ONLINE: ISSN 2075-7182

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2025)

Выпуск 23

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2025)

Issue 23

Редакционная коллегия: *В. П. Селегей (главный редактор), В. И. Беликов, И. М. Богуславский, Б. В. Добров, Д. О. Добровольский, И. М. Кобозева, Е. Б. Козеренко, М. А. Кронгауз, Н. В. Лукашевич, Д. Маккарти, П. Наков, Й. Нивре, В. Раскин, Э. Хови, Т. О. Шаврина, С. А. Шаров, Т. Е. Янко*

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 23. 2025. С. I–492.

Сборник включает 39 докладов международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2025», представляющих широкий спектр теоретических и прикладных исследований в области описания естественного языка, моделирования языковых процессов, создания практически применимых компьютерных лингвистических технологий.

Для специалистов в области теоретической и прикладной лингвистики и интеллектуальных технологий.

© Оформление. Редакционная коллегия, 2025

Предисловие

23-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 30-й международной онлайн-конференции «Диалог». В 2025 году для публикации в основном томе сборника редколлегией были отобраны 39 докладов. Работы, представленные в сборнике, отражают те направления исследований в области компьютерного моделирования и анализа естественного языка, которые по традиции представляются на Диалоге:

- **Интеллектуальный анализ документов;**
- **Лингвистический анализ текста;**
- **Глубокое обучение в компьютерной лингвистике:** методики и примеры применения в лингвистических исследованиях, содержательная интерпретация работы LLM;
- **Компьютерные лингвистические ресурсы:** включая новые датасеты и новые сценарии и типы разметки, Evaluation Benchmarks;
- **Компьютерный анализ Social Media;**
- **Корпусная лингвистика и корпусометрия:** методики создания, использования и оценки корпусов;
- **Компьютерная семантика:** аналитические и дистрибуционные модели, связь между ними;
- **Лингвистические онтологии и автоматическое извлечение знаний;**
- **Мультимодальная коммуникация:** аналитические и нейронные модели речевого акта;
- **Модели общения и диалоговые агенты;**
- **Компьютерная лексикография;**
- **Многоязычие:** language transfer, новые технологии работы с малоресурсными языками.

В соответствии с традициями «Диалога» отбор работ основывается на представлении о важности соединения новых методов и технологий анализа языковых данных с полноценным лингвистическим анализом. Диалог является де-факто крупнейшим форумом по проблемам создания современных компьютерных ресурсов, моделей и технологий для русского языка, поэтому ключевым событием «Диалога» является подведение итогов технологических соревнований между разработчиками систем лингвистического анализа русскоязычных текстов — *Dialogue Evaluation*. В сборнике представлены статьи участников и организаторов 3-х соревнований:

- **CoBaLD Parsing:** Соревнование по автоматической лингвистической разметке;
- **RuOpinionNE-2024:** Соревнование по извлечению мнений из новостных текстов;
- **RuTermEval-2024:** Соревнование по выявлению терминов в научных статьях на русском языке.

Статьи в сборнике публикуются на русском и английском языках. При выборе языка публикации действует следующее правило:

- доклады по компьютерной лингвистике подаются на английском языке. Это расширяет их аудиторию и позволяет привлекать к рецензированию международных экспертов;
- доклады, посвященные лингвистическому анализу русского языка, предполагающие знание этого языка у читателя, подаются на русском языке (с обязательной аннотацией на английском).

Несмотря на традиционную широту тематики представленных на конференции и отобранных в сборник докладов, они не могут дать полной картины направлений «Диалога». Ее можно получить с помощью сайта конференции https://dialogue-conf.org/, на котором представлены обширные электронные архивы «Диалогов» последних лет и все результаты проведенных тестирований Dialogue Evaluation.

Мы обращаем внимание авторов и читателей сборника, что с 2018 года Редсовет отказался от печати сборника на бумаге. Все сборники размещаются на сайте конференции. С 2014 года основной том индексируются Scopus.

Программный комитет конференции «Диалог»
Редколлегия сборника «Компьютерная лингвистика и интеллектуальные технологии»

Рецензенты

Баюк Александра Михайловна
Баюк Илья Сергеевич
Беликов Владимир Иванович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бухаров Ян Михайлович
Васильев Виталий Геннадьевич
Галицкий Борис Васильевич
Добров Борис Викторович
Добровольский Дмитрий Олегович
Жарков Андрей
Ильвовский Дмитрий Алексеевич
Инькова Ольга Юрьевна
Киосе Мария Ивановна
Киселева Ксения Львовна
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Лепехин Михаил
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Лукашевич Наталья Валентиновна
Ляшевская Ольга Николаевна
Мамонтова Ангелина
Митрофанова Ольга Александровна
Мичурина Мария Александровна
Никишина Ирина Юрьевна
Пазельская Анна Германовна
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Пиперски Александр Чедович
Подлесская Вера Исааковна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Смирнов Иван Валентинович
Смуров Иван Михайлович
Татевосов Сергей Георгиевич
Урысон Елена Владимировна
Федорова Ольга Викторовна
Феногенова Алена Сергеевна
Хохлова Мария Владимировна
Циммерлинг Антон Владимирович
Шаврина Татьяна Олеговна
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents[1]

Akhmadeeva I., Kononenko I., Sidorova E., Shestakov V.
Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts 1

Alibekov A., Migal A., Matenkov A., Muryshev A., Bolshakov V., Kozachenko A., Mukhtarova G., Mikhaylovskiy N.
RuTaR—A Dataset in Russian for Reasoning about Taxes 12

Baiuk I., Baiuk A., Petrova M.
CoBaLD Parser: Joint Morphosyntactic and Semantic Annotation 28

Баранов А. Н.
***На отскребись*: вложенная цитация как фактор идиоматичности** 38

Богуславский И. М.
Совместность и раздельность 45

Бонч-Осмоловская А. А., Гладилин С. А., Козеренко А. Д., Ляшевская О. Н., Морозов Д. А., Кузнецова Ю. Н., Махова А. А., Пискунова С. В., Буйлова Н. Н., Бородина Д. Г., Виноградова И. И., Сизов В. Г., Дьяченко П. В., Казенников А. О., Власова Н. А., Глазкова А. В., Столяров С. С., Гарипов Т. А., Смаль И. А., Губарькова Я. Н.
Национальный корпус русского языка 2.0: корпусная платформа, инструменты анализа, нейросетевые модели разметки данных 57

Буров Д. А., Панич М. Б., Садковский Ф. А., Федорова О. В., Шевелев С. М.
Задача лексического решения: модальность имеет значение 74

Chuikova O. Iu., Gorbova E. V.
Derivation of Aktionsarten from (non-)directional verbs of motion: posing restrictions on restrictions 86

Galitsky B., Ilvovsky D., Morkovkin A.
Enhancing RAG and Knowledge Graphs with Discourse 103

Gromenko E., Kalacheva D., Klokova K., Krongauz M., Moroz O., Shulginov V., Yudina T.
Cultural Evaluation of LLMs in Russian: Catchphrases and Cultural Types 117

Инькова О. Ю.
Неоднозначность коннекторов: некоторые уточнения 129

Ivleva I., Solovyev V.
An Algorithm for Genre Imbalance Correction in the Russian Subcorpus of the Google Books Ngram Corpus 137

Khomenko A., Komratova A., Isakov D., Balba D., Shishkovskaya T., Khudyakova M.
Computer tools in mental disorders diagnostics by oral speech 147

Киосе М., Ржешевская А., Петров А.
Локальная структура дискурса описания в речи взрослых, изучающих русский язык как иностранный 158

[*] The reports of each section are ordered by the surname of the first author in compliance with the English alphabet.

Коротаев Н. А.
Дискурсивные контексты наложений в диалоге на русском языке 169

Kotov A. A., Nosovets Z. A., Filatov A. A., Arinkin N. A.
The development of a natural language reasoning system for a companion robot 183

Kovalev G., Tikhomirov M., Kozhevnikov E., Kornilov M., Loukachevitch N.
Building Russian Benchmark for Evaluationof Information Retrieval Models 195

Lee G. F., Loukachevitch N. V., Khokhlov A. R.
Generating Encyclopedic Articles Based on a Collection of Scientific Publications 206

Lepekhin M., Sharoff S.
Causal Models and Adversarial Training: Selecting the right properties for robust non-topical text classification 224

Loukachevitch N. V., Tkachenko N. M., Lapanitsyna A. M., Tikhomirov M. M., Rusnachenko N. L.
RuOpinionNE-2024: Extraction of Opinion Tuples from Russian News Texts 234

Mamontova A., Ischenko R., Vorontsov K.
RuTermEval-2024: Cross-domain Automatic Term Extraction and Classification in Russian scientific texts 245

Morozov D., Glazkova A., Garipov T.
BERT-like Models for Automatic Morpheme Segmentation of the Russian Language 257

Подлесская В. И.
«Ой, нет! Не в среду! В четверг!»: корпусное исследование междометия в контексте речевого сбоя 267

Prokofyeva O., Kiose M., Leonteva A., Smirnova E.
Intensity and its manifestations in speech and recurrent gestures in spontaneous dialogue 281

Rossyaykin P.
Structured sentiment analysis using few-shot prompting of an ensemble of LLMs 291

Rozhkov I. S., Loukachevitch N. V.
Methods for Recognizing Nested Terms 299

Satdarov K., Kharlamova D., Yakuboy A., Lezina A.
Faroese Corpus Development: Strategies for Low-Resource Languages Corpora 312

Semak V. V., Bolshakova E. I.
Comparing Transformer-Based Approaches for Term Recognition in Russian texts 323

Шерстинова Т. Ю., Мельник А. Г., Петрова И. А., Азаревич К. И., Мелкозерова В. И., Чеповецкая С. В.
«Оки доки, вот вам без рофла!»: Повседневная речь молодежи в корпусном представлении (структура и назначение звукового корпуса КУРС/ESC) 331

Shulginov V., Şimşek H. B., Kudriashov S., Randautsova R., Shevela S. A.
Evaluating the Pragmatic Competence of Large Language Models in Detecting Mitigated and Unmitigated Types of Disagreement 345

Sidorova E. A., Ivanov A. I., Ilina D. V., Ovchinnikova K. A., Osmushkin N. M., Sery A. S.
An Approach to Information Extraction from Texts of a Limited Subject Domain Based on a Chain of Large Language Models 361

Sorokin A., Nasyrova R.
LORuGEC: the Linguistically Oriented Rule-annotated corpus for Grammatical Error Correction of Russian 374

Studenikina K. A., Lyutikova E. A., Gerasimova A. A.
Gradual Acceptability Judgments with LLMs: Evidence from Agreement Variation in Russian 384

Татевосов С. Г., Киселева К. Л.
Видовые пары глагольных идиом в русском языке: условия и механизмы аспектуальной редукции 393

Тимошенко С. П., Сердобольская Н. В., Кобозева И. М.
Статистические метрики как критерии отбора составных коннекторов (на материале базы Рускон) 407

Vatolin A.
Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024 416

Vatolin A., Gerasimenko N., Loukachevitch N., Ianina A., Vorontsov K.
ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian 435

Янко Т. Е.
Русский upstep: просодическая модель переспроса к плану выражения 460

Zimmerling A. V., Baiuk A. M.
Operator words versus MT systems and LLMs 471

Abstracts 481

Авторский указатель 489

Author Index 490

Using Rhetorical Structures to Analyze Argumentation in Scientific Communication Texts

Irina Akhmadeeva
A.P. Ershov Institute of Informatics Systems / Novosibirsk, Russia
i.r.akhmadeeva@iis.nsk.su

Irina Kononenko
A.P. Ershov Institute of Informatics Systems / Novosibirsk, Russia
irina_k@cn.ru

Elena Sidorova
A.P. Ershov Institute of Informatics Systems / Novosibirsk, Russia
lsidorova@iis.nsk.su

Vladimir Shestakov
A.P. Ershov Institute of Informatics Systems / Novosibirsk, Russia
shestakov@iis.nsk.su

Abstract

This article explores the role of rhetorical structures in the argument mining task on the material of scientific Internet communication texts in Russian. Two approaches are proposed and studied for argumentative relation prediction: the first one constructs segment vector representations using a Graph Neural Network (GNN) based on rhetorical structure, and the second one uses multitask learning that combines argumentation extraction with rhetorical relations prediction tasks. With proposed approaches three models were implemented: two variations of the model using GNN and one model employing the multitask approach. These models were compared with a simple baseline using the Lonformer model on a dataset annotated with both argumentative and rhetorical structures. Argumentative annotating was performed manually by four experts. Existing resources and tools were used to obtain rhetorical markup. The conducted experiments showed that the approaches using additional rhetorical information improve the quality of argumentative relation prediction, particularly for long-distance relations. The best performance, with an F1 score of 72.32%, was achieved by a model incorporating GNN-enhanced statement representations.

Keywords: argument mining; text corpus; argument annotation; argumentative relation prediction; rhetorical structure theory

DOI: 10.28995/2075-7182-2025-23-1-11

Использование риторических структур для анализа аргументации в текстах научной коммуникации

Ирина Ахмадеева
Институт систем информатики им. А.П. Ершова СО РАН / Новосибирск, Россия
i.r.akhmadeeva@iis.nsk.su

Ирина Кононенко
Институт систем информатики им. А.П. Ершова СО РАН / Новосибирск, Россия
irina_k@cn.ru

Елена Сидорова
Институт систем информатики им. А.П. Ершова СО РАН / Новосибирск, Россия
lsidorova@iis.nsk.su

Владимир Шестаков
Институт систем информатики им. А.П. Ершова СО РАН / Новосибирск, Россия
shestakov@iis.nsk.su

Аннотация

В данной работе исследована роль риторических структур в задаче извлечения аргументации на материале текстов научной интернет-коммуникации на русском языке. Предложены и изучены два подхода для

предсказания аргументативных отношений: первый использует риторическую структуру при построении векторных представлений сегментов с помощью GNN, а второй использует многозадачное обучение, совмещая извлечение аргументации с предсказанием риторических связей. В рамках предложенных подходов были реализованы три модели: две вариации модели с использованием GNN и одна модель, использующая многозадачный подход. Эти модели сравнивались с простой базовой с использованием модели Lonformer на наборе данных, аннотированном как аргументативными, так и риторическими структурами. Аргументационная разметка выполнялась вручную четырьмя экспертами. Для получения риторической разметки применялись существующие ресурсы и инструменты. Проведенные эксперименты показали, что подходы, использующие дополнительную риторическую информацию, улучшают качество предсказания аргументативных отношений, особенно для дальних связей. Лучшее качество с F1 = 72,32% показала модель, в которой векторные представления утверждений модифицировались с помощью графовой нейронной сети и дополнялись информацией о типе риторической связи.

Ключевые слова: извлечение аргументации; корпус текстов; аргументативная разметка текста; предсказание аргументативных отношений; теория риторических структур

1 Introduction

The analysis of reasoning structures presented in natural language texts has emerged relatively recently as a field of computational linguistics that is attracting increasing attention from researchers. Text fragments are sequences of functional steps that contribute to the presentation of the author's idea, supporting its understanding and/or acceptance. In applied linguistics, the structure of reasoning has been considered from a functional point of view based on the theory of rhetorical structure [1–4] and the theory of argumentation [5].

The study of discourse involves describing its structure in the form of discourse units connected by various relations. One of the most famous models used to solve this problem is the Rhetorical Structure Theory (RST) and its modifications [1, 2]. Under RST, simple sentences, clauses and some collapsed propositions (represented by nominalizations and prepositional groups) are connected by symmetric (multinuclear) and asymmetric (mononuclear) relations. In this case, larger discourse units are formed, thereby creating an integral tree-like structure.

Argumentation presented by the author to convince the audience of a particular position plays a special role in discourse analysis. The most famous study of argumentation that has found application in practical argumentation analysis is the model of Douglas Walton [6], in which a structured argument is defined as a set of statements consisting of premises, conclusion (thesis), and inference from premises to conclusion. Walton introduced the concept of an argument scheme, a form representing a stereotypical model of reasoning. The work [5] provides a compendium containing 60 basic argumentation schemes.

Practical application of argumentation theory is complicated by a number of factors, including the lack of annotated material, the non-obviousness of identifying argumentation zones, the presence of functionally similar arguments that complicate the choice of a scheme in a specific situation, and insufficiency of the set of schemes when it comes to new genres of texts. Such factors lead to ambiguity in text annotation, disagreement between annotators, which worsens the quality of the created datasets to be applied in machine learning methods and, as a consequence, leads to low quality of automation of argumentation analysis in general.

There exist methods to overcome the lack of argumentatively annotated material. In particular, corpora of texts with multi-level annotation are created to explore the correlation between rhetorical structure presented in RST-markup and argumentative structure shown by A-markup, in order to use existing resources and tools of discourse analysis to extract arguments. Thus, in [7] the development and use in experiments of a two-level corpus of 112 argumentative essays is described, in [8] the material is scientific articles from the field of computational linguistics, and [9] is devoted to the comparison of RST-markup and A-markup of popular science texts. In the presence of such a parallel corpus, neural network methods of machine learning can be applied both based on multitasking and considering RST-markup as additional features.

The purpose of this work is to develop methods for automatic construction of argumentation structure based on RST structure for Russian-language texts. Existing resources and tools for rhetorical analysis of Russian-language texts are used to obtain RST-markup.

The following research questions were formulated within the framework of this work.

Q1. How can language models aimed at analyzing argumentation use information about the rhetorical structure of a text?

Q2. To what extent does the quality of argumentation analysis improve when using rhetorical information obtained with the help of rhetorical parsers?

Experimental studies are conducted on a Russian-language corpus of texts related to the field of scientific communication, equipped with A-markup in accordance with the model of D. Walton.

2 Review

In the last few years, research has emerged examining the potential use of tools for automatically constructing the rhetorical structure of the text to solve the problem of argument mining. Most of the researchers below use the formalism proposed by RST, and parsers that allow you to build a rhetorical structure in the appropriate format.

The idea of using the text rhetorical structure to present argumentation is based on the fact that both illustrate the functional structure of displayed thoughts. More precisely, argumentation reflects the structure of reasoning to prove a thought, and the RST structure is a representation of this reasoning.

In both cases, the elements of the structure that form the vertices of the graph representation are text segments expressing the steps of reasoning, which are considered elementary. The function of the segment is displayed in the final structure as a relation linking it with the structure of the entire text. The final structure in both cases is a directed graph, its vertices forming elementary segments, and the arcs being marked with the names of the relations that link the segments. The coincidence of rhetorical and argumentative segmentations is observed for the rhetorical relations Cause-effect, Condition, Contrast, Concession, Purpose and partially for Attribution. In the "classical" set of rhetorical relations, some authors [10, 11] distinguish relations close to those used in argument analysis: Motivation, Evidence, Justify, Antithesis, Concession.

Two methods of presenting the structure of discourse demonstrate not only common properties, but also significant differences. Therefore A-markup cannot be considered as a special case of RST-markup [9, 11, 12]. Differences are revealed at all levels: a) not all text segments are included in the argumentative structure, but only those that relate to the argumentation zone; b) rhetorical segmentation is more fragmented, e.g. segments connected by Joint, Same-unit and Elaboration rhetorical relations are usually embedded in argumentative segments; c) the opposite situation in which a single rhetorical segment is presented in A-markup by several segments can be caused by pragmatic aspects of the context; d) when constructing the argumentative structure, there is no procedure for enlarging units by combining adjacent segments – the argument and the thing being argued (theses) can be located far from each other, which is quite natural for the structure of many scientific texts.

Several ways of using information about the rhetorical structure of text to extract argumentation have been proposed.

Early approaches used handcrafted features based on the rhetorical structure of text, such as the distance between elementary discourse units, the type of rhetorical relation, the presence of children/parents, etc. [13] used expert RST markups and compared three models: (1) a simple tree-transformation model based on heuristics; (2) an RST and argumentation graph alignment model; and (3) a modification of the Evidence Graph model. [14] also trained the Evidence Graph model, where discourse annotation of text was performed by automatic parsers with use of not only RST annotations but also Penn Discourse Treebank ones.

[15] modified the approach based on the biaffine dependency parser [16] and proposed a model in which the edge weights are refined by the corresponding discourse coefficients obtained from the rhetorical tree. In addition, the authors use several variants of RST annotations obtained for paraphrased statements to account for the ambiguity of the discursive interpretation of the text. The highest Unlabeled Attachment Score (UAS) on the Microtexts corpus [17] reached 64.6%.

And in [18], information about the rhetorical structure is used to form prefixes that guide the generation of argumentation structure using the BART model. BART-Encoder and Relational Graph Convolutional Network are used to generate the prefix. The model proposed by the authors achieved a 58.51% F1 score on the AAEC dataset [19] and a 40.46% F1 score on the AbstRCT dataset [20] for argumentative relation classification subtask.

Another approach was considered in [21], where the authors applied transfer learning on the Discourse Dependency TreeBank for Scientific Abstracts (SciDTB) [22] with added argumentation annotation layer. In this approach, a pre-trained model designed to recognize rhetorical relations was adapted

on a limited training set for the task of recognizing argumentative relations. This model achieved a 40% F1 score for the argumentative attachment task on the SciDTB.

3 Methods

This paper proposes two methods for using rhetorical structure information in the argument mining process. They are compared with a baseline model that does not take such information into account.

In the first approach, the vector representations of statements were modified using a Graph Neural Network (GNN) to account for structural dependencies, and each type of rhetorical relation was assigned a trainable embedding which was combined with the corresponding segment vectors. In the second approach, the classifier was trained in a multitask mode, simultaneously solving the main task of extracting arguments and the auxiliary task of predicting rhetorical relations. This approach allowed the model to build better representations by transferring knowledge between similar tasks.

3.1 Baseline classifier

Let a text fragment consist of m Elementary Discourse Units (EDUs) EDU_i, where $i \in \{1,2,\cdots,m\}$, each including n_i tokens, and the total number of tokens in the text fragment is n. The contextual vector representations of the tokens obtained using encoder Enc are denoted as $e_i^{(1)}, e_i^{(2)}, \cdots, e_i^{(n_i)}$.

Vector representations of EDUs $h_1, h_2, \cdots, h_m$ are formed by aggregating the vector representations of tokens corresponding to each elementary segment:

$$h_i = AGG\left(e_i^{(1)}, e_i^{(2)}, \cdots, e_i^{(n_i)}\right)$$

The vector representations of EDU_i and EDU_j, are concatenated: $h_{ij} = \left[h_i; h_j\right]$, an MLP (Multi Layer Perceptron) classifier is applied to the resulting vector:

$$y_{ij}^{baseline} = Classify\left(h_{ij}\right)$$

where $y_{ij}^{baseline}$ – the predicted probability of a relation between EDU_i and EDU_j. Cross-Entropy Loss (CE Loss) is used in the training.

3.2 GNN based approach

The main peculiarity of this approach is that the vector representations of the EDUs obtained using Enc are further transformed using a Graph Convolutional Network [23] according to the rhetorical structure. The evolution of vector representations at each layer of the GNN is defined as follows:

$$h_i^{(k+1)} = W_k^T \sum_{j \in \mathcal{N}(i) \cup \{i\}} \frac{h_i^{(k)}}{\sqrt{d_i d_j}}$$

where W_k – trainable weight matrices, $N(i)$ – the neighbors for EDU_i in the graph of rhetorical structure, $d_i = 1 + deg(i)$, and $deg(i)$ – vertex degree.

The vector representations $h_1^{(K)}, h_2^{(K)}, \cdots, h_m^{(K)}$ obtained at the last Kth layer of GNN are final and are used in the classifier: $y_{ij}^{graph} = Classify\left(\left[h_i^{(K)}; h_j^{(K)}\right]\right)$, and the initial representations $h_i^{(0)}$ are set equal to h_i for all $i \in \{1,2,\cdots,m\}$.

Figure 1 shows the architecture of the proposed model with the parameter K is configured equal to one, and the number of EDUs in the text is equal to four. The rhetorical structure graph of the text is depicted by dotted lines.

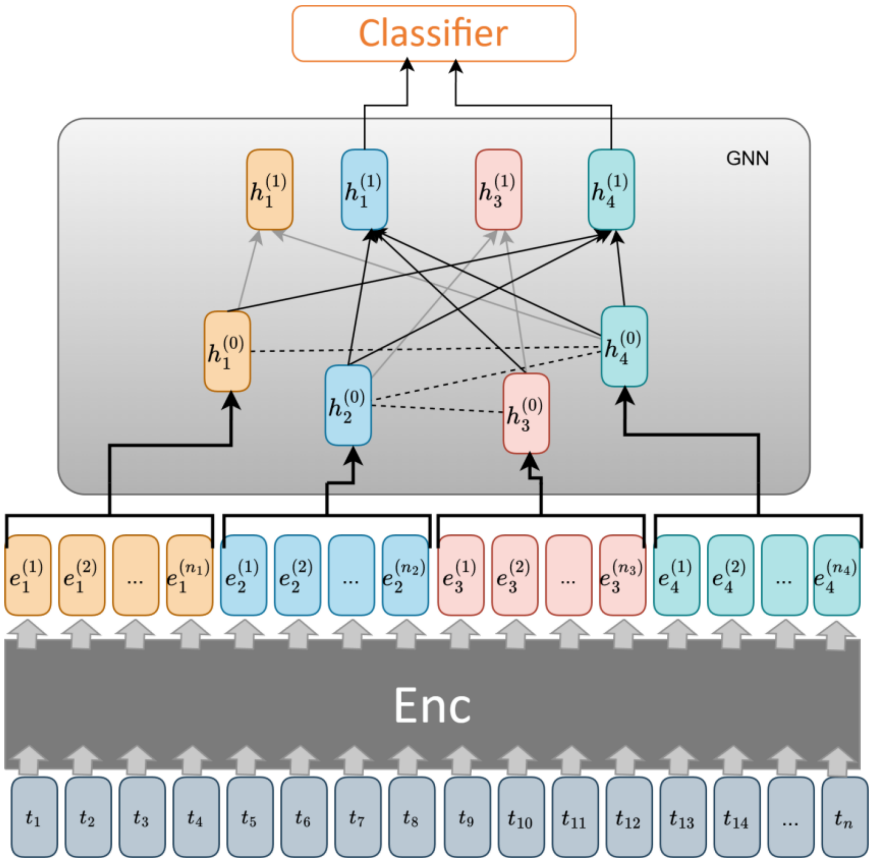


Figure 1: Architecture of the GNN-based model

This model uses only rhetorical structure and not relation types, however, not all rhetorical relations have correspondences in the argumentation structure. To account for the relation type, a trainable embedding e_r of dimension d_r is introduced for each rhetorical relation type r, as well as for an additional type none, denoting the absence of a relation. The updated vector representation for an EDU pair is defined as: $h'_{ij} = \left[h_i^{(K)}; h_j^{(K)}; e_r\right]$. Thus, $y_{ij}^{rel-graph} = Classify(h'_{ij})$.

3.3 Multitask based approach

The next approach involves using multitask learning [24, 25], where a single model is trained to solve several related tasks simultaneously. The main idea is that through joint training, the model can build generalized representations improving performance on each task and reducing overfitting [25].

The second (auxiliary) task here is predicting the presence of a rhetorical relation between two EDUs. Thus, the vector representations of EDUs $h^1, h^2, \cdots, h^m$ are used to predict both argumentative and rhetorical relations, contributing to the creation of more informative and generalized representations.

The loss function in this case is the sum of the losses for each task: $L = L^{arg} + \alpha L^{rhet}$, where L^{arg} is the CE Loss for predicting argumentative relations, L^{rhet} is the CE Loss for predicting rhetorical relations, and α is a weighting coefficient that controls the contribution of the rhetorical component to the overall loss function.

4 Annotated Data

The material for the study is a parallel corpus in which texts are provided with RST- and A-markups.

Figure 2 shows an example of parallel markup of a text fragment. In this example, there is a one-to-one correspondence between the relation pairs <*Adversative, Logical_Conflict*> and <*Casual, NegativeConsequences_Inference*>.

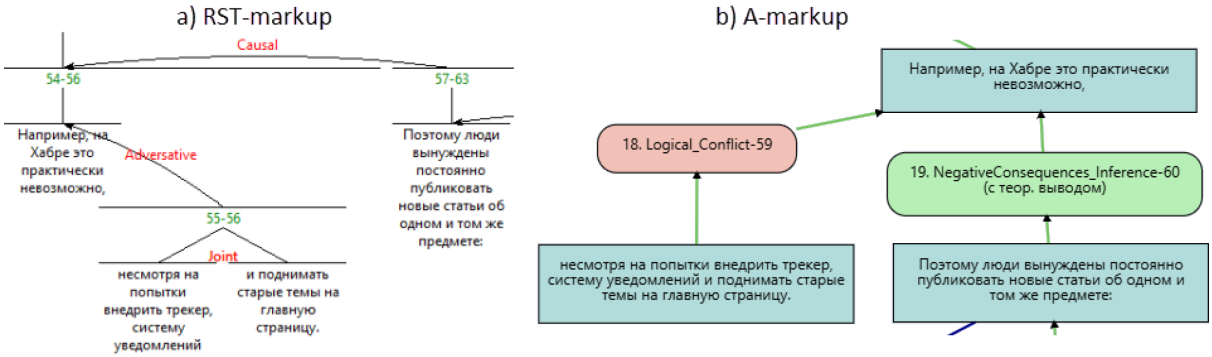


Figure 2: Example of correlation between a) RST-markup and b) A-markup

The corpus consists of 100 texts of Internet communication in several scientific and popular science genres: 1) scientific articles (10 articles, the average volume of the article is 3571 words); 2) scientific reviews (30 reviews, the average volume of the review is 346 words); 3) short articles about science news (30 articles from the website poisknews.ru , the average length of the article is 506 words); 4) long articles (30 articles from the site habr.com/ru , the average length of an article is 1912 words).

A-annotating was performed manually by four experts on the ArgNetBank Studio platform[1] [26], which presents the ontology of argumentation and description of schemes within the framework of D. Walton's approach used in annotating texts. In total, the corpus contains about 10 thousand marked-up arguments.

RST annotating was performed automatically. To build RST structures, the IsaNLP RST Parser analyzer was used [27], which is based on a classifier trained on a bilingual corpus, including annotated texts in Russian and English. The classifier is trained for segmentation, building RST structure and classifying the relations included in it.

A comparative analysis of RST- and A-segmentation statistics conducted for each text and on average by genre shows that there is no text in which the number of A-segments would be equal to the number of RST-segments. Moreover, in most cases this indicator for RST is 2–4 times greater than for A-segmentation. This shows both the presence of non-argumentative fragments in texts and larger granularity of RST-segmentation.

Comparison with the ¾ similarity threshold on the dataset shows that the proportion of A-segments that match RST segments is on average 42.23%. This once again confirms the fact that RST-markup provides significant information for constructing A-markup [12].

Based on RST- and A-markups, a dataset was constructed for experimental comparison of the proposed models, including positive and negative examples of pairs of statements (premise, conclusion). EDUs from RST-markups serve as statements. A pair is considered a positive example if there is an argumentative relation between premise and conclusion in at least one of the expert A-markups (i.e. markups of different experts are combined). Negative examples were formed as follows: for each positive pair of statements, a pair of statements from the same or adjacent paragraphs was selected (in the absence of suitable candidates) between which there is no path in the argumentation graph.

In total, the dataset used for the experiments contains 2722 pairs of statements, a half of them being positive examples and another half negative. 28.14% of positive pairs, i.e. connected by argumentative relation, are also connected by rhetorical relation, and only 12.05% of rhetorical relations are among negative examples. The distribution of types of rhetorical relations among positive and negative examples is presented in Figure 3.

1 https://uniserv.iis.nsk.su/arg/

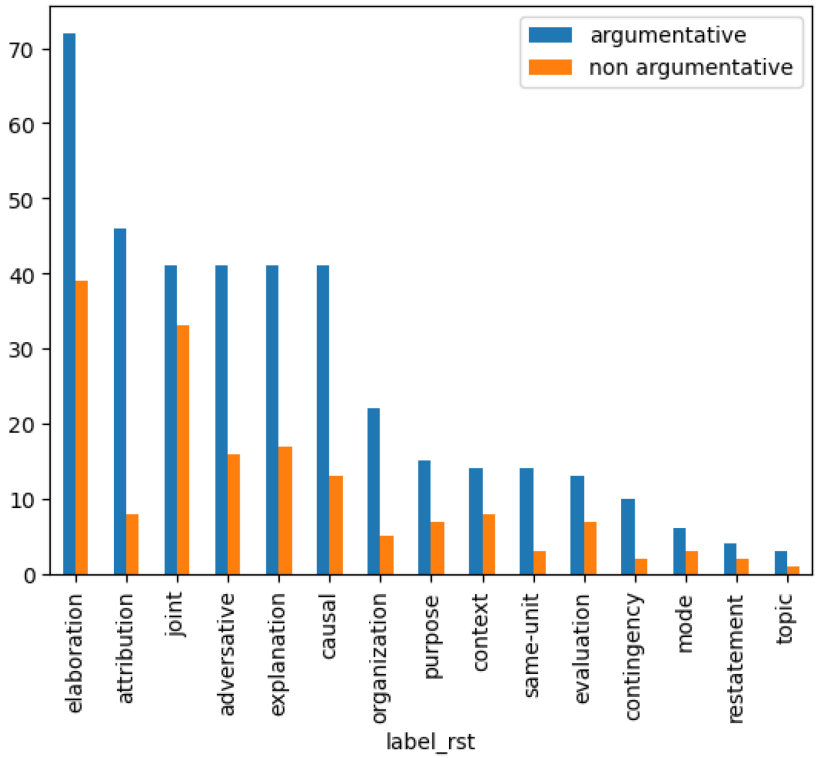


Figure 3: Rhetorical relations between positive pairs of statements (connected by argumentative relations) and negative ones

Additionally, each pair is provided with its paragraph-sized context and a graph of the context's rhetorical structure, constructed on the basis of RST-markup. The graph's vertices are EDUs. If a rhetorical relation in RST structure connects two discourse segments, then the graph connects two vertices that are the roots of these statements. The root of an EDU is itself, the root of a mononuclear relation is the root of the nucleus, and the root of a multinuclear relation is the root of the leftmost segment.

5 Experiments

Experiments were performed using the Longformer model [28] as an encoder (Enc) for long sequences, with all parameters remaining trainable. The max function was used as the aggregation function AGG. To mitigate the impact of data inconsistency, the label smoothing technique [29] with a coefficient of 0.1 was used. The baseline results were obtained with the model proposed in [30], which uses this Longformer model and the label smoothing technique.

The dataset was split into training, development, and test sets (57/20/23 texts), with the test subset including texts from all subcorpora in the dataset. Hyperparameter tuning was performed on the development subset. The size of the hidden layer in the two-layer classifier was set to 256, and the size of the embeddings for rhetorical relations was 32. The GNN module consisted of two layers. All models were optimized using AdamW with an initial learning rate of 5×10^{-6} and a batch size of 4.

Table 1 summarizes the results of the experiments on the use of rhetorical information.

- Baseline is a base model for predicting argumentative relations.
- ArgRhetRelGraph is a model that uses a GNN utilizing the rhetorical structure of a text fragment and trainable embeddings for each type of rhetorical relation.
- ArgRhetGraph is a modification of the previous model that does not use embeddings for rhetorical relations.
- ArgRhetRel is a modification of ArgRhetRelGraph that does not use a GNN.
- ArgRhetMT is a model trained in multitask mode on the tasks of predicting argumentative and rhetorical relations.

Model	P	R	F1
Baseline	74.46	63.97	68.82
ArgRhetRelGraph	**77.37**	**67.90**	**72.32**
ArgRhetGraph	77.05	65.13	70.59
ArgRhetRel	77.29	64.43	70.28
ArgRhetMT	77.09	63.74	69.79

Table 1: Results obtained by the proposed models

Experimental results demonstrated that the use of rhetorical information improves the quality of argument retrieval compared to the baseline version. The ArgRhetRelGraph model showed the best results and achieved F1 = 72.32%.

An analysis of examples where the predictions of the ArgRhetRelGraph differ from those of the Baseline has shown that the most frequent rhetorical relation which decreases the model's confidence in the presence of an argumentative relation is Elaboration (the satellite presents more detailed information about the nucleus). Conversely, among those in which the model became more confident the Attribution relation (the source, the author of the information presented in the nucleus) is more frequent. The Joint relation (a conjunctively related sequence of segments) can either decrease or increase the model's confidence in the presence of an argumentative relation, depending on the specific context.

A comparison of the prediction quality for short- and long-distance relations (see Table 2) showed that, in general, all models predict argumentative relations significantly better for statements within the same paragraph (short-distance relations) than for long-distance relations. Meanwhile, the use of rhetorical information improves prediction quality more strongly for long-distance relations, which can be explained by the fact that, despite the local nature of rhetorical relations, they contribute to the text's cohesion at the global level. Thus, the average linear distance in segments between pairs of statements in the test set is 12.8, while the average distance in terms of transitions between nodes of the rhetorical structure graph is 3.9.

Model	Same paragraph			Long distance		
	P	R	F1	P	R	F1
Baseline	76.34	70.30	73.20	68.82	49.23	57.40
ArgRhetRelGraph	**79.49**	**71.62**	**75.35**	71.96	**59.23**	**64.98**
ArgRhetGraph	79.25	69.31	73.94	71.29	55.38	62.34
ArgRhetRel	78.95	69.31	73.81	**72.63**	53.08	61.33
ArgRhetMT	79.23	67.99	73.18	71.43	53.85	61.40

Table 2: Comparison of the prediction quality of argumentative relations for segments located at different distances

6 Discussion

The most common errors resulting in incorrect argument predictions are categorized below.

Segmentation errors. RST segmentation being too fractional is inherited by the model, although, as the expert analysis given in [12] shows, argumentative segments (ADUs) are often concatenations of rhetorical segments (EDUs). As a correlating parameter, it is necessary to take into account the presence of *joint* type relations in the context: for example, in (1), two EDUs should be combined into a single ADU.

(1) Однако исследователи не могут назвать точный возраст костей, [поскольку это место является объектом Всемирного наследия,] <joint> [а окаменелости охраняются лаосским законодательством.]

However, researchers cannot give an exact age for the bones [since the place is a World Heritage site] <joint> [and the fossils are protected by Laotian law.]

Zoning errors are closely related to segmentation ones: there are examples in which an argumentative relation is based on segments that actually relate to a non-argumentation zone, as evidenced by the rhetorical context. In example (2) below, an argumentative relation is predicted that holds between the purpose segment and the following segment, which is incorrect: the corresponding rhetorical relation is *context* and annotators do not see any area of argumentation here. In the RST, subject matter relations are contrasted with textual presentational relations that facilitate the presentation process. The textual relations such as *elaboration, context, joint* are highly correlated with a larger argumentative segmentation and/or with belonging to the non-argumentation zone.

(2) Исследователи продолжают искать способ [вылечить или хотя бы смягчить симптомы болезни.] <context> [Команда ученых с участием НИУ ВШЭ детально изучила около 40 исследований за последние 20 лет] и предположила, как стимуляция мозга может влиять на целостность ГЭБ, в том числе при болезни Альцгеймера.

Researchers continue to look for a way [to cure or at least alleviate the symptoms of the disease.] <context> [A team of scientists with the participation of the Higher School of Economics has reviewed in detail about 40 studies over the past 20 years] and suggested how brain stimulation can affect the integrity of the BBB, including in Alzheimer's disease.

Errors in predicting the direction of argumentative relations which in the rhetorical representation correspond to subject matter relations such as *causal, purpose, explanation* linking adjacent segments or groups of segments. Most false-positive examples are found for adjacent segments within the same paragraph, when the direction of the predicted argumentative relation is erroneously opposite to the direction of the rhetorical link (in the examples below main segments predicted by the model are underlined).

*(3) [И вот на этом этапе кольцо "обиды" замыкается,] [**так как** публичное выражение обиды и оскорблённости направлено на вызов сопереживания других людей конкретно к вам и вашей ситуации,]*

*[And at this stage the circle of "resentment" closes,] [**since** the public expression of resentment and insult is aimed at evoking empathy from other people specifically towards you and your situation,]*

*(4) [«Во всем мире было проделано огромное количество теоретических и экспериментальных работ,] [**направленных на то, чтобы** понять, что происходит после поглощения фотона.]>*

*["A huge amount of theoretical and experimental work has been done around the world] [**aimed at** understanding what happens after a photon is absorbed.]>*

*(5) [но некоторые данные о нейронах и глиальных клетках доступны.] [**Так**, [в [[исследовании]] китайских ученых сообщалось,] что магнитные импульсы могут оказывать нейрозащитный эффект.]*

*[but some data on neurons and glial cells are available.] [**For example**, [a [[study]] by Chinese scientists reported that magnetic pulses can have a neuroprotective effect.]*

In the examples given, *causal* (3)*, purpose* (4)*, and explanation* (5) subject matter dependencies are marked by the presence of corresponding discourse markers (highlighted in bold), which are characterized by certain positional characteristics (as indicator of argumentation causal *так как 'since'* introduces a premise, and *поэтому 'therefore'* introduces a conclusion, etc.). A combination of the described model with the indicator method (see [31]) would allow us to avoid most of these types of errors.

Errors in predicting long-distance relations. The segments are in different paragraphs or even in different sections of the article. Long-distance relations should not contradict short-distance relations, which is the case in (6).

(6) Результаты подтверждают оценки возраста окаменелостей, найденных ранее в пещере Там Па Линг, [но увеличивают хронологию этого места примерно на 10 000 лет.]

[Пещера находится более чем в 300 километрах от моря,] поэтому открытие предполагает, что наши мигрирующие предки не просто следовали вдоль побережья и островов в своем путешествии из Африки, но пересекали лесные районы и долины рек.

The results confirm age estimates for fossils previously found at Tam Pa Ling Cave, [but push the site's chronology back by about 10,000 years.]

[The cave is more than 300 kilometres from the sea,] so the discovery suggests that our migrating ancestors did not simply follow coastlines and islands on their journey out of Africa, but crossed forested areas and river valleys.

Errors in determining the role of the title. In view of the text organization characteristics, these errors can be considered a special case of the previous ones.

(7) [В лаосской пещере найдена кость человека возрастом 86 тыс. лет]

[В недрах лаосской пещеры ученые обнаружили самые ранние известные свидетельства того,] что наши человеческие предки пробирались через материковую часть Юго-Восточной Азии по пути в Австралию около 86 000 лет назад.

[86,000-year-old human bone found in Laotian cave]

[In the depths of a Laotian cave, scientists have discovered the earliest known evidence] that our human ancestors made their way across mainland Southeast Asia on their way to Australia about 86,000 years ago.

It can be assumed that one of the parameters that must be taken into account is the genre characteristic. Thus, in scientific articles, the last paragraph/heading is more likely to represent the main thesis - the conclusion, to which segments from other paragraphs should be connected as premises. In news articles, on the contrary, the headline may well be the premise – this is how annotators see it in (7).

7 Conclusion

In this paper, we investigate the role of rhetorical structures in the task of argument extraction. Two approaches are proposed and studied: the first one uses rhetorical structure in constructing vector representations of segments using GNN, and the second one uses multitask learning that combines argumentation extraction with rhetorical relations prediction.

The proposed solutions contribute to improving the quality of predicting argumentative relations, including connections between statements located at a significant distance from each other. The best quality was demonstrated by a model in which vector representations of statements were modified using a GNN and supplemented with information about the type of rhetorical relation, and achieved F1 = 72.32%.

This study was limited to predicting the relations between EDUs. In the future, it is planned to expand the approach by using methods that allow combining several EDUs into one argumentative statement. In addition, it is planned to study the role of rhetorical relations in the task of predicting argumentation schemes.

Acknowledgements

This study was supported by the Russian Science Foundation (grant No. 23-11-00261), https://rscf.ru/project/23-11-00261/.

References

[1] Mann W.C., Thompson S.A. Rhetorical structure theory: Towards a functional theory of text organization. Text, 8(3), 1988, pp. 243–281.

[2] Mann W.C., Thompson S.A. Toward a theory of reading between the lines: An exploration in discourse structure and implicit communication. Paper presented at the Seventh International Pragmatics Conference, Budapest, Hungary, 2000. Access mode: https://studylib.net/doc/7881702/toward-a-theory-of-reading-between-the-lines.

[3] Taboada M., Mann W. Rhetorical Structure Theory: Looking back and moving ahead – Discourse Studies. SAGE Publications. Vol. 8. No. 3, 2006. Pp. 423–459.

[4] Taboada M., Mann W.C. Applications of Rhetorical Structure Theory. Discourse Studies, 8(4), 2006, pp. 567–588.

[5] Walton D., Reed C., Macagno F. Argumentation schemes. Fundamentals of critical argumentation. New York: Cambridge University Press, 2008. 443 p.

[6] Walton D. Argumentation theory: A very short introduction. Eds. Simari G., Rahwan I. Argumentation in Artificial Intelligence, Springer, Boston, 2009, pp. 1–22.

