и..и</i>	Перед уходом Ира закрыла и книжный, и платяной шкафы.
	грамм. филлер	За минуту Лера угадала имена и фамилии.
неграмм. филлер	За полчаса Илья выяснили и район, и квартал.	
Эксперимент 12	ед. ч., <i>и</i>	Профессор купит черный и серый пиджак.
	мн. ч., <i>и</i>	Профессор купит черный и серый пиджаки.
	ед. ч., <i>или</i>	Профессор купит черный или серый пиджак.
	мн. ч., <i>или</i>	Профессор купит черный или серый пиджаки.
	грамм. филлер	Врач назначит лекарства от анемии и гастрита.
неграмм. филлер	Мама и папа купил зимнюю куртку.	

Продолжение таблицы 5

Идентификатор эксперимента	Экспериментальные условия	Пример предложения
Эксперимент 13	нет сущ., м. р. вершины., м.р. глагола	Каждый, кто жил в центре города, ненавидит ночной шум.
	есть сущ., м. р. вершины., м.р. глагола	Каждый москвич, кто жил в центре города, ненавидит ночной шум.
	нет сущ., ж. р. вершины., м.р. глагола	Каждая, кто жил в центре города, ненавидит ночной шум.
	есть сущ., ж. р. вершины., м.р. глагола	Каждая москвичка, кто жил в центре города, ненавидит ночной шум.
	нет сущ., ж. р. вершины., ж.р. глагола	Каждая, кто жила в центре города, ненавидит ночной шум.
	есть сущ., ж. р. вершины., ж.р. глагола	Каждая москвичка, кто жила в центре города, ненавидит ночной шум.
	грамм. филлер	Каждый, о ком рассказал древний эпос, имеет реальный прототип.
неграмм. филлер	Каждый, на кого глядела рассерженная мать, чувствует дикому стыд.	
Эксперимент 14	нет сущ., ед. ч. гл.	Те, кто ответит на экзамене, получают пятерки.
	есть сущ., ед. ч. гл.	Те студенты, кто ответит на экзамене, получают пятерки.
	нет сущ., мн. ч. гл.	Те, кто ответят на экзамене, получают пятерки.
	есть сущ., мн. ч. гл.	Те студенты, кто ответят на экзамене, получают пятерки.
	грамм. филлер	Многие абитуриенты, поступающие в вуз, сдадут экзамены.
	неграмм. филлер	Многие хоккеисты, которые играют в сборной, закончат карьеру.
Эксперимент 15 (порядок SV)	<i>мы</i> , элект., 1 л. мн. ч.	Семеро из нас возьмёмся за новые проекты.
	<i>мы</i> , элект., 3 л. ед. ч.	Семеро из нас возьмётся за новые проекты.
	<i>мы</i> , элект., 3 л. мн. ч.	Семеро из нас возьмутся за новые проекты.
	<i>мы</i> , номин., 1 л. мн. ч.	Мы семеро возьмёмся за новые проекты.
	<i>мы</i> , номин., 3 л. ед. ч.	Мы семеро возьмётся за новые проекты.
	<i>мы</i> , номин., 3 л. мн. ч.	Мы семеро возьмутся за новые проекты.
	<i>они</i> , элект., 3 л. ед. ч.	Семеро из них возьмётся за новые проекты.
	<i>они</i> , элект., 3 л. мн. ч.	Семеро из них возьмутся за новые проекты.
	<i>они</i> , номин., 3 л. ед. ч.	Они семеро возьмётся за новые проекты.
	<i>они</i> , номин., 3 л. мн. ч.	Они семеро возьмутся за новые проекты.
грамм. филлер	Двое часовых будут стоять у парадного входа.	
неграмм. филлер	Только две фигуристка исполнят этот прыжки.	
Эксперимент 16 (порядок VS)	<i>мы</i> , элект., 1 л. мн. ч.	За новые проекты возьмёмся семеро из нас.
	<i>мы</i> , элект., 3 л. ед. ч.	За новые проекты возьмётся семеро из нас.
	<i>мы</i> , элект., 3 л. мн. ч.	За новые проекты возьмутся семеро из нас.
	<i>мы</i> , номин., 1 л. мн. ч.	За новые проекты возьмёмся мы семеро.
	<i>мы</i> , номин., 3 л. ед. ч.	За новые проекты возьмётся мы семеро.
	<i>мы</i> , номин., 3 л. мн. ч.	За новые проекты возьмутся мы семеро.
	<i>они</i> , элект., 3 л. ед. ч.	За новые проекты возьмётся семеро из них.
	<i>они</i> , элект., 3 л. мн. ч.	За новые проекты возьмутся семеро из них.
	<i>они</i> , номин., 3 л. ед. ч.	За новые проекты возьмётся они семеро.
	<i>они</i> , номин., 3 л. мн. ч.	За новые проекты возьмутся они семеро.
	грамм. филлер	Двое грабителей предстанут перед судом.
неграмм. филлер	Только двое девочка будут сдавать биологией.	

Продолжение таблицы 5

Идентификатор эксперимента	Экспериментальные условия	Пример предложения
Эксперимент 17	ед. ч. премод., ед. ч. гл.	Этот миллион рукописей хранился в библиотеке.
	ед. ч. премод., мн. ч. гл.	Этот миллион рукописей хранились в библиотеке.
	мн. ч. премод., ед. ч. гл.	Эти миллион рукописей хранился в библиотеке.
	мн. ч. премод., мн. ч. гл.	Эти миллион рукописей хранились в библиотеке.
	грамм. филлер	Пара позвала тысячу человек на свадьбу.
Эксперимент 18	неграмм. филлер	Повар сварил дюжину яиц на завтрака.
	номинатив, неодуш.	В шкафах стояли эти сотня книг.
	генитив, неодуш.	Архив не досчитался этих сотни книг.
	датель, неодуш.	Коллекционер радуется этим сотне книг.
	аккузатив, неодуш.	В магазин привезли эти сотню книг.
	номинатив, одуш.	В приюте жили эти сотня собак.
	генитив, одуш.	В городе остерегались этих сотни собак.
	датель, одуш.	Прививку поставили этим сотне собак.
	аккузатив, одуш.	Волонтеры покормили этих сотню собак.
	грамм. филлер	Начальник руководит десятком отделов.
неграмм. филлер	Шум мешало паре отдыхающих.	
Эксперимент 19	ж. р., SV, ед. ч.	Три картины висело в гостиной, совмещенной с кухней.
	ж. р., SV, мн. ч.	Три картины висели в гостиной, совмещенной с кухней.
	м. р., SV, ед. ч.	Три портрета висело в гостиной, совмещенной с кухней.
	м. р., SV, мн. ч.	Три портрета висели в гостиной, совмещенной с кухней.
	ж. р., VS, ед. ч.	В гостиной, совмещенной с кухней, висело три картины.
	ж. р., VS, мн. ч.	В гостиной, совмещенной с кухней, висели три картины.
	м. р., VS, ед. ч.	В гостиной, совмещенной с кухней, висело три портрета.
	м. р., VS, мн. ч.	В гостиной, совмещенной с кухней, висели три портрета.
	грамм. филлер	Официанты, работавшие в кафе, принесли три чека.
неграмм. филлер	Машина, сжавшие по трассе, перевозили четырех людей.	
Эксперимент 20	числ., SV, ед. ч.	В аэропорту багаж ждет двадцать туристов
	числ., SV, мн. ч.	В аэропорту багаж ждут двадцать туристов
	сущ., SV, ед. ч.	В аэропорту багаж ждет пара туристов.
	сущ., SV, мн. ч.	В аэропорту багаж ждут пара туристов.
	числ., VS, ед. ч.	Двадцать туристов ждет багаж в аэропорту.
	числ., VS, мн. ч.	Двадцать туристов ждут багаж в аэропорту.
	сущ., VS, ед. ч.	Пара туристов ждет багаж в аэропорту.
	сущ., VS, мн. ч.	Пара туристов ждут багаж в аэропорту.
	грамм. филлер	В ресторане три стола бронирует владелец.
неграмм. филлер	В сборнике две статью пишет редактор.	

## Литература

- [1] Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. SuperGLUE: a stickier benchmark for general-purpose language understanding systems // *Advances in Neural Information Processing Systems (NeurIPS)*. 2019.
- [2] Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. Measuring Massive Multitask Language Understanding // *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Event, Austria, 2020.
- [3] Warstadt A., Singh A., Bowman S. R. Neural Network Acceptability Judgments. — *Transactions of the Association for Computational Linguistics*, 2019. Vol. 7. P. 625-641.
- [4] Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang S. F., Bowman S. R. BLiMP: The benchmark of linguistic minimal pairs for English // *Transactions of the Association for Computational Linguistics*. 2020. Vol. 8. P. 377-392.
- [5] Chomsky N. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press, 1965.
- [6] Sprouse J. Continuous acceptability, categorical grammaticality, and experimental syntax // *Biolinguistics*. 2007. Vol. 1. P. 123-134.
- [7] Sprouse J., Schütze C.T., Almeida D. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010 // *Lingua*. 2013. Vol. 134. P. 219-248.
- [8] Lau J.H., Clark A., Lappin S. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge // *Cognitive Science*. 2017. Vol. 41. No. 5. P. 1201-1241.
- [9] Likert R. A technique for the measurement of attitudes. // *Archives of Psychology*. 1932. Vol. 22. No. 140. P. 5-55.
- [10] Sprouse J., Schütze C. T., Almeida D. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010 // *Lingua*. 2013. Vol. 134. P. 219-248.
- [11] Герасимова А.А. Количественные методы исследования грамматических ограничений (на материале вариативного согласования в русском языке): дис. канд. филол. наук 5.9.8. М., 2023.
- [12] Герасимова А.А. Учебные материалы практикума по экспериментальному синтаксису. Отбор респондентов. 2021. URL: [https://agerasimova.com/wp-content/uploads/Gerasimova\\_Practice\\_Outliers.pdf](https://agerasimova.com/wp-content/uploads/Gerasimova_Practice_Outliers.pdf) (дата обращения: 12.03.2025).
- [13] Davidjuk T. Agreement with disjoint subjects in Russian // *Argumentum* 2024. Vol. 20. P. 322-335.
- [14] Гращенков П.В., Паско Л.И., Студеникина К.А., Тихомиров М.М. Параметрический корпус русского языка RuParam // *Научно-технический вестник информационных технологий, механики и оптики* 2024. Т. 24ю № 6. С. 991-998.

## Evaluating Linguistic Competence of LLMs and Human Based Corpus of Variable Agreement in Russian

K. Studenikina, E. Lyutikova, A. Gerasimova

Lomonosov Moscow State University

This study aims to evaluate similarities and differences between the linguistic competence of native speakers and large language models (LLMs). The comparison is based on our developed benchmark KVAs (*Korpus Variativnogo Soglasovanija* ‘Corpus of Variable Agreement’). It includes 6,803 sentences with scores on a scale from 1 to 7, obtained during 20 syntactic experiments on variable agreement. The corpus contains mean scores of Russian sentences with various agreement conditions received from native speakers and allows us to find out how LLMs copes with gradual acceptability judgment task.