[7] Musi E., Alhindi T., Stede M., Kriese L., Muresan S., Rocci A. A Multi-layer Annotated Corpus of Argumentative Text: From Argument Schemes to Discourse Relations, Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'2018), Miyazaki, Japan, 2018, pp. 1629–1636.
[8] Accuosto P., Saggion H. Discourse-driven argument mining in scientific abstracts, 24th International Conference on Applications of Natural Language to Information Systems NLDB 2019, Salford, UK, 2019, pp. 182–194.
[9] Kononenko I.S., Sidorova E.A., Akhmadeeva I.R. Comparative analysis of rhetorical and argumentative structures in the study of popular science discourse – Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference "Dialogue". 19 (26), 2020. – pp. 432–444. DOI: 10.28995/2075-7182-2020-19-432-444.
[10] Azar M. Argumentative Text as Rhetorical Structure: An Application of Rhetorical Structure Theory – Argumentation 13, 1999, pp. 97–114.
[11] Chistova E. et al. RST Discourse Parser for Russian: An Experimental Study of Deep Learning Models. In: van der Aalst W.M.P. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2020. Lecture Notes in Computer Science, vol. 12602. Springer, Cham, 2021. Pp. 105–119.
[12] Timofeeva M.K., Kononenko I.S., Shestakov V.K., Zorkaltsev A.V. Analysis of the possibilities of using markup based on the theory of rhetorical structures to clarify argumentative markup [in Russian]. Bulletin of Tomsk State University. 2024. № 508. C. 86-98. doi: 10.17223/15617793/508/9
[13] Peldszus A., Stede M. Rhetorical structure and argumentation structure in monologue text. In Proceedings of the Third Workshop on Argument Mining (ArgMining2016), pages 103–112, Berlin, Germany, 2016. Association for Computational Linguistics.
[14] Hewett F., Rane R.P., Harlacher N., Stede M. The utility of discourse parsing features for predicting argumentation structure. In Proceedings of the 6th Workshop on Argument Mining, pages 98–103, Florence, Italy, 2019. Association for Computational Linguistics.
[15] Chistova E. End-to-End Argument Mining over Varying Rhetorical Structures – Findings of the Association for Computational Linguistics: ACL 2023 / ed. Rogers A., Boyd-Graber J., Okazaki N. Toronto, Canada: Association for Computational Linguistics, 2023. Pp. 3376–3391.
[16] Dozat T., Manning C.D. Deep Biaffine Attention for Neural Dependency Parsing: arXiv:1611.01734. arXiv, 2017.
[17] Peldszus A., Stede M. An annotated corpus of argumentative microtexts //Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon. – 2015. – Vol. 2. – P. 801-815.
[18] Sun Y. et al. Discourse Structure-Aware Prefix for Generation-Based End-to-End Argumentation Mining – Findings of the Association for Computational Linguistics ACL 2024. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, 2024. Pp. 11597–11613.
[19] Stab C., Gurevych I. Parsing argumentation structures in persuasive essays //Computational Linguistics. – 2017. – Vol. 43. – №. 3. – P. 619-659.
[20] Mayer T., Cabrio E., Villata S. Transformer-based argument mining for healthcare applications //ECAI 2020. – IOS Press, 2020. – P. 2108-2115.
[21] Accuosto P., Saggion H. Transferring knowledge from discourse to arguments: A case study with scientific abstracts. In Proceedings of the 6th Workshop on Argument Mining, pages 41–51, Florence, Italy, 2019. Association for Computational Linguistics.
[22] Yang A., Li S. SciDTB: Discourse Dependency TreeBank for Scientific Abstracts //Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). – 2018. – P. 444-449.
[23] Kipf T.N., Welling M. Semi-Supervised Classification with Graph Convolutional Networks – 5th International Conference on Learning Representations, ICLR 2017 – Conference Track Proceedings. 2017.
[24] Caruana R. Multitask Learning – Machine Learning. 1997. Vol. 28, No. 1. Pp. 41–75.
[25] Crawshaw M. Multi-Task Learning with Deep Neural Networks: A Survey – ArXiv. 2020.
[26] E.A. Sidorova, I.R. Akhmadeeva, Yu.A. Zagorulko, A.S. Sery, V.K. Shestakov, "Research platform for the study of argumentation in popular science discourse" [In Russian], Ontology of Designing, 2020, 10(4), pp. 489–502, DOI: 10.18287/2223-9537-2020-10-4-489-502.
[27] Chistova E. Bilingual Rhetorical Structure Parsing with Large Parallel Annotations. In Findings of the Association for Computational Linguistics: ACL 2024, pages 9689–9706, Bangkok, Thailand, 2024. Association for Computational Linguistics.
[28] Beltagy I., Peters M.E., Cohan A. Longformer: The Long-Document Transformer: arXiv:2004.05150. arXiv, 2020.
[29] Lukasik M. et al. Does label smoothing mitigate label noise? – Proceedings of the 37th International Conference on Machine Learning. PMLR, 2020. Pp. 6448–6458.
[30] E.A. Sidorova, I.R. Akhmadeeva, D.V. Ilina, A.S.Sery Extracting Argumentation from Scientific Communication Texts in Russian Language [in Russian]. 1st Eurasian Congress of Linguists. Moscow, December 9–13, 2024: Book of Abstracts / Ed. by Andrej A. Kibrik, Julia V. Mazurova, Marina K. Raskladkina. Moscow: Institute of linguistics RAS, 2025. — pp.405-407. — ISBN 978-5-6049527-9-5
[31] Sidorova E.A., Akhmadeeva I.R., Kononenko I.S., Chagina P.M. Argument Extraction Based on the Indicator Approach // Pattern Recognition and Image Analysis, Volume 33, Issue 3, 498–505 (2023). https://doi.org/10.1134/S1054661823030410

RuTaR—A Dataset in Russian for Reasoning about Taxes

Andrey Alibekov
NTR Labs, Tomsk, Russia
Higher IT School, Tomsk State University, Tomsk, Russia
aalibekov@ntr.ai

Aleksandr Migal
NTR Labs, Moscow, Russia
HSE, Saint-Petersburg, Russia
amigal@ntr.ai

Andrey Matenkov
NTR Labs, Moscow, Russia
amatenkov@ntr.ai

Andrey Muryshev
NTR Labs, Moscow, Russia
amuryshev@ntr.ai

Vladislav Bolshakov
NTR Labs, Moscow, Russia
BMSTU, Moscow, Russia
vbolshakov@ntr.ai

Aleksander Kozachenko
NTR Labs, Tomsk, Russia
akozachenko@ntr.ai

Gyuli Mukhtarova
NTR Labs, Moscow, Russia
gmukhtatova@ntr.ai

Nikolay Mikhaylovskiy
NTR Labs, Moscow, Russia
Higher IT School, Tomsk State University, Tomsk, Russia
nickm@ntr.ai

Abstract

In 2024, reasoning have emerged as a new frontier for artificial intelligence and computational linguistics. Reasoning models are typically evaluated either on STEM-related datasets, or on synthetic datasets. This ignores a huge area of human thought—namely, humanitarian. To bridge this gap partially, we present a new open dataset, RuTaR (Russian Tax Reasoning). The dataset consists of modestly modified content of 199 select Ministry of Finances of Russia and Russian Federal Tax Service letters that typically reason to answer some taxpayer question. Despite apparent simplicity of yes/no questions, both off-the-shelf Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) systems struggle to achieve high results on the dataset, with top RAG system studied achieving 77% accuracy.

Keywords: reasoning; dataset; artificial intelligence
DOI: 10.28995/2075-7182-2025-23-12-27

RuTaR — набор данных рассуждений о налогах на русском языке

Андрей Алибеков
ООО «НТР Томск», Томск, Россия
Высшая ИТ-Школа Томского Государственного Университета, Томск, Россия
aalibekov@ntr.ai

Александр Мигаль
ООО «НТР», Москва, Россия
НИУ ВШЭ, Санкт-Петербург, Russia
amigal@ntr.ai

Андрей Матенков
ООО «НТР», Москва, Россия
amatenkov@ntr.ai

Андрей Мурышев
ООО «НТР», Москва, Россия
amuryshev@ntr.ai

Владислав Большаков
ООО «НТР», Москва, Россия
МГТУ им. Баумана, Москва, Россия
vbolshakov@ntr.ai

Александр Козаченко
ООО «НТР Томск», Томск, Россия
akozachenko@ntr.ai

Гюли Мухтарова
ООО «НТР», Москва, Россия
gmukhtatova@ntr.ai

Николай Михайловский
ООО «НТР», Москва, Россия
Высшая ИТ-Школа Томского
Государственного Университета,
Томск, Россия
nickm@ntr.ai

Аннотация

В 2024 году рассуждения стали ключевой точкой развития искусственного интеллекта и вычислительной лингвистики. Модели рассуждений обычно оцениваются либо на наборах данных, связанных с точными науками, либо на синтетических наборах данных. Это игнорирует огромную область человеческой мысли — гуманитарную. Чтобы частично восполнить этот пробел, представлен новый открытый набор данных, RuTaR (Russian Tax Reasoning). Набор данных состоит из слегка измененного контента 199 отобранных писем Министерства финансов России и Федеральной налоговой службы, которые с рассуждениями отвечают на вопросы налогоплательщиков. Несмотря на кажущуюся простоту бинарных вопросов, как стандартные большие языковые модели, так и системы генерации на основе поиска испытывают трудности с достижением высоких результатов на наборе данных, при этом лучшая из оцененных система достигает точности лишь в 77 %.

Ключевые слова: рассуждения; набор данных; искусственный интеллект

1 Introduction

In 2024, reasoning have emerged as a new frontier for artificial intelligence and computational linguistics [1-12]. The pivotal productization of o1 reasoning system by Open AI [12] have sparked a wave of research and replication efforts [14-21]. These reasoning models are typically evaluated either on STEM datasets [22-34] (which from our perspective include logics-specific datasets [35-44]), or on synthetic datasets [2, 45-48]. Exceptions are few [49-52]. This ignores the modes of reasoning typical of the humanities, including law and other similar reasoning-intensive fields.

In Russian, there are few original datasets specifically targeted at multi-turn reasoning. Russian SuperGLUE [52] was a fantastic effort for its time, but it is now largely solved [46], including the most relevant to our interests MuSeRC and RuCoS [53]. MultiQ from TAPE [46] is a synthetically constructed 2-hop reasoning benchmark. Therefore, we would like to have a dataset that will be challenging in 2025 and thus require multi-turn reasoning over a large natural corpus.

To address this twofold deficiency, we present a new open dataset, RuTaR (Russian Tax Reasoning). The dataset consists of modestly modified content of select Ministry of Finances of Russia and Federal Tax Service letters that typically answer some taxpayer question. The dataset is made available at https://github.com/rutar-anonymous/RuTaR under the terms of the Creative Commons Attribution License (CC-BY).

Figure 1 provides an example of an original letter of Russian Federal Tax Service. Figure 2 provides examples of data inferred from the same letter: inferred question that the letter answers, inferred short-form answer that can be used to automatically evaluate the reasoning systems and the legal references that can be used to evaluate retrieval systems.

The dataset comprises two distinct types of official letters, 131 of which were issued by the Ministry of Finance and 78 by the Federal Tax Service. Each letter is categorized into one of two possible response classes, with 120 records labeled as "0" and 89 as "1". Following a common binary classification logic, responses labeled as "0" correspond to a summarized answer of "No", while those labeled as "1" correspond to "Yes". On average, each letter contains 4.71 references to legal documents. It's also worth noting that two of the letters do not include any references at all.

Федеральная налоговая служба, рассмотрев обращение <...>, сообщает, что в соответствии с пунктом 3 статьи 29 Налогового кодекса Российской Федерации (далее – Кодекс) уполномоченный представитель налогоплательщика-организации осуществляет свои полномочия на основании доверенности, выдаваемой в порядке, установленном гражданским законодательством Российской Федерации, если иное не предусмотрено настоящим Кодексом. Порядок и правила совершения, а также причины прекращения действия доверенностей устанавливаются положениями раздела 10 Гражданского кодекса Российской Федерации, при этом применение доверенности в отношениях, регулируемых законодательством о налогах и сборах, в рамках Кодекса эти положения не нарушает.

Положениями статьи 188 Гражданского кодекса Российской Федерации установлены основания прекращения действия доверенности. При этом приведенный в указанной статье перечень оснований, по которым доверенность прекращает свое действие, является исчерпывающим и расширительному толкованию не подлежит.

Положениями Кодекса установлены возможность создания доверенности в форме электронного документа, подписанного электронной подписью доверителя (далее – электронная доверенность), – пунктом 3 статьи 29 Кодекса и полномочия ФНС России по утверждению формата и порядка направления электронной доверенности по телекоммуникационным каналам связи (далее – ТКС) – пунктом 5 статьи 80 Кодекса.

На основании вышеуказанных положений Кодекса издан Приказ ФНС России от 30.04.2021 № ЕД-7-26/445@ «Об утверждении формата доверенности, подтверждающей полномочия представителя налогоплательщика (плательщика сбора, плательщика страховых взносов, налогового агента) в отношениях, регулируемых законодательством о налогах и сборах, в электронной форме и порядка ее направления по телекоммуникационным каналам связи».

При этом электронная доверенность применяется как еще один способ создания (совершения) доверенностей, не оказывает влияния на процесс приема от уполномоченных представителей налогоплательщиков налоговой и бухгалтерской отчетности по ТКС, отличного от оказываемого доверенностями, совершенными на бумажных носителях, и не предполагает их отмену.

Внедрение электронной доверенности при осуществлении электронного документооборота с налоговыми органами также не предполагает оснований прекращения действия доверенностей, отличных от установленных в указанной выше статье Гражданского кодекса Российской Федерации.

Таким образом, обязанность, установленная Федеральным законом от 31.07.2023 № 389-ФЗ «О внесении изменений в части первую и вторую Налогового кодекса Российской Федерации, отдельные законодательные акты Российской Федерации и о приостановлении действия абзаца второго пункта 1 статьи 78 части первой Налогового кодекса Российской Федерации» путем внесения изменений в пункт 3 статьи 29 Кодекса, вступающих в силу с 01.03.2024, в соответствии с которыми уполномоченный представитель, указанный в пункте 5.1 статьи 23 Кодекса, осуществляет свои полномочия на основании электронной доверенности, распространяется на доверенности, которые будут совершаться и применяться для подтверждения полномочий уполномоченных представителей после 01.03.2024.

При этом прекращение использования квалифицированных сертификатов сотрудников юридических лиц или замена квалифицированной электронной подписи уполномоченного представителя в рамках действующей доверенности не являются основаниями для прекращения действия доверенности.

Настоящее письмо не является нормативным правовым актом, не влечет изменений правового регулирования налоговых отношений, не содержит норм, влекущих юридические последствия для неопределенного круга лиц, носит информационный характер и не препятствует налогоплательщикам руководствоваться нормами законодательства Российской Федерации о налогах и сборах в понимании, отличающемся от положений настоящего письма.

Figure 1: Original Federal Tax Service letter text (№ ЗГ-3-26/13425 dated 18.10.2023)

Inferred question	Inferred answer	References
Можно ли применять электронные доверенности уполномоченным представителям налогоплательщика-организации?	Да	1. п. 3 ст. 29 НК РФ 2. раздел 10 ГК РФ 3. ст. 188 ГК РФ 4. п. 5 ст. 80 НК РФ 5. п. 5.1 ст. 23 НК РФ 6. Приказ ФНС России от 30.04.2021 № ЕД-7-26/445@

Figure 2: Data inferred from the Federal Tax Service letter № ЗГ-3-26/13425

For the simplicity of evaluation we only consider questions that imply a binary answer (Yes/No). On the one hand, this way of binary classification offers us a very useful evaluation framework. However, it also imposes some significant simplifications on the complexity of a legal discourse. Legal responses, particularly in tax law, often contain a big amount of exceptions, temporal constraints or even contradictions. Therefore, reducing these complex texts to a binary label may cause the losing of semantic context and interpretive richness. In many cases, the final message of the legal authority may only be valid under a very specific set of assumptions, which the binary label does not reflect. Thus, from the overall corpus of over 5000 letters we only select 200 where we are confident enough in the meaningfulness of binary answer.

Legal opinions often involve highly complex reasoning: the introductory part of the text may outline general legal principles, followed by a specification towards the details of the taxpayer's query. And while we aim to associate the binary label with the most appropriate summary of the response, this decision is still interpretive. Therefore, assigning a "Yes" or "No" label may erase the conditional aspect of the opinion.

However, this simplification may favor systems that work well with minimal human supervision. Therefore, we consider our binary classification scheme a necessary abstraction to facilitate evaluation, but it should be modified in the future. Still, we believe that the importance and utility of the dataset we introduce overweighs the drawbacks of the approach.

The basic stats of the dataset are in the Table 1.

Attribute	Value
Number of letters from the Ministry of Finance	131
Number of letters from the Federal Tax Service	78
Answers that can be summarized as "Yes"	89
Answers that can be summarized as "No"	120
Average number of references per letter	4.71
Average number of words per letter	455
Maximum number of words in a letter	1929
Minimum number of words in a letter	89
Median number of words per letter	415
Average number of characters per letter	3580
Maximum number of characters in a letter	15,374
Median number of characters per letter	3255

Table 1. Core statistics of the dataset

2 Prior Datasets Used to Evaluate Reasoning

We have analyzed various papers related to reasoning, including [1-21], and have identified several groups of such datasets. Below is a brief review of such datasets.

2.1 STEM-related datasets

Cobbe et al. [23] introduced GSM8K, a dataset of 8.5K high quality linguistically diverse grade school math word problems. Shi et al. [21] introduced the Multilingual Grade School Math (MGSM) benchmark by manually translating 250 grade-school math problems from the GSM8K dataset into ten typologically diverse languages. Following that, Chen et al. [22] construct the multilingual math reasoning instruction dataset, MGSM8KInstruct, encompassing the same ten languages. Hendrycks et al. [24] introduced MATH, a dataset of 12,500 challenging competition mathematics problems. Each problem in MATH has a full step-by-step solution which can be used to teach models to generate answer derivations and explanations.

Clark et al. [33] introduced the ARC dataset that contains natural, grade-school science questions authored for human tests. Rein et al. [27] introduced GPQA, a challenging dataset of 448 high-quality and extremely difficult multiple-choice questions written by domain experts in biology, physics, and chemistry. He et al. [28] introduced OlympiadBench, an Olympiad-level bilingual multimodal scientific benchmark, featuring 8,476 problems from Olympiad-level mathematics and physics competitions, including the Chinese college entrance exam.

2.2 Programming datasets

Austin et al. [25] introduced two benchmarks, MBPP and MathQA-Python. The Mostly Basic Programming Problems (MBPP) dataset contains 974 programming tasks, designed to be solvable by entry-level programmers. The MathQA-Python dataset, a Python version of the MathQA benchmark, contains 23914 problems that evaluate the ability of the models to synthesize code from more complex text. Chen et al. [26] introduced HumanEval, an evaluation set intended to measure functional correctness for synthesizing programs from docstrings. Schuster et al. [31] propose a dataset of Python Programming Puzzles (P3). Each puzzle is defined by a short Python program f, and the goal is to find an input that makes f return True. In addition, SWE-bench [65] has recently been proposed as one of the most challenging programming benchmarks. It evaluates models on their ability to solve real-world GitHub issues using information from the associated codebase and documentation.

2.3 First-order logic-based datasets

Saparov and He [43] introduced a synthetic question-answering dataset called PrOntoQA, where each example is generated from a synthetic world model represented in first-order logic.

Tian et al. [41] introduced LogicNLI, an NLI-style dataset that effectively disentangles the target FOL reasoning from commonsense inference and can be used to diagnose LMs from four perspectives: accuracy, robustness, generalization, and interpretability. Yu et al. [36] introduced a Reading Comprehension dataset requiring logical reasoning (ReClor) extracted from standardized graduate admission examinations. Han et al. [34] introduced FOLIO, a human-annotated, logically complex and diverse dataset for reasoning in natural language (NL), equipped with first-order logic (FOL) annotations.

2.4 Synthetic datasets

Yao et al. [29] introduce three tasks requiring non-trivial planning or search: Game of 24, Creative Writing, and 5x5 Mini Crosswords. Game of 24 is a mathematical reasoning challenge, where the goal is to use 4 numbers and basic arithmetic operations (+-*/) to obtain 24. In the creative writing task, the input is 4 random sentences and the output should be a coherent passage with 4 paragraphs that end in the 4 input sentences respectively. Such a task is open-ended and exploratory, and challenges creative thinking as well as high-level planning.

Sinha et al. [35] introduced CLUTRR benchmark that requires that an NLU system infer kinship relations between characters in short stories. Liu et al. [37] introduced LogiQA dataset that is sourced from expert-written questions for testing human Logical reasoning. Liu et al. [38] then amended and re-

annotated it in LogiQA 2.0, increasing the data size, refining the texts with manual translation by professionals, and improving the quality by removing items with distinctive cultural features like Chinese idioms.

Mirzaee and Kordjamshidi [45] introduceв SpaRTUN, a synthetic dataset on spatial question answering (SQA) and spatial role labeling (SpRL) to provide a source of supervision with broad coverage of spatial relation types and expressions.

Sakaguchi et al. [47] introduced Winogrande, a large-scale pronoun resolution problem synthetic dataset as a benchmark for commonsense reasoning. Trivedi et al. [44] constructed MuSiQue-Ans, a synthetic multihop QA dataset with 25K 2–4 hop questions by systematically combining pairs of single-hop questions that are connected.

2.5 Datasets in Russian

Fenogenova et al. introduced two Russian machine reading comprehension (MRC) datasets, called MuSeRC and RuCoS, which require reasoning over multiple sentences and commonsense knowledge to infer the answer. The former follows the design of MultiRC [55], while the latter is a counterpart of the ReCoRD dataset [56]. Shavrina et al. [52] introduced a Russian general language understanding evaluation benchmark – Russian SuperGLUE, organized similarly to SuperGLUE [54] and including MuSeRC and RuCoS. Taktasheva et al. [46] proposed TAPE (Text Attack and Perturbation Evaluation), a benchmark that includes six more complex NLU tasks for Russian, covering multi-hop reasoning, ethical concepts, logic and commonsense knowledge.

2.6 Other datasets

Srivastava et al. [42] introduced BIG-bench that consists of 204 tasks, contributed by 450 authors across 132 institutions. Task topics are diverse, drawing problems from linguistics, childhood development, math, common-sense reasoning, biology, physics, social bias, software development etc. Suzgun et al. [30] select a suite of 23 challenging BIG-Bench tasks for which prior language model evaluations did not outperform the average human-rater and call it BIG-Bench Hard (BBH).

Another prominent benchmark is ARC AGI [66]. It emphasizes general intelligence and abstract reasoning skills through a set of a diverse tasks that for now remain unsolved by many modern systems.

Thorne et al. [51] introduced a dataset for verification against textual sources, FEVER: Fact Extraction and VERification. It consists of claims generated by altering sentences extracted from Wikipedia and subsequently verified without knowledge of the sentence they were derived from. Yang et al. [50] introduced HotpotQA, a dataset with Wikipedia-based question-answer pairs with four key features: (1) the questions require finding and reasoning over multiple supporting documents to answer; (2) the questions are diverse and not constrained to any pre-existing knowledge bases or knowledge schemas; (3) sentence-level supporting facts required for reasoning are provided, allowing QA systems to reason with strong supervision and explain the predictions; (4) factoid comparison questions are provided to test QA systems' ability to extract relevant facts and perform necessary comparison.

Bisk et al. [49] introduced the task of physical commonsense reasoning and a corresponding benchmark dataset Physical Interaction: Question Answering or PIQA. Cui et al. [48] introduced MuTual, a dataset for Multi-Turn dialogue Reasoning based on Chinese student English listening comprehension exams.

3 Data Processing Pipeline

In this section, we describe the steps performed to build the dataset. In short, the steps are:

- Downloading
- Candidate selection
- Answer binarity validation
- Question extraction
- Postprocessing
- Extracting legal sources

3.1 Downloading

Any document types mentioned below were downloaded from the locations detailed in the Table 2.

3.2 Candidate Selection and Answer Binarity Validation

Any Ministry of Finance letters from departments other than the tax policy department were removed from consideration. The remaining letters were run through GPT-4o mini with a prompt that determines whether the answer can be summarized as "yes" or "no" (see Appendix A).

3.3 Question Extraction and Manual Postprocessing

The letters that were determined by the prompt as implying a binary answer were run through GPT-4o mini with a prompt that created the questions implied by the answer given by the respective authority (see Appendix A). Since the ability to summarize a response as "Yes" or "No" largely depends on the phrasing of the question, two distinct questions were generated for each letter. One of them was generated in such a way that the answer could be expressed as a clear "Yes", and the other in such a way that the answer could be expressed only as "No". Then the letters and questions have been checked manually by the team members. During the review process, team members selected the most appropriate of the two generated questions, making adjustments if necessary. Where necessary, the question was completely rewritten. We deliberately allowed the balance to be shifted towards the questions that assume the answer "No", since according to our observations, refuting the assumption given in the question requires more bold reasoning from a LLM. It is also worth noting that some of the compiled question-answer pairs were then manually removed from the dataset, since their content was limited to links to some regulatory and legal acts of the Russian Federation, and therefore they were of little value to the dataset.

Source type	Download location
Письма Минфина, Письма ФНС, Приказы ФНС	https://www.audar-info.ru
Гражданский кодекс, Бюджетный кодекс, Трудовой кодекс	http://pravo.gov.ru
Налоговый кодекс	https://nalog.garant.ru/fns/nk
Постановления правительства, Распоряжения правительства, Указы президента, Распоряжения президента, Федеральные законы, Федеральные конституционные законы	http://government.ru/docs/all
Разъяснения с сайта buh.ru	https://buh.ru/articles
Разъяснения с сайта 1С	https://its.1c.ru/db/newscomm
Статьи kontur-extern.ru	https://www.kontur-extern.ru/info/

Table 2: Legal data sources

3.4 Extracting legal references

To benchmark the search quality in RAG etc. one needs to have a gold standard on mentions of legal sources in the answer (legal references). To create one, we automatically extracted any legal references with a script and then selectively checked the completeness of references extracted. Whenever the manual testing revealed incomplete list of references extracted, the script was modified and the cycle was repeated.

4 Baseline systems

4.1 LLM and RAG

We used Mixtral 8x7B [57], Llama 3.3 70B [58] and GPT-4o mini [62] as baseline LLMs. We have also built a baseline Retrieval-Augmented Generation (RAG) system [63] that combines vector search using E5 [59] dense vector embeddings with generation using the above LLMs.

4.2 Finetuning E5

To improve the search quality, we used contrastive pre-training of E5 [59]. We pre-train the model on positive examples in the form of pairs (user query - relevant documents), and on negative examples in the form of pairs (user query - random irrelevant documents). This is the same approach as for the pre-training stage of E5 [59]. We use InfoNCE [60], [61] as a contrastive loss function, and use in-batch negatives, i.e. when answers from all other pairs of the batch are taken as negative examples for a question. The quality of such training usually increases with the batch size.

We have created two types of pairs: question – text from legal reference, and heading – text. Based for the first one we have created the following types of pairs:

- Question title - text from reference
- Question text - text from reference
- Compressed question text - text from reference
- Question title - compressed text from reference
- Question text - compressed text from reference
- Compressed question text - compressed text from reference

Long texts exceeding the model context were summarized using IlyaGusev/rut5_base_sum_gazeta.

Heading – text included pairs of heading - article text, or paragraph text - article subparagraph text, if such a hierarchy exists in the document, the title of the law / regulation and its text from the following sources:

- Tax Code of the Russian Federation
- Civil Code of the Russian Federation
- Labor Code of the Russian Federation
- Letters from the Ministry of Finance
- Letters from the Federal Tax Service
- Federal laws
- Government decrees

This data set was split into training (23501) and validation (1201) samples randomly (ratio 0.95 / 0.05). Finetuning was performed on NVIDIA RTX 4090 GPU. We used the hyperparameters recommended for fine-tuning in the E5 [59]. Despite the mixed precision, the batch size more than 5 overflowed VRAM. To artificially increase the batch size, gradient accumulation was used.

Validation consists of:

- looking at InfoNCE loss on validation samples of task 1 (question titles / full texts of questions separately) and task 2;
- using the updated version of the model, we collect a database (faiss) from the link texts of the training sample of task 1, and perform a semantic search on questions from the validation sample of task 1:
  - looking at Recall@10, Recall@20, Recall@40;
  - for all types of documents together and for each separately - Tax Code of the Russian Federation, Civil Code of the Russian Federation, Letters of the Ministry of Finance, Letters of the Federal Tax Service, Orders of the Federal Tax Service, Federal Laws, Government Resolutions (except for the Labor Code of the Russian Federation and Courts).

Figures 3-5 demonstrate the finetuning process in the terms detailed above.

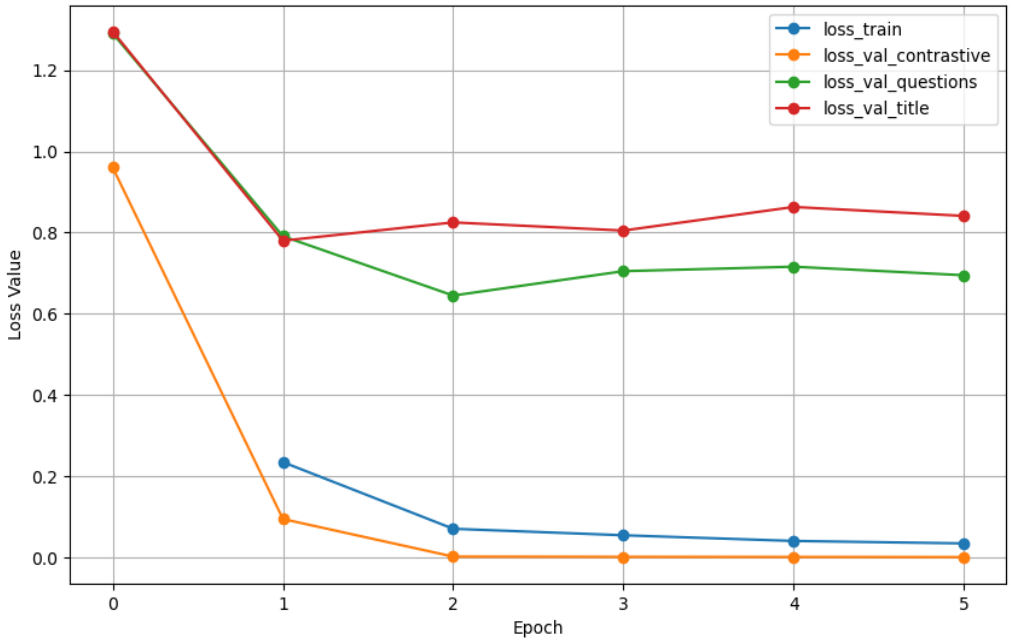


Figure 3: Losses for E5 finetuning

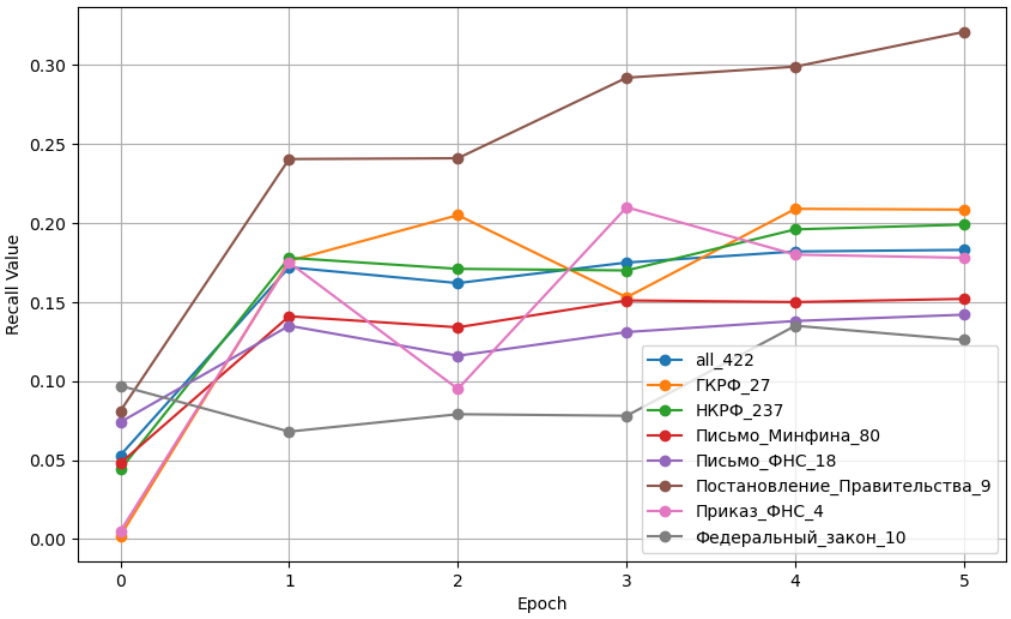


Figure 4: Recall @10 for the validation set of E5 finetuning

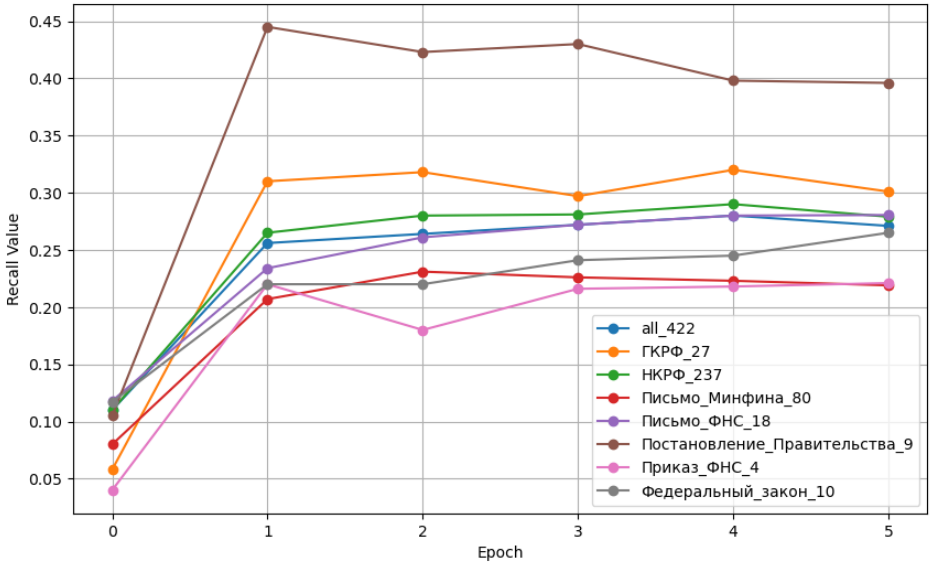


Figure 5: Recall @40 for the validation set of E5 finetuning

5 Evaluation results

We have evaluated four experimental setups:

- **Simple**: Questions from the dataset are presented to the model in their original form, accompanied by a straightforward prompt within a Chain-of-Thought framework [64]. A positive motivational tone is included to guide responses. Prompts are presented in Appendix A.
- **"Perfect" RAG**: Questions are supplemented with verbatim excerpts from the documents referenced in the corresponding Letters, ensuring that the model has access to the same information available to legal experts that wrote those Letters.
- **Base RAG:** This setup employs the E5-large model with a preconstructed vector database of relevant documents. The top 15 most relevant text chunks, retrieved based on the question, are provided to the model as supporting context alongside with the question.
- **Fine-tuned RAG:** Similar to the Base RAG setup, but the retrieval step utilizes our fine-tuned E5 model for improved relevance in retrieving supporting documents.

The accuracy scores for each setup are summarized in Table 3.

Model	Simple	"Perfect" RAG	Base RAG	Fine-tuned RAG
Mixtral-8x7B-Instruct-v0.1	0.566	0.684	0.658	0.606
Llama-3.3-70B-Instruct	0.625	0.775	0.766	0.746
GPT-4o mini	0.644	0.748	0.770	0.760

Table 3: Accuracy scores on RuTaR dataset in various experimental setups

These results highlight the consistent performance improvement achieved through the inclusion of RAG setups, while also pointing to potential areas for improvement in future, more advanced models. Interestingly, while the fine-tuned RAG setup demonstrates a significant enhancement in retrieval quality—evidenced by the substantially higher recall metrics shown in Figure 6 (e.g., recall@15 of 72% for fine-tuned versus 49% for Base RAG)—its accuracy is paradoxically slightly lower than that of the Base RAG setup. Even more intriguing is that the Perfect RAG does not help GPT-4o mini to improve over the Base RAG setup. This discrepancy indicates that retrieval quality alone does not fully account for downstream performance. Instead, factors such as the relevance ranking of retrieved chunks, the diversity of retrieved information, or the model's ability to integrate and reason over retrieved content may have played a more significant role.

Further research is needed to better understand this phenomenon. Potential avenues for analysis include a qualitative review of the retrieved chunks' characteristics and an examination of how these influence the reasoning and decision-making process of the model. Such insights would provide valuable guidance for enhancing fine-tuning approaches and optimizing the end-to-end RAG pipeline.

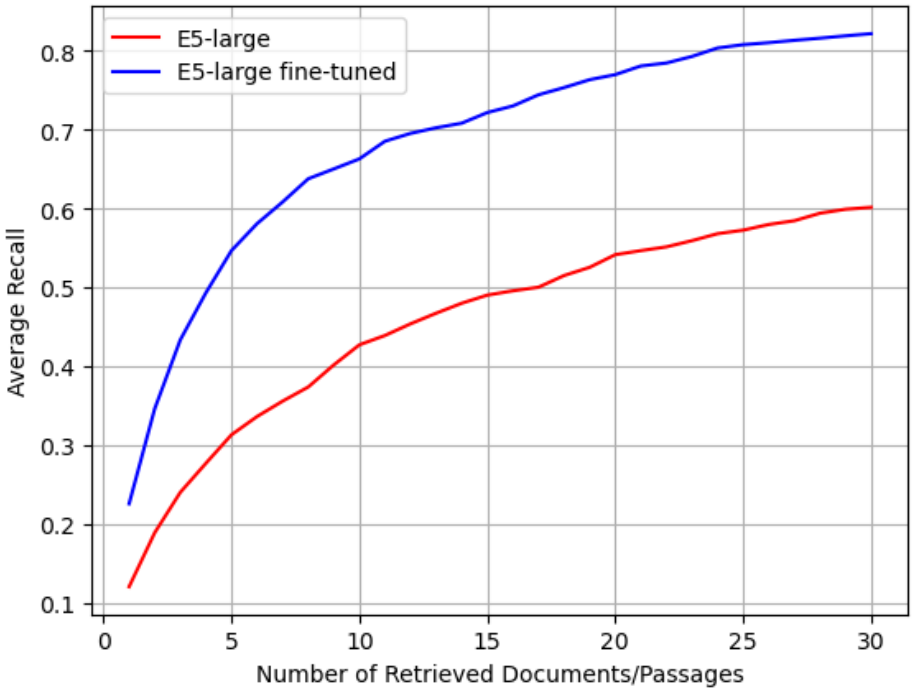


Figure 6: Average Recall of Retrieval Models on Questions from the RuTaR Dataset

6 Discussion

The presented dataset consists of questions that are hard for human professionals. That's why they come to government bodies for authoritative answers. The dataset is equally complicated for LLMs and vector search. Current popular LLMs built without specific attention to reasoning exhibit accuracy of 57% to 65% out-of-the-box, and 66% to 77% with the best RAG setup. Given the apparent simplicity of the setup that implies the accuracy of random choice of 50%, the dataset allows for validating significant improvement in both reasoning and retrieval capabilities of LLMs.

Across the evaluated systems, the most frequent error pattern was the systems failure to deliver a final binary verdict, despite partially correct or relevant reasoning. A likely explanation for that may be that, while the models demonstrate basic legal understanding, they often don't complete their reasoning processes when lacking the needed information from the retrieved documents. In cases where a RAG pipeline was employed (both base and ideal), some models still failed to reach a conclusion. This may be attributed to either retrieval noise (irrelevant or tangentially related documents), or the models' difficulty in synthesizing multiple legal references into a coherent output. Conversely, fine-tuned RAG pipelines significantly reduced such errors, suggesting that both retrieval quality and task-specific tuning play a critical role in guiding models toward conclusive legal reasoning.

Acknowledgements

The authors are grateful to their colleagues from NTR Labs computational linguistics and RL divisions for the support and discussions. This work is also largely inspired by discussions with our colleagues from Norilsky Nikel innovation and IT divisions including Alexey Testin, Anton Burkov, Kirill Maidanik, Sergey Greben and Gleb Volk.

References

[1] Janice Ahn, Rishu Verma, et al. Large Language Models for Mathematical Reasoning: Progresses and Challenges // Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop, P. 225–237, St. Julian's, Malta. Association for Computational Linguistics.

[2] Sohee Yang, Elena Gribovskaya, et al. Do Large Language Models Latently Perform Multi-Hop Reasoning? // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), P. 10210–10229, Bangkok, Thailand. Association for Computational Linguistics.

[3] Jie Huang, Xinyun Chen, et al. Large Language Models Cannot Self-Correct Reasoning Yet // The Twelfth International Conference on Learning Representations, 2024, https://openreview.net/forum?id=IkmD3fKBPQ

[4] Aviral Kumar, Vincent Zhuang, et al. Training Language Models to Self-Correct via Reinforcement Learning // Computing Research Repository. — 2024. — Vol. arXiv:2409.12917

[5] Pranav Putta, Edmund Mills, et al. Agent Q: Advanced Reasoning and Learning for Autonomous AI Agents // Computing Research Repository. — 2024. — Vol. arXiv:2408.07199 https://arxiv.org/abs/2408.07199

[6] Aditya Kalyanpur, Kailash Karthik Saravanakumar, et al. LLM-ARC: Enhancing LLMs with an Automated Reasoning Critic// Computing Research Repository. — 2024. — Vol. arXiv:2406.17663 https://arxiv.org/abs/2406.17663

[7] Chaojie Wang, Yanchen Deng , et al. Q*: Improving Multi-step Reasoning for LLMs with Deliberative Planning Computing Research Repository. — 2024. — Vol. arXiv:2406.14283 https://arxiv.org/abs/2406.14283

[8] Ye Tian, Baolin Peng, et al. Toward Self-Improvement of LLMs via Imagination, Searching, and Criticizing Computing Research Repository. — 2024. — Vol. arXiv:2404.12253 https://arxiv.org/abs/2404.12253

[9] Leonardo Ranaldi, Giulia Pucci, et al.. 2024. Empowering Multi-step Reasoning across Languages via Program-Aided Language Models. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 12171–12187, Miami, Florida, USA. Association for Computational Linguistics.

[10] Charlie Snell, Jaehoon Lee, , et al. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. Computing Research Repository. — 2024. — Vol. arXiv:2408.03314.

[11] Yoshua Bengio, From System 1 Deep Learning to System 2 Deep Learning https://youtu.be/T3sxeTgT4qc

[12] Open AI Team. Learning to Reason with LLMs. Technical Report'24. September 12, 2024 https://openai.com/index/learning-to-reason-with-llms/

[13] Yu Zhao, Huifeng Yin , et al. Marco-o1: Towards Open Reasoning Models for Open-Ended Solutions, Computing Research Repository. — 2024. — Vol. arXiv:2411.14405

[14] Zhen Huang, Haoyang Zou, et al. O1 Replication Journey -- Part 2: Surpassing O1-preview through Simple Distillation, Big Progress or Bitter Lesson Computing Research Repository. — 2024. — Vol. arXiv:2411.16489 https://arxiv.org/abs/2411.16489

[15] Yiwei Qin, Xuefeng Li, , et al. O1 replication journey: A strategic progress report–part 1. Computing Research Repository. — 2024. — Vol. arXiv:2410.18982.

[16] LLaMaO1 Team. 2024. Llamao1. Github https://github.com/SimpleBerry/LLaMA-O1

[17] OpenO1 Team. 2024. Openo1. Github https://github.com/Open-Source-O1/Open-O1

[18] kimi. 2024. k0math. website https://kimi.moonshot.cn/

[19] kunlun. 2024. skyworko1. website https://ai-bot.cn/skywork-o1/

[20] Yuhao Dong, Zuyan Liu, et al. Insight-V: Exploring Long-Chain Visual Reasoning with Multimodal Large Language Models Computing Research Repository. — 2024. — Vol. arXiv:2411.14432 https://arxiv.org/abs/2411.14432

[21] Freda Shi, Mirac Suzgun, et al. Language models are multilingual chain-of-thought reasoners The Eleventh International Conference on Learning Representations 2023, https://openreview.net/forum?id=fR3wGCk-IXp

[22] Nuo Chen, Zinan Zheng, et al. Breaking Language Barriers in Multilingual Mathematical Reasoning: Insights and Observations. // Findings of the Association for Computational Linguistics: EMNLP 2024, pages 7001–7016, Miami, Florida, USA. Association for Computational Linguistics.

[23] Karl Cobbe, Vineet Kosaraju, et al. Training Verifiers to Solve Math Word Problems Computing Research Repository. — 2021. — Vol. arXiv:2110.14168 https://arxiv.org/abs/2110.14168

[24] Dan Hendrycks, Collin Burns, et al. Measuring Mathematical Problem Solving With the {MATH} Dataset // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2), 2021, https://openreview.net/forum?id=7Bywt2mQsCe

[25] Jacob Austin, Augustus Odena, et al. Program Synthesis with Large Language Models 2021, Computing Research Repository. — 2021. — Vol. arXiv:2108.07732, https://arxiv.org/abs/2108.07732

[26] Mark Chen, Jerry Tworek, et al. Evaluating Large Language Models Trained on Code Computing Research Repository. — 2021. — Vol. arXiv: 2107.03374 https://arxiv.org/abs/2107.03374

[27] David Rein, Betty Li Hou, et al. GPQA: A Graduate-Level Google-Proof Q&A Benchmark Computing Research Repository. — 2023. — Vol. arXiv:2311.12022 https://arxiv.org/abs/2311.12022
[28] Chaoqun He, Renjie Luo, et al. OlympiadBench: A Challenging Benchmark for Promoting AGI with Olympiad-Level Bilingual Multimodal Scientific Problems Computing Research Repository. — 2024. — Vol. arXiv:2402.14008, https://arxiv.org/abs/2402.14008
[29] S. Yao, D. Yu, et al. Tree of thoughts: Deliberate problem solving with large language modelsm // Advances in Neural Information Processing Systems, vol. 36, 2024.
[30] M. Suzgun, N. Scales, et al. Challenging big-bench tasks and whether chain-of-thought can solve them // Findings of the Association for Computational Linguistics: ACL 2023, P. 13003–13051, 2023.
[31] T. Schuster, A. Kalyan, et al. Programming puzzles // Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021.
[32] A. T. K. Patrick Haluptzok, Matthew Bowers. Language models can teach themselves to program better // Eleventh International Conference on Learning Representations (ICLR), 2023.
[33] Peter Clark, Isaac Cowhey, et al. Think you have solved question answering? Try ARC, the AI^2 reasoning challenge. Computing Research Repository. — 2018. — Vol. arXiv:1803.05457.
[34] Simeng Han, Hailey Schoelkopf, et al. FOLIO: Natural Language Reasoning with First-Order Logic. // Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, P. 22017–22031, Miami, Florida, USA. Association for Computational Linguistics.
[35] Koustuv Sinha, Shagun Sodhani, et al. CLUTRR: A diagnostic benchmark for inductive reasoning from text. // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), P. 4506–4515, Hong Kong, China. Association for Computational Linguistics
[36] Weihao Yu, Zihang Jiang, et al. ReClor: A reading comprehension dataset requiring logical reasoning. // International Conference on Learning Representations.
[37] Jian Liu, Leyang Cui, et al. LogiQA: a challenge dataset for machine reading comprehension with logical reasoning // Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, P. 3622–3628
[38] Hanmeng Liu, Jian Liu, et al. LogiQA 2.0 — an improved dataset for logical reasoning in natural language understanding // IEEE/ACM Transactions on Audio, Speech, and Language Processing, P. 1–16, 2023. doi: 10.1109/TASLP.2023.3293046.
[39] Peter Clark, Oyvind Tafjord, and Kyle Richardson. Transformers as soft reasoners over language. // Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence, P. 3882–3890.
[40] Oyvind Tafjord, Bhavana Dalvi, and Peter Clark. ProofWriter: Generating implications, proofs, and abductive statements over natural language. // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, P. 3621–3634, Online. Association for Computational Linguistics.
[41] Jidong Tian, Yitian Li, et al. Diagnosing the firstorder logical reasoning ability through LogicNLI. // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, P. 3738–3747, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
[42] Aarohi Srivastava, Abhinav Rastogi, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Computing Research Repository. — 2024. — Vol. arXiv:2206.04615
[43] Abulhair Saparov and He He. Language Models Are Greedy Reasoners: A systematic formal analysis of chain-of-thought. // International Conference on Learning Representations 2023
[44] Harsh Trivedi, Niranjan Balasubramanian, et al. ♫ MuSiQue: Multihop Questions via Single-hop Question Composition. // Transactions of the Association for Computational Linguistics, 10:539–554.
[45] Roshanak Mirzaee and Parisa Kordjamshidi. Transfer Learning with Synthetic Corpora for Spatial Role Labeling and Reasoning. // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, P. 6148–6165, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
[46] Ekaterina Taktasheva, Alena Fenogenova, et al. TAPE: Assessing Few-shot Russian Language Understanding. // Findings of the Association for Computational Linguistics: EMNLP 2022, P. 2472–2497, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
[47] Keisuke Sakaguchi, Ronan Le Bras, et al. Winogrande: An adversarial Winograd schema challenge at scale. Computing Research Repository. — 2019. — Vol. arXiv:1907.10641.
[48] Leyang Cui, Yu Wu, et al. MuTual: A dataset for multi-turn dialogue reasoning. // Proceedings of the 58th Conference of the Association for Computational Linguistics. Association for Computational Linguistics, 2020.
[49] Yonatan Bisk, Rowan Zellers, et al. PIQA: Reasoning about physical commonsense in natural language. // Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020.