We present the results of testing four models: two predominantly Russian-speaking (YandexGPT 5 Pro and GigaChat 2 Max) and two multilingual (Llama 3.3 70B and Mistral Large). Two modes were tested for each model: zero-shot, which contains only instructions, and few-shot, where training examples and their scores are added. Since the data in KVAs shows different levels of acceptability depending on experimental conditions, calculating only mean absolute error (MAE) for the scores predicted by the models is not demonstrative enough. We have developed a metric that allows us to assess what proportion of contrasts between experimental conditions relevant to humans is detected by LLMs. The results show that the MAE value is less for the sentences with unambiguous agreement. Variable agreement is more complicated for LLMs. The least number of contrasts for humans and models match in the experiments on coordinated constructions. A much larger number of correctly predicted contrasts are revealed for constructions with quantitative noun phrases, quantified subjects and postpositive relative clauses. Mistral demonstrates the most accurate coincidence in detecting significant contrasts compared to native speakers. Llama reveals the least amount of contrast. Russian-speaking models occupy an intermediate position with YandexGPT outperforming GigaChat. The results show that the quality of solving the gradual acceptability judgment task varies for different classes of linguistic phenomena. A comparison of the models demonstrates that the most important factor is the number of parameters. Meanwhile, the amount of Russian data during training compensates for the smaller number of parameters.

**Keywords:** linguistic competence, syntax, agreement, LLM, benchmark

**Reference for citation:** Studenikina K., Lyutikova E., Gerasimova A. Evaluating Linguistic Competence of LLMs and Human Based Corpus of Variable Agreement in Russian // *Computational Linguistics and Computational Ontologies*. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 100-119. DOI: 10.17586/3033-5582-2025-9-100-119.

## Reference

- [1] Wang A., Pruksachatkun Y., Nangia N., Singh A., Michael J., Hill F., Levy O., Bowman S. R. SuperGLUE: a stickier benchmark for general-purpose language understanding systems // *Advances in Neural Information Processing Systems (NeurIPS)* 2019.
- [2] Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J. Measuring Massive Multitask Language Understanding // *Proceedings of the International Conference on Learning Representations (ICLR)*. Virtual Event, Austria, 2020.
- [3] Warstadt A., Singh A., Bowman S. R. Neural Network Acceptability Judgments // *Transactions of the Association for Computational Linguistics*. 2019. Vol. 7. P. 625-641.
- [4] Warstadt A., Parrish A., Liu H., Mohananey A., Peng W., Wang S. F., Bowman S. R. BLiMP: The benchmark of linguistic minimal pairs for English // *Transactions of the Association for Computational Linguistics*. 2020. Vol. 8. P. 377-392.
- [5] Chomsky N. *Aspects of the theory of syntax*. Cambridge, MA: MIT Press, 1965.
- [6] Sprouse J. Continuous acceptability, categorical grammaticality, and experimental syntax // *Biolinguistics*. 2007. Vol. 1. P. 123-134.
- [7] Sprouse J., Schütze C. T., Almeida D. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010 // *Lingua*. 2013. Vol. 134. P. 219-248.
- [8] Lau J. H., Clark A., Lappin S. Grammaticality, acceptability, and probability: A probabilistic view of linguistic knowledge // *Cognitive Science*. 2017. Vol. 41. No. 5. P. 1201-1241.
- [9] Likert R. A technique for the measurement of attitudes // *Archives of Psychology*. 1932. Vol. 22. No. 140. P. 5-55.

- [10] Sprouse J., Schütze C. T., Almeida D. A comparison of informal and formal acceptability judgments using a random sample from Linguistic Inquiry 2001–2010 // *Lingua*. 2013. Vol. 134. P. 219-248.
- [11] Gerasimova A.A. *Kolichestvennye metody issledovaniya grammaticheskikh ogranichenij (na materiale variativnogo soglasovaniya v russkom yazyke): dis. kand. filol. nauk 5.9.8 M., 2023.* (In Russian)
- [12] Gerasimova A. A. *Uchebnye materialy praktikuma po eksperimental'nomu sintaksisu. Otbor respondentov.* 2021. URL: [https://agerasimova.com/wp-content/uploads/Gerasimova\\_Practice\\_Outliers.pdf](https://agerasimova.com/wp-content/uploads/Gerasimova_Practice_Outliers.pdf) (accessed date: 12.03.2025).
- [13] Davidjuk T. Agreement with disjoined subjects in Russian // *Argumentum*. 2024. Vol. 20. P. 322-335.
- [14] Grashchenkov P.V., Pasko L.I., Studenikina K.A., Tikhomirov M.M. Parametricheskij korpus russkogo yazyka RuParam // *Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki*. 2024. Vol. 24. No. 6. P. 991-998. (In Russian)

## Разработка системы анализа разноплановых характеристик поэтического текста

А. И. Панкова, Е. В. Ягунова

Санкт-Петербургский государственный университет

arina.pnkv@yandex.ru, iagounova\_elena@mail.ru

### Аннотация

В статье рассматриваются аспекты разработки модулей для автоматизации анализа поэтических текстов средствами машинного обучения и компьютерной лингвистики. Мы рассматриваем особенности структуры поэтических текстов и подходов к их анализу, а также способы вычисления разноплановых характеристик: разработаны и оценены алгоритмические и нейросетевые модели для определения силлаботонических стихотворных размеров, выполнен семантический анализ на базе трансформера RuBERT, позволяющий автоматически выделять ключевые темы стихотворения на основе косинусного сходства эмбеддингов, проведено синтаксическое (подсчет доли частей речи, определение параллелизма) и лексическое (подсчет доли редких слов на основе созданного словаря частотных слов в поэзии) исследование поэтических текстов. Анализ проведен с помощью библиотек языка программирования Python. Материалом исследования послужили открытые русскоязычные поэтические корпуса. Реализованные модули мы интегрировали в разработанное веб-приложение для анализа разноплановых характеристик стихотворений. Полученное в результате работы приложение может быть использовано в образовательных учреждениях для наглядного демонстрация приемов стихосложения и анализа стихотворных форм, а также на платформах литературных сообществ для предоставления возможностей интерактивного анализа поэзии. В дальнейшем веб-приложение планируется масштабировать и расширить новыми моделями и корпусами.

**Ключевые слова:** компьютерная лингвистика, автоматический анализ, русская поэзия, машинное обучение

**Библиографическая ссылка:** Панкова А. И., Ягунова Е. В. Разработка системы анализа разноплановых характеристик поэтического текста // Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г. Сборник научных статей). – СПб.: Университет ИТМО, 2025. С. 120-129. DOI: 10.17586/3033-5582-2025-9-120-129.

### 1. Введение

Поэтический текст отличается от обычного необходимостью соблюдать дополнительные по отношению к языку ограничения: метроритмические нормы, организованность на фонологическом, рифмовом, лексическом и композиционном уровнях. Благодаря формализуемой структуре поэтические тексты становятся перспективным объектом для применения компьютерных методов анализа.

В связи с развитием информационных технологий в гуманитарных исследованиях и появлением больших размеченных корпусов поэтических текстов, набирает актуальность запрос на автоматизацию анализа различных уровней стихотворной структуры. Наиболее полным и описанным по нашей информации является веб-приложение «Анализ

поэтических текстов онлайн», разработанное исследователями Института вычислительных технологий СО РАН для определения метрики и стопности, а также подробного анализа рифмовки русскоязычных поэтических текстов [1; 2]. Целью данного приложения является автоматизация процесса создания метрических справочников и конкордансов, оно ориентировано в первую очередь на литературоведов и научных сотрудников.

Существует также ряд веб-приложений, направленных не на профессиональных исследователей, а скорее на начинающих авторов и любителей, например, «Fet.Online», функционал которого позволяет визуализировать ритмическую схему построчно, определить размер, ритм, рифму и количество слабоударных гласных [3].

Другие найденные нами веб-приложения тоже направлены в первую очередь на подробный анализ метrorитмических характеристик стихотворений. В рамках настоящего исследования мы представляем разработку веб-приложения, в котором можно будет исследовать не только размер и рифму стихотворения, но и другие характеристики — семантические, лексические, синтаксические. Это приложение ориентировано, прежде всего, на любителей поэзии и начинающих авторов, а также может помочь преподавателям в обучении школьников и студентов. Одновременно результаты предлагаемых вариантов анализа исследовательских корпусов поэтических текстов будут интересны лингвистам и литературоведам.

## 2. Определение стихотворного размера

### 2.1 Алгоритм определения ритмической схемы

Существуют готовые корпуса поэтических текстов, обращение к которым позволяет отложить вопрос проблематики выбора поэтических текстов для анализа. Нами были использованы данные трех доступных размеченных (вручную или автоматически) корпусов с русскоязычной поэзией.

Для определения ритмической схемы мы воспользовались поэтическим подкорпусом Национального корпуса русского языка [НКРЯ]. Это наиболее объемный корпус, состоящий из текстов стихотворений с вручную размеченной дополнительной информацией. Используемые данные включают в себя токены, состоящие из строки и её размерности.

Метром называется последовательность слабых и сильных (ударных) позиций. Силлабо-тоническое стихосложение предполагает регулярное чередование ударных и безударных слогов, а также постоянное количество ударений в строке. Силлабо-тонические размеры можно выразить данными схемами:

- ' хорей;
- -' ямб;
- ' -- дактиль;
- -' амфибрахий;
- --' анапест.

В данной работе мы будем обучать модель классифицировать только силлабо-тонические размеры у стихотворений, так как они являются основными и легко формализуются. Однако в современных поэтических текстах ритмическая организация нередко отклоняется от классических схем силлабо-тонического стихосложения. Отклонения проявляются в виде пропусков (пиррихий — метрическая позиция, где вместо ожидаемого ударного слога появляется безударный) или добавлений ударений (спондей — два ударных слога подряд вместо ожидаемого чередования), варьирования числа слогов в строке, а также использования синтаксических или интонационных пауз, компенсирующих нарушения метрической схемы. В результате строка может сохранять общее ритмическое ощущение, несмотря на формальные отклонения от регулярного метра,

что требует более гибкого подхода к автоматическому анализу метра, о чем будет сказано далее при описании алгоритма определения метра.

Для создания модели определения стихотворного размера воспользуемся данными поэтического подкорпуса НКРЯ, описанного выше. Информация о метре приводится к общему виду и переводится в метки. Далее текст проходит процесс обработки. Всего полученный корпус состоит из 50000 токенов, состоящих из строк стихотворений и их размерности, по 10000 строк каждого из пяти видов размерности. Часть получившегося корпуса представлена в таблице 1.