[50] Zhilin Yang, Peng Qi, et al. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. // Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, P. 2369–2380, Brussels, Belgium. Association for Computational Linguistics.
[51] James Thorne, Andreas Vlachos, et al. FEVER: a Large-scale Dataset for Fact Extraction and VERification. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), P. 809–819, New Orleans, Louisiana. Association for Computational Linguistics.
[52] Tatiana Shavrina, Alena Fenogenova, et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), P. 4717–4726, Online. Association for Computational Linguistics.
[53] Alena Fenogenova, Vladislav Mikhailov, and Denis Shevelev. Read and Reason with MuSeRC and RuCoS: Datasets for Machine Reading Comprehension for Russian. // Proceedings of the 28th International Conference on Computational Linguistics, pages 6481–6497, Barcelona, Spain (Online). International Committee on Computational Linguistics.
[54] Alex Wang, Yada Pruksachatkun, et al. SuperGLUE: a stickier benchmark for general-purpose language understanding systems. // Proceedings of the 33rd International Conference on Neural Information Processing Systems. Curran Associates Inc., Red Hook, NY, USA, Article 294, 3266–3280.
[55] Daniel Khashabi, Snigdha Chaturvedi, et al. Looking Beyond the Surface: A Challenge Set for Reading Comprehension over Multiple Sentences. // Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), P. 252–262, New Orleans, Louisiana. Association for Computational Linguistics.
[56] Zhang, Sheng, Xiaodong Liu, et al. Record: Bridging the gap between human and machine commonsense reading comprehension. arXiv preprint arXiv:1810.12885 (2018).
[57] Albert Q. Jiang, Alexandre Sablayrolles, et al. Mixtral of experts. https://arxiv.org/abs/2401.04088
[58] https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_3/
[59] Wang, Liang, Nan Yang, et al. Text Embeddings by Weakly-Supervised Contrastive Pre-training. Computing Research Repository. — 2022. — Vol. arXiv:/2212.03533
[60] Ting Chen, Simon Kornblith, et al. A simple framework for contrastive learning of visual representations. // Proceedings of the 37th International Conference on Machine Learning (ICML'20), Vol. 119. JMLR.org, Article 149, 1597–1607.
[61] Oord, Aäron van den, Yazhe Li and Oriol Vinyals. Representation Learning with Contrastive Predictive Coding. Computing Research Repository. — 2018. — Vol. arXiv:1807.03748
[62] Open AI Team. GPT-4o mini: advancing cost-efficient intelligence. Technical Report'24. July 18, 2024 https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
[63] Patrick Lewis, Ethan Perez, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. // Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20). Curran Associates Inc., Red Hook, NY, USA, Article 793, P. 9459–9474.
[64] Wei, Jason, Xuezhi Wang, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. Computing Research Repository. — 2022. — Vol. arXiv:2201.11903. https://arxiv.org/pdf/2201.11903.
[65] Jimenez, Carlos E., Yang, John, Wettig, Alexander, Yao, Shunyu, Pei, Kexin, Press, Ofir, Narasimhan, Karthik. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? Computing Research Repository. — 2024. — Vol. arXiv:2310.06770. URL: https://arxiv.org/abs/2310.06770
[66] Chollet, François. On the Measure of Intelligence. Computing Research Repository. — 2019. — Vol. arXiv:1911.01547. URL: https://arxiv.org/abs/1911.01547

Appendix A. Prompts used.

Prompt that determines whether the answer can be summarized as "yes" or "no"

```
system_prompt = (
    "Вы - юридический помощник, которому нужно определить, можно
ли суммаризировать представленный ответ на вопрос как 'да' или
'нет'. "
    "Если это возможно, ответьте 'Да' или 'Нет', в зависимости от
смысла. "
    "Если суммаризация невозможна, ответьте 'Недвоичный ответ'."
    "Возможные варианты вашего ответа: 'Да','Нет','Недвоичный от-
вет'. Пожалуйста, больше ничего не пишите.”

user_prompt = f"Вопрос: {question}\nОтвет: {answer}\nМожно ли
суммаризировать ответ как 'Да' или ‘Нет'?"
```

The letters that were determined by the prompt as implying a binary answer were run through GPT-4o with a prompt that created the questions implied by the answer given by the respective authority (see Appendix B).

```
Вариант 1.

system_prompt = (
    "Вы — квалифицированный юридический помощник, который генери-
рует бинарные вопросы (такие, на которые можно ответить 'да' или
'нет') на основе предоставленных ответов. "
    "Сгенерированные вопросы должны максимально напоминать во-
просы, задаваемые реальными пользователями и строго соответство-
вать содержанию ответа."
    "Каждый вопрос должен быть составлен таким образом, чтобы от-
вет на него можно было суммаризировать как 'да' или 'нет'."
    "Вопросы должны быть вежливыми, корректными и при необходимо-
сти включать в себя релевантную юридическую терминологию."
    "Вы не должны писать ничего, кроме текста самого вопроса. В
том числе, пожалуйста, избегайте приветствий, прощаний, вваодных
конструкций и комментариев."
)

user_prompt = f"Пожалуйста, сгенерируйте бинарный вопрос на ос-
нове следующего ответа:\n{answer}"
```

Вариант 2.

```
system_prompt = (
    "Вы — квалифицированный юридический помощник, который генерирует бинарные вопросы (такие, на которые можно ответить 'да' или 'нет') на основе предоставленных ответов. "
    "Сгенерированные вопросы должны максимально напоминать вопросы, задаваемые реальными пользователями и строго соответствовать содержанию ответа."
    "Каждый вопрос должен быть составлен таким образом, чтобы ответ на него можно было суммаризировать только как 'нет'."
    "Вопросы должны быть вежливыми, корректными и при необходимости включать в себя релевантную юридическую терминологию."
    "Вы не должны писать ничего, кроме текста самого вопроса. В том числе, пожалуйста, избегайте приветствий, прощаний, вводных конструкций и комментариев."
)

user_prompt = f"Пожалуйста, сгенерируйте бинарный вопрос на основе следующего ответа:\n{answer}. Сгенерируйте вопрос таким образом, чтобы ответ на него был отрицательным.”
```

CoBaLD Parser: Joint Morphosyntactic and Semantic Annotation

Ilia Baiuk
MIPT
Russia, Moscow
ilia.baiuk.hu@gmail.com

Alexandra Baiuk
RSUH
Russia, Moscow
alexandra.m.baiuk@gmail.com

Maria Petrova
RSUH, Smart Data Expert
Russia, Moscow
maria.ap.world@gmail.com

Abstract

Dependency parsing is a common task for modern NLP, and Universal Dependencies (De Marneffe et al., 2021) is widely acknowledged nowadays as a morphosyntactic annotation standard. Yet, its dependency relations are rather generalized, therefore, in order to take more syntactic details into account, the Enhanced UD standard was proposed. A newly developed CoBaLD annotation standard elaborates the E-UD principles by enriching it with the semantic level. It is aimed at structural simplicity and the compatibility with UD in all possible issues. Currently, there are several datasets annotated in CoBaLD standard, but until now, there has been no appropriate tool for automatic data parsing in CoBaLD format. In this paper, we present a neural-based joint parser capable of automatic annotation both in E-UD and in CoBaLD, including ellipsis restoration which is supposed by these standards. Additionally, we provide a qualitative analysis of automatic annotation errors.

Keywords: morphosyntactic parsing, semantic parsing, joint parsing, Enhanced Universal Dependencies, CoBaLD annotation standard

DOI: 10.28995/2075-7182-2025-23-28-37

Парсер CoBaLD: интегральная морфосинтаксическая и семантическая разметка

Илья Баюк
МФТИ
Россия, Москва
ilia.baiuk.hu@gmail.com

Александра Баюк
РГГУ
Россия, Москва
alexandra.m.baiuk@gmail.com

Мария Петрова
РГГУ, Smart Data Expert
Россия, Москва
maria.ap.world@gmail.com

Аннотация

Задача лингвистической разметки – одна из центральных задач в современной обработке естественного языка (NLP), а стандарт Universal Dependencies (De Marneffe et al., 2021) является наиболее распространенным стандартом морфосинтаксической разметки. Данный стандарт, однако, не содержит семантического уровня, а его синтаксические зависимости носят несколько обобщенный характер. В качестве продолжения UD был предложен стандарт Enhanced UD, позволяющий больше детализировать синтаксические отношения. Новый стандарт разметки CoBaLD развивает принципы E-UD, добавляя к ним уровень семантической аннотации. Его основные цели – простота использования, максимальная совместимость с UD и обогащение UD и Enhanced UD семантической разметкой. Существует ряд датасетов, размеченных в стандарте CoBaLD, однако до настоящего момента не существовало подходящего инструмента для автоматической разметки данных в данном формате. В настоящей работе мы представляем интегральный парсер для CoBaLD, выполняющий автоматическую разметку как в формате E-UD, так и в формате CoBaLD. Важным достижением является также способность восстанавливать эллипсис, предусмотренный этими стандартами. Кроме того, мы проводим качественный анализ ошибок автоматической аннотации.

Ключевые слова: морфосинтаксическая разметка, семантическая разметка, лингвистическая разметка, Enhanced Universal Dependencies, CoBaLD

1 Introduction

Linguistic annotation is an important NLP task. In addition to theoretical purposes, annotated text corpora can also be used for solving different downstream tasks, such as the creation of chat-bots or question-answering systems, for instance. Most annotation schemes provide morphological and syntactic markup, nevertheless, there are standards which suppose semantic or discourse markup as well.

One of the most popular standards nowadays is Universal Dependencies (UD, (De Marneffe et al., 2021)) – a project aimed to capture the so-called dependencies in a language. It includes morphological and syntactic levels, although the relations UD annotates are not purely syntactic due to its principles concerning linguistic universality. Moreover, the UD project offers its enhanced syntax version, Enhanced UD (E-UD, (Schuster and Manning, 2016)), which suggests a more detailed syntax annotation and ellipsis restoration.

On the other hand, there are a few standards of semantic annotation, though none of them is universally accepted as of now. Some models are purely semantic (such as Abstract Meaning Representations (Banarescu et al., 2013) or Universal Cognitive Conceptual Annotation (Abend and Rappoport, 2013)), others include several language levels (Prague Tectogrammatical Graphs (Mikulová et al., 2006) or the ETAP model (Apresian et al., 2003)). There are projects as well developed to be compatible with UD, namely, Universal Decompositional Semantics, or UDS (White et al., 2016) and CoBaLD Annotation Project (Petrova et al., 2023; Petrova et al., 2024).

UDS suggests only the semantic annotation for the UD morphosyntax, presented in the form of different semantic attributes rather than labels for definite semantic meanings, which may make its usage inconvenient for solving practical tasks and parser learning.

CoBaLD includes both morphosyntax and semantics, striving to be as compatible with UD/E-UD as possible in the description of morphosyntax and enriching it with the semantical pattern. Its semantics provides both lexical meanings and the relations between words. The former are presented in the form of semantic classes (SCs, or semantic fields) organized in a thesaurus-like structure, the latter – in the form of deep slots (DSs, or semantic roles). Their main feature is that the DSs in CoBaLD include not only arguments, but adjuncts, attributes and all other dependencies as well. The descriptions and lists of the DSs and the SCs can be found on our Github page[1].

In UD, the annotation is presented in the CONLL-U Plus file standard[2], which looks like a table, where each column corresponds to some annotated meaning and each row – to the annotated token. As the key purpose of the CoBaLD standard is to add the semantic domain to the UD/E-UD morphosyntax, it just adds two new columns with the semantic information in the CONLL-U Plus, that is, a column for the DSs and a column for the SCs.

The only difference in the syntactic level annotation is that CoBaLD restores more ellipsis cases than E-UD, namely, the elided subjects. As in E-UD, some referential connections are annotated here, too. Currently, there are two CoBaLD-annotated datasets[3] that have an E-UD level as well. Therefore, the parser developed for CoBaLD can also be used to parse E-UD.

Our contributions in this paper are as follows:

- we propose a joint morphosyntactic and semantic parser which can be used to parse both CoBaLD and E-UD, and publish our code and pre-trained models;
- we investigate syntax and semantics mutual influence that affects the quality of automatic annotation;
- we provide a qualitative analysis of semantic annotation errors.

[1] https://github.com/CobaldAnnotation

[2] https://universaldependencies.org/ext-format.html

[3] https://github.com/CobaldAnnotation/CobaldEng CoBaLD Eng for English
https://github.com/CobaldAnnotation/CobaldRus CoBaLD Rus for Russian

2 Experimental Setup

2.1 Architecture

The suggested model is a multi-task sequence tagger based on Joint Morpho-Syntactic Parser proposed in (Anastasyev, 2020). It consists of a Masked Language Model encoder paired with a set of tagging heads. The encoder transforms the sequence of tokens into contextual embeddings, then each head selects the features relevant to its tier and predicts the tags upon the embeddings (Fig. 1).

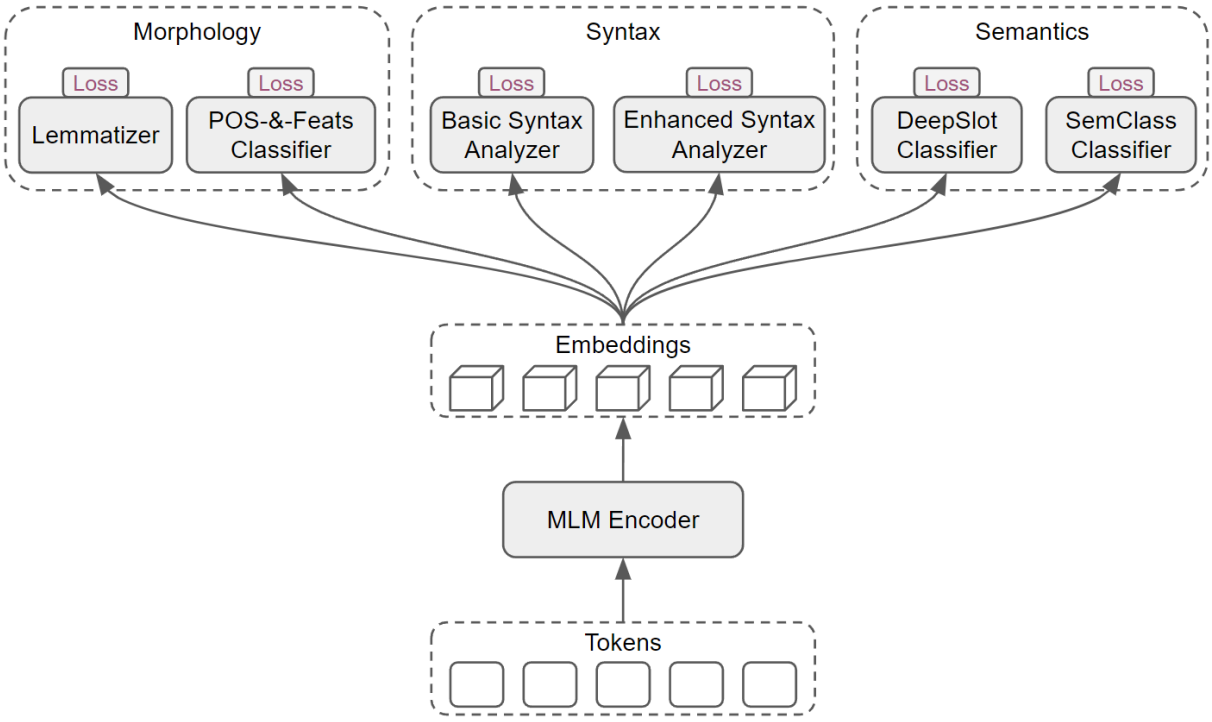


Figure 1: Parser architecture

There are six heads: a lemmatizer, a joint part-of-speech and grammatical features classifier, two syntax analyzers and two semantic classifiers.

Lemmatizer predicts *lemmatization rules* – a sequence of modifications that transform a word into a lemma. The three modifications are available:

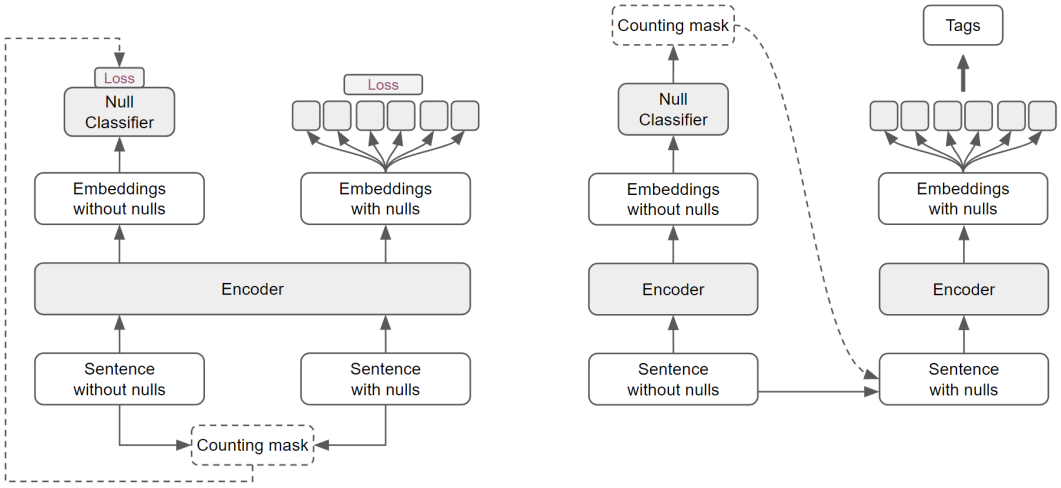
- cut N symbols at the prefix,
- cut N symbols at the suffix,
- append a specific suffix.

For example, in order to obtain *elf* from *elves*, one needs to leave prefix intact, cut suffix of length three and append "f", so the lemmatization rule for this case is represented as a `cut_pref=0|cut_suff=3|append_suff=f`.

POS-&-Feats head predicts joint part of speech and grammatical features tags. It ensures POS and features consistency, so that nouns never acquire a category of tense and cases never relate to verbs. Since the morphological tags are heavily correlated, the number of combinations remains small and tractable for the classification approach.

Basic syntax analyzer is a biaffine dependency classifier (Dozat and Manning, 2016) that predicts basic syntactic connections between tokens in a sentence. During the training, the arcs are chosen greedily; at the inference, we use Chu-Liu-Edmonds algorithm to find a spanning arborescence of a maximum probability.

As for the **enhanced syntax**, our model is based on the dependency graph parser (He and Choi, 2020). In contrast to the basic analyzer that builds a single distribution over a token's arcs by minimizing a multi-class cross-entropy loss, the graph parser estimates each arc probability independently using a



(a) Training pipeline. The null classifier is optimized alongside the tagging heads.

(b) Inference pipeline. The parser restores the nulls first (left), then tags the modified sentence (right).

Figure 2: Parser workflow

multi-label approach with a binary cross-entropy loss. This distributional relaxation allows the enhanced analyzer to predict multiple edges per token, accounting for the non-acyclic nature of enhanced syntactic graphs.

However, the dependency graph parser does not address the ellipsis, as the latter involves a more complex task of elided (null) tokens restoration. Up to this date, no attempt has been made to approach this problem in full — for example, at IWPT Shared Task (Bouma et al., 2020), the arcs involving null nodes are collapsed into multi-edges, which allows one to identify the presence of an elided node, but not its position. In the current version of CoBaLD, as opposed to E-UD, elided functional words such as auxiliary verbs or copulas may be restored, so the positions of the elided elements are important. For this reason, we target the original task and actually restore the elided tokens.

To achieve this, we introduce a *counting mask* that shows how many nulls follow a token. Formally, given a gold sentence $(s_1, \ldots, s_n)$ with nulls, denote $I = (i \in \{1, \ldots, n\} \mid s_i \neq \varnothing)$ as indices of non-null tokens, i.e. the subsequence $(s_i)_{i \in I}$ is a test sentence with null tokens removed. The mask is defined as $L = (I^+_{i+1} - I^+_i - 1)^m_{i=1}$, where $I^+ = I \cup \{n\}$ and $m = |I|$ is a length of the test sentence.

For example, suppose the gold sentence is `[Quick, brown, #NULL, #NULL, fox, #NULL]`. Then the test sentence without nulls would be `[Quick, brown, fox]`, $I = [1, 2, 5]$ and the counting mask $L = [0, 2, 1]$, for there is no null after `Quick`, two nulls after `brown` and one after `fox`.

Given the counting mask, ellipsis restoration becomes as simple as sequence labeling; we add another head that predicts how many nulls should be inserted after a token. It splits the pipeline into two stages: first, the null classifier retrieves the elided tokens and adds them to an input sentence, then the refined sentence is parsed as usual (Fig. 2b). During the training, sentences with and without nulls are passed through the encoder independently, so the tagging heads are always trained upon the well-formed sentences, and not the restored ones (Fig. 2a). It is similar to a teacher forcing approach used in sequence generation, where a ground truth token is fed to a model instead of the previously predicted word to avoid error accumulation and speed up the training.

We jointly optimize the null classifier with the tagging heads by adding up the losses. The encoder aggregates the gradients from all classifiers and updates the weights once per training step.

Note that the counting mask ignores the leading nulls, e.g. `[#NULL, On, that, date]`, as there is no token to follow. To account for this issue, we prepend an auxiliary token to the beginning of a test sentence prior to mask construction.

The **semantic classifiers** predict deep slots and semantic classes. As the semantic categories of CoBaLD are very simple from a technical point of view, the task of semantic parsing represents basic multinomial classification for both DSs and SCs which are predicted independently.

Pretrain dataset	Train dataset	Test quality, %			
		UAS	LAS	EUAS	ELAS
-	CoBaLD Eng	92.3	91.2	86.4	84.6
-	CoBaLD Eng + GUM + EWT	92.5	91.3	87.1	85.1
GUM + EWT	CoBaLD Eng	**93.3**	**91.6**	**89.3**	**87.6**

Table 1: Different training strategies (scores for the joint model)

All the heads (except for syntactic ones) are two linear layers separated by a ReLU activation.

2.2 Experiments

We have experimented with several setups for the parser, changing the data we trained it on, the encoder and the amount of tagging heads in order to analyze how joint morphosyntax and semantics parsing would affect overall quality.

The two datasets we experimented on are referred to as EWT and BBC. The former is a portion of English EWT dataset (Silveira et al., 2014) available on UD official Github[4], with manually annotated semantics level, containing around 80,000 tokens. The latter is the aforementioned CoBaLD Eng dataset which contains BBC news texts and is around 190,000 tokens. We also used a portion of English PUD dataset available on UD Github[5] for testing.

Since the syntax level, especially the enhanced one, is known to be a weak spot of the parsers (judging by the scores, e.g. in (Straka et al., 2016)), we experimented on enriching the training data with syntactic tags of public E-UD treebanks, namely GUM (Zeldes, 2017) and EWT[6], around 200,000 tokens each. We tried two strategies: enlarge the training set with the new data, or pre-train the model on the public data first and then finetune it on the original dataset. The Table 1 shows that the pre-training approach significantly improves the syntactic scores, so we adhere to this strategy in the subsequent experiments.

We have used two pre-trained MLMs as encoders: distilbert-uncased[7] and XLM-RoBERTa Base[8].

The experiments concerning syntax and semantics mutual influence included training of the following models: 1) a joint model; 2) a model with morphosyntax parsing only; 3) a model with semantics parsing only.

3 Results

The results for our experiments are presented in Table 2. The metrics we adopted for the morphology and base UD syntax are commonly used metrics which are described, for instance, in (Lyashevskaya et al., 2020) with slight improvements for lemmatization and grammatical features described in (Petrova et al., 2023).

Since the F_1-based enhanced attachment scores used in IWPT Shared Tasks are described quite loosely, we define enhanced labeled attachment score in a manner similar to a Jaccard coefficient, but use max instead of union to soften the penalty:

$$ELAS(test,\ gold) = \frac{|test_{deps} \cap gold_{deps}|}{max(|test_{deps}|, |gold_{deps}|)},$$

where $token_{deps}$ is a dictionary with keys representing heads and values representing relations, e.g. `{(24, nsubj), (26, nsubj:xsubj)}`. The EUAS is defined similarly, but the relations are ignored. These metrics equal to LAS and UAS in case of tree syntax, which makes base and enhanced tiers easy to compare to each other.

[4]https://github.com/UniversalDependencies/UD_English-EWT
[5]https://github.com/UniversalDependencies/UD_English-PUD
[6]The part of EWT annotated with semantics was excluded
[7]https://huggingface.co/distilbert/distilbert-base-uncased
[8]https://huggingface.co/FacebookAI/xlm-roberta-base

As for the semantics metrics, they are calculated according to the next formulas.
The *deep slot* (DS) score is a simple accuracy:

$$DeepSlot(test,\ gold) = [test_{DS} = gold_{DS}].$$

Semantic classes (SC) are scored based on the semantic hierarchy of CoBaLD standard:

$$SemClass(test,\ gold) = \frac{1}{1 + Distance(test_{SC},\ gold_{SC})}.$$

That is, the closer test and gold semantic classes are in hierarchy, the higher the score is.

Considering the fact that elided tokens are to be predicted, the number of tokens in test and gold sentences may differ. It causes certain problems, because all metrics operate on the token level and expect the latter to be aligned. To account for this issue, we add a special *empty token*, which fills the gaps caused by misplaced nulls. If a gold null is missed, the empty token is inserted into the test sentence; if an extra null is predicted, the empty token is added to the gold one. Thus, the empty token always faces a null in the opposite sentence. If one of the tokens in a test-gold pair is empty, the scores of such pair are zeroed.

Dataset	Model	Type	Lemma	POS	Feats	UAS	LAS	EUAS	ELAS	DeepSlot	SemClass
EWT	Distil	Joint	**93.21**	91.97	**93.79**	88.23	81.06	83.14	75.84	84.38	85.77
		NoSem	92.83	**92.05**	93.70	**88.36**	**81.31**	**83.89**	**76.62**	-	-
		Sem	-	-	-	-	-	-	-	**84.92**	**87.78**
	XLM	Joint	**93.73**	**93.12**	**93.98**	**89.21**	**82.20**	80.13	73.37	**84.85**	86.95
		NoSem	93.38	93.02	93.95	89.12	82.11	**81.39**	**74.67**	-	-
		Sem	-	-	-	-	-	-	-	83.80	**88.30**
BBC	Distil	Joint	**97.10**	**97.90**	**98.52**	93.41	91.84	89.77	88.07	89.63	92.06
		NoSem	97.06	97.76	98.47	**93.47**	**91.88**	**90.41**	**88.74**	-	-
		Sem	-	-	-	-	-	-	-	**90.00**	**93.04**
	XLM	Joint	97.23	**98.51**	**98.55**	92.93	91.10	84.84	82.83	88.71	91.93
		NoSem	97.23	98.45	98.47	**93.16**	**91.28**	**85.08**	**83.03**	-	-
		Sem	-	-	-	-	-	-	-	**88.79**	**92.83**
BBC + EWT	Distil	Joint	**97.15**	**97.75**	**98.41**	93.02	91.28	89.18	87.31	89.59	92.32
		NoSem	97.12	97.60	98.29	**93.30**	**91.54**	**90.05**	**88.16**	-	-
		Sem	-	-	-	-	-	-	-	**90.20**	**93.02**
	XLM	Joint	**97.25**	98.13	98.19	**92.83**	90.59	84.89	82.46	88.92	92.20
		NoSem	97.08	**98.22**	**98.23**	92.72	**90.64**	**84.94**	**82.62**	-	-
		Sem	-	-	-	-	-	-	-	**89.23**	**92.95**

Table 2: Model scores for various setups with different training data, backbone LM and tuned tagging heads. Joint stands for a joint morphosyntactic and semantic model, NoSem – a model with morphosyntax only, Sem – a model with semantics only

3.1 Morphosyntax and Semantics

Our results appeared quite stable and a little bit surprising as, first, a more complex XLM-RoBERTa Base (XLM in Table 2) steadily showed worse scores than distilbert-uncased, second, the quality of Enhanced UD syntax and semantics decreased a little for joint versions. Slight drops in quality for the models trained on combined datasets may be caused by different data distribution (EWT contains colloquial texts while BBC is comprised of news only). Also intriguing is the fact that base UD syntax prediction quality didn't suffer much (differences in scores are quite insubstantial).

A qualitative error analysis showed that there is no obvious connection between semantic and syntax levels concerning errors, but a joint model has a tendency of predicting higher probabilities to plausible arcs which sometimes leads to better results compared with the syntax-only model, whose probabilities are too low to overcome the threshold causing it to predict no arcs at all. As for the joint model, it ends up predicting more arcs than it should (so generally there are about 1% more arcs predicted by a joint model than by a syntax-only model). It must be noted that incorrectly predicted syntax heads are quite reasonable and usually do have both semantic and syntactic relationship, e.g., for a sentence 'Based on the Hellblazer comics, the film took $ 11.8 m (£ 6.1 m) on its second week of release' token *comics* is governed by *based*, which in turn has *took* as its head. XLM-R joint model predicts both *based* and *took* as heads with the same dependency type (which is correctly predicted as *obl:on*).

Unfortunately, we also found out that the private test dataset we used for validation purposes contained errors in syntactic annotation, but this fact does not seem to influence the quality drops much, as we also tested syntax prediction quality on English PUD (see Table 3), and it showed the same tendency, though overall quality worsened, most probably due to differences in data distribution and annotation principles (note that E-UD annotation level may be quite different for various E-UD datasets as there are no detailed manuals for it).

Model	Type	EUAS	ELAS
Distil	Joint	83.68	78.40
	NoSem	**83.87**	**78.66**
XLM	Joint	81.02	75.50
	NoSem	**81.33**	**76.06**

Table 3: English PUD testing scores for E-UD annotation level

Another interesting fact is that although the XLM-R-based models consistently show worse results than the distilbert-based ones, their performance drops for E-UD syntax are less, moreover, they are diminished when the training data amount (and diversity) increases. This tendency for XLM-R is less pronounced with semantics, though. Nevertheless, there may be another factor we should take into account, namely, there are some discrepancies in semantics and syntax level we are aware of. Most notable is the difference in copula annotation: as the semantics part of CoBaLD standard was inherited from another annotation model (for details see (Ivoylova et al., 2023)), copula is considered the head of its complement, while in UD, it is the dependent node. However, deep slots in CoBaLD are annotated according to the former principle, so that it may be confusing for the parser. There are also some minor discrepancies of this kind which we plan to obliterate, so the drops in semantics parsing quality may disappear.

In any case, it should be said that overall quality of the parser is quite high and the performance drops for joint versions that we have discussed are tolerable. Therefore, morphosyntax and semantics can be parsed simultaneously at almost no cost in quality, and there is a chance that if we increase the amount of training data and improve the compatibility of the syntactic and semantic annotation tiers, the quality of joint parsing will even grow compared to separate syntax and semantics prediction.

3.2 Semantic errors of CoBaLD parser

The semantic errors of the parser concern the definition of the word's SC, the DS which binds the dependent node with its parent node, or the combination of the items.

The scores for DSs and SCs are given in Table 2.

The mistakes in defining the SCs deal mostly with the following cases:

- the parser suggests a SC of a homonym instead of the proper class: for instance, the SCs BEING vs CH_REFERENCE_AND_QUANTIFICATION for the word 'couple' in *The couple now have four children together* ('couple' meaning 'a pair of people being together' vs 'an indefinite small number' as in *just a couple more questions*);

- the parser suggests a hyperonym with a more general meaning - e.g. ORGANIZATION instead of MILITARY_FORCES_AS_ORGANIZATION;
- the parser defines a SC of an unknown word (usually, proper name) or a pronoun other than the human annotator: as the SCs ORGANIZATION vs ARRANGEMENTS for 'CES' in *CES showcases 50,000 new gadgets that will be hitting the shelves in 2005*.

The mistakes dealing with the DSs can be divided in two groups: first, when the SC of the head token is correct but the DS of its dependent is improper (≈56% of all DS errors), and, second, when both the SC of the head (or the head itself) and the DS of its dependent are improper (≈44% respectively).

The first case can be illustrated by the example: *In my book this is cheating on the club, the supporters, the manager and his own team-mates.* Here the human annotator labelled 'my' as Possessor, while the parser marked it as Agent (as if 'I am the author of the book').

An example of the second case can be the sentence: *A number of fans questioned Gerrard's commitment and sarcastically branded his own goal in Liverpool's 3-2 defeat as his first goal for Chelsea.* The parser defines 'number' as Agent DS of the verb 'question', while the human annotator suggests the Experiencer DS here. The reason is that the parser and the annotator chose different SCs for 'question' as well: 'question' as VERBAL_COMMUNICATION ('to ask a question') which demands Agent for its subject and 'question' as STATE_OF_MIND ('feel or express doubt about') which demands the Experiencer subject.

Such errors seem natural as they mainly deal not with the parser's work itself, but with the semantic ambiguity.

Nevertheless, there are errors as well, where the parser suggests the annotation which is semantically impossible in terms of the given model, for instance, marks the subject of some active verb as Experiencer instead of Agent.

This problem mainly concerns actant slots (Agent, Object, Experiencer, and alike) – that is, slots with similar filling and syntactic realizations which differentiate through their semantic relations with their cores only. Circumstantial adjuncts (like locative, temporal, conditional, or concession adjuncts) are usually defined properly as their surface realizations, semantic filling and sense does not depend on the cores they are governed by. The number of such errors should reduce with enlarging the volume of the annotated data.

4 Related Work

There have been two shared tasks dedicated to the parsing of E-UD held on the International Conference on Parsing Technologies in 2020 (Bouma et al., 2020) and 2021 (Oepen et al., 2021). The winner for the former task is TurkuNLP (Kanerva et al., 2020), and for the latter one – TGIF (Shi and Lee, 2021).

There exists a hypothesis that joint parsing of several language levels should help to improve the quality for all of them. As for morphology and syntax, it has become a standard to parse them jointly. Concerning joint syntax and semantics parsing, some of the earliest attempts to build the complete formal semantic representations along with the syntactic ones automatically were made almost 20 years ago (Ge and Mooney, 2005), and the attempts to solve the task of Semantic Role Labeling simultaneously with syntactic dependency parsing even earlier than that, e.g., (Gildea and Jurafsky, 2002).

Normally, the task of syntactic parsing is regarded as a way to improve the quality of semantic parsing (Li et al., 2020), although modern research shows that semantic parsing can achieve impressive results without the support of syntax. Thus, syntax and semantic parsing can be now viewed as separate tasks for multi-task learning. There have been researches dedicated to such simultaneous parsing in recent years as well, e.g. (Zhou et al., 2019), where the authors attempt parsing constituents and syntax dependencies together with the Semantic Role Labeling. As shown in the mentioned paper, syntax dependencies don't improve semantics, while the constituents affect everything positively. The other notable work (Stengel-Eskin et al., 2021) concerns the aforementioned Universal Decompositional Semantics; the authors report that syntax dependencies slightly improve the semantic level parsing.

5 Conclusion

In this paper, we present a joint morphosyntactic and semantic parser for CoBaLD annotation standard which can be applied to Enhanced UD parsing as well and is capable of ellipsis restoration. Its scores across all linguistic levels show the quality comparable with that of SOTA solutions for E-UD and UD parsing. The code for the parser is available on Github[9] and the trained models are published on Huggingface.co[10]. We also provide a qualitative analysis of the parser's semantic errors and analyze the possible reasons of the slight performance drops for the joint syntax and semantics parsing. Our future goals include further investigation of the syntax and semantics mutual influence, as well as the improvements to the architecture and the annotation standard itself.

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). // *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 228–238.

DG Anastasyev. 2020. Exploring pretrained models for joint morpho-syntactic parsing of russian. // *Computational Linguistics and Intellectual Technologies*, P 1–12.

Jurij D Apresian, Igor M Boguslavsky, Leonid L Iomdin, Alexander V Lazursky, Vladimir Z Sannikov, Victor G Sizov, and Leonid L Tsinman. 2003. Etap-3 linguistic processor: a full-fledged nlp implementation of the meaning ótext theory1. *Institute for Information Transmission Problems, Russian Academy of Sciences*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. // *Proceedings of the 7th linguistic annotation workshop and interoperability with discourse*, P 178–186.

Gosse Bouma, Yuji Matsumoto, Stephan Oepen, Kenji Sagae, Djamé Seddah, Weiwei Sun, Anders Søgaard, Reut Tsarfaty, and Daniel Zeman. 2020. Proceedings of the 16th international conference on parsing technologies and the iwpt 2020 shared task on parsing into enhanced universal dependencies. // *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*.

Marie-Catherine De Marneffe, Christopher D Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational linguistics*, 47(2):255–308.

Timothy Dozat and Christopher D. Manning. 2016. Deep biaffine attention for neural dependency parsing. *CoRR*, abs/1611.01734.

Ruifang Ge and Raymond Mooney. 2005. A statistical semantic parser that integrates syntax and semantics. // *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL-2005)*, P 9–16.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Han He and Jinho D. Choi. 2020. Adaptation of multilingual transformer encoder for robust enhanced Universal Dependency parsing. // Gosse Bouma, Yuji Matsumoto, Stephan Oepen, Kenji Sagae, Djamé Seddah, Weiwei Sun, Anders Søgaard, Reut Tsarfaty, and Dan Zeman, *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, P 181–191, Online, July. Association for Computational Linguistics.

Alexandra Ivoylova, Darya Dyachkova, Maria Petrova, and Mariia Michurina. 2023. The problem of linguistic markup conversion: the transformation of the compreno markup into the ud format. // *International Conference on Computational Linguistics and Intellectual Technologies «Dialog*.

Jenna Kanerva, Filip Ginter, and Sampo Pyysalo. 2020. Turku enhanced parser pipeline: From raw text to enhanced graphs in the iwpt 2020 shared task. // *Proceedings of the 16th International Conference on Parsing Technologies and the IWPT 2020 Shared Task on Parsing into Enhanced Universal Dependencies*, P 162–173.

Zuchao Li, Hai Zhao, Shexia He, and Jiaxun Cai. 2020. Syntax role for neural semantic role labeling. *arXiv preprint arXiv:2009.05737*.

[9]https://github.com/CobaldAnnotation/CobaldParser
[10]https://huggingface.co/CoBaLD

ON Lyashevskaya, TO Shavrina, IV Trofimov, NA Vlasova, et al. 2020. Grameval 2020 shared task: Russian full morphology and universal dependencies parsing. // *Proc. of the International Conference Dialogue*, volume 2020, P 553–569.

Marie Mikulová, Alevtina Bémová, Jan Hajič, Eva Hajičová, Jiří Havelka, Veronika Kolářová, Lucie Kučová, Markéta Lopatková, Petr Pajas, Jarmila Panevová, et al. 2006. Annotation on the tectogrammatical level in the prague dependency treebank. *annotation manual. Technical Report*, 30:5–11.

Stephan Oepen, Kenji Sagae, Reut Tsarfaty, Gosse Bouma, Djamé Seddah, and Daniel Zeman. 2021. Proceedings of the 17th international conference on parsing technologies and the iwpt 2021 shared task on parsing into enhanced universal dependencies (iwpt 2021). // *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*.

Maria Petrova, Alexandra Ivoylova, Ilya Bayuk, Darya Dyachkova, and Mariia Michurina. 2023. The cobald annotation project: the creation and application of the full morpho-syntactic and semantic markup standard. // *Proceedings of the International Conference "Dialogue*, volume 2023.

Maria Andreevna Petrova, Alexandra M Ivoylova, and Anastasia Tishchenkova. 2024. Cobald annotation: The enrichment of the enhanced universal dependencies with the semantical pattern. // *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 3422–3432.

Sebastian Schuster and Christopher D Manning. 2016. Enhanced english universal dependencies: An improved representation for natural language understanding tasks. // *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, P 2371–2378.

Tianze Shi and Lillian Lee. 2021. Tgif: Tree-graph integrated-format parser for enhanced ud with two-stage generic-to individual-language finetuning. *arXiv preprint arXiv:2107.06907*.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Christopher D. Manning. 2014. A gold standard dependency corpus for English. // *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

Elias Stengel-Eskin, Kenton Murray, Sheng Zhang, Aaron Steven White, and Benjamin Van Durme. 2021. Joint universal syntactic and semantic parsing. *Transactions of the Association for Computational Linguistics*, 9:756–773.

Milan Straka, Jan Hajic, and Jana Straková. 2016. Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. // *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, P 4290–4297.

Aaron Steven White, Drew Reisinger, Keisuke Sakaguchi, Tim Vieira, Sheng Zhang, Rachel Rudinger, Kyle Rawlins, and Benjamin Van Durme. 2016. Universal decompositional semantics on universal dependencies. // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, P 1713–1723.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

Junru Zhou, Zuchao Li, and Hai Zhao. 2019. Parsing all: Syntax and semantics, dependencies and spans. *arXiv preprint arXiv:1908.11522*.

Embedded quotation as a factor of idiomaticity

Baranov A. N.
Vinogradov Russian Language Institute of the Russian
Academy of Sciences / Volkhonka 18/2, Moscow
Baranov_anatoly@hotmail.com

Abstract

The report examines one of the types of idiomaticity—embedded quotation. This type is peripheral and is implemented in a small set of idioms, slightly more than twenty elements. Idioms with embedded citations include, in particular, the expressions *po samoye ne baluysya* (lit. to the very don't play) ≈ 'to the maximum extent possible', *po samoye ne kho-chu* (lit. to the very don't want) 'to the maximum extent possible', *na otskrebis'* (lit. on fuck off) ≈ 'to do something quickly and badly', *prikazat' dolgo zhit'* (lit. to order to live long) ≈ 'to die', *derzhat'sya na chestnom slove* (lit. to keep one's word of honor) ≈ 'be poorly connected to the main part of the object', *pod chestnoye slovo* (lit. under a word of honor) ≈ 'to do something to someone or allow him to do something under the guarantee of this person', *za Khrista radi* (lit. for Christ's sake) ≈ 'to beg', *[i] pominay kak zvali* (lit. [and] remember his name) ≈ 'disappear without a trace', *vot tebe / te[i] na* (lit. here's for you) ≈ 'expression of surprise at failure', *za spasibo* (lit. for thanks) ≈ 'doing something for free', *za [prosto] tak* (lit. for no reason) ≈ 'doing something for free' and some others.

Embedded quotation is a relevant factor of idiomaticity, leading to the emergence of new idioms in the language.

Embedded quotation as a factor of idiomaticity is directly related to the phenomena of quotation (citation) and generalized quotation (citation). The existence of generalized quotation in speech provides the very possibility of forming idioms, the idiomaticity of which is based on embedded quotation. This is a necessary intermediate link in the sequence: quotation—generalized quotation—embedded quotation.

Keywords: idiomatics; phraseology; the phenomenon of quotation; embedded quotation
DOI: 10.28995/2075-7182-2025-23-38-44

На отскребись[1]: вложенная цитация как фактор идиоматичности

Баранов А. Н.
Институт русского языка им. В. В. Виноградова РАН /
Волхонка 18/2 Москва
Baranov_anatoly@hotmail.com

Аннотация

В докладе рассматривается один из типов идиоматичности — вложенная цитация. Данный тип относится к периферийным и реализуется в небольшом множестве идиом, чуть превышающем двадцать единиц. К идиомам с вложенной цитацией относятся, в частности, выражения *по самое не балуйся, по самое не хочу, на отскребись*, *приказать долго жить*, *держ ат ься на чест ном слове, под чест ное слово*, *за Христ а ради, [и] поминай как звали, вот т ебе / т е[и] на, за спасибо, за [прост о] т ак* и некоторые другие.

Вложенная цитация представляет собой актуальный фактор идиоматичности, приводящий к возникновению в языке новых идиом.

Вложенная цитация как фактор идиоматичности непосредственно связана с феноменами цитации и генерализованной цитации. Существование генерализованной цитации в речи обеспечивает саму возможность формирования идиом, идиоматичность которых основана на вложенной цитации. Это необходимое промежуточное звено в последовательности: цитация — генерализованная цитация — вложенная цитация.

Ключевые слова: идиоматика; фразеология; феномен цитации; вложенная цитация

[1] Имеется в виду обсценный глагол с приставкой *от-*.

1 Постановка задачи

Идиоматичность — фундаментальное свойство функционирования языка, возникающее в процессе речевой деятельности в результате действия принципа экономии. Действительно, повторяющиеся ситуации приводят к повторению речевого поведения, а повторяющееся в речи сокращается по форме из-за закона экономии. Сокращение формы, в свою очередь, приводит к нарушению регулярности в формировании смысла и, соответственно, к идиоматичности речи. Области идиоматичности речевых форм регулярно модифицируются, поскольку мир меняется, меняются и условия взаимодействия человека с действительностью. В то же время идиоматичность воспроизводится в новых областях реализаций языка, поскольку стремление к экономии усилий универсально.

Хотя идиоматичность подвижна в своих реализациях (ранее идиоматичное может стать неидиоматичным и наоборот), ее принципы остаются постоянными. Под идиоматичностью, или некомпозициональностью семантики, (далее используются оба термина как синонимы) в соответствии с общепринятой трактовой понимается формирование значения речевой формы, нарушающее правила, фиксированные в данной языковой системе. Хотя идиоматичность распространяется, как минимум, на морфологическую структуру слова, словосочетание и предложение, ниже обсуждаются исключительно сверхсловные образования — идиомы, то есть словосочетания и предложения, которым свойственна высокая степень идиоматичности и устойчивости [4].

В [3; 4] описаны базовые типы идиоматичности. Чаще всего идиоматичность возникает из-за **переинтерпретации**: значение словосочетания возникает как результат переосмысления значения языкового выражения с регулярной семантикой, вследствие чего возникает другое значение, сформированное не по регулярным правилам (ср. идиомы типа *работать спустя рукава*, *поезд ушел, бить баклуши* и под.). Еще один распространенный способ идиоматизации — **непрозрачность**, характеризующая идиомы, в которых присутствует уникальный компонент (*бить баклуши, не видно ни зги, охулки на руку не класть*). Наконец, третий тип идиоматичности, который широко представлен во фразеологии, — **усложнение способа указания на денотат** (*Новый Свет* — вместо *Америка*, *Город на Неве* — вместо *Петербург*, *Утренняя Звезда* — вместо *Венера*, *дать дуба* — вместо *умереть*). Перечисленные факторы (типы) идиоматичности являются базовыми для фразеологии в том смысле, что в подавляющем большинстве идиом обнаруживается один из перечисленных факторов или любая их комбинация.

Кроме базовых типов идиоматичности, есть переходные, периферийные случаи, которые, во-первых, в разных своих проявлениях не прямо связаны с базовыми и, во-вторых, характеризуют небольшие группы идиом. В [2; 4] описан один из таких периферийных типов — **онимизация**, то есть переход имен нарицательных в имена собственные или структурное преобразование имен нарицательных в имена собственные, приводящее к приобретению ими тех или иных семантических свойств имен собственных. Характерные примеры онимизации — идиомы *Софья/Степанида Властьевна* (‘советская власть’), *Манька Величкина* (‘мания величия’); *Голим Голимыч* и *Облом Обломыч* (‘неуспех’), *Дед Мороз, ванька мокрый, иван-да-марья*, *Михайло Иваныч* (‘медведь’) и др. Часто онимизация сочетается с языковой игрой и даже порождает ее. Так, игровая идиома советских времен *Александр Третий* возникла как контаминация наименования *тройной одеколон* с названием мужского одеколона *Саша* и обозначала коктейль из указанных ингредиентов. Персонификация с использованием имен собственных также относится к этому типу идиоматичности: *Кондратий хватил*.