**Таблица 1.** Пример обучающего корпуса

Строка	Размерность из НКРЯ	Класс размерности
в капиллярах ненастья и вереска	Ан3д 2*2*2*2	Анапест
где звёзды словно виноград	Я4м	Ямб
вещи нездешней формы	Дк3ж 0*2*1*1	Дактиль
рассеянно слушаю	Аф2д 1*2*2	Амфибрахий
прилетает птица вновь	Х4м 0*1*1*1*0	Хорей

На вход модели подается строка с текстом стихотворения, далее происходят этапы предобработки: входной текст разделяется на массив строк, удаляются лишние символы и знаки препинания. Производится акцентуация строк с помощью библиотеки «RUAccent», обученной расставлять ударение в том числе на поэтическом корпусе, ударения маркируются добавлением символа «+» перед ударной гласной. Далее для каждой строки генерируется схема ударности слогов из трех меток: основное ударение (P) — для слогов, размеченных «RUAccent» как ударные, вторичное ударение (S) — для односложных и не определенных библиотекой слов, безударный слог (U) — для всех остальных [4]. В таблице 2 приведен пример сгенерированной схемы.

**Таблица 2.** Пример сгенерированной схемы ударности

Строка	Схема ударности строки
Дубовый листок оторвался от ветки родимой	U P U U P U U P U S P U U P U
И в степь укатился, жестокою бурей гонимый;	S S U U P U U P U U P U U P U
Засох и увял он от холода, зноя и горя	U P S U P S S P U U P U S P U
И вот, наконец, докатился до Чёрного моря.	S S U U P U U P U S P U U P U

Далее на основе полученной схемы мы реализуем два варианта определения размерности стихотворения: с помощью сравнения ритмических схем исследуемого стихотворения с набором силлабо-тонических масок, а также с помощью обучения нейронной сети.

Маски пяти возможных ритмических схем генерируются из элементов P (ударные слоги) и U (безударные слоги) по количеству слогов заданной строки. Они сравниваются с полученной схемой, штрафуются при несовпадении меток слогов, после чего наиболее

похожая маска возвращается в качестве результата стихотворного размера строки. Отклонение увеличивается в двух случаях: если в маске слог обозначен как P (основной ударный), а в полученной схеме — U (безударный), и наоборот, если в маске слог обозначен как безударный, а в полученной схеме — основной ударный. При несовпадении со слогом, обозначенным как вторично ударный, отклонение не увеличивается. Чтобы определить, на какие веса увеличивать отклонение, мы перебрали комбинации весов от 1 до 10 и рассчитали с ними точность (precision) работы алгоритма. При тестировании алгоритма на размеченном корпусе точность (precision) по всем классам 0,84 достигается при весах 1 и 2 соответственно, результаты оценки качества работы алгоритма приведены в таблице 3.

**Таблица 3.** Оценка качества работы алгоритма

	Precision (точность)	Recall (полнота)	F1-score (F1-мера)
Амфибрахий	0,85	0,91	0,88
Анапест	0,83	0,90	0,86
Дактиль	0,90	0,58	0,71
Хорей	0,82	0,89	0,85
Ямб	0,80	0,90	0,85

## 2.2 Нейросетевая классификация

Для улучшения определения стихотворного размера мы применили нейросетевую классификацию. Простейшая полносвязная нейросеть, на вход которой подаются массивы ритмических схем строк, приведенные к общей длине, на тестовых данных продемонстрировала точность (precision) 0,90. Нейросеть на базе LSTM, архитектуре, позволяющей обрабатывать последовательные данные, показала точность (precision) 0,91.

Была протестирована нейросеть, на вход которых подавались не только схема ударности строки, но и дополнительный параметр — метка с результатом работы описанного выше алгоритма. Она показали аналогичную точность (precision), но обучалась быстрее. Также были протестированы жесткие и мягкие голосования нейросети и алгоритма, но это не дало существенного увеличения качества модели. Оценка качества итоговой модели приведена в таблице 4.

**Таблица 4.** Оценка качества работы модели

	Precision (точность)	Recall (полнота)	F1-score (F1-мера)
Амфибрахий	0,95	0,94	0,95
Анапест	0,92	0,93	0,93
Дактиль	0,86	0,86	0,86
Хорей	0,89	0,90	0,90
Ямб	0,89	0,90	0,90

На данном этапе интерес представляет анализ матрицы замен, в дальнейшем планируем подробнее исследовать данную тему. В процессе исследования мы заметили, что и при сравнении с масками ритмических схем, и при нейросетевой классификации наименьшие значения точности наблюдаются для класса «дактиль». Вероятно, это связано с тем, что в используемом корпусе данный размер соблюдается авторами менее точно: в некоторых строках наблюдается пропуск или добавление ударений, не соответствующих классической схеме дактиля. В таких случаях при разметке учитывались схемы других строк, чтобы все строки стихотворения имели одинаковую размерность. Примеры таких строк и их размер их поэтического корпуса НКРЯ приведены в таблице 5. Числа означают количество безударных слогов между \* — ударными слогами. Такие строки с неклассической для дактиля разметкой модель соотносит с другими классами.

Таблица 5. Примеры со сложными случаями дактиля

Строка	Размер из НКРЯ	Предсказание модели
не надо мне счастья не надо снов	Дк4м 1*2*2*1*0	Амфибрахий
пресмыканье страсти и взлет	Дк3м 2*1*2*0	Хорей
без тебя утверждать что ты	Дк3м 2*2*1*0	Анапест
твою жестокость тая	Дк3м 1*1*2*0	Ямб
в те края где я рос под кленом	Дк3ж 2*2*1*1	Хорей

### 3. Определение характеристик стихотворений

#### 3.1 Определение схемы рифмовки

Одним из способов поиска рифмующихся строк является использование веб-приложения «Большой словарь рифм», которое принимает слово и возвращает множество рифмующихся с ним слов. Такой способ используется в ряде исследований, например, в работе В. Н. Бойкова [5]. Однако отправка запросов веб-приложению для каждого интересующего слова для извлечения множеств рифмующихся слов требует больших ресурсов.