Идиоматизация — продуктивный и активно функционирующий процесс. Следует ожидать, что, наряду с онимизацией, существуют и другие типы идиоматичности, которые, будучи периферийными в количественном отношении, важны с точки зрения потенциала языковой системы. Здесь рассматривается еще один периферийный тип идиоматичности — вложенная цитация. Поскольку последняя связана по своим свойствам с цитацией, рассмотрим сначала цитацию как фактор идиоматичности.

2 Цитация как характеристика переинтерпретации

Постмодернизм органично сочетается с цитацией — явной или скрытой, а иногда становится основным приемом организации художественного текста. В литературе неоднократно отмечалось, что цитация используется для языковой игры и позволяет создать текст в тексте [6] и придает ему свойство диалогичности [5]. Во фразеологии цитация рассматривается как фактор переинтерпретации — один

из важнейших аспектов идиоматичности [4]. Однако сама по себе цитация далеко не всегда приводит к возникновению идиом. В традиционной лексикологии и лексикографии цитаты образуют довольно аморфный класс «крылатых слов». Для фразеологии значимыми оказывают только такие цитаты, которые общеизвестны и воспроизводимы в дискурсе. Именно сочетание этих двух свойств приводит к идиоматизации и вхождению цитаты в состав идиом.

Таким сочетанием свойств обладают идиомы — цитаты по происхождению — *Боливар не выдержит двоих* (из рассказа О. Генри); *вы не церкви, вас не обманут* (из «Двенадцати стульев» И. Ильфа и Е. Петрова); *давненько не брал я в руки шашек* (из «Мертвых душ» Н. В. Гоголя); *зеркало русской революции* (из статьи Ульянова-Ленина о Толстом); *ум, честь и совесть нашей эпохи* (о РКПб — из статьи «Политический шантаж» Ульянова-Ленина).

Множество крылатых слов неоднородно в том отношении, что авторство некоторых цитат теряет определенность. Так, идиома *сквозь тернии к звёздам* возникла в русском языке как калька латинского выражения *per aspera ad astra*. По некоторым источникам ее авторство приписывается Сенеке Младшему, однако точно установить авторство не представляется возможным. Идиома *матрос ребенка не обидит* представляет собой измененное название повести советского и российского писателя М. М. Попова — «Солдат ребенка не обидит». Таким образом, во многих случаях идиома воспринимается носителями языка как цитата, но авторство ее для значительной части носителей языка утеряно, а для форм типа *матрос ребенка не обидит* вообще под вопросом. Б. С. Швацкопф в связи с этим отмечал: «Часто в сознании говорящего крылатое выражение не ассоциируется с конкретным автором или конкретным произведением» [8]. Иными словами, общепринятое определение цитации как точное воспроизведение в одном тексте фрагментов другого текста, принадлежащих другому лицу, должно быть изменено в отношении «другого лица»: цитация в рассматриваемом случае предполагает «не какое-либо отдельное конкретное лицо <…>, а некую совокупность говорящих на данном языке» [8]. Назовем такую цитацию «**генерализованной**».

3 Генерализованная цитация в идиоматике

Указанный тип идиоматичности представлен в идиомах двумя группами выражений. С одной стороны, это формы *отец народов* и *лучший друг советский физкультурников* (о Сталине), *железный Феликс, рыцарь революции* (о Дзержинском), *вечно живой* (об Ульянове-Ленине). Идиоматичность этих выражений поддерживается сочетанием цитации и уникальности референта. Впрочем, идиома *вечно живой* в современном дискурсе относится отнюдь не только к Ленину. Согласно «Яндекс играм», на вопрос *Кого имеют в виду, говоря «он вечно живой»?* правильными признаются ответы: Ленин, Цой, Высоцкий, Бог, Герой, Вампир [9].

С другой стороны, генерализованная цитация представлена типичными репликами в типизированных речевых ситуациях. Это идиомы *от забора до обеда, полный вперед, от винта, сами мы не здешние, для белых, так держать!* и под. Авторство таких цитаций не известно, но они воспроизводят речевые акты, типичные или даже нормативные для определенных коммуникативных ситуаций. Так, идиома *от забора до обеда* представляет собой идиотический приказ, якобы отданный солдатам мелким армейским чином, путающим место и время:

(1) Уставы не надо понимать. Их надо знать! Сортир был вонюч, но мы перестали быть восприимчивыми к подобным мелочам и, посветив себе спичкой и запалив самокрутки, высаживались орлами над выгребной ямой, чтобы провести полчаса — час в интеллектуальной беседе, послушать анекдоты и рассказы очевидцев о своих сержантах и офицерах. Сидение в сортире, в связи с иррациональностью этого занятия, мы называли «копанием траншеи ***от забора до обеда***». [Б. Левин. Блуждающие огни]

Идиомы *от винта* и *полный вперед* исходно являются нормативными приказами в летном деле и в сфере мореплавания соответственно:

(2) **а**. Сзади сидел старый шаман и прерывисто дышал, словно поднимался на высокую гору. ***От винта***! — крикнул летчик. Мотор чихнул, глотнув холодного воздуха, выпустил облачко голубого дыма. [Ю. Рытхэу. Самые красивые корабли] **б**. Пилот закричал: «***От винта***!» — желтого кругляка солнца видно уже не было. [В. Маканин. Гражданин убегающий]

(3) Капитан, медленно-медленно опуская пистолет, рявкнул: ***Полный вперед***! Держать прежний курс! [А. Бушков. Дверь в чужую осень (сборник)]

В приведенных примерах регламентированные приказы все равно остаются идиомами, основывающимися на переинтерпретации: буквально семантика словосочетания *от винта* указывает пространственное расположение относительно винта. Еще один семантический шаг от буквального значения: ‘побуждение отойти от начавшего работать винта самолета’. Только на третьем шаге возникает регламентированный приказ к началу движения самолета. Словосочетание *полный вперед* неидиоматичного прочтения не имеет, и его семантика изначально связана с приказом к началу движения корабля, парохода, катера и пр. плавсредств, оснащенных мотором. В переосмыслениях эти формы сохраняют связь с речевым актом приказа, хотя и передают совершенно другую семантику. Идиома *от винта* используется в (4а) значении ‘Отказ кому-либо в чем-либо, *осмысляемый как приказ уйти от опасно вращающихся лопастей винта, сопровождающий начало движения самолета*’.[2] [1], а идиома *полный вперед* представлена в примере (4б) в значении: ‘Разрешение делать что-либо, не ограниченное какими-либо условиями, *осмысляемое как приказ уполномоченного лица — например, капитана — к началу движения корабля, парохода, катера и пр. плавсредств, оснащенных мотором*’.

(4) **а**. Костя с трудом понимал Валерку. Деньги нужны. Денег нет. — Денег-то, Валер... — Ну, здрасте, приехали! — Валерка хлопнул себя руками по ляжкам. — Как бухать — есть, как земелю выручить — ***от винта***! Хреновый ты земеля! [С. Каледин. Стройбат] **б**. Алкаш? Тоже в диспансере, или вообще не жить! Старик? ***Полный вперед***, шуруй в свой дом престарелых и живи там! [А. Грачев. Ярый-3. Ордер на смерть (2000)]

4 Вложенная генерализованная цитация

Генерализованная цитация в небольшой группе идиом обрамляется другими компонентами фразеологизма, замещая валентности предикатов. Таковы идиомы *откладывать / оставлять на потом* (ср. идиому *откладывать в долгий ящик*, близкую по семантике, но без вложенной цитации), *по самое не балуйся, по самое не хочу, по самое не могу*, *по самое некуда*, *на отскребись*, *приказать долго жить*, *держ ат ься на чест ном слове, под чест ное слово*, *за Христ а ради, [и] поминай как звали, вот т ебе / т е[и] на, за спасибо, за [прост о] т ак, взят ь на слабо, за здорово ж ивёшь*, *сказат ь последнее «прост и» (кому-л.), на вы (с кем-л. / чем-л.), на т ы с кем-л. / чем-л.)*, *[– Не могу.] — [А ты] через не могу*, *[– Чест ное слово] — [Чест ное слово] врат ь гот ово.* К этой же группе идиом относятся и некоторые относительно новые сверхсловные образования (например, *давай без давай*), указывающие на актуальность вложенной цитации как фактора возникновения новых идиом.

В количественном отношении эта группа идиом очень мала: чуть больше двадцати идиом из восьми тысяч идиом, представленных в довольно полном словарном источнике [7], однако степень идиоматичности этих выражений интуитивно ощущается как весьма значительная. Все дело в образе, в котором заложено цитирование речевого акта, с обобщенным авторством. Так, во внутренней форме идиомы *откладывать / оставлять на потом* фиксирован речевой акт, в составе которого фигурирует наречие *потом*: *сделаю потом*; *потом разберемся; потом станет яснее* и под. Очевидно, что образ связан с перемещением того, что необходимо сделать (*откладывать, оставлять*), на некую поверхность (предлог *на*), осмысляемую как будущее, представителем которого является слово *потом*.

Группа идиом *по самое не балуйся, по самое не хочу, по самое не могу* связана с достижением предела, за которым находится нечто запретное (не будем уточнять). Во внутренней форме данных идиом процитированы прототипические реплики, представляющие соответствующие ситуации: *Не балуйся!; Не хочу!, Не могу!* Контексты употребления показывают, что носители языка хорошо ощущают, о чем идет речь:

[2] Здесь и далее используется нотация, принятая в [7]. В толковании компонент, передающий внутреннюю форму, выделен курсивом.

(5) **а**. Народу — пруд пруди, ведь пятница, и в ресторанах, барах нет свободных мест, и открывается их воспаленным взорам не то Аркадия, не то Валгалла, столько девушек вокруг — раздетых, полуоголенных, открывших грудь по самые «***не могу***» и ноги, — соответственно, по самые «***не балуйся***». [С. Самсонов. Кислородный предел] **б**. Но ведь и там, где форма обязательна встречаются девочки с блузкой, расстёгнутой почти до пупка. Кому очень хочется, тот любую форменную юбку отрежет под «самое ***не балуйся***». [Форум: Школьная форма. За и против] **в**. Была красивая девчонка, нравилась, а теперь вот она уезжает насовсем, и герою Российской армии ничего больше не светит. Выходит, его, героя, кинула какая-то сучка, да? Променяла на городскую жизнь? И герой догнал сучку, завалил и впёр ***по самое не хочу***. [Алексей Иванов. Псоглавцы] **г**. Здесь буквально все сочетается со всем. Из-под розовой кисейной юбчонки имеют право торчать, просвечивая ***по самое не могу***, толстые кривоватые ноги в вязаных черных рейтузах. <…> Ничто не слишком и в смысле цветовых сочетаний, а также размеров и форм бижутерии. [И. Васюченко. Хромые на склоне // «Ковчег»]

Разумеется, употребление идиом данной группы совсем не обязательно связано со сферой сексуального. Образ переосмысляется в различных направлениях. Так, идиома *по самое не балуйся* имеет, как минимум два, вполне невинных значения: ‘(Погрузиться в жидкость или густую вязкую среду или оказаться в ней) на значительную глубину, так, что дальнейшее движение (обычно человека или автомобиля) затруднено или невозможно *как бы из-за того, что уже достигнут запретный уровень*’ и ‘В максимально возможной степени, так, что некоторое действие или состояние при этой степени проявления может влечь неприятные последствия, *что осмысляется как достижение запретного уровня*’ [1], ср. примеры (6) и (7) соответственно:

(6) В городе Шарыпово неизвестный утопил экскаватор в болоте. Видео с места событий снял случайный прохожий, который не смог равнодушно пройти мимо. Автор видеоролика отмечает, что технику засосало «***по самое не балуйся*** *вообще*» [«Запад 24»]

(7) – Да, — произнес Плотник. — Конец, и Богу слава. И вовремя, вовремя, друзья мои. Мы в этом маленьком дельце уже увязли ***по самое «не балуйся»***. [А. Стругацкий. Дьявол среди людей]

Переосмысляются в невинном духе и внутренние формы идиом *по самое не хочу, по самое не могу*:

(8) **а**. — Банки и компании в середине 2000-х умудрились влезть во внешние долги ***по самое не хочу***. [Долговой карман: худой и еще худее // «Эксперт», 2013] **б**. И он, довольный ***по самое не могу***, быстрым шагом пошел на радиостанцию, докладывать наверх. [А. Бушков. Дверь в чужую осень]

Внутренняя форма идиомы *на отскребись* совершенно прозрачна: она основана на цитации распространенной обсценной реплики с семантикой отказа. Неудивительно, что актуальное значение этой идиомы связано с плохо сделанной работой: ‘сделать что-либо плохо и быстро, не заботясь о качестве результата работы, *как бы отказываясь от нее в неприличной форме*’, ср. (9):

(9) [Ivan Sugak, муж] Читай отзывы о авто… думаю лучше взять Ланос с нуля чем ВАЗ, мне лично сборки ВАЗов не нравятся вообще… Лучше подержанную иномарку (3-5 лет)… <…>
[Andrey Ponomarenko, муж] Согласен, к ланосу хоть применимо слово качество, ваз вообще сделан «***на отскребись***, и так сойдёт». [Форум: Посоветуйте. Хочу купить машину]

Внутренняя форма идиомы *приказать долго жить* апеллирует к ситуации благословления умирающим близких родственников: ‘Умереть, как правило, естественной смертью (от болезни, старости и т. п.), *что описывается как благословление — ритуализованный акт прощания в форме пожелания умирающего долгой и счастливой жизни близким*’ (несколько уточненный

вариант толкования идиомы из [1]). В переносных значениях речь идет о прекращении существования какой-либо организации, институции (10а) и прекращении функционирования технического устройства (10б):

(10) **а**. Банк, в котором нынешний советник президента по культуре хранил свои деньги, ***приказал долго жить***. [Московский комсомолец] **б**. Когда в девять тридцать утра я попытался завести мотор, оказалось, что батарея ***приказала долго жить***, и был полдень, когда наконец я покинул Паркингтон. [В. Набоков. Лолита]

Во внутренней форме идиом *держ ат ься на чест ном слове, под чест ное слово* фиксирована отсылка к речевому акту, передаваемому идиомой *чест ное слово!*, которая в современном русском языке используется в значениях: ‘Выражение искренности *в форме гарант ии собеседнику, чт о сказанное соот вет ст вует т ому, чт о дейст вит ельно думает по эт ому поводу говорящий*’ и ‘Выражение уверенности в сказанном *в форме клят вы, основанной на гарант ии соот вет ст вия сказанного т ому, чт о дейст вит ельно думает говорящий*’ [1]. Эта семантика, выступая как образ, мотивирует значения идиом *держ ат ься на чест ном слове, под чест ное слово*:

(11) **а**. Есть подозрение, что эпоксидная смола не затвердела. Цыпа бухнул слишком много растворителя. Короче, эта мраморная фига держится ***на чест ном слове***. Поэтому, когда начнется митинг, отойди в сторонку. [С. Довлатов. Чемодан] **б**. Так или иначе, но ***под честное слово*** дворянина Осоргина выпустили из камеры смертников на свидание с женой, обязав не говорить ей, что его ожидает. [Д. Лихачев. О русской интеллигенции] **в**. В Лос-Анджелесе Булату стало плохо, и ему потребовалась операция на сердце, а деньги на такую операцию — около шестидесяти пяти тысяч долларов — отсутствовали. Булат лег в клинику «***под честное слово***». Средства на операцию достал живший в Кельне Лева Копелев. [Б. Мессерер. Промельк Беллы. Булат Окуджава // «Октябрь»

Значение идиомы *под чест ное слово* тесно связано с речевым актом *чест ное слово!* В примерах (11б,в) семантику выражения *под чест ное слово* можно передать так: ‘сделать что-то кому-то или разрешить ему что-то под гарантию этого лица, *осмысляемую как клят ва, выраж енная репликой «чест ное слово», сделат ь чт о-т о в будущем как компенсацию или оправдание сделанного*’. Значение идиомы *держ ат ься на чест ном слове* соотносится с речевым актом *чест ное слово!* в меньшей степени: ‘(о какой-то части целого объекта) быть плохо присоединенным к основной части объекта, *как бы исключит ельно на обещании не от креплят ься, данном неизвест но кем и неизвест но кому*’.

Похожим образом устроена семантика идиом *за Христ а ради, [и] поминай как звали, вот т ебе / т е[и] на, за спасибо, за [прост о] т ак, за здорово ж ивёшь, сказат ь последнее «прост и» (кому-л.)* — внутренняя форма этих выражений отсылает к речевым актам *Христ а ради!, Как ж е его / ее звали?; На!; Спасибо!; Прост о т ак; Здорово ж ивешь!; Прост и!* В структуре идиом они образуют вложенную генерализованную цитацию.

План содержания идиом *на вы (с кем-л. / чем-л.), на т ы с кем-л. / чем-л.)* устроен несколько иначе в том смысле, что цитируются формы *вы* и *т ы* в функции обращения. Актуальное значение основывается в данном случае на том, что обращение с использованием местоимения *т ы* предполагает более тесное знакомство говорящих между собой, чем при использовании в этой функции местоимения *вы*. Соответственно, *быт ь с т ехникой на вы* значит нечто вроде ‘плохо разбираться в технических устройствах’, а *быт ь с т ехникой на т ы* — ‘хорошо разбираться в технических устройствах’.

Идиомы *[– Не могу.] — [А ты] через не могу*, *[– Чест ное слово] — [Чест ное слово] врат ь гот ово* образуют пары реплик, причем реплика-идиома иллокутивно вынуждается предшествующей репликой, состоящей из цитируемого речевого акта или включающей его:

(12) **а**. — ***Не могу*** я. Сил окончательно нет. — Нельзя ***не могу***, горбун. Я погибну, ты погибнешь, бойцы погибнут — Советам конец! Айда, горбун, ***через не могу***... [А. Алдан-Семенов. Красные и белые] **б**. — Слышь ты, а он весь в пупырьях, инда как жаба, — сказал Тимофей, проводя по вздутой груди убитого. — ***Честно слово***. Он зачем-то ответил вслух по-детски: — ***Честное слово врать готово***. [Н. Кононов. Магический бестиарий]

Как показывает анализ текстов современной прозы, идиома *[– Не могу.] — [А ты] через не могу* имеет недиалогический вариант, в котором конструкция с вложенной цитацией сохраняется:

(13) **а**. Утром встала рано, ***через не могу***, хотя смысла в этом подъеме не было. [Екатерина Завершнева. Высотка (2012)] **б**. Как же он себя на разворотах, в першпективах Петербурга чувствовал? Или заставлял себя — просыпаться, вставать, шагать — ***через*** «***не могу***»? [А. Битов. Русский устный и русский письменный // «Звезда»]

В примере (13б) помещение словосочетания *не могу* в кавычки подчеркивает его статус как генерализованной цитации.

В модели вложенной цитации цитата оказывается непрозрачной для каких-либо преобразований. В ряде случаев она меняет исходные грамматические и семантические характеристики, приобретая иной категориальный статус. Так, речевой акт запрета *не балуйся* в составе идиомы *по самое не балуйся* обнаруживает черты существительного. Речевой акт отказа *не могу* в идиоме *через не могу* также теряет иллокутивные свойства, становясь частью другого речевого акта. Речевой акт клятвы *чест ное слово!* персонифицируется и становится субъектом в идиоме *[Чест ное слово] врат ь гот ово*. Невложенная генерализованная цитация в идиомах в значительной степени сохраняет свойства исходной формы, хотя и там некоторые изменения происходят.

5 Заключение

Вложенная цитация относится к уникальным типам идиоматичности, более редким, чем даже онимизация. Она может сочетаться с переинтерпретацией и непрозрачностью — факторами идиоматичности, существенно более частотными и в некотором смысле более универсальными. Исследование периферийных типов идиоматичности, тем не менее, важно для как для полного описания фразеологический системы, так и для исследования ее потенциала развития.

Категория цитации, как показывает проведенное исследование, оказывается намного более сложным феноменом, чем это представлялось ранее. Б. С. Шварцкопф — лингвист, живо интересовавшийся сферой непознанного в речи — не случайно обратил на это внимание [8]. Цитация неоднородна, как минимум, с точки зрения авторства. Формы бытования цитации в языке требуют дальнейшего изучения.

References

[1] Academic dictionary of Russian phraseology (2020) [Akademicheskiy slovar' russkoy frazeologii], 3th ed. / Ed. by A.N. Baranov, D.O. Dobrovol'skij. Moscow: Lexrus. 892 p.

[2] Baranov A.N. (2010), Once again about the factors of idiomaticity: tautology and onymization [Yeshche raz o faktorakh idiomatichnosti: tavtologiya i onimizatsiya], Proceedings of the international conference Dialogue 2010 [Materialy mezhdunarodnoy konferentsii Dialog 2010]. Moscow, pp. 20-24.

[3] Baranov A.N., Dobrovol'skij D.O. (1996), Idiomaticity and idioms [Idiomatichnost' i idiomy], Topics in the study of language [Voprosy jazykoznanija], № 5, pp. 51-56.

[4] Baranov A.N., Dobrovol'skij D.O. (2013), The foundations of phraseology [Osnovy frazeologii], Moscow: Flinta-Nauka, 312 p.

[5] Zemskaya E.A. (1996), Newspeak clichés and citations in the language of post-Soviet society [Klishe novoyaza i tsitatsiya v yazyke postsovetskogo obshchestva], Topics in the study of language [Voprosy jazykoznanija], № 3, pp. 23-31.

[6] Lotman Yu.M. (1992), Text in another text [Tekst v tekste], Lotman Yu.M. Selected Articles [Lotman Yu.M. Izbrannyje statji]. Vol. 1. Tallinn, pp. 148-160.

[7] Dictionary-thesaurus of modern Russian idioms [Slovar'-tezaurus sovremennoy russkoy idiomatiki] (2007), Ed. by A.N. Baranov, D.O. Dobrovol'skij. Moscow: Avanta+.

[8] Schwarzkopf B.S. (1970) About some linguistic problems associated with citation [O nekotorykh lingvisticheskikh problemakh, svyazannykh s tsitatsiyey], Sign, Language, Culture. The Hague — Paris, pp. 658-673.

[9] Yandex games [Yandeks igry]. Access mode: https://gamepedia.su/3153/100-k-1-kogo-imeyut-v-vidu-govorya-on-vechno-zhivoj-yandeks-igry

Collectivity & distributivity

Boguslavsky I. M.
IITP (Kharkevich Institute), FRC CSC, Moscow, Russia
Universidad Politécnica de Madrid, Madrid, Spain
bogus@iitp.ru

Abstract

The opposition ‘separately’ / ‘together’ manifests itself in different parts of the linguistic system—in the interpretation of noun phrases (distributive vs. collective), in syntax (*200 roubles a piece, 100 kilometres per hour*), in the lexicon (*together, jointly, collectively, in aggregate, common, collective; apart, separate, apart, separately, each,* etc.) in word formation (*coauthor*). We discuss these meanings in the context of developing a computational model of semantic analysis that highlights the following issues: 1) What information about the meaning of distributive/collective NPs should be available to the parser for reasoning? 2) What valency slots are characteristic of the meanings ‘together’ / ‘separately’ and how are they instantiated in syntactic constructions and lexical items?

Keywords: distributivity, collectivity, distributive construction, semantic valencies
DOI: 10.28995/2075-7182-2025-23-45-56

Совместность и раздельность

Богуславский И. М.
Институт проблем передачи информации РАН
им. А. А. Харкевича, Москва, Россия
Universidad Politécnica de Madrid, Madrid, Spain
bogus@iitp.ru

Лёне Иомдину (1947–2024)

Аннотация

Противопоставление ‘по отдельности’ / ‘вместе’ проявляется в разных участках языковой системы — в интерпретации именных групп (дистрибутивная vs. коллективная), в лексике (*вместе, совместно, в совокупности, общий, коллективный; врозь, раздельно, отдельно, по отдельности, каждый* и др.), в синтаксисе (*200 рублей штука, 100 километров в час*), в словообразовании (*соавтор*). Мы рассматриваем эти значения в контексте разработки компьютерной модели семантического анализа SemETAP, что выводит на первый план следующие вопросы: 1) какая информация о значении дистрибутивных и коллективных ИГ должна быть доступна анализатору для задачи извлечения следствий? 2) какие валентности имеют значения совместности и раздельности и как эти валентности реализуются в синтаксических конструкциях и в лексических единицах, выражающих эти значения?

Ключевые слова: совместность, раздельность, дистрибутивность, коллективность, дистрибутивная конструкция, семантические валентности

1 Постановка задачи

Эта статья предназначена для конференции серии Dialogue, одна из важных задач которой состоит в том, чтобы строить и укреплять мосты между теоретической лингвистикой и автоматической обработкой языка (NLP). Представляется, что со стороны лингвистики это удобно делать, излагая лингвистические факты и находки таким образом, чтобы читателю было понятно, какое место они могут занять в компьютерной системе, моделирующей те или иные языковые способности человека. Нашим предметом в этой статье будут разные типы языковых выражений, в

которых присутствует значение совместности или раздельности. Эти значения проявляются в разных участках языковой системы — в интерпретации именных групп (дистрибутивная vs. коллективная), в лексике (*вместе, совместно, в совокупности, общий, коллективный; врозь, раздельно, отдельно, по отдельности, каждый* и др.), в синтаксисе (*200 рублей штука, 100 километров в час*), в словообразовании (*соавтор, совладелец*). Мы постараемся сделать акцент на таких свойствах этих выражений, которые могут быть востребованы в компьютерных системах. В качестве такой системы мы возьмем семантический анализатор SemETAP.

В статье должны быть получены ответы на следующие вопросы:

1) Понятие дистрибутивности фигурирует в литературе при описании различных явлений. Оно участвует в оппозиции дистрибутивность -коллективность, которая интенсивно изучается логической семантикой, оно формирует специфические дистрибутивные (=распределительные) конструкции (*100 рублей килограмм, каждому по конфете*) и изучается синтаксисом, оно образует дистрибутивный (=распределительный) способ действия (*перебить посуду, повыходить замуж*) и описывается в морфологии. В чем различие между этими видами дистрибутивности?
2) Какие противопоставления, характерные для дистрибутивных и коллективных ИГ, существенны для задачи извлечения выводов из текста и поэтому должны быть отражены в семантической структуре SemETAP?
3) Является ли класс коллективных ИГ однородным с точки зрения их идентификации в тексте?
4) Какие синтаксические позиции занимают в предложении актанты предиката дистрибутивности в разных типах дистрибутивной конструкции?
5) Как заполняются валентности наречий совместности/раздельности?
6) Какое место в классификации наречий по ориентации занимают наречия совместности/раздельности?

Статья строится по следующему плану. Сначала мы коротко расскажем о семантическом анализаторе SemETAP (раздел 2), а затем сможем перейти к языковому материалу. В разделе 3 мы будем говорить о дистрибутивной и коллективной интерпретации именных групп (ИГ), в разделе 4 — о дистрибутивных синтаксических конструкциях, а в разделе 5 — о некоторых лексических единицах со значением совместности и раздельности. В Заключении будут подведены некоторые итоги.

2 Семантический анализатор SemETAP

SemETAP представляет собой семантический модуль лингвистического процессора ETAP-4, предназначенного для анализа и синтеза текста на разных языках (см. Boguslavsky, Dikonov et al. 2016, 2020; Boguslavsky, Frolova et al. 2018; Богуславский 2021 и ссылки в этих работах). Задача семантического модуля состоит в том, чтобы, получив от предыдущего — синтаксического — модуля синтаксическую структуру предложения, построить его семантическую структуру. Основное положение, положенное в основу концепции SemETAP, состоит в том, что главный критерий понимания текста — это те выводы, которые мы сможем из него извлечь. Чем больше этих выводов и чем они качественнее, тем лучше мы поняли текст. Поэтому семантическое представление, получаемое SemETAP-ом, содержит два слоя — базовую семантическую структуру (БСемС), которая отражает непосредственное содержание предложения, и расширенную семантическую структуру (РСемС), в которой к базовой структуре добавляются все выводы, которые система смогла извлечь. Для целей настоящей работы существенны два обстоятельства:

1) Одна из главных задач БСемС — эксплицировать семантические актанты всех предикатов предложения независимо от того, как они выражены. В частности, должны быть обнаружены и зафиксированы такие актанты, которые не имеют непосредственной синтаксической связи со своим предикатом. Такая ситуация возникает, как известно, в первую очередь, для слов с пассивными и разрывными валентностями — прилагательных, наречий, частиц и союзов (Boguslavsky 2003). В полной мере это относится к наречиям со значением совместности и раздельности, а также в известной степени к дистрибутивной синтаксической конструкции, которая кодирует предикат дистрибутивности, имеющий две валентности.

2) В РСемС, как сказано выше, входят разнообразные следствия, которые можно извлечь из базовой структуры. Дистрибутивные и коллективные именные группы различаются в первую очередь именно тем, допускает ли предложение «вывод по дистрибутивности».

Отсюда видно, что свойства выражений со значением совместности и раздельности, о которых мы будем говорить ниже, имеют непосредственную значимость для семантического анализатора и должны быть учтены при его разработке.

3 Дистрибутивная и коллективная интерпретация именных групп

В естественном языке нередко описывается ситуация, когда предикат применяется не к единичному объекту, а к группе объектов, принадлежащих некоторой совокупности. Здесь возникает важное противопоставление между дистрибутивностью и коллективностью. Конкретнее, пусть в некотором предложении предикат P имеет в качестве аргумента объект X, выраженный множественной именной группой, например, существительным во множественном числе (*мальчики*), сочинительной или комитативной конструкцией (*Петя и Коля, Петя с Колей*) или существительным, обозначающим группу (*класс 5«Б»*). Если из данного предложения можно вывести, что P справедливо относительно каждого элемента множественного объекта X, то принято говорить, что X имеет дистрибутивную интерпретацию относительно P. Например, из предложения *Мальчики бегут* можно вывести, что бежит каждый отдельный мальчик, и поэтому ИГ *мальчики* дистрибутивна относительно глагола *бегут*. Такой «вывод по дистрибутивности» представляет ценность, и система логического вывода должна уметь его делать. Однако он возможен далеко не всегда. Дистрибутивной интерпретации ИГ противопоставлена *недистрибутивная*, которую часто называют *коллективной*. В этом случае нельзя сказать, что действие выполняют отдельные элементы группы, а только группа в целом. Таково, например, предложение *Мальчики собрались во дворе*. Бывают и предложения, допускающие обе интерпретации, например *Мальчики решают задачу*. В этом предложении непонятно, идет ли речь о нескольких актах решения, когда каждый участник решает свою задачу, например, на контрольной (дистрибутивная интерпретация), или же об одном акте решения задачи, в котором участвует несколько человек (коллективная интерпретация).

Недистрибутивные ситуации бывают двух типов. Первый тип образуют ингерентно-множественные предикаты, которые обозначают ситуацию, с обязательностью предполагающую множественный субъект. В таком случае попытка заменить множественную группу на группу, обозначающую единичный объект, приводит к аномальному предложению.

(1а) *После завтрака Петя, Коля и Маша собрались во дворе.*

(1б) **Петя собрался во дворе.*

Второй тип недистрибутивной ситуации реализуется тогда, когда предикат обозначает ситуацию, которая в принципе может иметь и единичный субъект, но в данном предложении субъектом является группа в целом. В таком случае предложение с единичным субъектом окажется грамматически правильным, но его значение не будет следовать из значения исходного предложения. Например, из (2а) не следует (2б), а только (2в):

(2а) *Петя и Коля весят 150 кг.*

(2б) *Петя весит 150 кг.*

(2в) *Петя весит 150 кг вместе с Колей.*

Аналогично, (3б) категорически не следует из (3а):

(3а) *Ильф и Петров написали роман «Двенадцать стульев».*

(3б) *Ильф написал роман «Двенадцать стульев».*

Заметим, что если глагол в (3а) стоит в несовершенном виде, то вывод, касающийся единичного объекта, становится несколько более приемлемым:

(4а) *В этот период Ильф и Петров писали роман «Двенадцать стульев».*

(4б) *В этот период Ильф писал роман «Двенадцать стульев».*

В большинстве подобных случаев результирующее предложение хоть и не является явно ложным, но все же семантически не вполне корректно. Так, если в ситуации, когда Петя и Коля поднимают рояль по лестнице, на вопрос (5а) ответить предложением (5б), то это, возможно, и допустимо для некоторых говорящих, но несколько некорректно, поскольку предпочтительная интерпретация предложения (5б) состоит в том, что Петя поднимает рояль в одиночку. Корректнее было бы сказать что-то вроде (5в):

(5а) *Что сейчас делает Петя?*

(5б) *Петя поднимает рояль по лестнице.*

(5в) *Петя вместе с другом поднимает рояль по лестнице* (участвует в переноске рояля).

Аналогичную некорректность можно было встретить в некоторых СМИ после недавних президентских выборов в Грузии. Например, в интернет-издании https://www.newkaliningrad.ru/ от 14.12.2024 помещен заголовок (6):

(6) *Президентом Грузии стал чемпион России по футболу Михаил Кавелашвили.*

Некорректность (6), очевидным образом, обусловлена тем, что в футболе побеждают (и становятся чемпионами) не отдельные игроки, а команды в целом. Более аккуратной формулировкой, которая в том или ином варианте использовалась в большинстве изданий, является (7):

(7) *Чемпион России сезона-1995 в составе владикавказской «Алании» Михаил Кавелашвили стал президентом Грузии* (https://www.championat.com/).

Как видно из приведенных примеров, можно выделить три группы предикатов: (а) предикаты, которые порождают только дистрибутивную интерпретацию своего субъекта; это действия, которые не может выполнять группа без того, чтобы их выполнял и каждый индивид: *бежать, идти, улыбаться, дышать, пить* и др., (б) предикаты, которые совместимы только с недистрибутивной интерпретацией: *собираться (во дворе), разъехаться (в разные стороны), различаться (по величине), образовывать (дружную команду)* и др., и (в) предикаты, допускающие оба типа интерпретации: *решать (задачу), нести (рояль), открывать (окно)* и др.

Обращаясь к вопросу о том, какова значимость рассмотренных данных для семантического анализатора SemETAP, отметим следующее. Дистрибутивные ИГ допускают «вывод по дистрибутивности», а недистрибутивные его не допускают. Однако чтобы воспользоваться этим свойством, анализатор должен уметь автоматически отделять первые от вторых, но сделать это можно далеко не всегда. Легко идентифицировать в тексте ингерентно-множественные предикаты, поскольку их можно предварительно задать в словаре. Что касается прочих ситуаций недистрибутивности, то для их идентификации необходимо привлечение экстралингвистических сведений, как, например, в случаях (2а)–(2б) или (3а)–(3б).

Изучению явления дистрибутивности посвящена огромная литература, обзор которой не может быть выполнен в короткой статье. Отметим лишь две группы работ, затрагивающих эту тематику.

Работы, выполняемые в рамках формальной семантики (Verkuyl 1994, Champollion 2019, Zimmermann 2002, Glass 2018, Glass 2020, de Vries 2017), обсуждают в первую очередь такие вопросы, как операционное определение дистрибутивности, различия между атомарной и неатомарной

дистрибутивностью, различия между коллективной интерпретацией и выводом на основе семантических ролей, соотношение между дистрибутивностью и коллективностью и между дистрибутивностью и кумулятивностью, а также то, как возможность дистрибутивной или коллективной интерпретации связана с лексическим значением глагола.

В отечественном языкознании проблематика дистрибутивности обсуждается в другом ключе. Главным образом, изучаются свойства дистрибутивных конструкций (*Мальчикам дали по яблоку, Яблоки стоят 100 рублей килограмм, Выдавали сто грамм на человека в день*), их связь с диагностикой неаккузативности, а также маркеры дистрибутивной конструкции — предлог по, местоимения *каждый, какой-то, тот или иной* и др. (Иомдин 1991, Кузнецова 2010, Падучева 2015, Борик 1995, Дудина 2002).

Анализ перечисленных работ показывает, что описания дистрибутивности в рамках этих двух традиций имеют между собой мало общего. Дистрибутивность в смысле формальной семантики — это отношение между предикатом и связанной с ним множественной сущностью. В предложении *Мальчики бегут* ИГ *мальчики* дистрибутивна относительно глагола *бегут*, поскольку из этого предложения можно сделать вывод о том, что бежит каждый мальчик. О дистрибутивности / недистрибутивности в этом смысле можно говорить всегда, когда в предложении есть множественная ИГ. Дистрибутивная конструкция в понимании российской школы — явление более специфическое. На нем мы остановимся в следующем разделе.

4 Дистрибутивная конструкция

В традиции российской лингвистической школы дистрибутивное значение — это значение ‘для каждого X-а свой Y’, то есть отношение между двумя множественными сущностями[1]. Будем называть их аргументами дистрибутивного значения. Типовой пример — (8):

(8) *Каждый студент знает по два языка.*

Один из этих аргументов (назовем его X) имеет значение всеобщности (‘каждый из’) и играет роль «распределяющего множества» (в терминологии Падучева 2015). Второй (назовем его Y) представляет собой множество «распределяемых» сущностей, которые ставятся в соответствие элементам первого множества. В примере (8) распределяющее множество X — это студенты. Каждому из студентов соответствует два языка, то есть множество Y представляет собой множество пар языков. Третий аргумент (P) — это предикат, с которым связаны X и Y. Значение дистрибутивной конструкции удобно представить в виде логической формулы:

(9) ∀x∈{студент} ∃y∈{2 языка} [x знает y] (‘для каждого элемента множества студентов существует элемент множества пар языков, такой, что этот студент знает эти языки’).

Весьма детальное описание дистрибутивной конструкции с предлогом *по* содержится в основательной работе Кузнецова 2010. В ней рассматриваются разнообразные ограничения на ее использование и показывается, что они естественным образом выводятся из ее значения. В задачу этой работы, однако, не входил один важный аспект: исследование того, как выражаются в предложении аргументы X и Y, между которыми устанавливается дистрибутивное отношение. Без прояснения этого вопроса, очевидным образом, нельзя будет построить для каждого предложения с дистрибутивной конструкцией формулу типа (9). Ниже мы наметим основные типы синтаксической реализации дистрибутивной конструкции.

Пример (8) демонстрирует самый простой случай: аргументы X и Y соподчинены (прямо или косвенно) одному и тому же глаголу P, причем ИГ, обозначающая распределяющее множество X, содержит местоимение *каждый* или другое местоимение со значением всеобщности, а ИГ, обозначающая Y, вводится предлогом *по*. Однако столь простая маркировка имеется далеко не всегда.

Нередко ни X, ни Y не имеют эксплицитной маркировки, или же ее имеет только один из этих аргументов.

[1] Мы оставляем здесь в стороне дистрибутивный способ действия глагола (*старики поумирали*).

(10) *Науки имеют свои названия* (‘каждая наука [X] имеет свое название [Y]’).

(11) *У игроков в начале игры имеется по одинаковому набору фигур, условно называемых «белыми» и «чёрными»* (‘каждый игрок [X] имеет набор фигур [Y], и они одинаковые’).

Может быть и так, что маркер дистрибутивности в предложении имеется, но он синтаксически не присоединен к маркируемому аргументу. Например, в предложениях (12а)–(12б) местоимение *каждый* не входит в ИГ, обозначающую X, а является копредикативом, связанным с глаголом. Правильная интерпретация дистрибутивной конструкции в этих предложениях обеспечивается в два шага. Прежде всего, падежная форма местоимения *каждый* показывает, что в (12а) и (12б) разные копредикативные конструкции. В соответствии с синтаксической моделью, реализованной в системе SemETAP и в синтаксически размеченном корпусе SynTagRus, в (12а) представлена субъектно-копредикативная, а в (12б) — объектно-копредикативная конструкция. Это означает, что местоимение *каждый* семантически связано в первом случае с субъектом предложения, а во втором — с его объектом. На этом основании можно заключить, что в (12а) распределяющим множеством дистрибутивной конструкции является субъект *девочки*, а в (12б) — косвенный объект *мальчикам*. В английском языке формальная маркировка субъектно- и объектно-копредикативных конструкций отсутствует, и поэтому предложение (13) неоднозначно и может соответствовать как (12а), так и (12б).

(12а) *Девочки рассказали мальчикам по две истории каждая.*

(12б) *Девочки рассказали мальчикам по две истории каждому.*

(13) *The girls told the boys two stories each (Champollion 2019).*

Еще одна синтаксическая реализация дистрибутивного значения — это так называемая дистрибутивная (=распределительная) конструкция. Она чаще всего реализуется сочетанием количественной группы в роли Y с существительным в именительном падеже или группой с предлогами *в*, *на* или *за*, указывающей на единицу измерения в роли X. В качестве аргумента P выступает обычно параметр с валентностями на объект и значение параметра:

(14) *Колбаса стоит 1000 рублей (за) килограмм* (‘каждый килограмм колбасы [X] (стоит) 1000 рублей [Y]’).

Несколько более сложный случай — (15):

(15) *Снаряд летит со скоростью 1000 метров в секунду.*

Здесь перед нами тоже распределительная конструкция с единицей измерения: каждой секунде соответствует 1000 метров полета. Сложность состоит в том, как соотносятся элементы конструкции с аргументной структурой параметра *скорость*. Сравним ситуации, описываемые примерами (14) и (15). Цена, выступающая в качестве параметра в (14), стандартно измеряется в денежных единицах (ср. *Цена билета — 70 рублей*). В (15) параметром является скорость, измеряемая в составных единицах, например, в метрах в секунду. Представим оба предложения в виде «характеризуемый объект + параметр + значение параметра»:

(14а) *Килограмм колбасы стоит 1000 рублей.*

(15а) *Полет снаряда имеет скорость 1000 метров в секунду* (но не: **Секунда полета имеет скорость 1000 метров*).

Отсюда видно, что дистрибутивные конструкции 1000 рублей за килограмм и *1000 метров в секунду* по-разному соотносятся с аргументной структурой соответствующих параметров. Если в первом случае элементы дистрибутивной конструкции распределены по разным аргументам

параметра цены, то во втором — они оба соответствуют составной единице измерения, служащей одним из аргументов параметра скорости.

Следующая, весьма интересная разновидность дистрибутивной конструкции — такая, в которой аргумент X реализован посредством не именной, а глагольной группы. Сравним предложения (16а) и (16б):

(16а) *Я ждал его час.*

(16б) *Я ждал его по часу.*

В (16а) описывается единичное событие ожидания, длившееся час. В (16б) глагол явно многократный, и именно серия ожиданий образует то множество X, относительно которого распределяются отрезки времени Y: ‘каждое ожидание длилось по часу’.

Аналогичное явление можно видеть в предложении (17):

(17) *В пятницу к парому выстраивается гигантская многочасовая очередь по 40-50 машин.*

Здесь также описывается серия однотипных событий — выстраивание многочасовых очередей из машин по пятницам. Каждому такому событию соответствует длина очереди в 40-50 машин. Очевидно, что употребление предлога *по* обусловлено именно множественностью этих событий. Если бы речь шла о единичном событии, предлог *по* был бы невозможен:

(18) **В пятницу к парому выстроилась гигантская многочасовая очередь по 40 машин.*

Заметим также, что несмотря на то, что в (17) описывается серия событий и тем самым упоминается несколько многочасовых очередей, ИГ стоит в единственном числе. Сфера действия этого единственного числа не вся серия событий, а каждое событие по отдельности. Это видно из того, что при замене его на множественное число, значение предложения несколько изменяется. Предложение (19) может обозначать не только ту же ситуацию, что и (17) (каждый раз одна гигантская очередь), но и ситуацию, когда одновременно выстраивалось несколько очередей. Этот второй вариант для предложения с единственным числом, по-видимому, невозможен — см. (20).

(19) *По пятницам в городе выстраивались гигантские многочасовые очереди.*

(20) *По пятницам в городе выстраивалась гигантская многочасовая очередь.*

Приведем еще один контрастный пример того, как дистрибутивное по вносит представление о множественности ситуаций.

(21а) *Они жили в комнате ввосьмером* (восемь человек жили в комнате совместно).

(21б) *Они жили по восемь человек в комнате* (имеется более одной ситуации, в каждой из которых люди жили в комнате ввосьмером).

В этом разделе мы ввели в рассмотрение предикат дистрибутивности (‘для каждого X-а свой Y’) и рассмотрели ряд синтаксических конструкций, в которых реализуется это значение. Особое значение уделено тому, какие позиции занимают в предложении элементы X и Y, и каким образом они маркируются. Эта информация существенна для идентификации дистрибутивной конструкции в предложении и построении ее семантической структуры.

5 Наречия со значением совместности и раздельности

В русском языке имеется целая серия слов со значением совместности и раздельности. Это, например, наречия *вместе*, *совместно*, *коллективно*, *сообща*, *хором*, *бок о бок*, *плечом к плечу*, *отдельно*, *раздельно*, *по отдельности*, *врозь*; прилагательные *общий*, *совместный*, *коллективный*; *отдельный*, *раздельный*, *индивидуальный*, *одиночный*, *личный* и др. Мы не имеем

возможности подробно разбирать их все, тем более, что некоторые их особенности мы описывали в работе (Богуславский, Иомдин 2012). Мы остановимся в первую очередь на наречиях *вместе* и *отдельно*.

Концепт совместности подробно рассмотрен в статье (Зализняк-Шмелев 2021: 32-30 (1999)). В ней подробно проанализировано значение этого концепта, некоторые типы совместных действий и способы их представления в русском языке. С близких нам позиций интересующие нас слова описывались в словарях, выпущенных под руководством Ю.Д. Апресяна — в «Новом объяснительном словаре синонимов» (2004) (статья «Совместный, общий, коллективный») и в Активном словаре (т.2, 2014) (статья «Вместе»). Однако вопрос о том, как выражаются в текстах актанты «что совместно с чем?», «что отдельно от чего?» привлекал мало внимания. Это позволяет нам сосредоточиться на вопросе об актантной структуре этих наречий и способах заполнения их валентностей.

Идея совместности/раздельности предполагает наличие некоторого множества объектов Q и ситуации P, в которой они совместно или раздельно участвуют. Ситуация P реализуется, как правило, в виде главного предиката клаузы, а множество Q может быть представлено в предложении либо посредством множественной ИГ, либо в виде двух соподчиненных ИГ (Q1, Q2):

(20а) *Дети* [Q] *пришли вместе.*

(20б) *Мальчики и девочки* [Q] *учатся отдельно.*

(21а) *Коля* [Q1] *пришел вместе с Машей* [Q2].

(21б) *Девочки* [Q1] *учатся отдельно от мальчиков* [Q2].

Таким образом, предикаты совместности/раздельности могут иметь две диатезы — двухместную (с актантами P и Q) и трехместную (с актантами P, Q1 и Q2). Есть основания считать, что для предикатов совместности исходной диатезой является двухместная, а трехместная получается в результате расщепления актанта Q на Q1 и Q2. Для предикатов раздельности дело обстоит обратным образом: исходная диатеза трехместна, а двухместная представляет собой объединение актантов Q1 и Q2 (Богуславский, Иомдин 2012).

Тот факт, что значения интересующих нас наречий включают упоминание не только ситуации P, но и объектов Q, которые в ней участвуют, помещает эти наречия в класс так называемых ориентированных наречий (oriented adverbs), которым посвящено большое количество работ. Упомянем некоторые из них: Jackendoff 1972, Ernst 2002, Ernst 2003, Geuder 2002, Garcia Nuñez 2002, Taverniers & Rawoens 2010, Kelepouris 2012. Среди ориентированных наречий выделяются два больших класса — наречия, ориентированные на говорящего (speaker-oriented) и ориентированные на субъект (subject-oriented). Имеется в виду, что предикация, вводимая наречием, содержит отсылку к говорящему, либо к субъекту клаузы, в которую входит данное наречие. Например, в предложении *Петя, к сожалению, опоздал* наречие *к сожалению* приписывает чувство сожаления говорящему, а в предложении *Маша благоразумно осталась дома* наречие благоразумно описывает свойство Маши. Если наречие содержит отсылку к актанту, участвующему в клаузе, к которой относится данное наречие, то такой актант *обычно* является субъектом этой клаузы. Так, например, в предложении *Родители благоразумно оставили Машу* дома благоразумие приписывается родителям, а никак не Маше. Поэтому этот класс наречий был обозначен как subject-oriented.

Между тем, в языке существуют и такие наречия, которые ориентированы не только на субъект, но и на прямое дополнение и другие позиции в предложении. Например, такие адвербиалы, как *группой*, *классом*, *толпой*, *стаей* и т.п. легко могут характеризовать ИГ, находящиеся в позиции как субъекта, так и прямого дополнения. Можно сказать *Они ушли с урока всем классом* (они — подлежащее), *Сюда нас часто водили всем классом после уроков истории* (нас — прямое дополнение), но нельзя **Детям раздали яблоки всем классом* (детям — косвенное дополнение).

Наречия совместности/раздельности обладают еще большей свободой. Они могут иметь актант, являющийся подлежащим, прямым или косвенным дополнением и даже обстоятельством. Приведем несколько примеров:

(22) *Петр тренировал сына вместе с женой* (*Петр* — подлежащее; ‘Петр с женой тренировали сына’)

(23) *Петр тренировал сына вместе с дочерью* (*сын* — прямое дополнение; ‘Петр тренировал сына и дочь’)

(24) *Им вместе со стариками разрешили вернуться* (*им* — косвенное дополнение; ‘разрешили им и старикам’).

(25) *Выбор препарата осуществляется для каждой особи* [Q, обстоятельство] *отдельно.*

Как говорилось выше, валентность Q предикатов совместности / раздельности могут заполнять множественные ИГ типа *Петя и Коля* или *мальчики*. При этом в случае предикатов совместности речь идет о единой валентности (Q), а в случае предикатов раздельности — о слиянии в одной ИГ двух валентностей (Q1 и Q2). Такая ИГ может быть выражена существительным во множественном числе, количественной или кванторной группой, существительным со значением совокупности, сочинительной или комитативной конструкцией. Здесь наблюдаются различные интересные эффекты; одни из них свойственны большой группе рассматриваемых предикатов, а другие характеризуют конкретные лексические классы или даже отдельные лексические единицы.

Сравним предложения (26) и (27):

(26) *Напитки оплачиваются отдельно.*

(27) *Напитки оплачиваются раздельно <по отдельности>.*

Они, очевидным образом, понимаются по-разному. В случае (26) предпочтительное понимание состоит в том, что, помимо напитков, надо платить за что-то еще, например, за еду или за гостиничный номер. Утверждается, что оплата этих подразумеваемых услуг должна быть осуществлена отдельно от оплаты напитков. В противоположность этому, предложение (27) отсылки к контексту не содержит. Здесь речь идет о том, что нельзя объединять вместе оплату всех напитков. Каждый напиток должен быть оплачен индивидуально. Таким образом, роль одной и той же ИГ *напитки* в этих предложениях разная. В (26) она заполняет только валентность Q1 наречия *отдельно*, предоставляя слушающему самостоятельно найти материал для заполнения валентности Q2 («отдельно от чего?») в контексте. В предложении (27) эта ИГ совмещает в себе обе валентности Q1 и Q2.

Существенно, что наречие *отдельно* допускает употребления со слитным выражением валентностей Q1 и Q2 и в этих употреблениях синонимично наречиям *по отдельности* и *раздельно* (ср. примеры (28)–(29)), а наречия *раздельно* и *по отдельности* употреблений типа (26) не допускают: предложение (27) не может пониматься в значении (26).

(28) *Смывать покрытие следует отдельно с каждой плоскости предмета* (= ...с каждой плоскости предмета по отдельности)

(29) *Застройка левой части острова не была регулярной: на плане мы видим кривые улицы, тупики и стоящие отдельно, не «в линию», дома* (= дома, стоящие отдельно друг от друга).

Тем самым, можно утверждать, что в случае *отдельно* валентность Q2 является синтаксически факультативной, в то время как в случае *раздельно* и *по отдельности* она синтаксически обязательна и должна быть насыщена в пределах предложения, пусть даже это насыщение происходит синкретично, в пределах множественной ИГ. Отметим важный момент: синтаксическую обязательность валентности не следует понимать узко — как обязательность выражения данной валентности исключительно группой, непосредственно синтаксически зависящей от валентного слова. Следует принимать в расчет и другие регулярные способы заполнения валентности, в частности, — слитное выражение двух валентностей в составе одной ИГ.

Замечание: Следует подчеркнуть методологически важное обстоятельство. Наличие непосредственной синтаксической связи с валентным словом не является для актанта обязательным требованием. Для того чтобы признать некоторое выражение актантом данного слова, достаточно, чтобы это выражение соответствовало обязательному участнику ситуации и систематическим образом выражалось в предложении вместе с данным словом, причем не обязательно в качестве его синтаксического зависимого [Mel'čuk 2015: 18]. Например, таким систематическим выражением может служить присоединение актанта через лексико-функциональный глагол группы Oper-Func-Labor, например, *Вратарь составляет половину команды*.

Характерно, что при всей обязательности наличия в предложении актанта Q2, он не может быть выражен каноническим способом в виде самостоятельной ИГ; сочетания **по отдельности* от и **раздельно с* невозможны.

К предикатам совместности / раздельности с синтаксически обязательной валентностью Q2 относятся такие слова, как *вместе*, *совместно*, *врозь*, *порознь*, *раздельно*, а за пределами наречий — глаголы *совпадать*, *различать*, *различаться*, *сравнивать*, *пожениться*, *переглянуться*, прилагательные *различный*, *равный*, *тождественный*, *одинаковый*, *синонимичный*, *противоположный*, и др.

Выше мы видели, что что валентность Q2 не является синтаксически обязательной для наречия *отдельно*. Любопытно, что для наречия *вместе* это не так. Если предложения типа (30) вполне правильны, то (31) вряд ли допустимы:

(30) *Он хочет, чтобы она жила отдельно* (от него, или еще от кого-нибудь)

(31) *??Он не любит писать работы вместе* (в значении: *писать работы в соавторстве, совместные работы*)

На первый взгляд, утверждению об обязательности валентности Q2 наречия вместе противоречат предложения (32) и (33):

(32) *Хочешь, пойдем туда вместе?*

(33) *Он предложил мне написать статью вместе.*

Между тем, противоречия здесь нет. В обоих случаях мы имеем дело с синкретичным выражением валентностей Q1 и Q2 в составе множественной именной группы. Но только эта именная группа выражена в этих предложениях не словами, а анафорическим нулем. В первом случае опущен субъект личного глагола во множественном числе (*пойдем* — 'ты и я'). Во втором случае множественным субъектом ('он и я') обладает инфинитив, и источником этой множественности является глагол *предлагать*. Субъект инфинитива при глаголе *предлагать* допускает три интерпретации. Им может быть: а) субъект самого глагола *предлагать* (*Он предложил маме сходить в магазин* — предложил, что он сходит), б) адресат *предлагать* (*Я предлагаю вам все мне рассказать* — вы все расскажете) и группа субъект+адресат (*Я предлагаю сходить в кино* — предлагаю, чтобы мы сходили). В предложении (33), очевидным образом, реализована третья из этих возможностей: 'он предложил мне, чтобы он и я написали статью вместе'.

Наконец, стоит упомянуть еще одну очень специфическую возможность заполнения валентностей Q1 и Q2 наречия *отдельно*. Она реализуется в конструкциях с параллельно организованными частями и состоит в том, что элементы одной из частей фактически реализуют семантические валентности предиката, содержащегося в другой части. Подробно это явление мы разбирали в (Богуславский, Иомдин 2012), а здесь ограничимся только парой примеров:

(34) *Я люблю, чтобы котлеты были отдельно, а мухи отдельно* ('котлеты отдельно от мух')

(35) *Старая болезнь русской армии — офицеры отдельно, рядовые отдельно, и они играют друг против друга, — цветет пышным цветом* [Александр Будберг Военная тайна (2003)] (‘... офицеры отдельно от рядовых’)

Интересно, что такие конструкции лексически очень ограничены. Практически никакие симметричные предикаты, кроме *отдельно*, в этих конструкциях невозможны: немыслимо сказать что-либо вроде **Сумма квадратов катетов равна*, и *квадрат гипотенузы равен*, **Отец похож, а сын похож*, и даже близкие синонимы и антонимы отдельно тут неуместны: **Отец живет вместе, и мать вместе*, **Я завтракаю врозь, а дочь врозь*.

6 Заключение

Концепт совместности / раздельности выражается в языке многими разными способами. Он участвует в интерпретации именных групп при предикатах разных типов, порождая их дистрибутивное/коллективное осмысление, входит в состав дистрибутивной синтаксической конструкции, а также составляет ядро большого круга адвербиальных и адъективных выражений. Подводя итоги, суммируем ответы на вопросы, поставленные в разделе 1.

1. Различия между разными типами дистрибутивности. Дистрибутивность, противопоставленная коллективной интерпретации ИГ представляет собой отношение между предикатом и его множественным аргументом, означающее, что предикат применим к каждому элементу множественного аргумента. Близкое, хотя и не тождественное, значение выражается в глаголах дистрибутивного способа действия (Все подруги повыходили замуж). Значение, реализуемое в дистрибутивной конструкции типа *Студенты знают по два языка* или *Картофель стоит 100 рублей килограмм* есть отношение между двумя множественными сущностями, участвующими в одной ситуации: каждому студенту соответствует некоторая пара языков, каждому килограмму картофеля соответствует цена 100 рублей.
2. Релевантна ли оппозиция дистрибутивных и коллективных ИГ для задачи извлечения выводов из текста? Да, дистрибутивные ИГ допускают вывод «от множества к элементу», а коллективные его не допускают.
3. Является ли класс коллективных ИГ однородным с точки зрения их идентификации в тексте? Нет, предикаты с ингерентной множественностью (типа *собраться (во дворе)*) могут быть заданы словарно, и поэтому связанные с ними множественные ИГ однозначно интерпретируются как коллективные. Если же предикат может применяться как к группам, так и к отдельным индивидам (например, *Ильф и Петров написали роман «Двенадцать стульев»*), то для квалификации связанной с ним множественной ИГ как коллективной может потребоваться экстралингвистическая информация.
4. Какие синтаксические позиции занимают в предложении актанты предиката дистрибутивности в разных типах дистрибутивной конструкции? Возможные синтаксические реализации актантов этого предиката зависят от того, в какой конструкции он участвует. Разные их типы проиллюстрированы в разделе 4.
5. Как заполняются валентности наречий совместности/раздельности? Эти наречия обнаруживают весьма необычные актантные свойства, которые описываются в разделе 5.
6. Какое место в классификации наречий по ориентации занимают наречия совместности/раздельности? Эти наречия относятся к классу ориентированных наречий, поскольку их значение включает упоминание не только ситуации P, но и объектов Q, которые в ней участвуют. Однако они не входят в основные классы этой классификации (speaker-oriented и subject-oriented), поскольку могут иметь актант, являющийся подлежащим, прямым или косвенным дополнением и даже обстоятельством.

Благодарность

Эта работа выполнялась при финансовой поддержке гранта РНФ «Разработка семантического корпуса SemOntoCor на базе семантического анализатора SemETAP» №24-18-00988.

Литература

[1] Богуславский, И. М. Семантический анализ с опорой на умозаключения в функциональной модели языка. Вопросы языкознания, no. 1, 2021, pp. 29–56.

[2] Богуславский И.М., Иомдин Л.Л. О некоторых необычных типах заполнения семантических валентностей предикатов, или еще раз об отделении мух от котлет // Смыслы, тексты и другие захватывающие сюжеты. Сб. статей в честь 80-летия Игоря Александровича Мельчука. М.: Языки славянской культуры, 2012. С. 87-102.

[3] Борик, О. М. 1995. Синтаксический признак неаккузативности глагола (на материале русского языка). Дипломная работа. ОТиПЛ, филологический факультет, МГУ.

[4] Дудина, М. С. 2002. Признак неаккузативности глагола в итальянском и русском языках. Дипломная работа. Институт лингвистики, РГГУ.

[5] Зализняк Анна А., Шмелев А.Д. (2021). Исследования по руссской и компаративной семантике. Издательский дом ЯСК. Москва.

[6] Иомдин Л.Л. 1991. Словарная статья предлога ПО. В кн. Семиотика и информатика. Материалы к Интегральному словарю современного русского литературного языка (образцы словарных статей). Вып. 32. М., стр. 94-120.

[7] Кузнецова Ю. Л. 2010. Русская конструкция с предлогом по и проблема неаккузативности. В кн. Лингвистика конструкций. Отв. ред. Е.В.Рахилина. Азбуковник. М., стр. 184-218.

[8] Падучева Е.В. 2015. Нереферентные местоимения на -нибудь. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.

[9] Boguslavsky I. On the Passive and Discontinuous Valency Slots (2003). // First International Conference on Meaning — Text Theory (MTT'2003). June 16-18, 2003. Paris: Ecole Normale Supérieure.

[10] Boguslavsky, Igor, Vyacheslav Dikonov, Tatiana Frolova, et al. (2016). 'Plausible Expectations-Based Inference for Semantic Analysis'. Proceedings on the International Conference on Artificial Intelligence (ICAI), The Steering Committee of The World Congress in Computer Science, 2016, p. 477-483.

[11] Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S. (2018). Semantic analysis with inference: high spots of the football match. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2018", Moscow, May 30—June 2, 2018.

[12] Boguslavsky I.M., Dikonov V.G., Frolova T.I., Iomdin L.L., Lazursky A.V., Rygaev I.P., Timoshenko S.P. Full-fledged semantic analysis as a tool for resolving Triangle-COPA□social scenarios // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2020). Issue 19 (26). Pp. 106-118. ONLINE: ISSN 2075-7182.

[13] Champollion L. 2019. Distributivity in Formal Semantics. Annu. Rev. Linguist. 2019. 5:289–308

[14] Champollion L. 2020. Distributivity, Collectivity and Cumulativity. In Wiley's Companion to Semantics, ed. D Gutzmann, L Matthewson, C Meier, H Rullmann, TE Zimmermann. Hoboken, NJ: Wiley.

[15] Ernst, Thomas. 2002. The syntax of adjuncts. Cambridge: Cambridge University Press.

[16] Ernst, Thomas. 2003. Semantic features and the distribution of adverbs. In Lang, Ewald,Claudia Maienborn & Cathrine Fabricius-Hansen (eds.), Modifying adjuncts. Berlin: de Gruyter. 307-334.

[17] García Nuñez, José María. 2002. Adverb orientation: Semantics and pragmatics. Estudios de Lingüística Inglesa Aplicada 3. 299-315.

[18] Geuder, Wilhelm. 2002. Oriented adverbs: Issues in the lexical semantics of event adverbs. Tübinen: Universität Tübingen PhD thesis.

[19] Glass, L. 2018. Distributivity, lexical semantics, and world knowledge. PhD diss. Stanford University.

[20] Jackendoff Ray. 1972. Semantic Interpretation in Generative Grammar. Cambridge, MA: MIT.

[21] Kelepouris, S. 2012. The syntax and semantics of subject-oriented adverbs. A proposal for a new classification. Master dissertation at the University of Ghent.

[22] Kratzer, Angelika. 2007. On the plurality of verbs. In Johannes Dölling, Tatjana Heyde-Zybatow & Martin Schäfer (eds.), Event structures in linguistic form and interpretation, 269–300. Berlin, Germany: de Gruyter.

[23] Mel'čuk 2015 — Mel'čuk I. Semantics. From meaning to text. Vol. 3. Amsterdam: John Benjamins, 2014.

[24] Taverniers, Miriam & Gudrun Rawoens. 2010. Three types of oriented adjuncts in English and Swedish. Moderna Språk 104(1). 1-14.

[25] Verkuyl, Henk J. 1994. Distributivity and collectivity: a couple at odds. In MakotoKanazawa & Christopher Piñón (eds.), Dynamics, polarity, and quantification, 49–80. Stanford, CA: CSLI Publications.

Russian National Corpus 2.0: corpus platform, analysis tools, neural network models of data markup

Bonch-Osmolovskaya A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
abonch@gmail.com

Gladilin S. A.
IITP (Kharkevich Institute), FRC CSC
gladilin@iitp.ru

Kozerenko A. D.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
akozerenko@mail.ru

Lyashevskaya O. N.
HSE University
olesar@yandex.ru

Morozov D. A.
NSU
morozowdm@gmail.com

Kuznetzova Y. N.
MSU, Institute of Linguistics of the Russian Academy of Sciences
kuznetsova.yn@gmail.com

Makhova A. A.
Vinogradov Russian Language Institute of the Russian Academy of Sciences
discourse@yandex.ru

Piskounova S. V.
saruwatari.lara@gmail.com

Bujlova N. N.
Lopukhin Federal Research and Clinical Center of Physical-chemical Medicine of Federal Medical Biological Agency
bnn@rcpcm.ru

Borodina D. G.
St. Petersburg State University
daria-borodina2001@yandex.ru

Vinogradova I. I.
Prosveshchenie Publishers
irinaivinogradova@yandex.ru

Sizov V. G.
IITP (Kharkevich Institute)
victor.sizov@gmail.com

Dyachenko P. V.
IITP (Kharkevich Institute)
pavelvd@iitp.ru

Kazennikov A. O.
IITP (Kharkevich Institute)
kazennikov@gmail.com

Vlasova N. A.
A.K. Ailamazyan Institute of Program Systems of the Russian Academy of Sciences
nathalie.vlassova@gmail.com

Glazkova A. V.
University of Tyumen
a.v.glazkova@utmn.ru

Stolyarov S. S.
NSU
s.stolyarov@g.nsu.ru

Garipov T. A.
NSU
garipov154@yandex.ru

Smal I. A.
NSU
vanasmal@mail.ru

Gubar'kova Ya. N.
Yandex
karmastina-ya@yandex-team.ru

Abstract

The Russian National Corpus has existed for over 20 years and is a unique linguistic tool. However, the technical limitations of the software platform on which it was implemented significantly narrowed its development prospects. In 2020, work was launched on a comprehensive update of the RNC software platform, as a result of which the National Corpus switched to a new generation 2.0 platform. The implemented deep changes concerned both the development of functionality that meets modern approaches to corpus linguistics, and a fundamental restructuring of the platform architecture as a whole, from data preparation and indexing systems to the user interface. A separate area of development of the capabilities of the RNC was associated with the implementation of neural network models used for metadata tagging, disambiguation, word-formation markup, etc.

This article provides a short description of the new corpus platform as of 2024. The description includes key parameters of changes in the architecture of the RNC platform and its user interface, descriptions of new corpus data analysis services and the specifics of their implementation, as well as a description of the experience of using neural network models for tasks related to corpus data markup.

The purpose of the article is to describe the technological layer of changes implemented in the National Corpus of the Russian Language as part of a large-scale update carried out in recent years.

Keywords: corpus linguistics; Russian National Corpus language; architecture of software platforms; markup of language data

DOI: 10.28995/2075-7182-2025-23-57-73

Национальный корпус русского языка 2.0: корпусная платформа, инструменты анализа, нейросетевые модели разметки данных

Бонч-Осмоловская А. А.
ИРЯ им. В.В. Виноградова РАН
abonch@gmail.com

Гладилин С. А.
ИППИ им. А.А. Харкевича РАН,
ФИЦ ИУ РАН
gladilin@iitp.ru

Козеренко А. Д.
ИРЯ им. В.В. Виноградова РАН
akozerenko@mail.ru

Ляшевская О. Н.
НИУ ВШЭ
olesar@yandex.ru

Морозов Д. А.
НГУ
morozowdm@gmail.com

Кузнецова Ю. Н.
МГУ, ИЯз РАН
kuznetsova.yn@gmail.com

Махова А. А.
ИРЯ им. В.В. Виноградова РАН
discourse@yandex.ru

Пискунова С. В.
saruwatari.lara@gmail.com

Буйлова Н. Н.
Федеральный научно-клинический центр физико-химической медицины
bnn@rcpcm.ru

Бородина Д. Г.
СПбГУ
daria-borodina2001@yandex.ru

Виноградова И. И.
Издательство Просвещение
irinaivinogradova@yandex.ru

Сизов В. Г.
ИППИ им. А.А. Харкевича РАН
victor.sizov@gmail.com

Дьяченко П. В.
ИППИ им. А.А. Харкевича РАН
pavelvd@iitp.ru

Казенников А. О.
ИППИ им. А.А. Харкевича РАН
kazennikov@gmail.com

Власова Н. А.
Институт программных систем им. А.К. Айламазяна РАН
nathalie.vlassova@gmail.com

Глазкова А. В.
Тюменский государственный университет
a.v.glazkova@utmn.ru

Столяров С. С.
НГУ
s.stolyarov@g.nsu.ru

Гарипов Т. А.
НГУ
garipov154@yandex.ru

Смаль И. А.
НГУ
vanasmal@mail.ru

Губарькова Я. Н.
Яндекс
karmastina-ya@yandex-team.ru

Аннотация

Национальный корпус русского языка существует уже более 20 лет и представляет собой уникальный лингвистический инструмент. Однако технические ограничения программной платформы, на которой он был реализован, существенно сужали перспективы его развития. В 2020 году были запущены работы по комплексному обновлению программной платформы НКРЯ, в результате которого Национальный корпус перешел на платформу нового поколения 2.0. Реализованные глубинные изменения касались как развития функционала, отвечающего современным подходам корпусной лингвистики, так и фундаментальной перестройки архитектуры платформы в целом, начиная от систем подготовки и индексации данных и заканчивая пользовательским интерфейсом. Отдельное направление развития возможностей НКРЯ было связано с внедрением нейросетевых моделей, использующихся для разметки метаданных, снятия омонимии, словообразовательной разметки и др.

В настоящей статье представлено краткое описание новой корпусной платформы по состоянию на 2024 г. Описание включает в себя ключевые параметры изменений архитектуры платформы НКРЯ и его пользовательского интерфейса, описания новых сервисов анализа корпусных данных и специфики их реализации, а также описание опыта использования нейросетевых моделей для задач, связанных с разметкой корпусных данных.

Цель статьи заключается в описании технологического пласта изменений, реализованных в Национальном корпусе русского языка в рамках масштабного обновления, проведенного в последние годы.

Ключевые слова: корпусная лингвистика; Национальный корпус русского языка; архитектура программных платформ; разметка языковых данных

1 Введение

Национальный корпус русского языка был открыт для публичного доступа 29 апреля 2004 года. В этот момент объем единственного корпуса насчитывал 30 миллионов словоупотреблений. За более чем двадцать лет своего развития Национальный корпус не только заметно увеличился по объему и разнообразию данных, достигнув объема в 2,2 млрд словоупотреблений, но и претерпел концептуальную эволюцию. Изначальная задумка «Русского Стандарта» (Сичинава 2005) состояла в подготовке представительного собрания русских текстов, снабженных морфологической разметкой и предназначенных для удобного поиска при лингвистическом исследовании. НКРЯ в его современном состоянии охватывает тысячелетнюю историю развития русского языка и «представляет как язык предшествующих эпох, так и современный, в разных социолингвистических вариантах — литературном, разговорном, просторечном, диалектном»[1]. Традиционные поисковые инструменты расширяются сервисами статистического анализа и визуализации корпусных данных, принципиально изменились внутренние возможности управления корпусом, начиная от подготовки корпусных данных и кончая гибкими настройками интерфейса. Выход Национального корпуса русского языка на текущий уровень развития потребовал технологической трансформации корпуса как электронного ресурса, происходившей в одном русле с основными процессами современной корпусной лингвистики. В настоящей статье представлено краткое описание новой корпусной платформы по состоянию на 2024 г. Лингвистический аспект обновлений был подробно рассмотрен в статье (Савчук и др., 2024).

1.1 НКРЯ 2.0 в свете основных тенденций развития современных корпусов

Ключевыми параметрами развития современных корпусных технологий, задающими тот научный контекст, в котором формируется стратегия технологических изменений Национального корпуса русского языка, на наш взгляд, являются (1) критическое увеличение объемов корпусов (от

[1] https://ruscorpora.ru/page/corpora-about/

сотен тысяч словоупотреблений к миллиардам), (2) переход на стандартизованные системы лингвистической разметки, (3) внедрение инструментов статистического анализа корпусных данных и, как результат технологического развития, (4) расширение областей практического применения корпусов. Подробный обзор работ в этих направлениях можно увидеть в расширенной версии настоящей статьи.

С учетом этих параметров новая платформа по сравнению со старой приобрела следующие принципиальные свойства.

Объемы корпусов

Старая платформа: К 2020 году общий объем всех корпусов НКРЯ составляет около 1 млрд словоупотреблений. При этом грамматическая омонимия снята менее чем в 1% от всех словоупотреблений. Дальнейшее расширение корпуса затруднено в связи с архитектурными ограничениями платформы.

Новая платформа: К 2024 году общий объем всех корпусов НКРЯ составляет более 2,2 млрд словоупотреблений, грамматическая омонимия снята примерно в 65% от всех словоупотреблений. Новая платформа спроектирована для поддержки данных НКРЯ в объеме до 100 млрд словоупотреблений.

Разметка данных

Старая платформа: Корпусные данные снабжаются морфологической разметкой, осуществленной с помощью программы MyStem (Зобнин, Носырев 2015). Алгоритм позволяет строить гипотетические разборы для слов, которых нет в грамматическом словаре Зализняка, но не позволяет с достаточной точностью снимать омонимию.

Новая платформа: Сохраняется морфологическая разметка алгоритмом MyStem, к ней добавлена разметка данных с помощью нейросетевой модели Rubic. Модель размечает не только морфологические, но и синтаксические характеристики словоформ, а также снимает омонимию не только по леммам, но и по словоизменительным признакам.

Развитие инструментов корпусного анализа

Старая платформа: Основным инструментом корпусного анализа является выдача по поисковому запросу в формате конкорданса или KWIC (key word in context). Пользователь имеет возможности сортировки результатов по дате создания текста и другим релевантным параметрам, а также по правому/левому контексту в формате KWIC. Обобщенная информация об изменениях частотностей слова представлена в виде диахронического графика по конкретной словоформе. Пользователь имеет доступ к n-граммам, предпосчитанным по словоформам.

Новая платформа: Инструменты корпусного анализа существенно расширены как на уровне запроса, так и на уровне представления выдачи. На уровне поисковых запросов появился поиск по коллокациям, на уровне выдачи — анализ частотности запроса, допускающий разные способы сортировки и представления данных. Появилась возможность получить предпосчитанную информацию о слове в корпусе в целом — «Портрет слова», куда входят его скетчи, контекстуально близкие слова, выявленные на основе расчета семантических векторов, однокоренные слова в корпусе.

Целевая аудитория НКРЯ

Старая платформа: Платформа ориентирована на подготовленного пользователя-лингвиста, который использует корпус как источник материала для лингвистических исследований.

Новая платформа: Новая платформа ставит своей задачей расширить аудиторию пользователей, в том числе привлекая менее подготовленных пользователей, не работавших ранее с языковыми корпусами. Корпус существует в мобильной версии, имеет богатейшую документацию, логика интерфейса минимизирует усилия пользователя по получению информации.

Ниже будут более подробно рассмотрены три аспекта технической реализации новой корпусной платформы. Во-первых, это концептуально новые подходы к архитектуре корпуса, корпусному ядру и веб-интерфейсу. Во-вторых, разработанные сервисы для корпусного анализа данных. В-третьих, нейросетевые модели, использованные для разметки данных.

2 Корпусная платформа нового поколения: примененные подходы и решения

2.1 Общая архитектура системы

К корпусной платформе нового поколения предъявлялись требования не только соответствия современным стандартам сервисов, предоставляемых крупными лингвистическими корпусами, но и обеспечения гибкости для последующей модификации и развития в соответствии с перспективными подходами, которые могут возникнуть в будущем. На момент разработки имелась существенная неопределенность, в целом характерная для ИТ-проектов: у нас не было представления о полном функционале, который в будущем потребует поддержки в НКРЯ. С развитием корпусных технологий возникают новые потребности, и корпусная платформа должна быть готова к их реализации. Поэтому требовалось организовать программную систему так, чтобы в будущем по возможности облегчить добавление нового функционала. Для этого мы стремились обобщить различные требования (уже имеющиеся или же потенциально возможные) к функционалу в однородные с точки зрения технической реализации группы. Таким образом, перед корпусной платформой ставилась задача поддержки не конкретных видов функционала, а целых функциональных групп; конкретные виды функционала рассматривались как представители этих групп. Такой подход потребовал унификации корпусных данных: атрибуты одинакового типа обрабатываются одними и теми же алгоритмами, а значит должны быть единообразно представлены.

Структурно программная система разделена на три независимые части. Вычислительное ядро реализует универсальный функционал для целой функциональной группы, лингвистическое ядро обеспечивает поддержку конкретных функций, используя реализованный функционал вычислительного ядра, а интерфейсный модуль осуществляет взаимодействие лингвистического ядра с пользователями. Таким образом, например, добавление новой функциональности, касающейся приписанных к токенам атрибутов, выполняется в вычислительном ядре и влияет одновременно на все атрибуты, и таким образом может не затронуть или незначительно затронуть лингвистическое ядро, а поддержка новых атрибутов легко реализовывается в лингвистическом ядре и не требует изменений вычислительных алгоритмов.

Такое разделение позволило использовать в корпусной платформе разные вычислительные ядра в зависимости от размера и структурной сложности корпуса. Так, например, для больших корпусов применяется вычислительное ядро, построенное на базе поисковой системы ElasticSearch, в то время как для сложно структурированной разметки небольшого Синтаксического корпуса лучше подошло вычислительное ядро на базе реляционной базы данных MySQL.

Выделение отдельного интерфейсного модуля важно, поскольку подходы к построению графических интерфейсов пользователя быстро меняются, а отдельный модуль легче заменить.

Было выделено 12 различных видов разметки, поддерживаемой платформой, и охватывающих структурные единицы разного уровня: от морфемной и акцентологической разметки внутри токена до метаразметки целых текстов. Все поддерживаемые НКРЯ виды разметки были преобразованы к этим видам. В дальнейшем при возникновении новых атрибутов, попадающих в один из выделенных видов, не потребуется изменения вычислительного ядра.

Например, в корпусной платформе нового поколения тексты, состоящие из слов, и аннотации жестикуляции в видеозаписи, состоящие из отдельных жестов, представляются при помощи одного и того же программного механизма. Таким образом, выровненный с текстом видеоряд внутренне представляется и обрабатывается тем же способом, что и два параллельных текста, выровненных между собой.

2.2 Автоматизированное взаимодействие корпусной платформы с другими лингвистическими системами

В настоящее время корпуса, доступные для пользователей через сеть Интернет, являются одним из важнейших инструментов корпусной лингвистики. Однако такой подход позволяет применять к корпусным данным только набор инструментов, реализованный в интерфейсе веб-сайта. Частично преодолеть это ограничение можно за счет предоставления пользователю возможности автоматизировать выполнение запросов к корпусу.

Автоматизация запросов к корпусной платформе осуществляется при помощи программного интерфейса приложений (англ. Application Programming Interface, API). Разработанный для нужд НКРЯ API обеспечивает возможность выполнения произвольных запросов, доступных через интерфейс. Однако в настоящее время API не доступен стороннему пользователю, а используется только самим графическим интерфейсом системы. Такой подход позволил нам отделить реализацию интерфейса пользователя от непосредственно поискового сервера. Хотя в настоящее время нельзя утверждать, что API решает задачу автоматизации, успешное применение в архитектуре системы подтверждает его универсальность: поскольку любая операция с корпусом преобразуется интерфейсом в запрос к API, можно сделать вывод, что API обеспечивает все необходимые возможности. После определенной доработки планируется сделать API общедоступным.

API поддерживает запросы, соответствующие основным видам использования корпусной платформы:

- запрос основных статистических данных о корпусах НКРЯ;
- запрос данных о конфигурации конкретного корпуса;
- запрос набора доступных в конкретном корпусе поисковых форм;
- запрос типов атрибутов, имеющихся в корпусе, и их допустимых значений;
- поисковый/аналитический запрос;
- запрос «Портрета» конкретного слова или конкретного корпуса.

2.3 Новая концепция интерфейса корпусной платформы, ориентированная на широкий круг пользователей

Интерфейс новой корпусной платформы обеспечивает возможность её использования с самых различных устройств: от смартфонов до настольных компьютеров с большим размером экрана. Для этого применяется подход «первичности мобильной версии» (англ. mobile first), в соответствии с которым самые первыми в списке вариантов стилевых таблиц находятся таблицы для самых миниатюрных мобильных устройств (обладающих не только самыми маленькими размерами экрана, но и наименьшими вычислительными ресурсами), что позволяет им прекратить ресурсоемкий для них дальнейший перебор вариантов. Пользователю автоматически открывается версия, наиболее подходящая для размера устройства, с которого он выходит в интернет (Рис. 1).

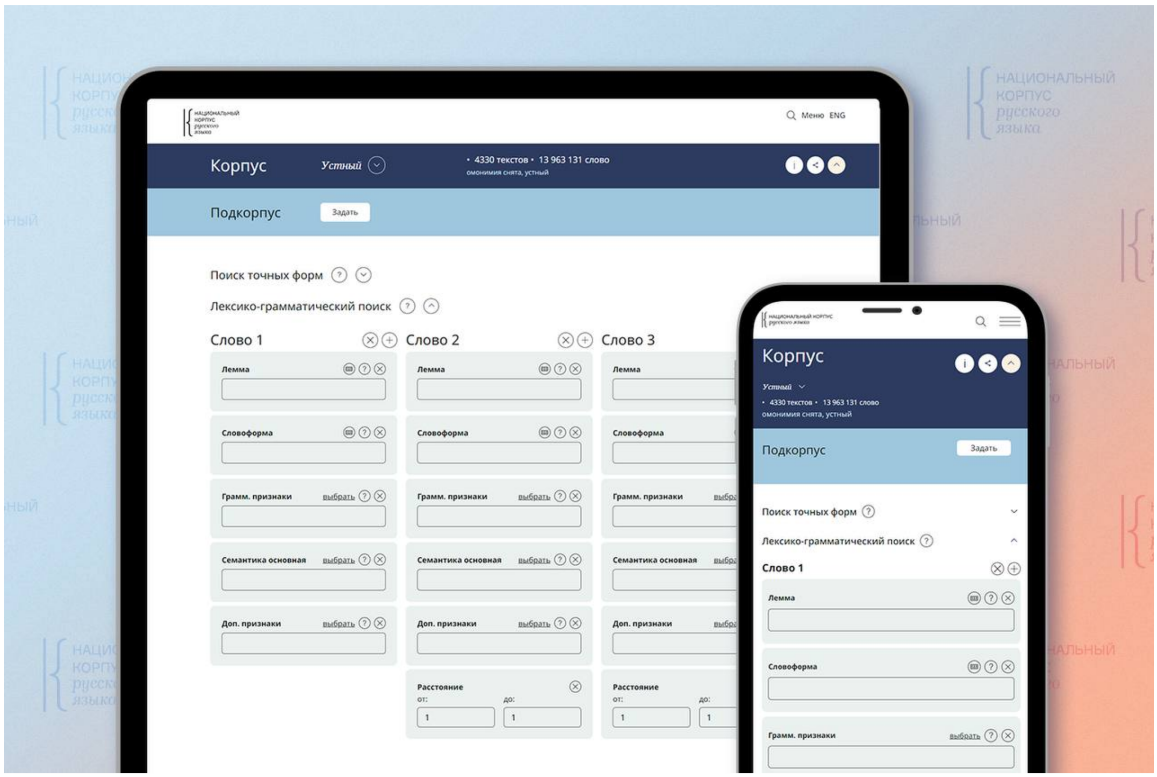


Рис. 1: Слева — версия сайта для ПК и планшета, справа — оптимизированная для мобильных устройств.

Реализован ряд комплексных лингвистических инструментов, позволяющих на одном экране получать разностороннюю лингвистическую информацию. Так, сервис «Портрет слова» в визуально компактной и понятной форме предоставляет разнообразную информацию для всех имеющихся разборов заданной леммы.

Там, где это возможно, результаты представляются с помощью круговых, столбчатых диаграмм, географических карт и графиков. В инструменте «Частотность» показываются доверительные интервалы для рассчитанной частотности (Рис. 2). При отображении графиков указываются временные границы, за пределами которых данных слишком мало для достоверных выводов. Под графиком отображается тепловая шкала, описывающая количество текстов, в которых найдены результаты, позволяющая оценить, насколько рост частотности слова является случайным выбросом в конкретном тексте или же объективно наблюдаемым явлением (Рис. 3).

Слово 1 Словоформа	← [1..2] →	Слово 2 Лемма	Вхождения	Доля	Доля	ipm	Конкорданс
какая	1	разница	2	50%	[15%, 85%]	28.21	Примеры
какая	1	жалость	1	25%	[4.56%, 69.94%]	14.11	Примеры

Рис. 2: Вид выдачи «Частотность»

Рис. 3: Вид выдачи «График»

«Похожие слова» отображаются в виде облака тегов, в котором размер букв и удаленность слов друг от друга характеризуют степень близости контекстов употребления слов (Рис. 4). Для морфемного разбора использована наглядная нотация, принятая в школьном преподавании русского языка (Рис. 5).

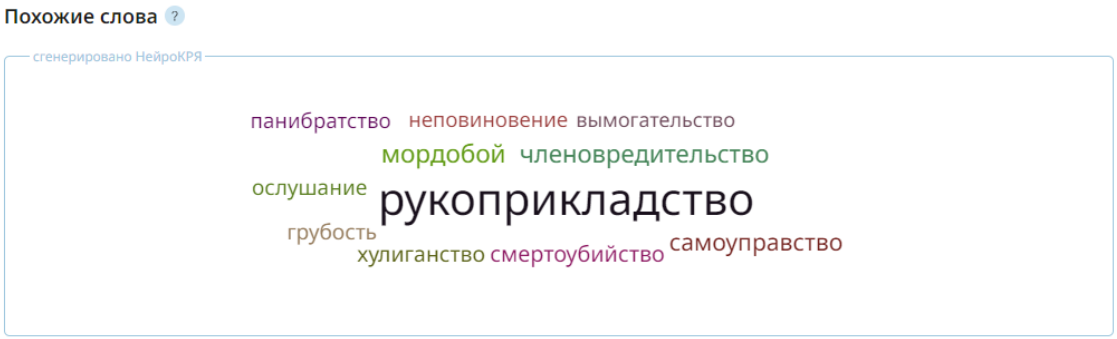


Рис. 4: Виджет «Похожие слова»

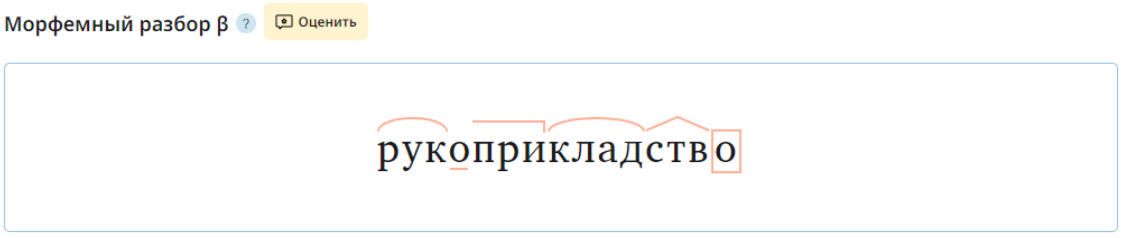


Рис. 5: Виджет «Морфемный разбор»

В нескольких сценариях поиска по Корпусу элементы интерфейса были намеренно перегруппированы по сравнению со старой версией корпусной платформы. Так, в интерфейсе задания условий лексико-грамматического поиска группы условий на искомые слова в словосочетании теперь расположены в одну строку слева направо (Рис. 6 и 7). Такой подход визуально более интуитивен для пользователей и соответствует расположению слов в тексте.

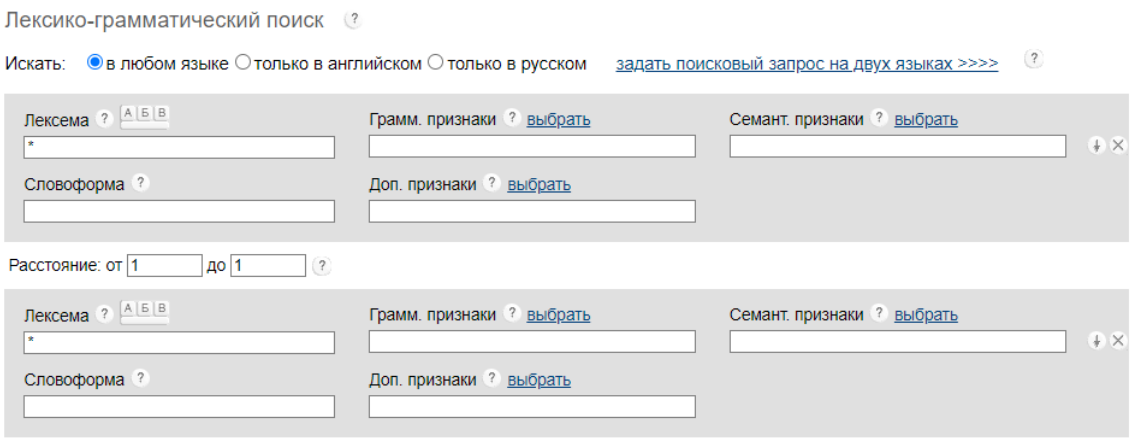


Рис. 6: Старое расположение условий лексико-грамматического поиска

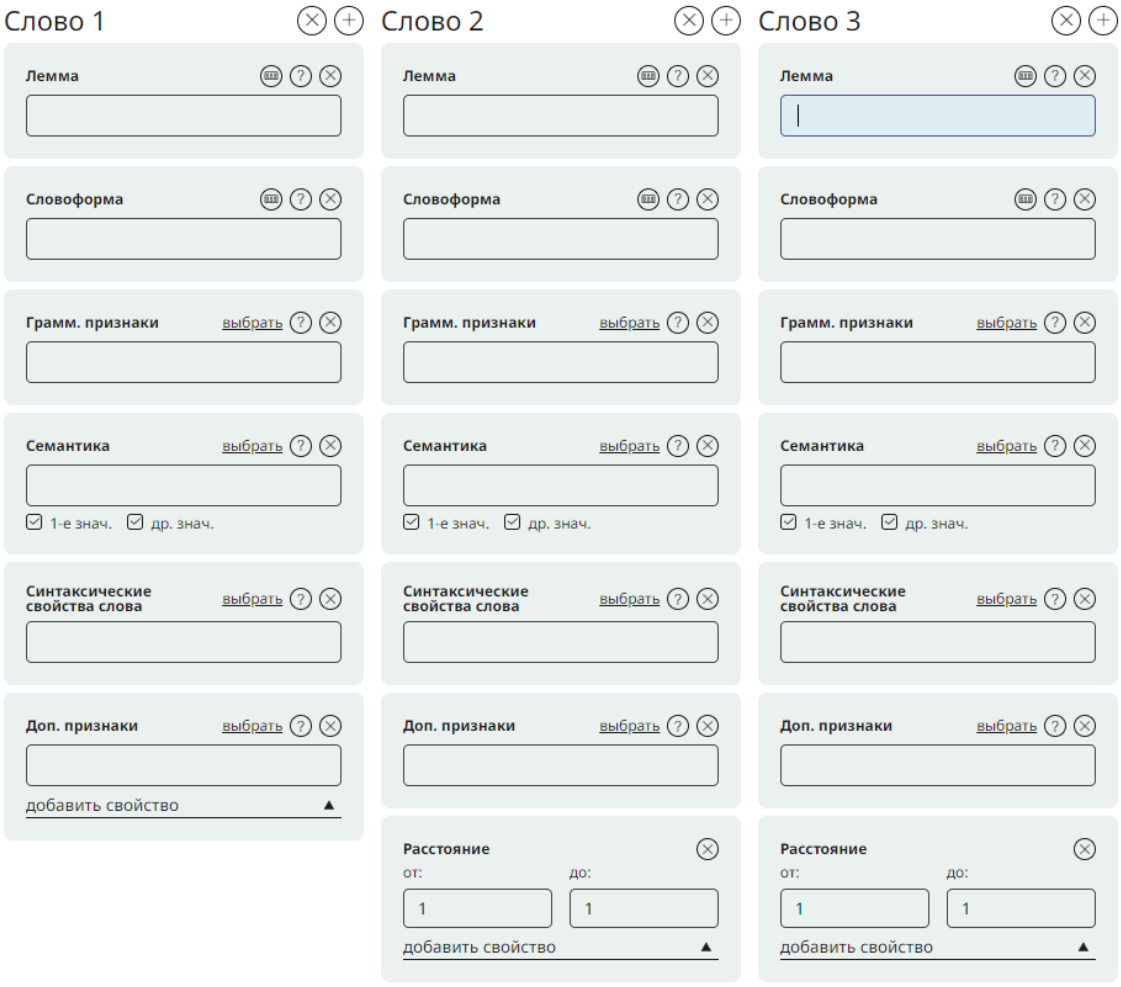


Рис. 7: Новое расположение условий лексико-грамматического поиска

Новый интерфейс ориентирован на уменьшение количества лишних действий пользователя. Так, выбор настроек, предпочтительный вид поиска в каждом корпусе, вид выдачи, открывающийся по умолчанию, а также режим отображения информации о запросе в шапке корпуса запоминаются в браузере пользователя и применяются для следующих поисковых запросов. В шапке

корпуса отображаются параметры не только искомого слова, но и заданного пользователем подкорпуса. Возможность в любой момент вернуться к форме и откорректировать любые параметры запроса сокращает усилия в сравнении с заданием параметров заново.

В новом интерфейсе поддерживается регулярно обновляемое руководство пользователя и поиск по нему. Система управления контентом позволяет в онлайн режиме структурировать, тегировать и редактировать все сопроводительные тексты.

Для иностранных пользователей весь интерфейс сайта переведен на английский язык.

3 Инструменты анализа корпусных данных

3.1 Основные направления развития статистико-аналитической компоненты НКРЯ

Современные методы корпусной лингвистики в наибольшей степени ориентированы на использование количественного анализа распределения языковых единиц. Именно поэтому современные корпусные платформы не ограничиваются лишь конкордансами для отображения словоупотреблений, но включают дополнительные сервисы, которые позволяют систематизировать, обобщать и статистически оценивать результаты анализа корпусных данных.

Программная платформа НКРЯ нового поколения реализует широкий спектр аналитических инструментов обработки корпусных данных. Эффективность их реализации обеспечена за счет предварительных вычислений на этапе индексации текстов, использования (при необходимости) приближенных вычислений по рандомизированной подвыборке и эффективных по времени доступа, но затратных по памяти механизмов кэширования. В вычислительном ядре алгоритмы реализации указанных статистических инструментов подразделяются на:

- алгоритмы, выполняемые разово в процессе индексации: результат работы такого алгоритма сохраняется в базе данных и готов к использованию при обработке запроса пользователя;
- алгоритмы, выполняемые в процессе пользовательского запроса на основе текстов корпуса и/или показателей, вычисленных и сохраненных в процессе индексации;
- алгоритмы, выполняемые в процессе пользовательского запроса на основании случайного подмножества результатов поиска. Результаты работы такого алгоритма являются приблизительными, поэтому они применяются в случае, когда точное вычисление в процессе запроса невозможно из-за ограничений на время ожидания.