В данной работе используется готовая модель из библиотеки чешского ученого Петра Плехача RhymeTagger [6], которая предназначена для поиска рифмы в стихотворениях. Модель работает с русским языком и возвращает массив меток строк в зависимости от их рифмы. Пример работы приведен в таблице 6.

Таблица 6. Пример работы RhymeTagger

Текст	Схема рифмовки
На чешуе жестяной рыбы прочёл я зовы новых губ. А вы ноктюрн сыграть могли бы на флейте водосточных труб?	[1, 2, 1, None, 1, 2] (первая по порядку строчка рифмуется с третьей и пятой, а вторая — с шестой, None означает, что у строки отсутствует рифмующаяся пара)

Для тестирования качества библиотеки мы воспользовались корпусом «RIFMA: Русская поэзия с акцентуационными аннотациями» [7]. Это корпус Ильи Козиева с 3834 токенами: тексты стихотворений с размеченными ударениями и схемы рифмовки в этих текстах. Протестировав работу модели на данных корпуса с акцентуационными аннотациями, мы получили правильное определение схемы на 92 % материала. До решения использовать готовую обученную модель мы пытались обучить собственную модель, однако столкнулась с нехваткой обучающих данных. В свободном доступе есть корпуса с размеченной рифмой, а также множество веб-сервисов, позволяющих генерировать рифму к словам, однако нам не удалось найти достаточное количество качественного вручную размеченного материала со строго не рифмующимися строками. По этой причине наша модель давала точность (precision) меньше, чем RhymeTagger.

### 3.2 Семантический анализ текста

Семантический анализ текста — процесс выявления и интерпретации смысловых связей между словами, фразами, предложениями и частями текста. В отличие от формального анализа, семантический анализ направлен на понимание глубинного смысла, который может зависеть от контекста и структуры текста.

Таблица 7. Темы и ключевые слова

Тема	Ключевые слова
Любовь	любовь, сердце, страсть, поцелуй, чувства
Разлука	расставание, прощание, слезы, утрата, боль
Природа	лес, ветер, река небо солнце дождь
Война	битва, пуля, кровь, герой, враг, солдат
Свобода	свобода, неволя, цепи, независимость, выбор
Смерть	смерть, могила, тьма
Жизнь	жизнь, дорога, судьба, путь, движение
Вера	бог, молитва, храм, душа, вера
Родина	родина, страна, флаг, народ, земля
Дружба	друг, поддержка, верность, плечо, забота
Воспоминания	воспоминание, прошлое, память
Одиночество	одиночество, тишина, пустота
Надежда	надежда, свет, рассвет, мечта
Детство	детство, игрушка, мама, школа, беззаботность
Искусство	музыка, поэзия, картина, художник, вдохновение



- синтаксический анализ и подсчет частотности частей речи;
- поиск параллелизма, анафоры и эпитифоры.

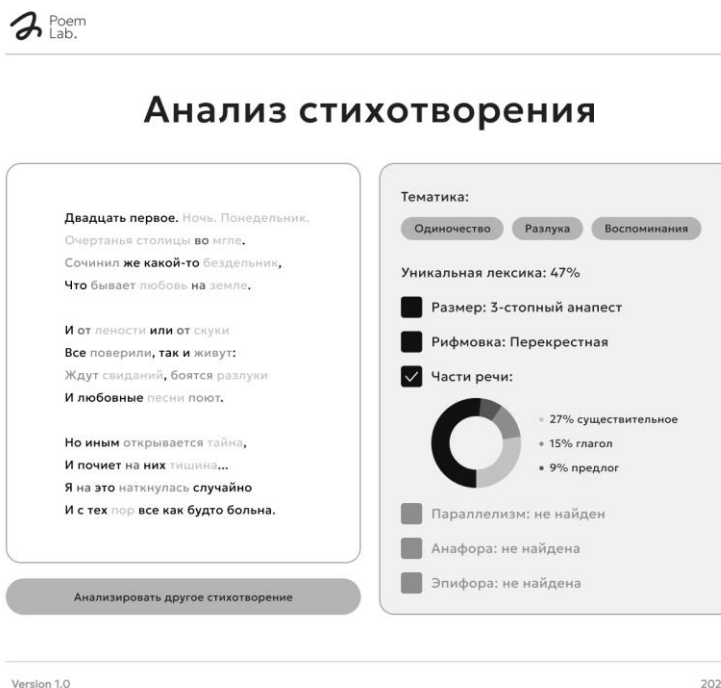


Рис. 2. Интерфейс веб-приложения

Результаты выводятся пользователю в визуально-интерактивной форме (рис. 2).

## 5. Заключение

В ходе исследования был разработан алгоритм определения стихотворного размера, сочетающий классический подход сопоставления ритмических шаблонов с обученной нейросетью. Алгоритмический метод продемонстрировал точность (precision) 0,84, а LSTM-модель — до 0,91. Реализован модуль семантического анализа на базе трансформера RuBert, позволяющий автоматически выделять ключевые темы стихотворения на основе косинусного сходства эмбедингов. Также для приложения были разработаны и внедрены модули лексического, синтаксического анализа, была интегрирована готовая модель RhymeTagger для поиска рифмующихся строк, которая продемонстрировала точность (precision) 0,92 на обучающем корпусе.

В результате была разработана и реализована веб-платформа для автоматизированного анализа поэтических текстов, объединяющая различные модули анализа стихотворений. Веб-приложение реализовано на базе Flask и React. Разработанное приложение применимо в образовательной среде для наглядного демонстрация приемов стихосложения и анализа стихотворных форм, на платформах электронных библиотек и литературных сообществ для предоставления читателям возможностей интерактивного анализа поэзии.