3.1.1 Инструменты статистической характеризации корпусов и подкорпусов

Статистические инструменты этого типа позволяют строить портрет корпуса и подкорпуса, получать статистическое распределение текстов по значениям мета-атрибутов и строить диахронические графики. Поскольку корпусная платформа позволяет рассматривать произвольный набор текстов, заданный пользовательскими условиями, как подкорпус, количество гипотетически возможных подкорпусов очень велико. Таким образом, статистические характеристики подкорпусов не могут быть предварительно рассчитаны на этапе индексации.

3.1.2 Инструменты статистической характеризации результатов поиска словосочетаний

Статистические характеристики этого типа позволяют строить распределение, удовлетворяющее поисковым условиям словоформ и лемм и получать наиболее распространенные в результатах поиска n-граммы. Это наиболее ресурсоемкие вычисления, выполняемые непосредственно в момент запроса, поскольку количество результатов поискового запроса может превышать размер всего корпуса целиком (в случае, если одно и то же слово входит в несколько разных словосочетаний, удовлетворяющих условию поиска). При отображении результатов поиска вычисления могут быть прекращены, как только нужное количество примеров сформировано, но при подсчете статистики должны быть учтены все результаты или их репрезентативная подвыборка. В случае если количество результатов поиска превышает миллион, в качестве такой выборки рассматривается случайное подмножество в миллион результатов и производится расчет относительных показателей только на основе них.

3.1.3 Инструменты статистической характеризации лемм

Инструменты этой категории позволяют, например, находить похожие слова, то есть слова, встречающиеся в одинаковых контекстах. В виджете «Похожие слова» отображаются ближайшие семантические ассоциаты слова; коэффициент близости слов подсчитывается с помощью моделей дистрибутивной семантики, построенных на актуальных материалах основного корпуса НКРЯ. Вычисления характеристик этого типа требует предрасчетов на этапе индексации текстов с сохранением информации, привязанной к каждой лемме. В момент пользовательского запроса происходит статистический расчет на основе сохраненной информации, а не самих текстов корпуса, что критично снижает вычислительную сложность.

3.1.4 Статистические коллокации

Статистические коллокации для произвольного запроса могут быть вычислены только в момент пользовательского запроса, но скетчи (заранее фиксированные для каждой части речи наборы коллокаций с учетом синтаксических связей) эффективнее вычислять на этапе индексации и сохранять для каждой леммы, поскольку количество лемм в корпусе ограничено и составляет не более нескольких сотен тысяч (с порогом встречаемости хотя бы 3 раза в 3 различных текстах). Для остальных лемм такая информация не сохраняется.

3.2 Аналитические инструменты для сервиса «Портрет слова»

В рамках рассмотрения аналитических инструментов новой корпусной платформы НКРЯ мы уделим особое внимание двум нейросетевым моделям, используемым в сервисе «Портрет слова»: модели словообразовательного разбора и семейству векторных word2vec-моделей, используемых в виджете «Похожие слова».

3.2.1 Модель словообразовательного разбора

В НКРЯ словообразовательная разметка лемм присутствует в двух корпусах: Основном и Обучающем. С точки зрения архитектуры системы морфемный разбор является атрибутом, приписанным структурной единице «разбор». Для каждого разбора указан список морфем и тип каждой из них.

В основе разметки в Основном корпусе лежит специально разработанный для корпуса словарь морфемного анализа **Morphodict-K**, где по состоянию на май 2023 года даны разборы для 75 тыс. лексем. Этот словарь составлялся на основании идеологии «Словаря морфем русского языка» (Кузнецова, Ефремова 1986). Принципы этой идеологии — значительная дробность выделения морфем и соотносимость с другими лексемами аналогичного строения.

Разметка морфем в Обучающем корпусе опирается на разработанный на основе «Морфемно-орфографического словаря» (Тихонов 2002) словарь морфемного анализа **Morphodict-T**. Этот словарь содержит около 100 тыс. лексем, морфемный состав которых определяется в соответствии с практикой морфемного анализа в средней школе. При этом используется более жесткий подход к определению того, какие смысловые связи являются прозрачными в современном языке, и, как правило, выделяется меньшее число морфем, чем в Morphodict-K.

Словари Morphodict-K и Morphodict-T сравнительно малы по отношению к многообразию всех лемм Основного и Обучающего корпусов: в совокупности в этих корпусах содержится более 300 тыс. уникальных лексем. Для слов, не входящих в словари, морфемные разборы строятся автоматически при помощи алгоритма на базе ансамбля сверточных нейронных сетей (Sorokin, Kravtsova 2018). В ходе тестирования доля полностью верных разборов составила 88.5% для словаря Morphodict-T и 90.8% для словаря Morphodict-K (Garipov, Morozov, Glazkova 2023). В настоящий момент модель, обученная на Morphodict-K, интегрирована в сервис «Портрет слова» в Основном корпусе НКРЯ: из 314 935 различных лемм, представленных в сервисе, для 255 821 леммы разбор сгенерирован моделью. Обе модели размещены в открытом доступе и доступны для выгрузки в разделе «Нейросетевые модели» на сайте НКРЯ.[2]

[2] https://ruscorpora.ru/license-content/neuromodels

3.3 Векторные модели в «Портрете слова» (сервис «Похожие слова»)

Так как НКРЯ содержит весьма разнообразные по домену (типу, тематике, жанру и т.д.) и времени создания корпусы, одни и те же слова могут употребляться в них в несовпадающих значениях и наборах контекстов. Для того чтобы обнаружить и визуализировать особенности использования слов в различных корпусах, могут быть использованы модели векторного представления слов. Статические векторные модели на базе Основного корпуса строились и раньше (Kutuzov, Kuzmenko 2017), однако в ходе нашей работы модели обучались с использованием лемм, сгенерированных моделью Rubic или, при их наличии, указанных вручную. Это позволило существенно уменьшить количество ошибочно сгенерированных или неправильно токенизированных лемм в словаре модели.

Мы построили модели семантических векторов для существительных, глаголов, прилагательных и наречий для Основного, Обучающего, Газетных корпусов, Древнерусского, Старорусского, корпуса «От 2 до 15» и корпуса «Русская классика», ср. разницу ассоциатов для слова *трубить* в разных корпусах (Рис. 8–9):

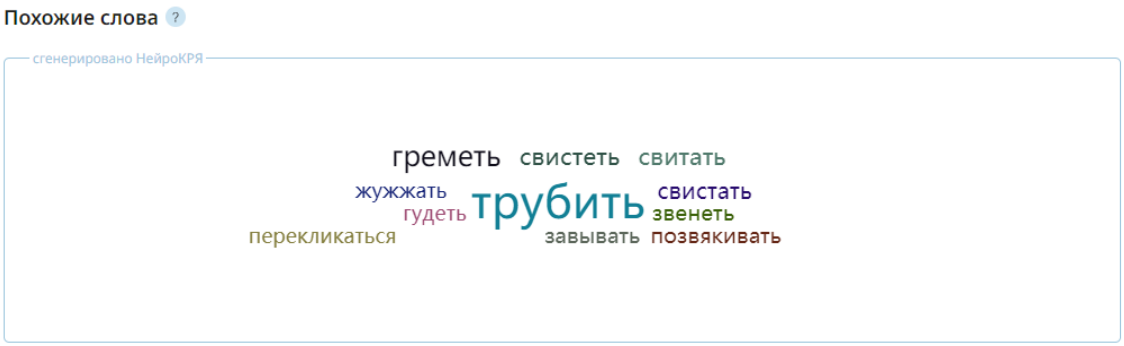


Рис. 8: Похожие слова для слова *трубить* в корпусе «Русская классика»

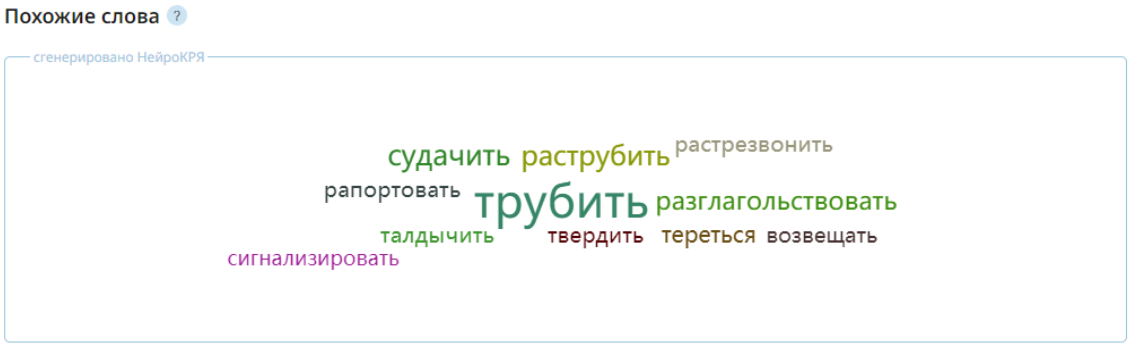


Рис. 9: Похожие слова для слова *трубить* в корпусе Центральных СМИ

Все используемые на сегодняшний день модели имеют одинаковую архитектуру (CBoW из библиотеки (Rehurek, Sojka 2011)) и схожие параметры обучения: окно размером 5, порог встречаемости 5-10 в зависимости от корпуса. Все семь моделей размещены в открытом доступе на странице «Нейросетевый модели» сайта НКРЯ[3] и могут быть использованы для научных и прикладных исследований.

[3] https://ruscorpora.ru/license-content/neuromodels

4 Нейросетевые модели в разметке данных и метаданных НКРЯ

Наиболее ресурсно затратным этапом при подготовке корпуса является этап корпусной разметки. Это касается как собственно внутритекстовой лингвистической разметки, так и разметки текстовых метаданных. В прошлом алгоритмы автоматической разметки морфологии показывали недостаточно высокое качество, поэтому центральным решением старой платформы НКРЯ был отказ от снятия омонимии — словоформе приписывались все возможные разборы, выбор релевантного разбора оставался на стороне пользователя. Такая неопределенность блокировала развитие статистических и аналитических сервисов, поскольку омонимичные разборы существенно зашумляют любые подсчеты. Применение интеллектуальных моделей позволяет значительно повысить качество автоматической разметки и исключить это ограничение. Ниже будут кратко представлены модели, которые сейчас используются при подготовке данных НКРЯ. Это, во-первых, нейросетевая модель морфосинтаксической разметки Rubic, а во-вторых, комплекс моделей для разметки метаданных: жанров в корпусе «Социальные сети» и ключевых слов в текстах Газетного корпуса. Использование этих моделей уже изменило экосистему НКРЯ, открыв новые возможности для значительного облегчения наиболее трудоемкого этапа подготовки корпусных данных.

4.1 Нейросетевая модель морфосинтаксической разметки для русского языка Rubic

Задачи, которые решает нейросетевая модель Rubic, — это автоматическая лемматизация, морфологическая характеристика для всех токенов, включая определение части речи, и построение дерева синтаксической зависимости предложения. Таким образом, Rubic представляет собой альтернативу морфологическому анализатору MyStem (Зобнин, Носырев, 2015), ранее применявшемуся для обработки текстов НКРЯ и основанному на грамматическом словаре. При разработке модели ставились задачи улучшения обработки «несловарных» разборов (например, *ажник, летось, сподтишка*), просторечных и грамматически аномальных форм и конструкций (например, *силов*, *хоцца, подумашь, ефту*), словоформ, записанных в нестандартной орфографии (в том числе петровской эпохи, в дореволюционной орфографии, а также в советской орфографии до реформы 1956 года). Модель также должна корректно обрабатывать архаичные формы из церковнославянского языка (например, *бысть, быша, многая лета*) согласно соглашениям, принятым для исторических корпусов НКРЯ.

4.1.1 Принципы работы Rubic

Архитектура Rubic (Lyashevskaya et al., 2023) основана на архитектуре модели qbic, победившей в соревновании по обработке русского языка GramEval2020 (Anastasyev 2020): однослойном LSTM-энкодере, комбинирующем векторизованные представления слов, получаемые из BERT-подобной модели (в текущей реализации, sberbank-ai/ruBert), и морфологические пометы, приписываемые анализатором PyMorphy2. Полученное представление анализируется тремя декодерами, выполняющими задачи классификации для выбора: а) части речи и грамматических признаков, б) леммы и в) дерева зависимостей. Для обучения модели использовались специально подготовленные обучающие данные на основе корпусов СинТагРус, UD-Taiga (Droganova, Zeman 2018; Droganova, Lyashevskaya, 2018) и НКРЯ (Lyashevskaya et al., 2020). Они охватывают тексты различных временных эпох и жанров общим объемом свыше 2,4 миллиона токенов. Все данные приводятся в расширенном формате морфологической и синтаксической разметки UD-ext (Lyashevskaya 2019). Целевая синтаксическая разметка представляется в формате CONLL-U, который затем сохраняется в синтаксисе XML, принятом в НКРЯ. Это позволяет, с одной стороны, сохранить подход к синтаксической разметке, принятый в Universal Dependencies, а с другой — использовать морфологическую разметку в стандарте НКРЯ.

4.1.1.1 Токенизатор

Практически любой анализ текста начинается с его разбиения на фрагменты (токенизации). Анализ существующих алгоритмов токенизации (udpipe (Straka, Hajic, Straková 2016), razdel[4], spacy[5],

[4] https://github.com/natasha/razdel
[5] https://github.com/explosion/spaCy

nltk (Bird, Loper, Klein 2009), PyMorphy2 (Korobov 2015), MyStem[6], rusenttokenize[7]) показал, что качество сегментации текста на предложения у таких подходов достаточно невелико (F1 на тестовой выборке сложных предложений GOLD не превысило 0.55). В связи с этим было принято решение обучить на имеющихся данных собственную модель для токенизации. В качестве архитектуры была выбрана модель Stanza (Qi P. et al. 2020). Мы обучили модель со стандартными параметрами обучения на открытых датасетах Тайга (Shavrina, Shapovalova 2017) и СинТагРус[8], а также на внутренних данных из корпусов прозы XX–XXI веков, поэзии, корпусов со старой орфографией XVIII века, а также текстах новостей XXI века. Полученная модель на выборке GOLD продемонстрировала высокое качество по отдельным словам (F1=0.94) и значительно лучшее качество по предложениям (F1=0.63). Для дополнительного сравнения новой модели с предыдущим решением была подготовлена расширенная тестовая выборка, схожая по составу с реальными данными (выборка TEST). На материале этой выборки обученная модель значительно превзошла лучший из рассмотренных алгоритмов, достигнув метрики F1=0.95 по предложениям и 0.99 по отдельным словам. Модель токенизатора доступна для скачивания на странице «Нейросетевые модели» сайта НКРЯ.[9]

4.1.1.2 Классификатор морфологических признаков

После токенизации текста происходит приписывание каждому выделенному слову морфологических признаков. Классификатор морфологических признаков работает на принципе полного морфологического тега, иными словами, входом и выходом модели служит набор, состоящий из частеречного и грамматических тегов вида «NOUN|Animacy=Anim|Case=Nom|Gender=Fem|Number=Sing». В отличие от классификаторов, в которых каждая грамматическая категория определяется независимо, данное решение позволяет избежать потери части разбора (например, одушевленности, в случае если вероятность ее определения в контексте низкая) и свести требуемый набор грамматических помет для частеречных классов и подклассов к грамматическому стандарту корпуса.

4.1.1.3 Лемматизатор

Размеченное морфологическими пометами слово попадает в лемматизатор. В основе лемматизации лежат правила преобразования словоформы в лемму вида «удалить последовательность символов в конце строки длины N, удалить последовательность символов в начале строки длины M > добавить последовательность D в конце строки > применить маску капитализации / декапитализации». Используются преобразования, встретившиеся в обучающих данных более 3 раз, чтобы исключить влияние несистемных опечаток и другого шума. В зависимости от объема обучающих данных, такой подход дает от 1000 до 2000 правил. Для улучшения качества лемматизации предусмотрена проверка наиболее вероятных гипотез лемм с данными словаря, составленного вручную.

4.1.1.4 Разметка синтаксического дерева

Наконец, на последнем этапе разметки происходит анализ синтаксического дерева каждого предложения. При построении гипотез синтаксических деревьев используется подход Т. Дозата и К. Маннинга (Dozat, Manning, 2016) на основе глубокого биаффинного внимания для определения пар связанных словоформ и метки синтаксического отношения.

4.1.2 Результаты

Качество работы модели Rubic'a оценивалось на коллекции тестовых данных, представляющих разные сферы употребления языка, см. Таблицу 1.

[6] https://yandex.ru/dev/mystem/
[7] https://github.com/deeppavlov/ru_sentence_tokenizer
[8] https://ruscorpora.ru/page/corpora-datasets/
[9] https://ruscorpora.ru/license-content/neuromodels

	fiction	news	poetry	social	wiki
Часть речи	0.9922	0.9893	0.9923	0.9777	0.9808
Леммы	0.9930	0.9923	0.9846	0.9848	0.9780
Морфологические признаки	0.9591	0.9517	0.9654	0.9528	0.9423
Неименованные синтаксические связи	0.9599	0.9563	0.9106	0.9296	0.9457
Именованные синтаксические связи	0.9530	0.9425	0.8942	0.9153	0.9231

Таблица 1: Результаты работы Rubic'а на тестовом множестве

Модель хорошо справляется с определением частей речи, морфологическим разбором некоторых грамматических категорий, таких как сравнительная степень, переходность, вид. При лемматизации модель демонстрирует высокое качество обработки для слов продуктивных парадигм. Результаты синтаксического парсинга показывают, что модель чаще всего правильно анализирует большинство часто употребляемых конструкций — вводные и сочиненные конструкции, предложные и атрибутивные группы и т.п. Также хорошо определяются дальние связи (например, субъект на расстоянии 5-10 слов от предиката). Более подробно особенности генерируемой разметки рассмотрены, например, в работе (Lyashevskaya et al., 2023).

4.2 Разметка жанров в корпусе «Социальные сети»

Корпус «Социальные сети» содержит тексты из открытых интернет-источников и включает в себя записи в блогах и сообщения в мессенджерах (подробнее о балансе источников корпуса «Социальные сети» см. в статье (Савчук и др., 2024)). Поскольку понятие «социальные сети» в этом случае трактуется максимально широко, а также в связи с большим объемом корпуса (почти 160 млн словоупотреблений), появилась необходимость в автоматической разметке жанров для систематизации текстов корпуса.

Для разметки жанров была выбрана предварительно обученная модель RuRoBERTa [10] (Zmitrovich et al., 2023). Модель была дообучена на подготовленном вручную наборе данных объёмом около 3 тысяч текстов, размеченном по 16 жанрам, с использованием следующих параметров: скорость обучения — 5e-6, количество эпох обучения — 3, максимальная длина входной последовательности — 256 токенов. При обучении жанрам назначались веса в зависимости от их доли в наборе данных. Качество модели было проверено с помощью кросс-валидации для десяти фолдов. Значение F-меры с макроусреднением составило 54,42% для 16 жанров, доля правильных ответов (accuracy) составила 71,16%.

Для итоговой разметки жанров в корпусе «Социальные сети» был использован ансамбль из трех моделей с усреднением предсказанных вероятностей жанров (soft voting).

4.3 Разметка ключевых слов в Корпусе региональных СМИ

Корпус Региональных СМИ в основном состоит из коротких информационных текстов, опубликованных в газетах различного уровня (Савчук 2015). Каждый текст, как правило, посвящен единственной теме. Для описания тематики и упрощения поиска текстов в корпусе выполнена автоматическая разметка ключевых слов. Одно ключевое слово может состоять из одного существительного в именительном падеже в единственном или множественном числе (*праздник, переломы*) либо из двусловного сочетания (биграммы) с главным словом-существительным (*таяние снега, обычные дни*).

[10] https://huggingface.co/ai-forever/ruRoberta-large

Извлечение ключевых слов из текстов Региональных СМИ выполнено с помощью библиотеки RuTermExtract[11] и набора правил постобработки. Он построен на анализе морфологических характеристик слов и словосочетаний и набора правил для извлечения ключевых слов. Для морфологического анализа в русскоязычной версии используется библиотека PyMorphy2 (Korobov 2015).

5 Заключение

В статье представлено краткое описание обновленной платформы НКРЯ с технологической точки зрения. Это обновление является важнейшим этапом 20-летнего развития Национального корпуса русского языка и является результатом внедрения комплексного подхода, соответствующего современным практикам и стандартам развития корпусных ресурсов, а также общим тенденциям цифрового развития общества. Более подробное описание произошедших изменений можно увидеть в расширенной версии настоящей статьи.

Литература

[1] Зобнин А. И., Носырев Г. В. (2015). Морфологический анализатор MyStem 3.0 // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 300–310.

[2] Кузнецова А. И., Ефремова Т. Ф. (1986). Словарь морфем русского языка. Москва: Рус. яз., 1986.

[3] Савчук С. О. (2015). Корпус региональных газет России и зарубежья // Труды Института русского языка им. В. В. Виноградова. 2015. № 6. С. 163—193.

[4] Савчук С. О. и др. (2024). Национальный корпус русского языка 2.0: новые возможности и перспективы развития //Вопросы языкознания. Т. 2. С. 7-34.

[5] Сичинава Д. В. (2005). Национальный корпус русского языка: очерк предыстории // Национальный корпус русского языка: 2003—2005. М.: Индрик. С. 21—30.

[6] Сичинава Д. В. (2022). Корпус берестяных грамот как параллельный // Труды Института русского языка им. В. В. Виноградова. 2022. № 2 (32), 92-106.

[7] Тихонов А. Н. (2002). Морфемно-орфографический словарь. Москва: Астрель: АСТ, 2002.

[8] Anastasyev D., (2020). Exploring pretrained models for joint morphosyntactic parsing of Russian // Computational Linguistics and Intellectual Technologies: Papers from the Annual Conference “Dialogue”, volume 19. P. 1–12.

[9] Bird, St., Loper E., Klein E. (2009). Natural Language Processing with Python. O’Reilly Media Inc.

[10] Dozat T., Manning Ch. D. (2016). Deep Biaffine Attention for Neural Dependency Parsing https://arxiv.org/abs/1611.01734

[11] Droganova, K, Lyashevskaya O. (2018). Cross-tagset parsing evaluation for Russian // Digital Transformation and Global Society Third International Conference, DTGS 2018, St. Petersburg, Russia, May 30 – June 2, 2018, Revised Selected Papers, Part I / Ed. by Daniel A. Alexandrov, A. V. Boukhanovsky, A. V. Chugunov, Y. Kabanov, O. Koltsova. Issue 858. P. 380-390.

[12] Droganova, K, Lyashevskaya O., Zeman D. (2018). Data Conversion and Consistency of Monolingual Corpora: Russian UD Treebanks // Proceedings of TLT 2018 International Workshop on Treebanks and Linguistic Theories, 13-14 November 2018, Oslo, Norway. NEALT Proceedings Series. Linköping University Electronic Press, 2018. P. 52-65.

[13] Garipov T., Morozov D. Glazkova A. (2023). Generalization Ability of CNN-Based Morpheme Segmentation // 2023 Ivannikov Ispras Open Conference (ISPRAS), Moscow, Russian Federation. P. 58-62.

[14] Korobov M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. P. 320-332.

[15] Kutuzov, A., Kunilovskaya, M. (2018). Size vs. Structure in Training Corpora for Word Embedding Models: Araneum Russicum Maximum and Russian National Corpus. In: van der Aalst, W., et al. Analysis of Images, Social Networks and Texts. AIST 2017. Lecture Notes in Computer Science, vol 10716. Springer, Cham.

[16] Kutuzov A., Kuzmenko E. (2017) WebVectors: A Toolkit for Building Web Interfaces for Vector Semantic Models. In: Ignatov D. et al. (eds) Analysis of Images, Social Networks and Texts. AIST 2016. Communications in Computer and Information Science, vol 661.

[17] Lyashevskaya O. (2019). A reusable tagset for the morphologically rich language in change: A case of Middle Russian // Komp’juternaja Lingvistika i Intellektual’nye Tehnologii. P 422–434.

[18] Lyashevskaya O. et al., (2023). Disambiguation in context in the Russian National Corpus: 20 years later // Computational Linguistics and Intellectual Technologies Papers from the Annual International Conference “Dialogue” (2023). Issue 22. P. 307-318.

[11] https://github.com/igor-shevchenko/rutermextract

[19] Rehurek, R., Sojka, P. (2011). Gensim–python framework for vector space modelling. NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic, 3(2)).
[20] Shavrina T., Shapovalova O. (2017) TO THE METHODOLOGY OF CORPUS CONSTRUCTION FOR MACHINE LEARNING: «TAIGA» SYNTAX TREE CORPUS AND PARSER. in proc. of "CORPORA2017", international conference, Saint-Petersbourg.
[21] Sorokin, A., Kravtsova, A. Deep Convolutional Networks for Supervised Morpheme Segmentation of Russian Language // Ustalov, D., Filchenkov, A., Pivovarova, L., Žižka, J. (eds) Artificial Intelligence and Natural Language. AINL 2018. Communications in Computer and Information Science, vol 930. Springer, Cham.
[22] Zmitrovich, D., et al. (2023). A Family of Pretrained Transformer Language Models for Russian. ArXiv, abs/2309.10931.

Lexical Decision Task: Modality matters

Daniil A. Burov
Lomonosov Moscow State University,
Moscow, Russia
duurovd@mail.ru

Maria B. Panich
Lomonosov Moscow State University,
Moscow, Russia
shmak1280@gmail.com

Fedor A. Sadkovskii
Lomonosov Moscow State University,
Moscow, Russia
RAS Institute of Linguistics,
Moscow, Russia
feudor987@mail.ru

Olga V. Fedorova
Lomonosov Moscow State University,
Moscow, Russia
olga.fedorova@msu.ru

Sergey M. Shevelev
Lomonosov Moscow State University,
Moscow, Russia
sh-sm01@yandex.ru

Abstract

Lexical decision task is one of the most wide-spread methods used is psycholinguistic experiments, performed in either visual or auditory modality. The comparability of results acquired in the two modalities, however, has only recently received attention. In this paper, we present the results of two parallel experiments, one in the visual modality and one in the auditory modality, run on the same lexical material (words and pseudo-words). The reaction time distributions differ between the two modalities: mean reaction time is longer in the auditory modality. Mean reaction time for pseudo-words in significantly longer the mean reaction time for real words.

Keywords: lexical decision; visual modality; auditory modality; reaction time; pseudo-words
DOI: 10.28995/2075-7182-2025-23-74-85

Задача лексического решения: модальность имеет значение

Буров Д. А.
МГУ имени М. В. Ломоносова,
Москва, Россия
duurovd@mail.ru

Панич М. Б.
МГУ имени М. В. Ломоносова,
Москва, Россия
shmak1280@gmail.com

Садковский Ф. А.
МГУ имени М. В. Ломоносова,
Москва, Россия
Институт языкознания РАН,
Москва, Россия
feudor987@mail.ru

Федорова О. В.
МГУ имени М. В. Ломоносова,
Москва, Россия
olga.fedorova@msu.ru

Шевелев С. М.
МГУ имени М. В. Ломоносова, Москва, Россия
sh-sm01@yandex.ru

Аннотация

Задача лексического решения — один из самых распространенных методов проведения психолингвистических экспериментов, проводящийся в одной из двух модальностей: зрительной или слуховой. Сопоставимость результатов, полученных в разных модальностях, однако, начала последовательно изучаться только недавно. В настоящей работе описываются результаты параллельных экспериментов в зрительной и слуховой модальности на одном и том же лексическом материале (слов и псевдослов). Результаты показывают, что распределения времени реакции по словам и псевдословам отличается между модальностями: среднее время реакции в слуховой модальности оказывается больше. Кроме того, в обеих модальностях среднее время реакции на псевдослова значимо превышает среднее время реакции на слова.

Ключевые слова: задача лексического решения; зрительная модальность; слуховая модальность; время реакции; псевдослова

1 Введение

Задача лексического решения (ЗЛР, Lexical Decision Task) – один из самых распространенных методов проведения психолингвистических экспериментов. В психолингвистике этот метод используется для изучения устройства ментального лексикона человека, а также процесса распознавания слов. Однако сфера применения результатов этих тестов намного шире и не ограничивается областью психолингвистики. На успешность выполнения ЗЛР, то есть на скорость распознавания слов, влияет множество разнообразных свойств этих слов, начиная от частотности и заканчивая тональностью. Интегрированным индексом, который вбирает в себя все значимые свойства слов, служит **время реакции** испытуемых: чем сложнее доступ к некоторому слову, тем дольше испытуемые делают правильный выбор. Таким образом, время реакции на слова в ЗЛР является важным фактором, который стоит учитывать при любой лингвистической работе со словами, как в теоретических, так и в прикладных исследованиях.

В последние десятилетия для подобных нужд стали проводиться **мегаисследования** на разных языках, содержащие от 10 до 40 тысяч слов, на основе которых создаются соответствующие базы данных (БД). В частности, существуют БД для английского [1], французского [2], малайского [3], нидерландского [4] и китайского [5] языков; в работе [6] авторы собрали данные сразу для семи языков: английского, французского, нидерландского, малайского, баскского, каталанского и испанского[1]. Имея подобные данные о времени реакции, исследователи могут как проводить свои собственные исследования, изучая роль того или иного фактора, так и выбирать из этих БД слова с определенным временем реакции.

Однако до самого последнего времени исследователи обращали недостаточное внимание на важные различия в методике проведения экспериментов на ЗЛР. Дело в том, что ЗЛР может проводиться в двух модальностях: как в зрительной (ЗМ), когда стимулы подаются в письменной форме, так и в слуховой (СМ), когда подача стимулов происходит в звуковой форме. Метод ЗЛР в ЗМ появился раньше [7] и в психолингвистике распространен заметно шире, чем его слуховой аналог. При прохождении подобного эксперимента задача испытуемого состоит в том, чтобы как можно быстрее классифицировать появляющиеся на экране стимулы на слова и неслова родного языка. В первых же исследованиях в 1970-х годах было установлено, что многие свойства слов (например, их частотность или наличие нескольких не связанных друг с другом значений) влияют на скорость принятия решения. Так, оказалось, что более частотные слова или омонимы опознаются быстрее и правильнее, чем менее частотные слова или слова, имеющие только одно значение.

Метод ЗЛР в СМ был введен на десятилетие позже, в работе [8], но с тех пор тоже стал достаточно популярен, особенно при работе с такими группами испытуемых, для которых невозможна ЗЛР в зрительной модальности – например, маленькими детьми или пациентами с расстройствами аутистического спектра (РАС). Необходимо было получить ответ на вопрос, насколько сопоставимы результаты при прохождении ЗЛР в разных модальностях. В статье [9] авторы впервые, насколько нам известно, последовательно описали различия между этими двумя методиками. Оказалось, в частности, что частотность является самым важным фактором для ЗЛР в ЗМ, но не СМ, для которой наиболее важна длительность звучания слова. Другим важным открытием стал тот факт, что испытуемые в ЗЛР в ЗМ осуществляют свой выбор еще до того, как идентифи-

[1] Подобных БД русского языка пока не существует; более того, насколько нам известно, настоящее исследование является первым исследованием в методологии ЗЛР, выполненным на материале русского языка.

цируют само слово, а испытуемые в ЗЛР в СМ обычно дожидаются конца звучания слова. Наконец, авторы статьи показали, что испытуемые в ЗЛР в ЗМ тратят больше времени на те слова, в которых так называемая «точка уникальности» слова[2] появляется рано, а испытуемые в ЗЛР в СМ, наоборот, анализируют такие слова быстрее; этот результат говорит о том, что, по-видимому, зрительный анализ происходит параллельно, а слуховой – последовательно.

Подытоживая вышесказанное, можно заключить, что за более чем полвека исследований был накоплен большой экспериментальный материал, который востребован во многих лингвистических исследованиях. Тем не менее, оказывается, что решены еще не все важные методологические вопросы. Так, насколько нам известно, еще не проводились параллельные эксперименты в обеих модальностях с использованием одних и тех же слов. Основной вопрос, на который необходимо получить ответ в таком исследовании, состоит в том, насколько в двух экспериментах (не) совпадут упорядоченные по времени реакции списки слов. Другими словами, если исследователь имеет возможность провести эксперимент в любой модальности, на чем должен основываться его выбор. Ответу на этот вопрос и посвящена данная работа.

2 Эксперимент 1. Зрительная модальность

2.1 Испытуемые

В эксперименте приняли участие 26 человек в возрасте от 17 до 55 лет (17 женщин), все – носители русского языка с нормальным или скорректированным до нормального зрением.

2.2 Стимульный материал

Все 144 слова русского языка, использованные в эксперименте, являлись стилистически нейтральными и общеупотребительными, согласно частотности по НКРЯ (ruscorpora.ru); все слова состояли из двух слогов. Кроме того, было сконструировано 144 «псевдослова» – последовательности букв, не являющиеся словами русского языка, однако напоминающие их по слоговой структуре и фонотактике и также двусложные[3]. Таким образом, был использован основной набор из 288 стимулов, а также 17 тренировочных стимулов (12 слов и 5 псевдослов).

2.3 Процедура

Эксперимент был запрограммирован в программе PsychoPy.Builder (версия 2024.1.5). Слова и псевдослова предъявлялись испытуемым на экране компьютера в случайном порядке. Испытуемым предлагалось как можно быстрее и точнее определить, является ли последовательность букв, которую они видят, словом русского языка или нет. После принятия решения испытуемый нажимал на соответствующую клавишу, и на экране появлялось следующее слово. Основному блоку эксперимента предшествовал тренировочный блок. Прохождение эксперимента занимало в среднем 8–10 минут. Зависимые переменные – правильность ответа и время реакции.

Материалы экспериментов выложены по адресу https://github.com/feudor2/LD-Russian-Auditory-Visual-2025.

[2] Точка уникальности слова — это позиция элемента слова (буквы или звука в зависимости от модальности), в момент достижения которой при восприятии символ за символом слева направо слово становится однозначно определяемым.

[3] Псевдослова подбирались следующим образом: у существующих русских слов заменялись одна или несколько букв. Согласные заменялись на другие согласные, гласные – на другие гласные. При этом в получившемся слове сохранялись инициали/финали, типичные для отдельных групп русских лексем. Так, псевдослово *пестце* образовано с опорой на слово *сердце*, при этом у слова сохранилась финаль, присутствующая в таких русских словах, как *блюдце, рыльце* и т.п. Кроме того, в процессе подбора псевдослов мы старались сделать так, чтобы в инициали слов были представлены большинство букв русского алфавита.

2.4 Результаты и обсуждение

Перед тем, как перейти к анализу результатов, мы исключили из выборки испытуемых с большим количеством выбросов и ошибок. Для определения выбросов мы руководствовались критерием (1), а исключение испытуемых из выборки по количеству ошибок происходило по критерию (2).

(1) Время реакции не должно превышать межквартильный размах более чем в 1.5 раза (метод IQR). Выбросами мы считали те ответы испытуемых, которые попадали под этот критерий хотя бы по одному параметру — для слова или для испытуемого. Исключались те испытуемые, количество выбросов в ответах которых превышало 20%[4].

По этому критерию мы исключили двоих испытуемых. Максимальное количество выбросов после удаления – 17.4%.

(2) Количество ошибок в ответах не должно превышать 15%. Под этот критерий попал один испытуемый. Его результат также был максимальным по числу выбросов после исключения испытуемых в соответствии с первым критерием, поэтому финальный максимум по количеству выбросов составил 11.5%, самая низкая точность — 92%.

После удаления выбросов выборка содержала результаты 21 испытуемого (средний возраст = 33.8, SD = 15.1). Среднее количество выбросов составило 16.3 (5.7%, SD = 7.8), средняя точность испытуемых — 98.3% (SD = 1.8%). Распределения времени реакции по испытуемым приводятся на рис. 1 (для слов) и рис. 2 (для псевдослов).

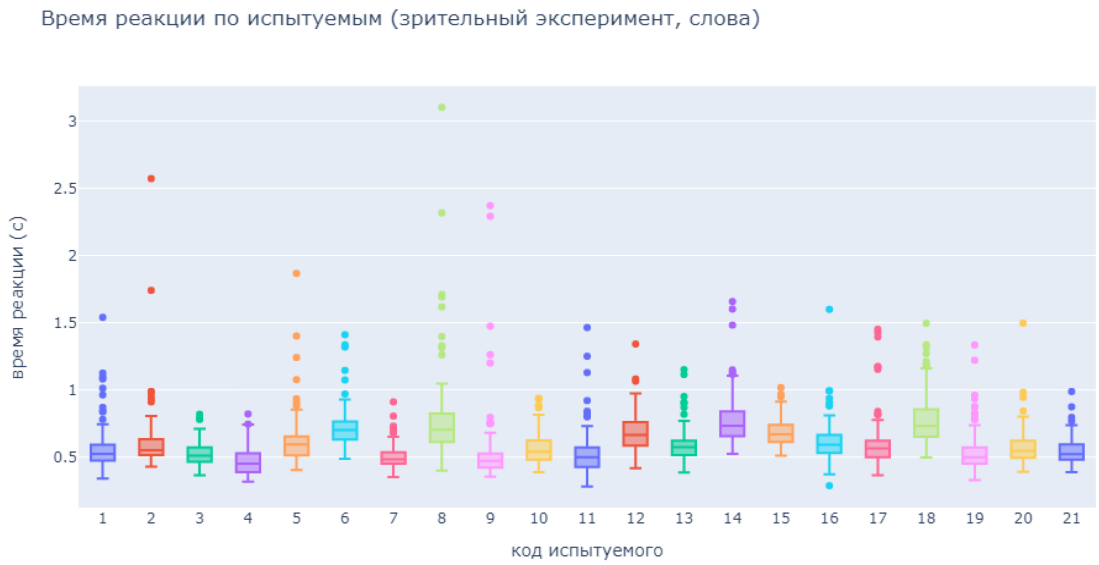


Рисунок 1: распределение времени реакции на слова по испытуемым

[4] Межквартильный размах, равный 1.5, а также количество выбросов в 20% – это стандартные пороговые значения в подобных исследованиях.

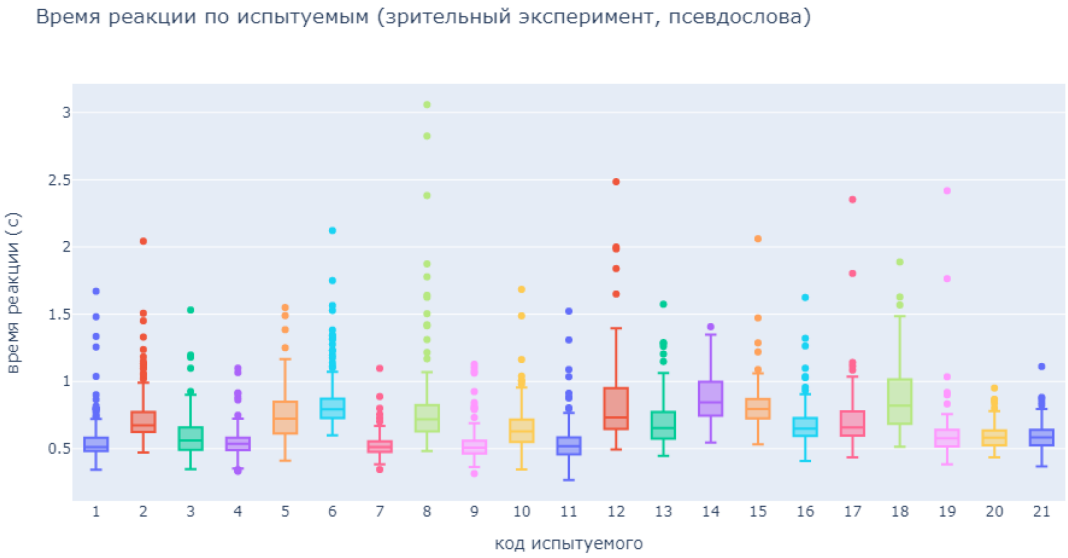


Рисунок 2: распределение времени реакции на псевдослова по испытуемым

Ошибки. Всего было сделано 42 ошибки на слова (14 испытуемых), в среднем испытуемые делали ошибки в 3 словах (SD = 2.4, med = 2.5, mod = 1 (4), min = 1, max = 10). Количество слов с ошибками — 37 (25.7%), максимальное число ошибок допущено в словах *йогурт* (3), *фонарь, ванна, зебра* (2).

На псевдослова 17 испытуемых сделали 60 ошибок, в среднем 3.5 псевдослова с ошибками (SD = 3.3, med = 2, mod = 1 (6), min = 1, max = 13). Число псевдослов с ошибками — 50 (34.7%), максимальное число допущено в псевдослове *овог* (4), по 2 ошибки в псевдословах *плинка, урул, бекня, барнак, вогой, рукса* и *лоча*.

Выбросы. Выбросы имеют место в 72 (50%) словах, всего 97, не более 3 на испытуемого; и в 60 (41.7%) псевдословах, всего 73, не более 2 на испытуемого. Слова с максимальным числом выбросов (3): *язык, зебра, доска, доктор*.

Сравнение времени реакции на слова и псевдослова. Для уменьшения шума в данных мы удалили все выбросы в соответствии с определением выброса из критерия (1) выше. Размер выборки слов составил 2893, псевдослов — 2813 измерения. Значения времени реакции мы стандартизовали по испытуемому для уменьшения влияния индивидуальных особенностей. Тем не менее, даже после удаления выбросов и стандартизации распределения времени реакции на слова и псевдослова значимо отличались от нормального: по критерию Шапиро-Уилка p-value $<< 0.001$ в обоих случаях. Для проверки гипотезы об однородности применялись два непараметрических критерия — U-критерий Манна-Уитни и критерий Колмогорова-Смирнова. Использовались односторонние версии критериев, и за нулевую гипотезу принимались предположения $\mathbb{P}(x < y) < 0.5$ (где $x \in X \sim F_{words}$ и $y \in Y \sim F_{pseudo}$) в критерии Манна-Уитни и $\forall x$: $F_{words}(x) <= F_{pseudo}(x)$ (где F_{words} и F_{pseudo} — функции распределения слов и псевдослов). Таким образом, при начальном предположении о стохастически большем времени реакции на псевдослова, чем на слова, мы ожидаем, что первый критерий покажет необходимость отклонить нулевую гипотезу, а второй — недостаточность оснований ее отклонить. Поскольку значения, стандартизованные по словам, нивелируют различие между словами и псевдословами, осмысленно сравнить изначальные значения и значения по стандартизации только по испытуемому. В обоих случаях мы получили результаты, согласующиеся с нашими ожиданиями: при сравнении выборок изначальных значений $D_n = 0.001$, p-value = 0.9957 (H_0 принимается); U = 5080037, p-value $<< 0.001$ (H_0 отклоняется) и $D_n = 0.0004$, p-value = 0.9992 (H_0 принимается); U = 5486942, p-value $<< 0.001$ (H_0 отклоняется) при сравнении значений, стандартизованных по испытуемому. Гистограммы плотности и графики эмпирической функции распределений нестандартизованных значений времени реакции приведены на рис. 3 и 4 соответственно.

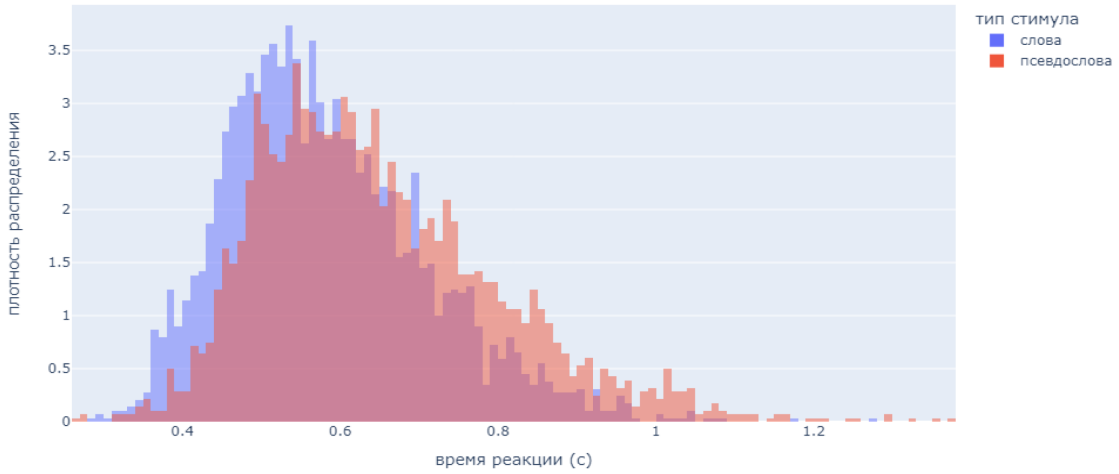


Рисунок 3: гистограммы плотности распределения времени реакции на слова (синий) и псевдослова (красный)

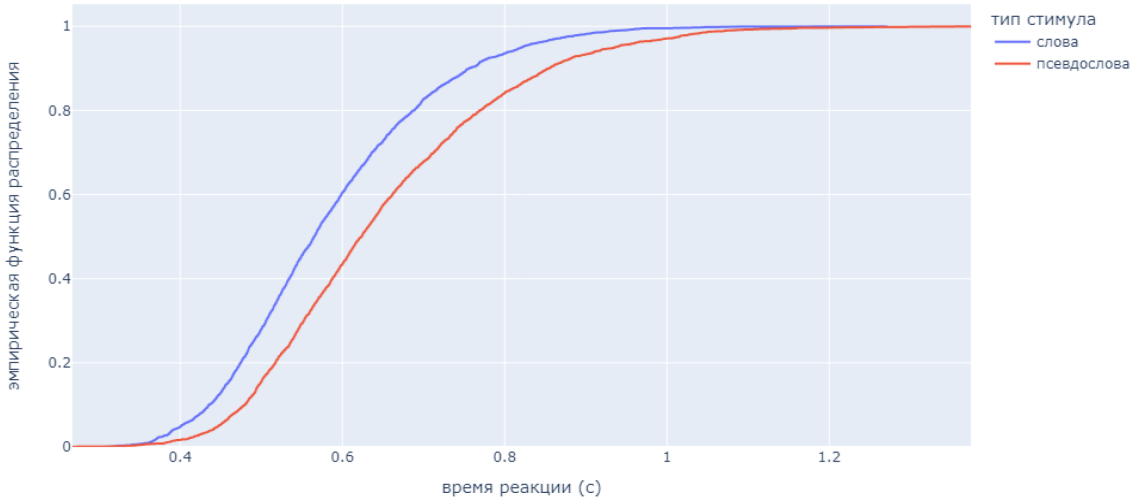


Рисунок 4: графики эмпирических функций распределения времени реакции на слова (синий) и псевдослова (красный)

Таким образом, в эксперименте в ЗМ среднее время реакции на псевдослова оказывается значимо больше, чем среднее время реакции на слова. Кроме того, на псевдословах испытуемые делают больше ошибок.

3 Эксперимент 2. Слуховая модальность

3.1 Испытуемые

В эксперименте приняли участие 27 человек в возрасте от 18 до 69 лет (24 женщины), все — носители русского языка с нормальным слухом.