В дальнейшем мы планируем продолжить исследование и расширить веб-приложение новыми методами и корпусами. В частности, мы планируем дообучить модель ruBERT на поэтическом материале для улучшения определения семантических тем, провести тематическое моделирование, улучшить алгоритм определения размеров стихотворений, добавить фонетический анализ (поиск аллитерации и ассонанса).

## Литература

- [1] Барахнин В.Б., Кожемякина О.Ю., Забайкин А.В., Хаятова В.Д. Автоматизация комплексного анализа русского поэтического текста: модели и алгоритмы // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2015. № 3. С. 5-18.
- [2] Барахнин В.Б., Кожемякина О.Ю., Борзилова Ю.С. Проектирование информационной системы представления результатов комплексного анализа поэтических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2019. Т. 17. № 1. С. 5-17.
- [3] Морозов А.Ю. Использование троичной системы для измерения стихотворного размера // Образовательные ресурсы и технологии. 2024. № 3(48). С. 18-24. DOI: 10.21777/2500-2112-2024-3-18-24
- [4] Agirrezabal M., Astigarraga A., Arrieta B., Hulden M. ZeuScansion: A tool for scansion of English poetry // Journal of Language Modelling, 2016. Vol. 4. No. 1. P. 3-28. DOI: 10.15398/jlm.v4i1.102
- [5] Бойков В.Н., Каряева М.С., Соколов В.А., Пильщиков А.И. Об автоматической спецификации стиха в информационно-аналитической системе // CEUR Workshop Proceedings. 2015. Vol. 1536. P. 144-151. URL: <https://ceur-ws.org/Vol-1536/paper22.pdf> (дата обращения: 21.03.2025).
- [6] Plecháč P. A collocation-driven method of discovering rhymes (in Czech, English, and French poetry) // Taming the Corpus: From Inflection and Lexis to Interpretation. Cham: Springer, 2018. P. 163-175. (Quantitative Methods in the Humanities and Social Sciences). DOI: 10.1007/978-3-319-98041-7\_10
- [7] Козиев И. Автоматическая оценка метра и рифмы в сгенерированной и авторской русской поэзии // arXiv preprint. 2025. URL: <https://arxiv.org/abs/2502.20931> (дата обращения: 21.03.2025).

### Development of a System for Multidimensional Analysis of Poetic Texts

A. I. Pankova, E. V. Yagunova

Saint Petersburg State University

This paper discusses the development of modules for the automated analysis of poetic texts using methods from machine learning and computational linguistics. We examine the structural features of poetic texts and approaches to their analysis, as well as techniques for computing various textual characteristics.

Algorithmic and neural network models have been developed and evaluated for identifying syllabo-tonic poetic meters. Semantic analysis was conducted using the RuBERT transformer model, which enables automatic extraction of key poem themes based on cosine similarity of embeddings. Additionally, syntactic analysis (including part-of-speech ratio calculation and detection of parallelism) and lexical analysis (e.g., quantifying rare word usage based on a custom frequency dictionary for poetry) were performed. The analysis was implemented using Python programming libraries. The research material consisted of publicly available Russian-language poetry corpora.

The implemented modules were integrated into a web application designed for the multidimensional analysis of poems. The resulting application can be used in educational institutions for demonstrative teaching of versification and poetic form analysis, as well as on literary community platforms to provide tools for interactive poetry analysis. The application is planned to be scaled and extended with new models and corpora.

**Keywords:** computational linguistics, automatic analysis, Russian poetry

**Reference for citation:** Pankova A. I., Yagunova E. V. Development of a System for Multidimensional Analysis of Poetic Texts // Computational Linguistics and Computational Ontologies. Vol. 9 (Proceedings of the XXVIII International Joint Scientific Conference «Internet and Modern Society», IMS-2025, St. Petersburg, June 23–25, 2025). — St. Petersburg: ITMO University, 2025. P. 120-129. DOI: 10.17586/3033-5582-2025-9-120-129.

## References

- [1] Barakhnin V.B., Kozhemyakina O.Yu., Zabaykin A.V., Khayatova V.D. Avtomatizatsiya kompleksnogo analiza russkogo poeticheskogo teksta: modeli i algoritmy // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii. 2015. No. 3. P. 5-18. (In Russian)
- [2] Barakhnin V.B., Kozhemyakina O.Yu., Borzilova Yu.S. Proektirovanie informatsionnoy sistemy predstavleniya rezultatov kompleksnogo analiza poeticheskikh tekstov // Vestnik Novosibirskogo gosudarstvennogo universiteta. Seriya: Informatsionnye tekhnologii. 2019. Vol. 17. No. 1. P. 5-17. (In Russian)
- [3] Morozov A.Yu. Ispol'zovanie troichnoy sistemy dlya izmereniya stikhotvornogo razmera // Obrazovatel'nye resursy i tekhnologii. 2024. No. 3 (48). P. 18-24. DOI: 10.21777/2500-2112-2024-3-18-24 (In Russian)
- [4] Agirrezabal M., Astigarraga A., Arrieta B., Hulden M. ZeuScansion: A tool for scansion of English poetry // Journal of Language Modelling. 2016. Vol. 4. No. 1. P. 3-28. DOI: 10.15398/jlm.v4i1.102
- [5] Boykov V.N., Karyaveva M.S., Sokolov V.A., Pilshchikov A.I. Ob avtomaticheskoy spetsifikatsii stikha v informatsionno-analiticheskoy sisteme // CEUR Workshop Proceedings. 2015. Vol. 1536. P. 144-151. URL: <https://ceur-ws.org/Vol-1536/paper22.pdf> (accessed date: 21.03.2025). (In Russian)
- [6] Plecháč P. A collocation-driven method of discovering rhymes (in Czech, English, and French poetry) // Taming the Corpus: From Inflection and Lexis to Interpretation. Cham: Springer, 2018. P. 163-175. (Quantitative Methods in the Humanities and Social Sciences). DOI: 10.1007/978-3-319-98041-7\_10
- [7] Koziyev I. Avtomaticheskaya otsenka metra i rifmy v sgenerirovannoy i avtorskoj russkoj poezii // arXiv preprint. 2025. URL: <https://arxiv.org/abs/2502.20931> (accessed date: 21.03.2025). (In Russian)