3.2 Стимульный материал

В эксперименте был использован тот же набор из 288 основных и 17 тренировочных стимулов, озвученный вторым автором статьи. Средняя длительность слов — 0.943 сек, псевдослов — 1.002 сек.

3.3 Процедура

В ходе эксперимента в программе PsychoPy.Builder записи проигрывались в случайном порядке, испытуемым предлагалось определить, что именно они только что услышали – слово или несуществующее псевдослово. Важно отметить, что переслушивать слово было нельзя, и фактически испытуемые воспринимали слово лишь один раз, в то время как в эксперименте с ЗМ они имели возможность его перечитывать, пока принимали решение. Все остальные условия совпадали с ЗМ.

3.4 Результаты и обсуждение

Перед анализом результатов мы проделали ту же процедуру с двумя критериями, что и при оценке результатов эксп. 1.

По первому и второму критериям было исключено по одному испытуемому. После удаления выбросов выборка содержала результаты 24 испытуемых (средний возраст = 27.5, SD = 15.6). Среднее количество выбросов составило 16.3 (21.3%, SD = 11.7), средняя точность испытуемых — 95.5% (SD = 3%). Максимальное количество выбросов достигло 53 (18.4%), минимальная точность — 88.5%. Распределения времени реакции по испытуемым приводятся на рис. 5 (для слов) и рис. 6 (для псевдослов).

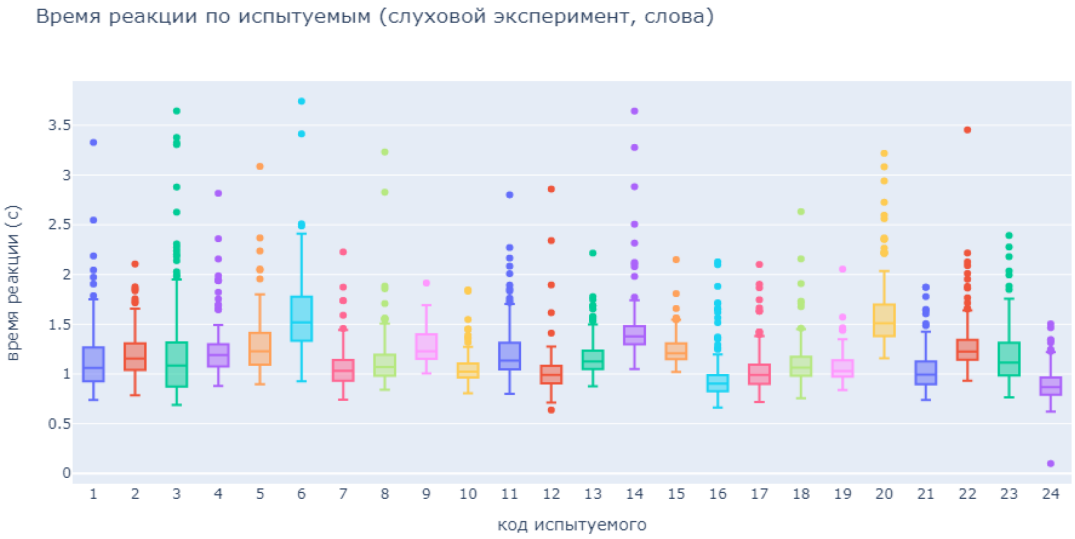


Рисунок 5: распределение времени реакции на слова по испытуемым

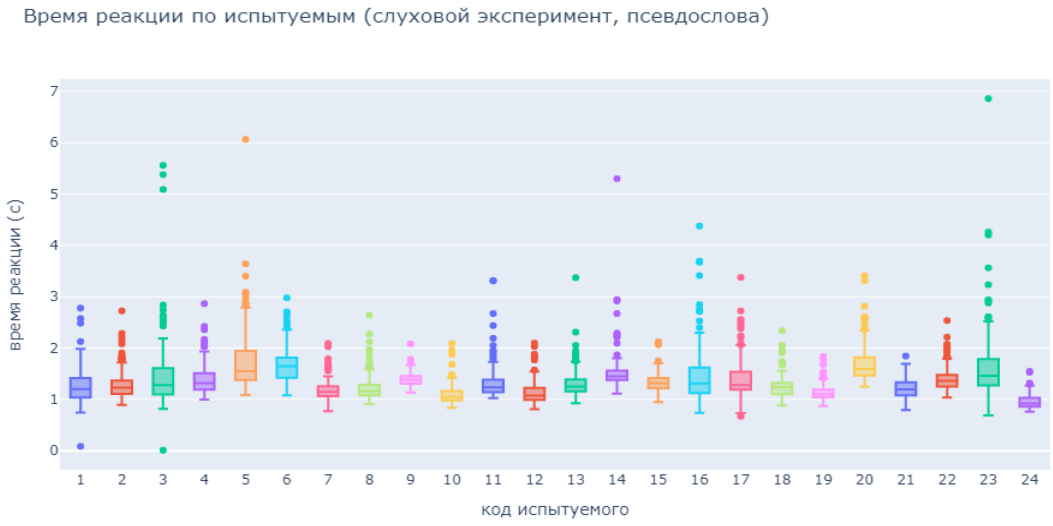


Рисунок 6: распределение времени реакции на псевдослова по испытуемым

Ошибки. Всего была сделана 101 ошибка на слова (21 испытуемый), в среднем испытуемые делали ошибки в 4.8 словах (SD = 3.3, med = 5, mod = 1 (4), min = 1, max = 13). Количество слов с ошибками — 59 (41%), максимальное число ошибок допущено в словах *вата, ведро* (7), *санки, пакет* (4), *пицца, краска, перо, этаж, точка, вагон, шорты* (3).

На псевдослова 23 испытуемых сделали 208 ошибок, в среднем 9 псевдослов с ошибками (SD = 9.4, med = 5, mod = 3 (4), min = 1, max = 33). Число псевдослов с ошибками — 98 (68%), максимальное число допущено в псевдослове *кулун* (10), *бодор* (8), *пленто*, *смилок* (7), *глека* (6).

Выбросы. Выбросы имеют место в 98 (68%) словах, всего 168; и в 104 (72.2%) псевдословах, всего 165, от 1 до 4 на испытуемого. Слова с максимальным числом выбросов: *перо* (6) и *щетка, город, земля, кольцо* (4).

Сравнение времени реакции на слова и псевдослова. После удаления всех выбросов размер выборок слов составлял 3159, псевдослов — 3243 измерений. В обоих случаях распределение стандартизованных данных нельзя было считать близким к нормальному (по критерию Шапиро-Уилка p-value << 0.001). Оба непараметрических критерия однородности показали значимый сдвиг значений распределений времени реакции в выборке псевдослов по сравнению с выборкой слов: $D_n << 0$, p-value = 1 (H_0 принимается); U = 6874521, p-value << 0.001 (H_0 отклоняется) для нестандартизованных значений (в соответствии с гипотезами, определенными при анализе результатов эксперимента 1) и $D_n << 0$, p-value > 0.9999 (H_0 принимается); U = 7268319, p-value << 0.001 (H_0 отклоняется) для значений, стандартизованных по испытуемому. Гистограммы плотности и графики эмпирической функции распределений нестандартизованных значений времени реакции приведены на рис. 7 и 8 соответственно.

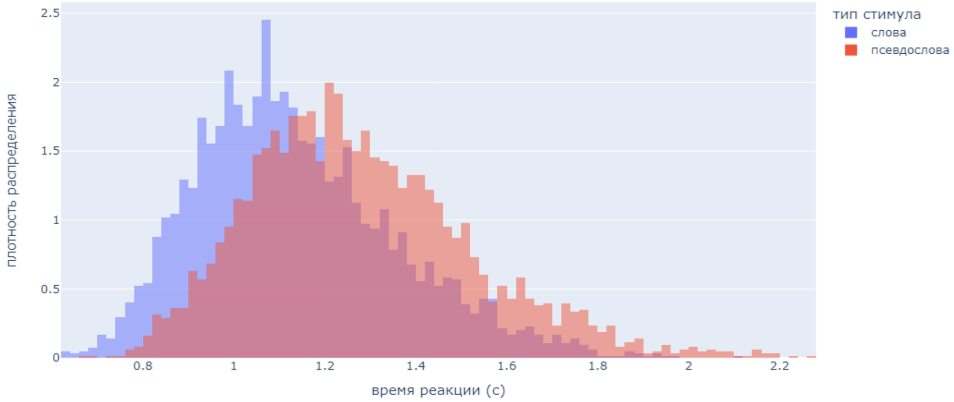


Рисунок 7: гистограммы плотности распределения времени реакции на слова (синий) и псевдослова (красный)

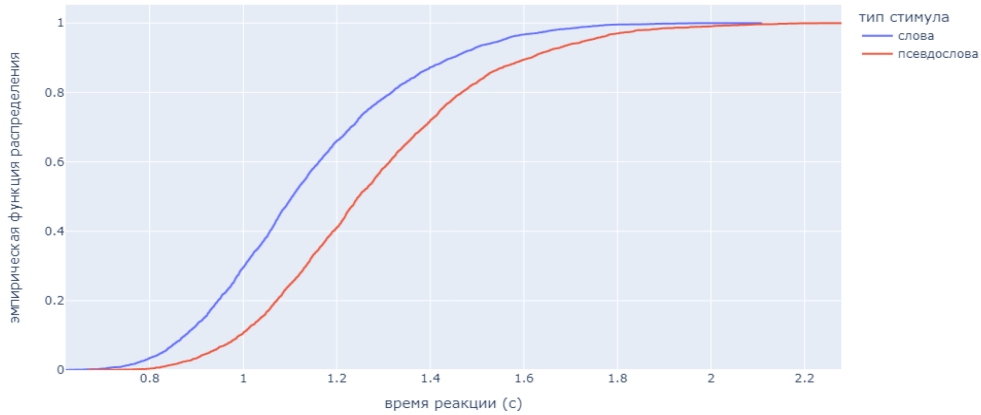


Рисунок 8: графики эмпирических функций распределения времени реакции на слова (синий) и псевдослова (красный)

Таким образом, как и в случае ЗМ, в СМ среднее время реакции на псевдослова значимо превышает среднее время реакции на слова. Закономерность с бóльшим числом ошибок в псевдословах также сохраняется, причем разница становится ощутимее: ошибок на псевдословах в среднем почти в два раза больше.

4 Сравнение двух модальностей. Общее обсуждение результатов

В разделах 2 и 3 мы сравнили между собой выборки слов и псевдослов и выявили, что распределение времени реакции на псевдослова смещено в большую сторону относительно распределения времени реакции на слова. В этом разделе мы покажем, что распределения времени реакции по словам и псевдословам также различаются и между модальностями.

В табл. 1 приводится информация о центральных значениях для двух условий в ЗМ и СМ, на рис. 9-10 показаны соответствующие гистограммы плотности распределения, а на рис. 11-12 — эмпирические функции распределения.

	ЗМ		СМ	
	среднее	медиана	среднее	медиана
слова	0.583	0.565	1.136	1.104
псевдослова	0.648	0.624	1.278	1.247

Таблица 1: центральные значения распределений в двух экспериментах

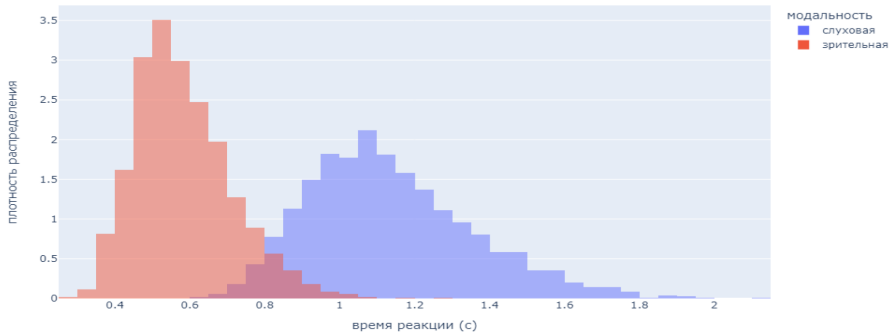


Рисунок 9: гистограммы плотности распределения времени реакции на слова в ЗМ (красный) и СМ (синий)

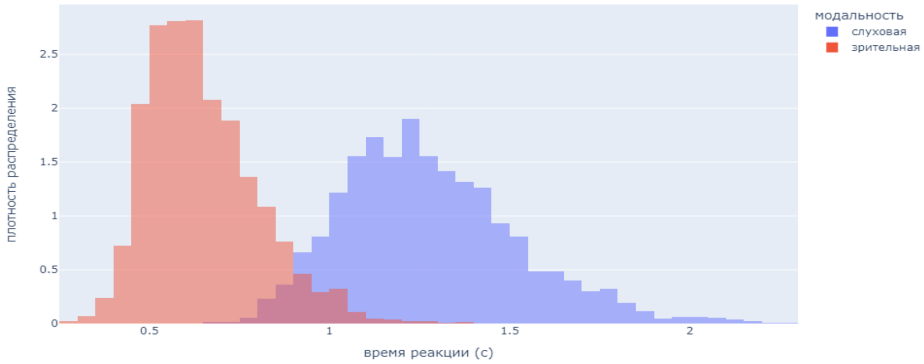


Рисунок 10: гистограммы плотности распределения времени реакции на псевдослова в ЗМ (красный) и СМ (синий)

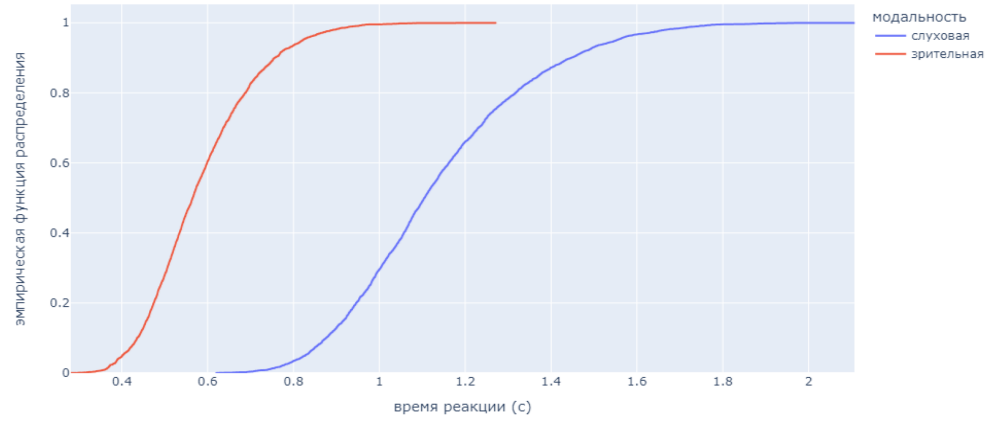


Рисунок 11: графики эмпирических функций распределения времени реакции на слова в ЗМ (красный) и СМ (синий)

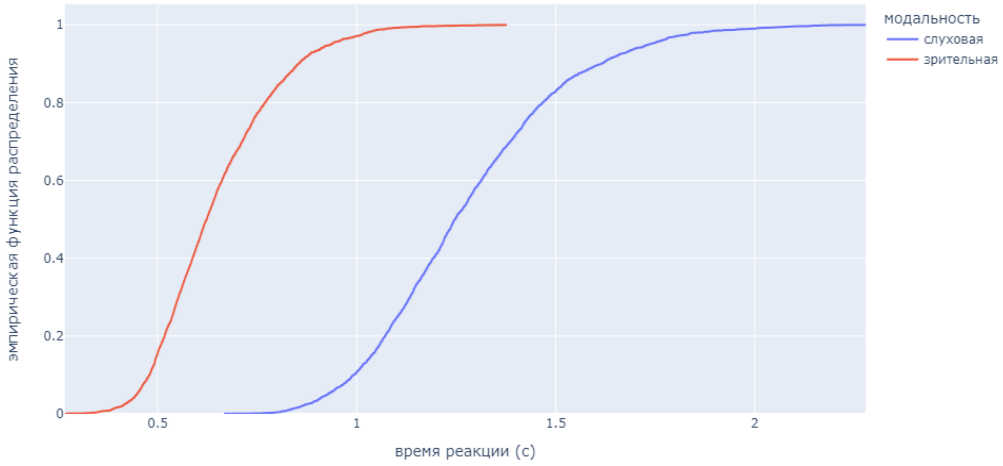


Рисунок 12: графики эмпирических функций распределения времени реакции на псевдослова в ЗМ (красный) и СМ (синий)

Время реакции в данных эксперимента в СМ сконцентрировано вокруг значения >1 и с большей концентрацией значений времени реакции в направлении их роста.

Сравнение выборок значений времени реакции однозначно указывает на неоднородность и значимость сдвига распределения в СМ (слова: $D_n << 0$, p-value = 1; U = 9296450, p-value << 0.001; псевдослова: $D_n << 0$, p-value = 1; U = 8794094, p-value << 0.001).

Чтобы сказать, насколько сложность условий в разных модальностях сопоставима, нужно оценить, насколько одинаково упорядочиваются стимулы по среднему времени реакции на них. Чтобы среднее значение отражало также и упорядочивание слов по сложности для каждого испытуемого, мы рассчитали средние значения времени реакции, стандартизованного по испытуемому, для всех слов и псевдослов. Полученные средние значения мы сравнили с помощью рангового критерия Уилкоксона. Для выборок слов в двух модальностях разница оказалась значима (T = 3763, p-value = 0.004), но не значимой для псевдослов (T = 5130, p-value = 0.858).

Кроме этого, мы оценили влияние на время реакции ряда признаков стимулов, таких как частотность, валентность [10], расстояние Левентштейна до ближайших слов (графическое и фонетическое) по [11], а также точки уникальности слова (графические и фонетические). Но, в отличие от более ранних работ, например, [9], нам не удалось выявить значимых корреляций (возможно, из-за небольшого размера наших выборок).

Таким образом, и в зрительной, и в слуховой модальности среднее время реакции на псевдослова значимо превышает среднее время реакции на слова. При этом само по себе среднее время реакции в СМ оказывается больше: поскольку слуховой анализ, в отличие от зрительного, происходит не параллельно, а последовательно, испытуемым приходилось прослушивать каждый звуковой стимул до конца, то есть на каждое (псевдо)слово в среднем тратилась по меньшей мере секунда, в то время как на среднее время восприятия зрительных стимулов никаких дополнительных ограничений наложено не было. Однако именно результаты в СМ характеризуются большим разбросом значений – возможно, это связано с тем, что слуховой стимул проигрывается только единожды, и принятие решения о его принадлежности к (псевдо)словам принимается уже после этого (и может, например, включать в себя вспоминание прослушанного слова), в то время как ЗМ характеризуется постоянным доступом к стимулу.

5 Заключение

В данной работе мы впервые провели два параллельных эксперимента на задачу лексического решения в зрительной и слуховой модальности с одним и тем же набором стимулов. Проанализировав полученные результаты, мы пришли к выводу, что большее распространение ЗМ при проведении подобных исследований оказывается не случайным. ЗМ удобнее сразу по нескольким параметрам, начиная от заметно более простого процесса подготовки стимулов и заканчивая постоянным доступом испытуемых к стимулам во время принятия решения, а также меньшим средним значением времени реакции. Однако выбор той или иной модальности влечет за собой и содержательные различия: так, в данном исследовании при сравнении списков слов и псевдослов, ранжированных по времени реакции, оказалось, что иерархии (упорядочивание по времени реакции) псевдослов в обеих модальностях сопоставимы, в то время как иерархии слов, напротив, значимо отличаются друг от друга. И хотя выявление факторов, влияющих на разное восприятие слов в разных модальностях, – предмет дальнейших исследований, принципиальная важность выбора модальности при работе с реальными словами налицо: обращение к слуховой модальности позволяет получить значимо иные результаты.

Результаты данной работы, на наш взгляд, имеют и несколько практических последствий, важных для будущих исследований. Сейчас мы остановимся на одном, актуальном для продолжения настоящей работы, имеющей отношение к разработке методов исследования слухо-речевых нарушений при расстройствах аутистического спектра (РАС). Как уже было отмечено во введении, люди с РАС, особенно дети, могут проходить тест на ЗЛР только в слуховой модальности. Более того, для подобных групп испытуемых важнейшую роль играет правильно составленный датасет. Прежде чем проводить экспериментальные исследования с детьми с РАС, необходимо тщательно протестировать стимульный материал на выборке здоровых взрослых людей, а затем провести экспериментальные исследования на выборке здоровых детей, проверив и скорректировав при необходимости стимульный материал. Именно ради возможности проведения в будущем таких крупномасштабных исследований необходимо развивать отдельные базы данных непосредственно для СМ, так как для подготовки исследований в СМ нельзя использовать данные, полученные в ЗМ – модальность имеет значение.

Благодарности

Исследование выполнено в рамках проекта РНФ №25-18-00739 «Низкоуровневые слухо-речевые нарушения при аутизме: мозговые механизмы, последствия для развития речевых функций, инновационные методы диагностики дефицита». Авторы выражают благодарность нашим испытуемым, принявшим участие в эксперименте, коллегам по проекту РНФ, а также анонимным рецензентам, высказавшим ряд важных замечаний.

Литература

[1] Balota, D. A., Yap, M. J., Cortese, M. J., Hutchison, K. I., Kessler, B., Loftis, B. et el. The English Lexicon Project // Behavior Research Methods, 39. — 2007. — P. 445–459.

[2] Ferrand, L., New, B., Brysbaert, M., Keuleers, E., Bonin, P., Méot, A. et al. The French Lexicon Project: Lexical decision data for 38,840 French words and 38,840 pseudowords // Behavior Research Methods, 42. — 2010. — P. 488–496.

[3] Yap, M. J., Rickard Liow, S. J., Jalil, S. B., & Faizal, S. S. B. (2010). The Malay Lexicon Project: A database of lexical statistics for 9,592 words // Behavior Research Methods, 42. — 2010. — P. 992–1003.

[4] Keuleers, E., Diependaele, K., & Brysbaert, M. Practice effects in large-scale visual word recognition studies: A lexical decision study on 14,000 Dutch mono- and disyllabic words and nonwords // Frontiers in Language Sciences, 1, 2010. — P. 174.

[5] Tse, C.-S., Yap, M. J., Chan, Y.-L., Sze, W. P., Shaoul, C., & Lin, D. The Chinese Lexicon Project: A mega-study of lexical decision performance for 25,000+ traditional Chinese two-character compound words // Behavior Research Methods, 49. — 2016. — P. 1503–1519.

[6] Dufau, S., Duñabeita, J. A., Moret-Tatay, C., McGonigal, A., Peeters, D., Alario, F. X. et al. Smart phone, smart science: How the use of smartphones can revolutionize research in cognitive science. PLoS ONE, 6(9), e24974. — 2011.

[7] Rubenstein H., Garfield L., Millikan J. A. 1970. Homographic entries in the internal lexicon // Journal of Verbal Learning and Verbal Behavior 9. — 1970. — P. 487–494.

[8] McCusker, L.X., Hillinger, M.L., & Bias, R.G. Phonological recording and reading // Psychological Bulletin, 89(2). — 1981. — P. 217–245.

[9] Ferrand, L., Méot, A., Spinelli, E. et al. MEGALEX: A megastudy of visual and auditory word recognition // Behavioral Research Methods, 50. — 2018. — P. 1285–1307.

[10] Sylvester, T., Liebig, J., & Jacobs, A. M. Neural correlates of affective contributions to lexical decisions in children and adults // Scientific Reports, 11(1). — 2021. — P. 945.

[11] Yarkoni, T., Balota, D., & Yap, M. Moving beyond Coltheart's N: A new measure of orthographic similarity // Psychonomic bulletin & review, 15(5). — 2008. — P. 971–979.

Derivation of Aktionsarten from (non-)directional verbs of motion: posing restrictions on restrictions

Chuikova O. Iu.
Herzen State Pedagogical University of Russia
oxana.chuykova@gmail.com

Gorbova E. V.
independent researcher
elenagorbova12@gmail.com

Abstract

The paper examines prefixal derivatives of Russian motion verbs that represent perfective Aktionsarten. The focus is on the statistical analysis of motion verbs as a specific subset of Aktionsarten characterized by their derivational features. Both sets (verbs of motion and Aktionsarten) are included in the database of Russian prefixed verbs developed by the authors based on the Dictionary of Russian language. One of the aims of the study is to verify the statement defended in previous works that the stem of a motion verb imposes restrictions on the possible Aktionsarten derived from it. This verification will be based on both dictionary data and the data from the Russian National Corpus and the Runet.

Keywords: verbs of motion; Aktionsarten; derived perfectives; prefixal and circumfixal derivation
DOI: 10.28995/2075-7182-2025-23-86-102

Деривация способов действия от (не)однонаправленных глаголов движения: ограничения и условия их смягчения

Чуйкова О. Ю.
РГПУ им. А. И. Герцена
oxana.chuykova@gmail.com

Горбова Е. В.
независимый исследователь
elenagorbova12@gmail.com

Аннотация

В статье рассматриваются приставочные дериваты русских глаголов движения – перфективные представители (морфемно-характеризованных) способов действия. В фокусе внимания находится статистический анализ глаголов движения как специфического по своим деривационным характеристикам подмножества способов действия. Оба упомянутых множества являются составной частью базы данных русских приставочных глаголов, созданной авторами по Словарю русского языка (Малому академическому словарю). Одной из задач исследования явилась также верификация известных по литературе вопроса ограничений на вхождение дериватов от глаголов движения в способы действия с опорой как на словарные данные, так на данные из Национального корпуса русского языка и Рунета.

Ключевые слова: глаголы движения; способы действия; производные перфективы; префиксальная и циркумфиксальная деривация

1 Preliminary notes

Russian verbs of motion (hereinafter, MVs) comprise a compact group of 15 or 14 pairs (see [7] for a review of interpretations) when treated narrowly as non-derivative verbs (simplexes). We accept the list of 14 pairs from [12]: *бежать – бегать, брести – бродить, везти – возить, вести – водить, гнать(ся) – гонять(ся), ехать – ездить, идти – ходить, катить(ся) – катать(ся), лезть – лазить, лететь – летать, нести(сь) – носить(ся), плыть – плавать, ползти – ползать, тащить(ся) – таскать(ся)*. The specificity of this group is that both members of a pair function as imperfectives, and their semantic distinction is interpreted through the binary opposition of (uni)directionality / multi/non-directionality and/or singularity / plurality, e.g., *идти*^{Dir} *– ходить*^{NonDir}, *ехать*^{Dir} *–*

ездить^{NonDir}, *бежать*^{Dir} – *бегать*^{NonDir}[1]. The subject of this study is prefixal and circumfixal derivatives from 14 pairs of MVs provided they are perfective and represent one of the Aktionsarten (hereinafter, As), e.g. *избегаться*^{PFV} (excessive-multiple A) from the simplex *бегать*^{NonDir} or *забродить*^{PFV} (inchoative A) from *бродить*^{NonDir}.

The peculiarities of derivation of a number of As from MVs were mentioned in the literature.

In [9: 319-322], a series of observations is presented regarding the relationship between *Dir/NonDir* derivatives and As (*совершаемости* in Isachenko's terms). From *Dir*, an ingressive *совершаемость* is formed with the prefix *по-*, as in *Он встал и пошёл к выходу* 'He stood up and went to the exit'. In certain forms, the corresponding derivatives express a resultative meaning, as in *Отца нет дома, он пошёл в город* 'The father is not at home, he went to the city' [9: 320]. For *NonDir* derivatives, [9: 323] notes the realization of the following As: inchoative (*заходить*); delimitative (*походить*); perdurative (*проходить*); evolutive (or ingressive-intensive; *расходиться*); and saturative (*находиться вдоволь*).

The interaction of *Dir*/*NonDir* with perfectivizing prefixes is discussed in [15: 100–103; 16: 6–11]. It is noted that *Dir* stems are compatible with all prefixes, while *NonDir* stems are not. When a prefix is compatible with both, their meanings differ: *Dir* typically conveys a spatial meaning (except for *по-*, which serves as an ingressive marker), whereas *NonDir* conveys a non-spatial meaning. Many prefixed derivatives of *NonDir* represent a specific A [15: 101]. An intersection between the semantics of *NonDir* prefixes and external prefixes, which are considered markers of As in the literature [14], is also noted [16: 11]. The authors observe that with *Dir* stems, the prefix *по-* has an ingressive meaning, while *NonDir* stems can take *за-* (inchoative), *по-* (delimitative), *от-* (finitive), *про-* (perdurative), and *на-* (cumulative). In fact, limitations regarding the compatibility of *Dir* and *NonDir* with certain prefixes[2] and their semantics have been formulated.

However, as far as we know, no comprehensive study of the peculiarities of the formation of As from 28 MVs along with the analysis of the system of perfective As (in any known edition) has been carried out so far.

This study aims to systematically examine the formation of perfective As from MVs using a continuous sample from the Small Academic Dictionary. The sample will be supplemented at certain stages with the data on derivatives not included in the dictionary but available in other sources, such as corpora and the Runet. Additionally, the study seeks to verify existing notions regarding limitations on forming perfective As based on the stem type of the root MVs. The ultimate goal is to develop a more accurate model of the intersection between MVs and aspectual grouping of As in Russian verb grammar.

Following the preliminary notes in Section 1, Section 2 describes the research material and outlines the methodology used in this study. Section 3 discusses the productivity of MVs as a basis for forming As, addressing limitations related to two types of stems and the impact of expanding linguistic sources on the findings. Finally, Section 4 summarizes the results of the study.

2 Materials and methods

The material of the study is a continuous sample of prefixal perfectives of 19952 lexemes (lexeme as "a word considered in one of its available meanings" [1: 55]), compiled according to the Small Academic dictionary, also known as MAS [4]. The sample includes 1378 MV-derivatives[3], 304 of which represent As. The derivation from the MVs was determined on formal grounds, so the sample also includes derivatives from non-literal meanings of MVs (e.g., *отбродить* (about wine), *провести* <an event>). The issue of semantic annotation of MVs and their derivatives as non-literal, or metaphoric, is covered in Appendix D.

We utilize the As system, which is based on specific principles (see [5; 3]). As denote "different types of semantic modifications of the verb expressed by certain formal means" [15: 110]. An A is defined as a linguistic sign comprising a signifier (a word-formation marker, i.e. an affix or a combination of

[1] Here we use the notations *Dir* and *NonDir* for (uni)directional (like *идти*) and non-directional (like *ходить*) MVs, borrowed from [16]; one of the important questions for [16] about the direction of derivation in a number of aspectual pairs of MVs derivatives: whether *приходить*^{PFV} is derived from *прийти*^{PFV}, as argued in [16] following [9], or from *ходить*^{NonDir}, as in [8], [10] or [6], is not addressed here, as it is irrelevant for the purely perfective material of the study.

[2] As for circumfixal As, the derivation from *Dir* only is postulated for durative-negative *до-...-ся* in [11].

[3] The principle we established in the annotation is that a derivative at any stage of its derivation is considered an MV-derivative if its original stem is one of the 28 simplex verbs (14 pairs of *Dir* and *NonDir* from [12]).

affixes) and a signified, which represents an aspectually relevant semantic modification. Ambiguity in either component is analyzed through the identity of the linguistic sign. Multiple meanings associated with a single marker are classified as polysemy if there is non-trivial semantic similarity, or as homonymy if such similarity is absent. When different signifiers share similar semantics, synonymy is posited. Consequently, combining derivatives with different affixes within a single A is prohibited. Using this framework and classifications from [15: 110-135; 13: 141-167], we compiled a working list of 22 As (see Appendix A) based on prefix perfectives in MAS. While this list may not be theoretically ideal, it suffices for our research objectives. For perfective MV-derivatives, the list is reduced to 20, excluding the least productive *в*(*о*)*з*-delimitative (e.g., *взгрустнуть*) and *при*-attenuative (*приоткрыть*).

Further, a few composites (*снизойти*), denominal (*занавозить*) and polyprefixal derivatives like *по-вы-лезти, по-вы-таск-ивать* derived at the second or third derivation step from the prefixal perfective or the secondary imperfective (27 lexemes of *по*-distributive and 4 lexemes of *по*-attenuative) were withdrawn from the material. Since all the above derivatives are not directly formed from initial *Dir* or *NonDir*, their analysis in terms of the derivational stem is more complex than that of other MV-derivatives, so it is reasonable to consider them separately. For comparability of MV-derivatives with other prefixal perfectives, composites, nominal and polyprefixal derivatives were also removed from the original sample. As a result, the sample of prefixal perfectives was reduced to 18226 lexemes, of which MV-derivatives amounted to 1304 units, with 270 of them representing 19 As.

3 Productivity of As derived from MVs

3.1 A general overview of productivity of As derived from MVs

We will start by examining the overall picture of As productivity from MVs, then move on to the peculiarities of individual As.

First, we note that the proportion of verbs representing As in the MVs subset does not differ significantly from that for the whole set of prefixal verbs: about 20% (3771 out of 18226 in the whole sample and 270 out of 1304). However, individual As derived from MVs show higher or lower productivity than those derived from the rest of the verbs in the sample.

Let us consider the As for which this difference is statistically significant (see Appendix B). The increased level of productivity in the set of MVs derivatives is observed for the following As: *до*-completive[4] (*долезть*, *добродить*), *о(бо)*-distributive-summary (*облазить*, *обежать*), *по*-ingressive (*побежать*, *полететь*), *раз(о)-/рас*-distributive-summary (*растащить*, *разбежаться*) and *у-...-ся*-excessive-intensive (*убегаться*, *уходиться*). The following As are characterized by a lower level of productivity among the MV-derivatives: *за*-inchoative (*забегать*, *занестись* (about a chicken), *пере*-distributive (*перетаскать*, *перевестись* ‘disappear’), *пере*-repetitive (*переходить* <when playing chess>), *по*-delimitative (*полазить*, *поводить*), *по*-distributive (*потаскать* <everything, much>) and *под*-attenuative (*подвезти* ‘to have a bit of luck’).

Comment

1. Most of the above-mentioned As demonstrate the ability to derive from both *Dir* and *NonDir*. Thus, we can question the statement in the literature regarding the “fixation” of a number of As exclusively to one or another type of MV stem.

2. Ingressive *по-* shows increased productivity for MVs, while inchoative *за-* demonstrates decreased productivity; this indicates a more complex relationship between the two markers of the beginning of a situation than their complementary distribution over *Dir* and *NonDir* stems suggests. Overall, ingressive *по-* is characterized by limited productivity, with 46 lexemes according to MAS. Outside of MVs, this A is formed from verbs with the semantics of movement (see [15: 115-116]), often synonyms of *Dir* (*потопать*, *поскакать*, *пошлепать*, *повлечь*), or, in isolated cases, from verbs of state (*полюбить*, *поверить*). In contrast, inchoative *за-* forms one of the most numerous A (after the delimitative *по-*, see [2]) As, with 480 lexemes according to MAS.

It is also interesting that synonymous As with different prefixes show the same tendency to productivity within MV-derivatives: the distributive-summary *о(бо)-* and *раз(о)-/рас-* are characterized by increased productivity, distributive *по-* and *пере-*, attenuative *по-*, *под*(*о*)- and *pri-* – by decreased

[4] Here, the completive A is the result of combining the final-completive, completive-partitive, and terminative-locative As identified in [13].

productivity (up to the lack of attenuatives with *no-* and *при-* among MVs). The observed picture may be a consequence of both formal (at least it concerns *no*-distributive and attenuative As, mostly forming a subset of polyprefixal MVs, which are left out of consideration) and semantic reasons. The latter require further investigation, which is beyond the scope of this study.

3.2 Productivity of As from *Dir* и *NonDir* in MAS

We will now turn our attention to the peculiarities of the derivation of As from *Dir* and *NonDir*.

Overall, derivation from *Dir* is more productive, yielding 955 derivatives compared to 349 from *NonDir*. However, within the subset of As, *NonDir* produce more derivatives: 147 from *NonDir* versus 123 from *Dir*. The difference in the distribution of derivatives from the two stem types within the subset of As and beyond is statistically significant (see Table 1). This finding aligns with existing literature, which suggests that deriving As from *NonDir* is more common, while forming As from *Dir* is more challenging.

	Dir	***NonDir***	**Total**	**% *NonDir***
non As[5]	833	202	1034	19,54%
As	123	147	270	54,44%
Total	955	349	1304	26,76%
The difference is statistically significant, $\chi 2 = 131{,}332$, $p << 0{,}01$				

Table 1: Distribution of derivatives from *Dir* and *NonDir* in and beyond the As subset

We will now examine the set As derived from *Dir* and *NonDir* in more detail. Table 2 presents the number of lexemes for each A formed from *Dir* and *NonDir*, first sorted in descending order by the "*Dir*" column and then in ascending order by the "*NonDir*" column. Fig. 1 presents this data in a more visually accessible format.

	As	***Dir***	***NonDir***	**Total**
1	*до-*: completive	30	7	37
2	*раз*(*о*)-/*рас-*: distributive-summary	24	2	26
3	*по-*: ingressive	23	0	23
4	*на-*: cumulative	19	12	31
5	*о(бо)-:* distributive-summary	8	8	16
6	*раз(о)/с-...(-ся):* ingressive-intensive	7	7	14
7	*про-:* perdurative	3	24	27
8	*пере-:* distributive	2	8	10
9	*за-:* inchoative	2	13	15
10	*под*(*о*)-: attenuative	1	0	1
11	*за-...-ся*: excessive-durative	1	2	2
12	*до-...-ся*: durative-negative	1	4	5
13	*от*(*о*)-: finitive	1	16	17
14	*по-*: delimitative	1	23	24
15	*пере-*: repetitive	0	1	1
16	*по-*: distributive	0	1	1
17	*из*(*о*)/*с-…-ся*: excessive-multiple	0	2	2
18	*у-*: excessive-intensive	0	6	6
19	*на-*: saturative	0	11	11

Table 2: Number of As from *Dir* and *NonDir* according to MAS

[5] The MV-derivatives not included in the As are mainly perfectives with prefixes that realize their spatial meanings (*вбежать*, *забрести*, *отплыть*).

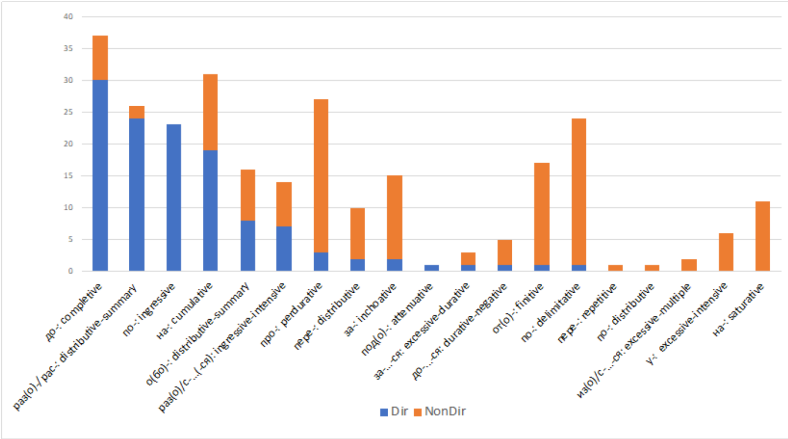


Figure 1: Distribution of As from *Dir* and *NonDir* according to MAS

Fig. 1 shows that there are more derivatives from *NonDir* than derivatives from *Dir*. According to MAS, lexemes belonging to 14 out of 19 As are found among derivatives from *Dir*, while representatives of 17 As are found among derivatives from *NonDir* (only the *по*-inchoatives and *под*-attenuatives are missing). At the same time, some of the derivatives from *Dir* belong to those As for which such derivation is postulated to be impossible in the literature. This applies for the *на*-cumulative (*набежать* 1. To congregate, accumulate, gather somewhere, *натащить* 1. Bring in some (usually large) amount of sth., *по*-delimitative (*понести* 5. Carry for a while sb, sth), *про*-perdurative (*проплыть* 3. Swim/sail for a period of time), *за*-inchoative (*занестись*[2] Start laying (about a chicken)).

The association plot[6] (Fig. 2) shows which As are significantly over- and underrepresented within *Dir* and *NonDir*. The As are labeled with numbers according to their numbering in Table 2.

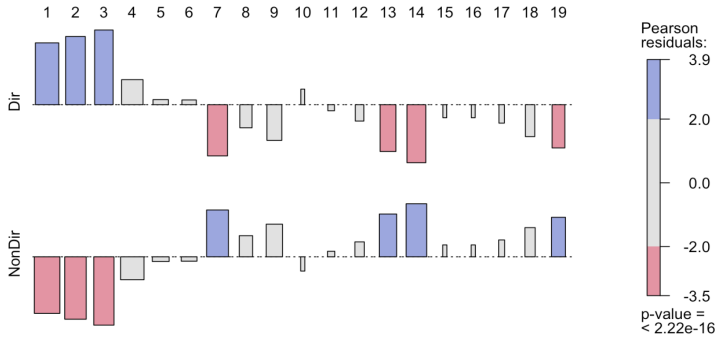


Figure 2: Association of *Dir/NonDir* and As according to MAS

According to Fig. 2, seven As show a significant association with the stem type. These are As that include non-unique lexemes (more than 10) with a strong tendency to derive from *Dir* or *NonDir*. The *до*-completive, *раз*(о)/*рас*-distributive-summary and *по*-ingressive are overrepresented when derived from *Dir* and underrepresented when derived from *NonDir*. In contrast, for the *про*-perdurative, *по*-delimitative, *от*(*о*)-finitive, and *на-...ся*-saturative, we observe an increase in productivity when derived from *NonDir*, while productivity decreases when derived from *Dir*.

To some extent, the data from MAS challenge the notion that certain As are strictly "fixed" to one of the stem types of MVs – either *Dir* or *NonDir*. Nevertheless, there remains a strong tendency for specific As to be formed predominantly from a single stem. Table 3 illustrates that when considering pairs of *Dir/NonDir* as the unit of analysis, instances of derivation from both types are relatively rare.

[6] Association plots are built with the assoc() function from the "vcd" package in R.

Specifically, among the 153 cases of As formation from a given *Dir/NonDir* pair documented in MAS, 79 cases involve derivatives exclusively from *NonDir*[7], 43 cases involve derivatives solely from *Dir*, and 31 cases include derivatives from both stem types.

As	***Dir* only**	***NonDir* only**	***Dir*&*NonDir***	**Total**
до-: completive	8	0	6	**14**
про-: perdurative	1	11	2	**14**
по-: ingressive	14	0	0	**14**
по-: delimitative	0	13	1	**14**
на-: cumulative	5	0	8	**13**
раз(*о*)-/*рас-*: distributive-summary	10	0	2	**12**
от(*о*)-: finitive	0	11	1	**12**
на-: saturative	0	11	0	**11**
за-: inchoative	0	8	2	**10**
раз(*о*)-/*рас-*…(*-ся*): ingressive-intensive	2	4	2	**8**
о(*бо*)-: distributive-summary	2	2	4	**8**
пере-: distributive	0	6	1	**7**
у-: excessive-intensive	0	6	0	**6**
до-....-ся: durative-negative	0	3	1	**4**
за-....-ся: excessive-intensive	0	1	1	**2**
пере-: repetitive	0	1	0	**1**
из(*о*)/с-…-*ся*: excessive-multiple	0	1	0	**1**
под(*о*)-: attenuative	1	0	0	**1**
по-: distributive	0	1	0	**1**
Total	**43**	**79**	**31**	**153**

Table 3: Derivatives from *Dir*/*NonDir* pairs within As in MAS

The existence of a strong relationship between the stem types (*Dir* or *NonDir*) and the As they produce is further corroborated through the application of machine learning algorithms.

First, the application of t-SNE (t-distributed stochastic neighbor embedding) to the As derived from *Dir* and *NonDir* reveals an almost unmistakable[8] division of 28 non-derivative stems into two distinct groups, as illustrated in Fig. 3.

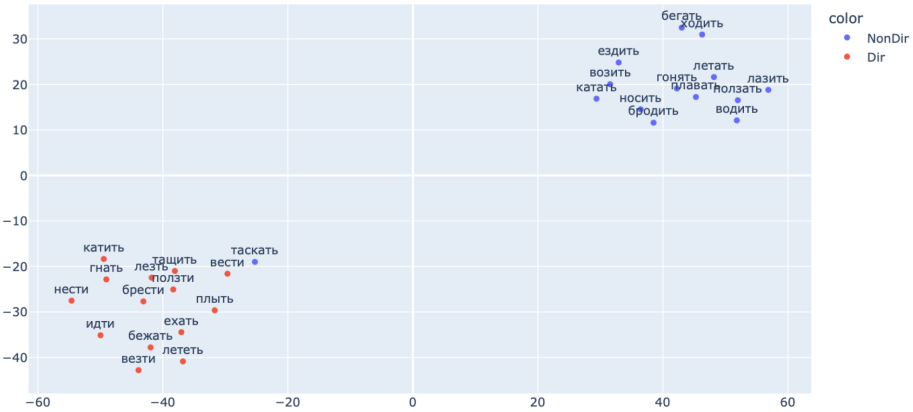


Figure 3: T-SNE projection for *Dir* and *NonDir* according to MAS

Furthermore, when training classification models (e.g., logistic regression and decision trees) on the data in which 270 lexemes recorded in MAS are treated as instances and As they belong to are used as features, the stem type (*Dir/NonDir*) is predicted with high metric values.

[7] Table 3 illustrates the presence of derivatives from paired *Dir* and *NonDir*, regardless of the number of lexemes involved.
[8] The only exception – *маскать* – requires further analysis.

	precision	recall	f1-score	support
NonDir	0.86	0.82	0.84	153
Dir	0.78	0.82	0.80	117
accuracy			0.82	270
macro avg	0.82	0.82	0.82	270
weighted avg	0.82	0.82	0.82	270

	precision	recall	f1-score	support
NonDir	0.86	0.83	0.84	152
Dir	0.79	0.82	0.80	118
accuracy			0.83	270
macro avg	0.82	0.83	0.82	270
weighted avg	0.83	0.83	0.83	270

Table 4: Classification reports for logistic regression (left) and decision tree (right) using As features with lexemes recorded in MAS as objects

Thus, the data exclusively from MAS confirm the relationship between the stem type and the derivation of As.

3.3 Sample expansion through As derived from MVs not recorded in MAS

The analysis in section 3.2 relies exclusively on verb lexemes recorded in MAS. However, examples of MV-derivatives MV – As absent from MAS – can be found in the Russian National Corpus (RNC)[9] and on the Runet (Yandex and/or Google), as shown (1)–(2). Analyzing these instances allows for preliminary conclusions about the actual productivity of As derived from MVs and the extent to which the representation in MAS and literature aligns with real speech practices.

(1) *А вот именно взять за руку и* ***повести*** *какое-то время не в сторону, а туда же, куда она сама шла ... не нашлось Мужика*.[10] (http://yanka.lenin.ru/stat/koblov.htm)

(2) *За ухи их брали и в воду тащили, они* ***поплывут-поплывут*** *и ворочаются.* [Светлана Василенко. Ген смерти (1997-2000)]

As a result of additional searches in the RNC and on the Runet for derivatives of MVs not recorded in MAS, we analyzed 153 instances of As formation, with the unit of analysis being the pair of *Dir* and *NonDir* stems. The distribution shows that most pairs allow for derivatives from both stems (100), while instances formed solely from *NonDir* are the least numerous (24), as illustrated in Table 4.