## Сведения об авторах

**Васильева Елена Викторовна**, Иркутский государственный университет, старший преподаватель, помощник директора Института филологии, иностранных языков и медиакоммуникации

**Герасимова Анастасия Алексеевна**, кандидат филологических наук, Московский государственный университет им. М. В. Ломоносова, научный сотрудник Научно-исследовательского вычислительного центра, ORCID 0000-0003-4686-5598

**Исаева Екатерина Владимировна**, кандидат филологических наук, доцент, Пермский государственный национальный исследовательский университет, заведующая кафедрой английского языка профессиональной коммуникации, ORCID 0000-0003-1048-7492

**Коган Марина Самуиловна**, кандидат технических наук, доцент, Санкт-Петербургский политехнический университет Петра Великого, доцент, ORCID 0000-0002-7519-2161

**Лаврентьева Екатерина Петровна**, Санкт-Петербургский политехнический университет Петра Великого, студент

**Лютикова Екатерина Анатольевна**, доктор филологических наук, Московский государственный университет им. М. В. Ломоносова, профессор кафедры теоретической и прикладной лингвистики филологического факультета, ведущий научный сотрудник Научно-исследовательского вычислительного центра, ORCID 0000-0003-4439-0613

**Митрофанова Ольга Александровна**, кандидат филологических наук, доцент, Санкт-Петербургский государственный университет, доцент, ORCID 0000-0002-3008-5514

**Панкова Арина Ильинична**, Санкт-Петербургский государственный университет, студент

**Сафарбеков Бехруз Зафарович**, Национальный исследовательский технологический университет «МИСИС», студент, ORCID 0009-0009-6021-7897

**Студеникина Ксения Андреевна**, Московский государственный университет им. М. В. Ломоносова, аспирант, младший научный сотрудник кафедры теоретической и прикладной лингвистики филологического факультета, ORCID 0000-0002-4098-7167

**Суров Илья Алексеевич**, кандидат физико-математических наук, доцент, Университет ИТМО, доцент, ORCID 0000-0001-5690-7507

**Черников Кирилл Михайлович**, Университет ИТМО, студент

**Шамаева Елена Денисовна**, Московский государственный университет им. М. В. Ломоносова, аспирант, ORCID 0009-0002-6233-8958

**Ягунова Елена Викторовна**, доктор филологических наук, Санкт-Петербургский государственный университет, профессор кафедры информационных систем в искусстве и гуманитарных науках

**Авторский указатель**

Васильева Е. В.	78	Панкова А. И.	120
Герасимова А. А.	100	Сафарбеков Б. З.	60
Исаева Е. В.	60	Студеникина К. А.	100
Коган М. С.	12	Суров И. А.	48
Лаврентьева Е. П.	12	Черников К. М.	48
Лютикова Е. А.	100	Шамаева Е. Д.	26
Митрофанова О. А.	93	Ягунова Е. В.	120

## Содержание

XXVIII Международная объединённая научная конференция «Интернет и современное общество» (IMS-2025).....	3
От редколлегии.....	10
Цифровое доверие как ключевой фактор в формировании датацентричного государственного управления	
Лаврентьева Е. П., Коган М. С. ....	12
Сравнение нейросетевых синтаксических анализаторов для русского языка	
Шамаева Е. Д. ....	26
Геометрия падежей в векторных моделях русского языка	
Черников К. М., Суров И. А. ....	48
Оптимизация обработки терминологии в беспилотной авиации: новый подход к извлечению терминов с использованием промт-инжиниринга	
Исаева Е. В., Сафарбеков Б. З. ....	60
Проблемы исследования словообразовательного потенциала с использованием современных поисковых систем: автоматизированный отбор дериватов через Яндекс	
Васильева Е. В. ....	78
Динамика тем научных статей в корпусе текстов по компьютерной и корпусной лингвистике	
Митрофанова О. А. ....	93
Оценка лингвистической компетенции больших языковых моделей на материале корпуса согласовательной вариативности	
Студеникина К. А., Лютикова Е. А., Герасимова А. А. ....	100
Разработка системы анализа разноплановых характеристик поэтического текста	
Панкова А. И., Ягунова Е. В. ....	120
Сведения об авторах.....	130
Авторский указатель.....	131

Компьютерная лингвистика и вычислительные онтологии. Выпуск 9 (Труды XXVIII Международной объединенной научной конференции «Интернет и современное общество», IMS-2025, Санкт-Петербург, 23–25 июня 2025 г.) Сборник научных трудов. — СПб.: Университет ИТМО, 2025. — 133 с.

**Компьютерная лингвистика и вычислительные онтологии**

**Выпуск 9**

Сборник научных трудов

Под редакцией А. В. Чижик и Д. Е. Прокудина  
Дизайн обложки С. Н. Ушаков  
Оригинал-макет Т.А. Новикова, Ю. В. Байкеева

Университет ИТМО, 197101, Санкт-Петербург,  
Кронверкский пр. 49, лит. А.