As	***Dir* only**	***NonDir* only**	***Dir*&*NonDir***	**Total**
до-: completive	1	0	13	**14**
про-: perdurative	0	2	12	**14**
по-: ingressive	14	0	0	**14**
по-: delimitative	0	0	14	**14**
на-: cumulative	0	0	13	**13**
раз(*о*)-/*рас-*: distributive-summary	10	0	2	**12**
от(*о*)-: finitive	0	3	9	**12**
на-: saturative	0	2	9	**11**
за-: inchoative	0	1	9	**10**
раз(*о*)-/*рас-*…(*-ся*): ingressive-intensive	2	4	2	**8**
о(*бо*)-: distributive-summary	1	0	7	**8**
пере-: distributive	0	5	2	**7**
у-: excessive-intensive	0	5	1	**6**
до-....-ся: durative-negative	0	0	4	**4**
за-....-ся: excessive- intensive	0	0	2	**2**
пере-: repetitive	0	0	1	**1**
из(*о*)/с-…*-ся*: excessive-multiple	0	1	0	**1**
под(*о*)-: attenuative	1	0	0	**1**
по-: distributive	0	1	0	**1**
Total	**29**	**24**	**100**	**153**

Table 5: Derivatives from *Dir*/*NonDir* pairs within As in the extended sample

[9] ruscorpora.ru

[10] We concur with the anonymous reviewer that non-codified usages arise by analogy with codified forms found in dictionaries, such as «надышаться», «отпеть (свои песни)», and «налетать (тысячу км)». We also agree that their acceptability can vary significantly among native speakers.

Considering data on derivatives from the RNC and the Runet, the quantitative ratio of derivatives from *Dir* and *NonDir* changes. Table 6 and Fig. 4 show a single preferred type of stem: *Dir* for the *no*-ingressive and *раз(о)-/рас*-distributive-summary As.

	As	***Dir***	***NonDir***	**Total**
1	*до-*: completive	30	14	44
2	*раз*(*о*)-/*рас-*: distributive-summary	24	2	26
3	*по-*: ingressive	23	0	23
4	*на-*: cumulative	19	18	37
5	*по-*: delimitative	19	23	42
6	*про-:* perdurative	13	25	38
7	*о(бо)-:* distributive-summary	11	9	20
8	*на-...-ся*: saturative	11	11	22
9	*за-:* inchoative	10	13	23
10	*от*(*о*)-: finitive	10	16	26
11	*раз(о)/с-...(-ся):* ingressive-intensive	7	7	14
12	*до-...-ся*: durative-negative	5	4	9
13	*пере-:* distributive	3	8	11
14	*за-...-ся*: excessive-durative	2	2	4
15	*под*(*о*)-: attenuative	1	0	1
16	*пере-*: repetitive	1	1	2
17	*у-...-ся*: excessive-intensive	1	6	7
18	*по-*: distributive	0	1	1
19	*из*(*о*)/*с-…-ся*: excessive-multiple	0	2	2

Table 6: Number of As from *Dir* and *NonDir* according to the extended sample

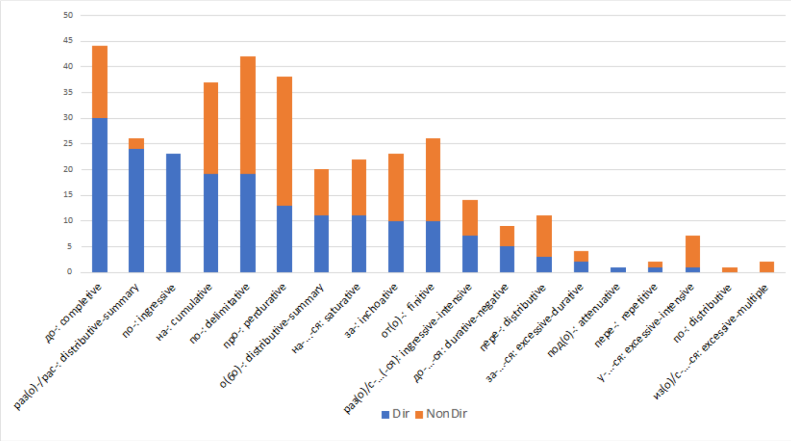


Figure 4: Distribution of As from *Dir* and *NonDir* according to the extended sample

Fig. 5 (with As numbered as in Table 6) confirms a significant association with the type of stem only for two As: the *no*-ingressive and *раз*(*о*)-/*рас*-distributive-summary. Both are overrepresented when derived from *Dir* stems and underrepresented when derived from *NonDir* stems. Notably, there are no As that demonstrate a high level of derivation from *NonDir*, contradicting the idea that many As are formed exclusively from *NonDir*.

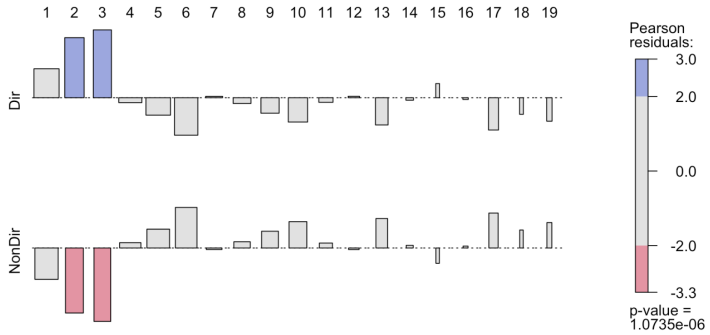


Figure 5: Association of *Dir/NonDir* and As according to the extended sample

Applying the same analytical tools used on the MAS data to the dictionary data supplemented with derivatives not recorded in MAS but found in the RNC and the Runet reveals a weakened dependence between the stem type and resulting As. Specifically, t-SNE analysis of the extended data shows that *Dir* and *NonDir* no longer exhibit a clear division into two groups (see Fig. 6).

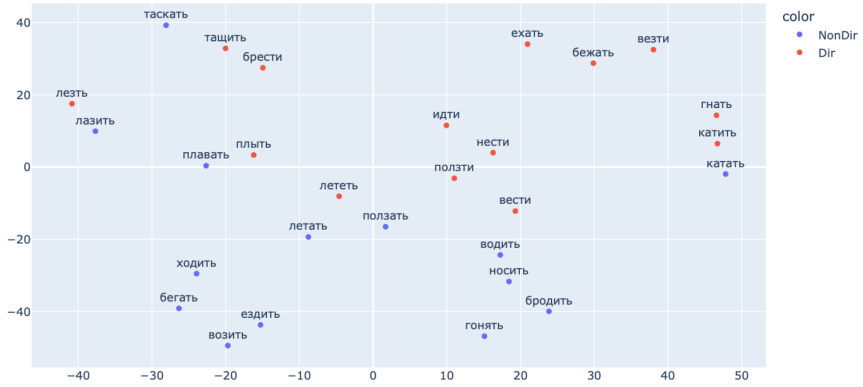


Figure 6: T-SNE projection for *Dir* and *NonDir* in the extended sample

The prediction quality of models trained on the extended data, considering only belonging to As, is significantly lower than that for the MAS data trained on the same features (see Table 7).

	precision	recall	f1-score	support		precision	recall	f1-score	support
NonDir	0.58	0.63	0.60	150	NonDir	0.71	0.60	0.65	192
Dir	0.71	0.66	0.68	202	Dir	0.59	0.71	0.65	160
accuracy			0.65	352	accuracy			0.65	352
macro avg	0.64	0.65	0.64	352	macro avg	0.65	0.65	0.65	352
weighted avg	0.65	0.65	0.65	352	weighted avg	0.66	0.65	0.65	352

Table 7: Classification reports for logistic regression (left) and decision tree (right) using As features with lexemes the from the extended sample as objects

The analysis of As derived from MVs found in MAS, supplemented with derivatives from the RNC and the Runet, reveals that relying solely on the stem type (*Dir* or *NonDir*) is insufficient to determine the formation of a specific A.

3.4 Consideration of extra features

Thus, when considering derivatives not recorded in MAS but found in the RNC and the Runet, the distribution of *Dir* and *NonDir* stems, as well as their association with certain As, differs from that

obtained through analysis of derivatives exclusively in MAS. Additionally, in some cases, it is evident that when forming As from lexemes with non-literal meanings of MVs, derivation from *Dir* is the only option, particularly noticeable in examples from the RNC and the Runet, e.g., (3)–(4).

(3) *Они решили* ***погнать*** *самогона ночью. Как-то заткнули неправильно, аппарат взорвался, муж весь обварился.* (https://culture.wikireading.ru/2408)

(4) *А эти последние три года я мужа упрашивала поработать, когда нас просили провести мероприятие. И Стас шел на это, чтобы я напелась и* ***навелась*** *праздника от души!* (https://vk.com/wall386557704_5459)

Let us analyze the relationship between the source of documentation and semantic features on the one side and the formation of As from *Dir* and *NonDir* on the other.

First, it is important to note that existing dictionary descriptions do not provide a strict tool for identifying the metaphoric meanings of MV-derivatives, as discussed in Appendix D. Consequently, the authors independently annotated based on intuition, achieving a Cohen's kappa agreement coefficient of 0.76. The first author adopted a broader view of semantic metaphoricity, with non-literal meanings comprising 30.68% (108 lexemes), while the second identified metaphoric meanings in 21.31% of derivatives (75 lexemes). It was decided not to unify the annotations but to treat both variants as distinct features, as their association with *Dir* and *NonDir* stems may differ.

Training that incorporates sources of derivatives (MAS, the RNC, the Runet) and their semantic features results in significantly better prediction quality compared to training solely on the derivatives' belonging to As. The best outcomes from the logistic regression and decision tree models are achieved when both features are included.

Logistic regression achieved the best results when trained on the source features (with the "Runet" binary feature removed due to strong linear dependence with "MAS") and the semantic characteristics based on the second author's annotations. The metric values are shown in Table 8.

	precision	recall	f1-score	support
NonDir	0.84	0.78	0.81	174
Dir	0.80	0.85	0.83	178
accuracy			0.82	352
macro avg	0.82	0.82	0.82	352
weighted avg	0.82	0.82	0.82	352

Table 8: Classification report for logistic regression using As, source and semantics as features with lexemes from the extended sample as objects

The decision tree shows the best prediction results when trained with the source and semantic annotations from both authors (see Table 9), with the source being the most significant feature at 28.25% (by Gini criterion).

	precision	recall	f1-score	support
NonDir	0.85	0.88	0.87	157
Dir	0.90	0.88	0.89	195
accuracy			0.88	352
macro avg	0.88	0.88	0.88	352
weighted avg	0.88	0.88	0.88	352

Table 9: Classification report for decision tree using As, source and semantics as features with lexemes from the extended sample as objects

Next, we examine the interaction between a stem type and semantic features in forming As from MVs. Among derivatives from *Dir*, the proportion of lexemes with non-literal meanings is higher in both annotation variants ("non-literal _1" and "non-literal _2"). This difference is statistically significant only for derivatives supplemented with verbs not recorded in MAS but found in the RNC and the Runet, and

marginally significant for As from MVs based on the dictionary data (see Table 10). The specific stems associated with literal / non-literal meanings also differ in the two annotation variants (see Appendix C).

non-literal_1							
	literal			non-literal			Statistical test
	Dir	*NonDir*	*% Dir*	*Dir*	*NonDir*	*% Dir*	
MAS	43	31	58.11	80	116	40.82	$\chi2$=5.798, p=0.016
Extended sample	75	33	69.44	115	129	47.13	$\chi2$=14.12, p<<0.01
non-literal_2							
	literal			non-literal			Statistical test
	Dir	*NonDir*	*% Dir*	*Dir*	*NonDir*	*% Dir*	
MAS	33	24	57.89	90	123	42.25	$\chi2$=3.827, p=0.0504
Extended sample	51	24	68.00	139	138	50.18	$\chi2$=6.844, p=0.008

Table 10: Distribution of literal and non-literal meanings of As derivatives from *Dir* and *NonDir*

Thus, the analysis of As derived from MVs, encompassing lexemes from MAS, the RNC, and the Runet, underscores the significance of the source of the verb and its semantic characteristics (literal vs. non-literal meaning) in evaluating As formation from *Dir* and *NonDir* stems.

4 Conclusion

Let us summarize the findings of the study on prefixal and circumfixal MV-derivatives in relation to morpheme-characterized As.

The analysis above shows that the restrictions on forming certain As from MVs, presented in [15: 102–103; 16: 6-9] as the complementary distribution of prefix meanings based on the stem type (*Dir*, *NonDir*), represent a tendency rather than an absolute rule. Both the comprehensive sampling of As from MAS and additional searches in the RNC and the Runet reveal perfective MV-derivatives that challenge established limitations on deriving As from these stems. Therefore, we can assert that there are no strict prohibitions on specific derivations, only tendencies that vary in intensity.

For MV-derivatives recorded in MAS, there is a strong tendency to form certain As from only one stem. However, when analyzing derivatives not recorded in MAS, this tendency weakens.

It is reasonable to assume that the stem type is not the sole factor influencing the derivation of As. The source of the derivative's documentation (MAS, the RNC, and/or the Runet) and its proximity to the literal meaning of the root MV also affect the derivational potential of *Dir* and *NonDir*.

Acknowledgements

The research was funded by a grant from the Russian Science Foundation, № 24-78-00172 (https://rscf.ru/en/project/24-78-00172/), with Oksana Chuikova as the grant recipient.

The authors would like to thank Ekaterina Enikeeva for her helpful advice on the application of machine learning methods.

References

[1] Apresyan Yu. D. Foundations of systemic lexicography [Osnovaniya sistemnoj leksikografii] // Linguistic picture of the world and systemic lexicography [Yazykovaya kartina mira i sistemnaya leksikografiya]. Ed. by Yu. D. Apresyan. — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul'tury], 2006. — P. 33–160.

[2] Chuikova Oksana Iu. (2020), On the secondary imperfectivation of *po*-perfectives in Russian [Ob osobennostyakh vtorichnoj imperfektivatsii glagolov s prefiksom *po-* v russkom yazyke], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue", Vol. 19 (26), pp. 160–176.

[3] Chuikova Oksana Iu. (2022), Partitive Genitive and Aktionsarten in Russian (based on dictionary and corpus data) [Roditel'nyj partitivnyj i sposoby glagol'nogo dejstviya v russkom yazyke (po slovarnym i korpusnym dannym)], Russian Linguistics, no. 46(1), pp. 25–54.

[4] Evgenieva Anastasija P. (ed.). Dictionary of Russian Language in 4 volumes [Slovar' russkogo jazyka v 4 tomah], 2nd ed. — Moscow: Academy of Science of USSR, Institute of Russian Language, 1981–1984. Access mode: http://feb-web.ru/feb/mas/mas-abc/14/ma239217.htm
[5] Gorbova Elena V., Chuikova Oksana Iu., Sharygina Sofya S. (2021), Imperfectivability of Russian prefixal perfectives: regularity and peculiarities [Imperfektiviruemost' russkikh pristavochnykh perfektivov: regulyarnost' i specifika], Topics in the study of language [Voprosy Jazykoznanija], no. 4, pp. 91–130.
[6] Gorbova Elena. V. (2023), Russian prefixed verbs of motion: aspectual correlativeness and aspect formation [Pristavochnye glagoly dvizheniia v russkom iazyke: vidovaia korreliativnost' i vidoobrazovanie], Scando-Slavica, no. 69(2), pp. 219–243.
[7] Gorbova Elena. V. (in press), Prefixal derivatives of Russian verbs of motion: word and aspect formation [Prefiksal'nye derivaty glagolov dvizheniia v russkom iazyke: slovo- i vidoobrazovaniie]. Vestnik of Saint Petersburg University. Language and Literature.
[8] Grammar of Russian language [Grammatika russkogo yazika]. Ed. by V. V. Vinogradov V. 1. — Moscow: Izdatel'stvo Akademii nauk SSSR [Publishing house of Academy of Sciences of the USSR], 1952.
[9] Isachenko A. V. The grammatical structure of the Russian language in comparison with Slovak. Morphology [Grammaticheskij stroj russkogo jazyka v sopostavlenii so slovackim. Morfologija]. Vol. 2. — Bratislava: Slovak Academy of Sciences Publ. [Izdatel'stvo Slovatskoi akademii nauk], 1960.
[10] Janda Laura. Perfectives from indeterminate motion verbs in Russian // Hasko V., Perelmutter R. (eds.). New Approaches to Slavic Verbs of Motion. — Amsterdam / Philadelphia: John Benjamins, 2010. — P. 125–139.
[11] Mel'cuk, I. (1990), Slovoobrazovanie v lingvisticheskix modelyax tipa "Smysl ⇔ Tekst" (Word formation in linguistic models of the type "Meaning ⇔ Text" (preliminary remarks)] // Metody formalne w opisie języków słowiańskich. Ed. by Z. Saloni. — Białystok: Dział Wydawnictw Filii UW w Białymstoku, 1990. — P. 47–74.
[12] Russian grammar [Russkaya grammatika]. Ed. by N. Yu. Shvedova. — Moscow: Nauka, 1980. Vol. 2.
[13] Shelyakin M. A. Category of aspectuality of the Russian verb [Kategoriya aspektual'nosti russkogo glagola]. — Moscow: URSS, 2008.
[14] Townsend, C. E. Russian Word Formation. — Ohio: Slavica Publishers Inc., 1975.
[15] Zalizniak Anna A., Mikaelyan Irina L., Shmelev Aleksey D. Russian aspectology: In defense of the aspectual pair [Russkaya aspektologiya: v zashchitu vidovoi pary] — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul'tury], 2015.
[16] Zalizniak Anna A. (2021), Russian Prefixed Verbs of Motion Revisited, Zeitschrift für Slawistik, no. 62(1), pp. 1–22.

Приложение A. List of As

As	Examples	Examples of MVs ГД
в(о)з(о)-/в(о)с-: delimitative	*всплакнуть, взгрустнуть*	———————
до-: completive	*дочитать, долить*	*долезть*, *добродить*
до-...-ся: durative-negative	*доиграться, долежаться*	*добегаться, доездиться*
за-: inchoative	*закричать, заплакать*	*забегать*, *занестись*
за-...-ся: excessive-durative	*засидеться, залюбоваться*	*забегаться, заходиться*
из-...-ся: excessive-multiple	*изголодаться, изленитьcя*	*избегаться*
на-: cumulative	*накупить, наловить*	*налетать, натащить*
на-...-ся: saturative/sative	*наесться, наиграться*	*набегаться, навозиться*
о(бо)-: distributive-summary	*опросить, обзвонить*	*обегать, облететь*
от(о)-: finitive	*отшуметь, отмучиться*	*отбродить, отплавать*
пере-: distributive	*перебудить, перестрелять*	*перевестись, перетаскать*
пере-: repetitive	*пересочинить, перезаказать*	*переходить*
по-: attenuative	*пообождать, попривыкнуть*	(*поразогнать*)
по-: delimitative	*погулять, поиграть*	*полазить, понести*
по-: distributive	*попрятаться, понаписать*	*потаскать* (*повылезти*)
по-: ingressive	*помчаться, полюбить*	*поехать*, *повести*
под-: attenuative	*подкрасить, подсократить*	*подвезти*
при-: attenuative	*припудрить, приоткрыть*	———————
про-: perdurative	*просидеть, проработать*	*проплавать, проездить*
раз(о)-/рас-: distributive-summary	*разворовать, растерять*	*растащить, разбежаться*
раз(о)/с-...(-ся): ingressive-intensive	*раскричаться, разволноваться*	*разбегаться, разойтись*
у-...-ся: excessive-intensive	*убегаться, упахаться*	*убегаться, уходиться*

Appendix B[11] The distribution of As within the subset of MVs and beyond it

	MVs			non-MVs			
	A	**non A**	**% A**	**A**	**non A**	**% A**	Statistical test
до-: completive	37	1267	2.84	209	16713	1.24	$\chi2=22.158$, $p<<0.01$
о(бо)-: distributive-summary	16	1288	1.23	15	16907	0.09	$\chi2=85.815$, $p<<0.01$
по-: ingressive	23	1281	1.76	46	16876	0.27	$\chi2=67.556$, $p<<0.01$
раз(о)-/рас-: distributive-summary	26	1278	1.99	117	16805	0.69	$\chi2=24.737$, $p<<0.01$
у-...-ся: excessive-intensive	6	1298	0.46	12	16910	0.07	$\chi2=14.853$, $p<<0.01$
за-: inchoative	15	1289	1.15	462	16460	2.73	$\chi2=11.245$, $p<<0.01$
пере-: distributive	10	1294	0.77	306	16616	1.81	$\chi2=7.108$, $p=0.0077$
пере-: repetitive	1	1303	0.08	211	16711	1.25	Fisher's exact test, $p<<0.01$
по-: delimitative	24	1280	1.84	532	16390	3.14	$\chi2=6.520$, $p=0.0107$
по-: distributive	1	1303	0.08	83	16839	0.49	Fisher's exact test, $p=0.0306$
под-: attenuative	1	1303	0.08	158	16764	0.93	Fisher's exact test, $p<<0.01$
до-...-ся: durative-negative	5	1299	0.38	36	16886	0.21	Fisher's exact test, $p= 0.21$
на-: cumulative	31	1273	2.38	368	16554	2.17	$\chi2=0.147$, $p=0.70$
от-: finitive	17	1287	1.30	130	16792	0.77	$\chi2=3.695$, $p=0.06$
про-: perdurative	27	1277	2.07	320	16602	1.89	$\chi2=0.124$, $p=0.72$
раз(о)-/рас-...-ся: ingressive-intensive	14	1290	1.07	180	16742	1.06	$\chi2=0$, $p=1$
из-...-ся: excessive-multiple	2	1302	0.15	27	16895	0.16	Fisher's exact test, $p=1$
за-...-ся: excessive-durative	3	1301	0.23	47	16875	0.28	Fisher's exact test, $p=1$
на-...-ся: saturative	11	1293	0.84	169	16753	1.00	$\chi2=0.160$, $p=0.69$

[11] Each row in the table of the Appendix B represents a transformation of a contingency table with positive and negative values for two features: belonging to a specific A (indicated in the first column) – "A" and "non-A", and belonging to MVs – "MVs" and "non-MVs". Thus, the representation of this A within the set of MVs is compared to similar data for this A outside of MVs (the sum of values in the "A" and "not A" columns within MVs is consistently 1304, representing the total volume of the analyzed MVs sample; the sum of values for "A" and "non-A" outside of MVs is consistently 16922, which equals the total number of analyzed prefixed perfectives minus the MVs (18226 minus 1304). The last column of the table presents the results of applying Pearson's chi-squared test (with Yates' correction, two-tailed variant) or Fisher's exact test (two-tailed variant; used when at least one value does not exceed 5). The color labeling indicates the presence or absence of statistical significance in the distribution differences (at $p < 0.05$) and the direction of the trend (increased or decreased productivity of the examined A within the group of MV-derivatives): in cases of statistical significance, this is determined by comparing the proportions of verbs for this A within the set of MVs and outside it (grey shading indicates no statistical significance, green indicates significant differences where the proportion of A within MVs is higher than that outside, and orange indicates significant differences where the proportion of A within MVs is lower than that outside).

Appendix C Frequency literal / non-literal (metaphoric) meanings of As derivatives in relation to root MVs

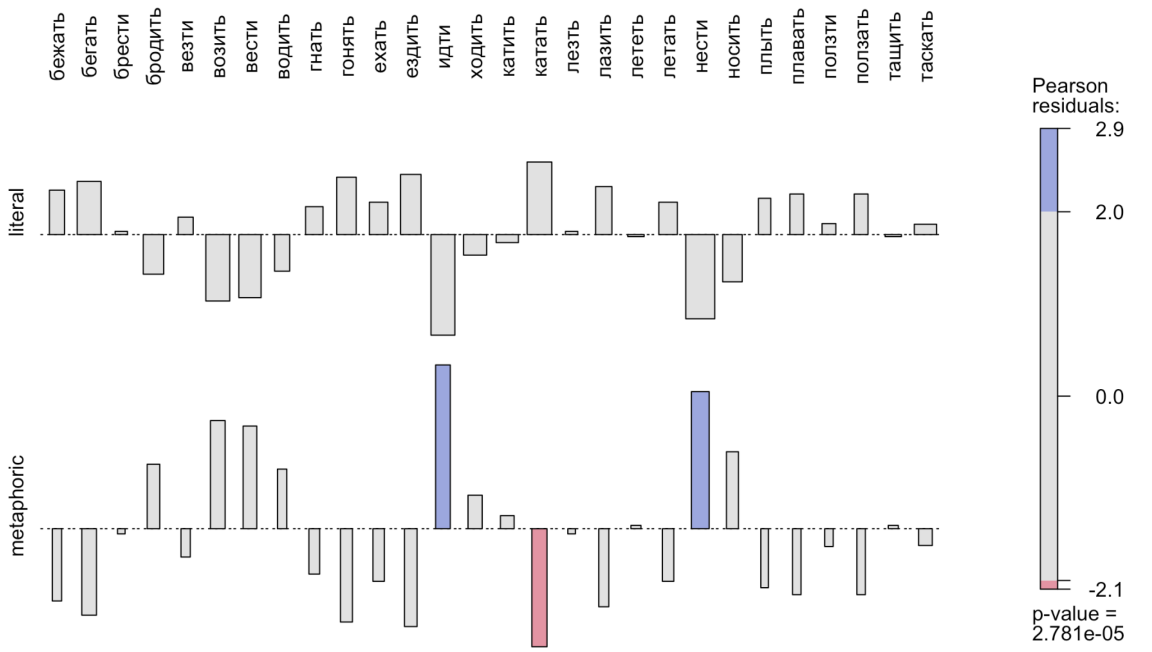


Figure C1: Association plot of root MVs and semantic characteristics based on MAS (Author 1)

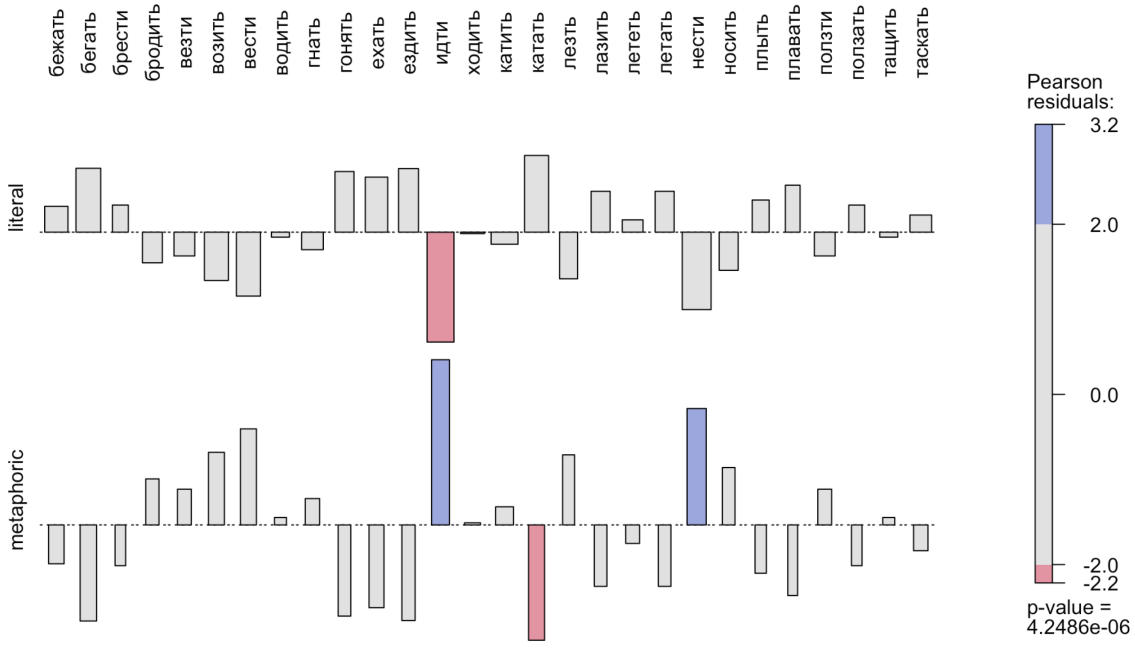


Figure C2: Association plot of root MVs and semantic characteristics based on the extended data (Author 1)

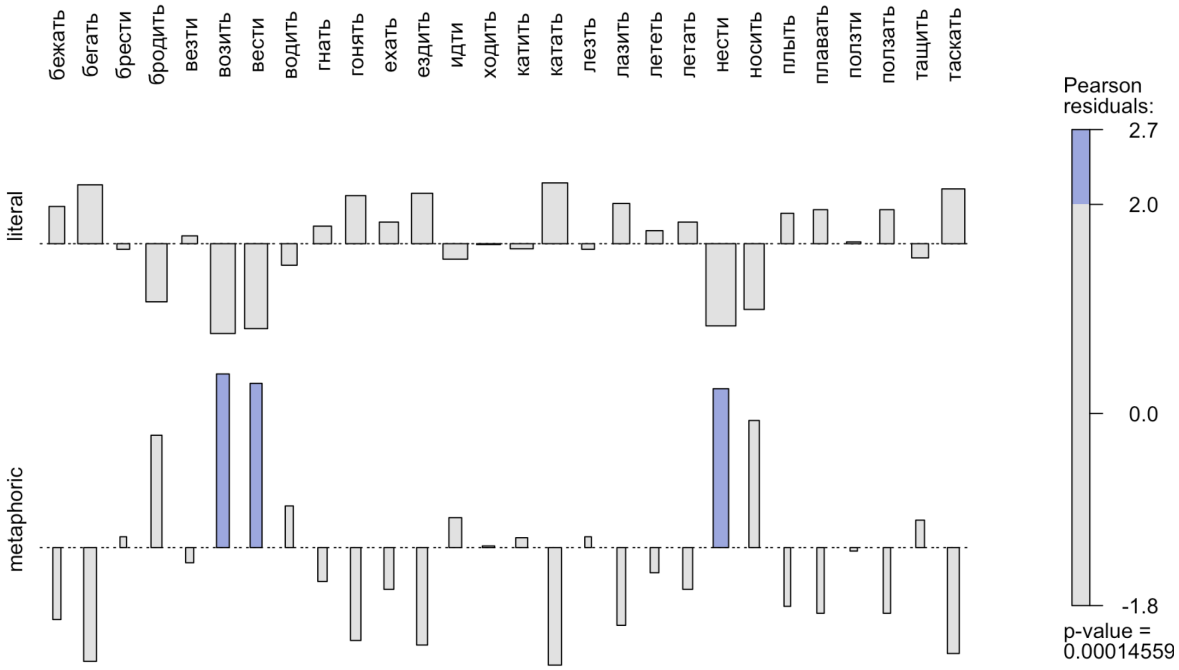


Figure C3: Association plot of root MVs and semantic characteristics based on MAS (Author 2)

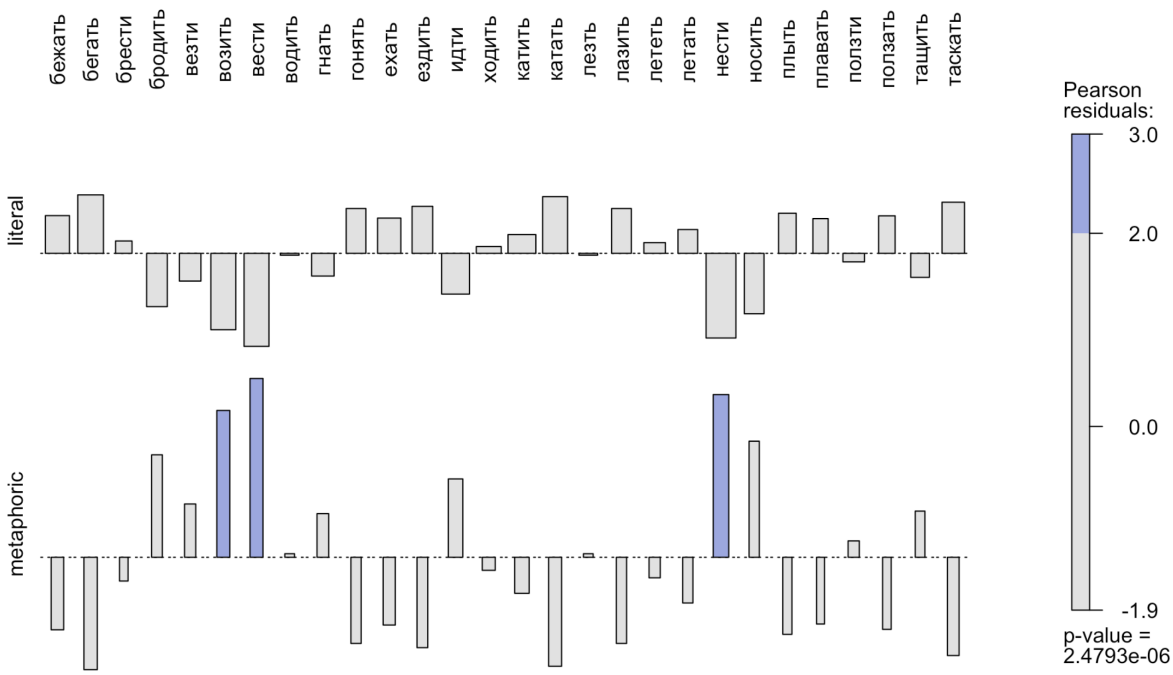


Figure C4: Association plot of root MVs and semantic characteristics based on the extended data (Author 2)

Appendix D The Issue of qualifying the meaning of the lexeme as non-literal (metaphoric)

When attempting to apply the traditional distinction in lexical semantics between literal and non-literal (metaphoric) meanings of a lexical unit, certain difficulties arise due to the following reasons.

1. The label "metaphoric" (*перен*.) is present in Russian language dictionaries created in the 20th century but is absent in more recent dictionaries. For example, it is used in "The Explanatory Dictionary of the Russian Language" by D. N. Ushakov (1935), "The Dictionary of the Russian Language" by S. I. Ozhegov (1949), and "The Dictionary of the Russian Language in 4 volumes", edited by A. P. Evgenyeva (MAS), but not in "The New Explanatory Dictionary of Synonyms of the Russian Language" edited by Yu. D. Apresyan or "The Large Dictionary of the Russian Language" by S. A. Kuznetsov. Additionally, the concept of metaphoric meaning is not relevant for semantic tagging in RNC, see https://ruscorpora.ru/page/instruction-semantic/.
2. Even in cases where the label "metaphoric" (*перен*.) is included in the dictionary's labeling system, the principles for its application are not specified in the relevant sections of the dictionaries (usually found in "How to Use the Dictionary"). This is particularly true for MAS, whose dictionary serves as the basis for this study, where it is stated only that: "§ 16. The metaphoric meaning (or shade of meaning) is accompanied by the label *перен*., which precedes all other labels (if any) and the definition of meaning".
3. Utilizing the decisions made by the dictionary compilers, particularly those of MAS, as a tool for annotation in this research proved to be unfeasible. This is because their decisions, lacking any explicit algorithm for their formulation, can only be regarded as intuitive judgments, which may not necessarily align with the intuitions of other researchers. This can be illustrated with two examples. First, one of our esteemed anonymous reviewers suggested that the derivative *занестись* from *нестись* 'to lay eggs (of birds)' in MAS and 'about birds: to lay eggs' in "The Dictionary of the Russian Language" by S. I. Ozhegov is derived from the metaphoric meaning of the original verb. However, this claim is not supported by either MAS or Ozhegov, as *нестись* does not carry the label *перен*. in either dictionary, even though it is listed as the third meaning of a polysemous word in MAS and appears as a separate entry in "The Dictionary of the Russian Language" by S. I. Ozhegov. The second example pertains to our subset of derivatives that represent As. According to MAS, only 9 of these derivatives are labeled *перен*., while Authors 1 and 2, in their independent assessments of metaphoric meanings within this subset, identified several dozen derivatives as metaphoric lexemes. However, their agreement coefficient during annotation (Cohen's kappa) was substantial but not perfect, measuring 0.76.

Enhancing RAG and Knowledge Graphs with Discourse

Boris Galitsky
Knowledge Trail, San Jose, CA, USA
bgalitsky@hotmail.com

Dmitry Ilvovsky
HSE University, Moscow, Russia
dilvovsky@hse.ru

Anton Morkovkin
HSE University, Moscow, Russia
ag.morkovkin@gmail.com

Abstract

We consider a number of Retrieval Augmented Generation (RAG) architectures to address a lack of specific information and hallucination issues of Large Language Models (LLM)—based question answering. We start with conformal prediction which acts on top of LLM and maintains a set of generations instead of a single one and attempts to find the best element of this set, which is assumed to be the "most average one". We then proceed to LLM self-reflection series of RAG architectures predicting the multi-hop question answering session before actual search for an answer. After that, we propose a mechanism for LLM to filter out answers inappropriate with respect to style. All these components need discourse-level analysis for more robust functioning. Knowledge graph (KG) and Abstract Meaning Representation (AMR)-based knowledge graph construction follow. We evaluate the contribution of all of these components to overall answer relevance and also zoom in on the role of discourse-based subsystem in each of these components. There is a substantial improvement of performance due to the four-component architecture introduced in this paper; the contribution of discourse-based subsystems is fairly modest.

DOI: 10.28995/2075-7182-2025-23-103-116

1 Introduction

Large Language Models (LLMs), such as GPT-4 and LLaMA, demonstrate strong performance in natural language understanding and generation, including in conversational search and question answering (QA). However, they often suffer from hallucinations and reduced accuracy when addressing queries involving rare, domain-specific, or long-tail entities. To address these limitations, Retrieval-Augmented Generation (RAG) approaches (Lewis et al., 2021) enhance LLMs by injecting retrieved textual information into the generation process. This combination of retrieval-based precision and generative fluency significantly improves the robustness of QA systems.

Recent work has extended RAG methods to broader domains and architectures. For instance, Feng et al. (2023) explore tighter integration between retrieval and generation components, improving LLM responsiveness to complex queries. Guo et al. (2023) demonstrate RAG's utility even in non-knowledge-intensive scenarios. More dynamic methods like active RAG [Jiang et al., 2023] adapt retrieval to evolving query intent in real time. Despite these advances, general-purpose RAG systems often fail in specialized domains. For example, RAG with GPT-4-Turbo struggles to answer financial questions derived from SEC filings, correctly handling only one out of five prompts. This indicates the necessity for more domain-aware methods, including fine-tuned LLMs and improved retrieval components.

Moreover, current RAG pipelines have difficulty performing complex reasoning, particularly in multi-hop QA tasks (Jeong et al., 2024; Zhang et al., 2024). Retrieval mechanisms often return noisy or irrelevant passages (Shi et al., 2023), and generated answers may contradict the retrieved evidence (Gao et al., 2023) or override LLMs' parametric knowledge (Parvez, 2024). While strategies like re-ranking (Nogueira and Cho, 2020) and conditional or active retrieval (Mallen et al., 2023) help alleviate some of these issues, they remain sensitive to annotation quality, may exclude useful information, and are computationally expensive. These limitations underscore the need for more robust integration of external knowledge.

To enhance the reasoning capabilities of RAG models, incorporating non-parametric knowledge is essential. Two primary sources are unstructured textual data (Izacard et al., 2022) and structured knowledge graphs (KGs). KGs provide compact and less noisy representations compared to full text and are more suitable for long-tail QA (Huang et al., 2024).

In this work, we also argue for the inclusion of discourse-level information. Discourse analysis examines how sentences relate to one another within a dialogue or text. It helps systems understand not only topical relevance, but also coherence, rhetorical structure, and intent. For example, rhetorical relations (e.g., Explanation, Cause, Elaboration) enable better matching between questions and answers. Discourse-level features are particularly beneficial for dialogue systems, where context extends beyond single utterances and includes stylistic and pragmatic dimensions.

We show that discourse analysis can:

- Improve the contextual understanding of dialogue history;
- Aid in structuring retrieved and generated content;
- Enhance answer filtering by identifying intent, tone, and rhetorical fit.

Our integrated system combines conformal prediction, self-reflective retrieval, discourse-aware QA coordination, knowledge graph integration, and AMR-based graph reasoning. Discourse elements enhance answer coherence and contextual appropriateness throughout the pipeline, contributing to the robustness of the overall architecture.

2 Conformal predictions for answer relevance improvement

Conformal prediction is a statistical technique for constructing prediction sets that are guaranteed to contain the semantically correct response with high probability. We apply it to improve the reliability of answers generated by RAG-based question answering systems. Instead of returning a single prediction, this approach constructs a set of candidates, such that the true answer is included in this set with a user-specified confidence level (e.g., 95%). This is especially important in RAG pipelines, where either the retrieved passage may be irrelevant or the LLM may produce an incorrect response even given a relevant passage (Fig. 1).

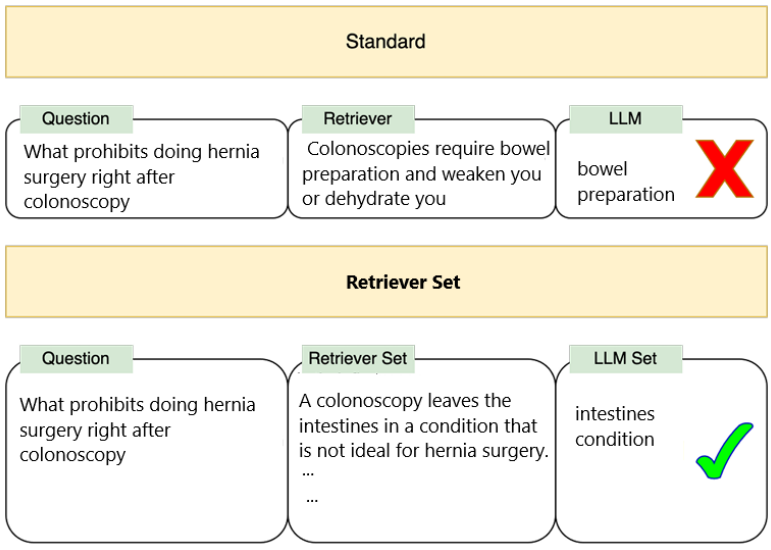


Figure 1. Proceeding from a single candidate to the candidate set for conformal prediction

The method builds two types of conformal prediction sets:

- Retriever set (CRet): ensures the relevant passage is included among those retrieved.
- LLM set (CLLM): ensures the generated answer is semantically correct with respect to the question and context.

These sets are then aggregated into an overall prediction set (C_{Agg}) that guarantees the presence of a correct answer with high probability. To construct the retriever set, we follow (Li et al., 2024) and use the negative inner product between the question and the annotated most relevant passage as the nonconformity measure. Given a calibration set and error tolerance α, a threshold τ_{Ret} is computed as the corresponding quantile over calibration scores, and all passages below this threshold are included in $C_{Ret}(q)$.

On the LLM side, prediction sets are constructed using a generation-and-clustering approach. For each <question, passage> pair, the LLM generates multiple responses (e.g., 30). These responses are clustered based on semantic similarity or entailment using an NLI model; responses are merged if they are semantically close or logically entailed. The confidence of each response is approximated as the ratio of its cluster size to the total number of generated responses. A threshold is applied to construct the prediction set $C_{LLM}(q, p)$from responses with sufficiently high confidence.

This methodology is extended to discourse-level representations. For each question, we construct prediction sets not only over generated answers but also over their discourse structures. The coordination between the question and the discourse structure of the answer is treated analogously to semantic coordination: a separate nonconformity score is defined for this dimension, and a threshold is applied to form a discourse-level prediction set. This allows the system to align both content and rhetorical structure with high confidence (Fig. 2).

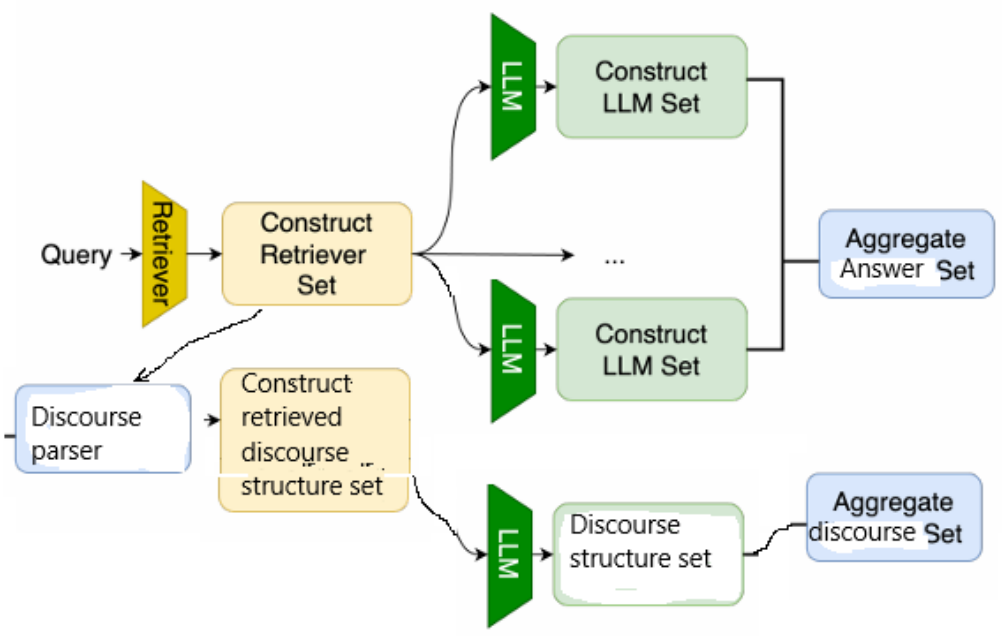


Figure 2. Conformal prediction for answers and discourse coordination of answers

To improve semantic coverage when constructing prediction sets, we apply Latin Hypercube Sampling, which ensures more uniform exploration of the confidence space than basic Monte Carlo sampling. Under mild assumptions such as data exchangeability, this conformal framework provides statistical guarantees for both semantic and discourse-level correctness of the predicted answers.

3 Self-reflection for RAGs

To address and control these behaviors such as retrieval frequency of the RAG model and guide the generation to be contextually consistent, Self RAG and its variants (Asai et al., 2024; Jeong et al., 2024) adopt a self-reflection based method. During training, these models learn to generate both task output and intermittent special reflection/critique tokens (e.g., *is_supported, is_relevant*, etc.), leveraging knowledge distillation from proprietary models like GPT-4. At inference, these generated tokens determine the usability of each candidate output. While these methods enable the model to effectively

rank candidate outputs from different retrievals and partially improve grounded/contextual generation, they struggle with navigating irrelevant or misleading information, especially when dealing with complex queries such as multi-hop retrieval tasks. This limitation arises since the models are not explicitly trained to contrast harder distractor passages and adhere to the facts from the retrievals.

To address this issue, Islam et al. (2024) reformulate the problem as a sparse mixture of experts (MoE) setup, where the model can selectively activate reasoning pathways that better differentiate subtle semantic contrasts in retrieved content. In our system, we extend this idea by incorporating discourse-aware self-reflection, enabling the model to not only evaluate semantic grounding but also analyze the rhetorical relation between the question and candidate answers.

In this paper, we enable self-reflection with discourse analysis of an occurrence of a potential answer in a passage obtained by retrieval. For different types of questions, the discourse structure of paragraphs containing answers significantly varies. So it is important not only to predict if a given question needs a retrieval session beyond what is encoded in LLM, but also to predict a discourse-level occurrence of a candidate answer in a text. A precise answer *A* occurs under a specific rhetorical relation to the phrase best matching the question (PBMQ).

During inference (Fig. 3), the model starts in a *no_retrieval* mode. It generates a preliminary answer and estimates its confidence. If no retrieval is needed, the model returns the answer based solely on parametric knowledge. Otherwise, for both single- or multi-hop queries from an external knowledge source $K = \{d_i\}_{i=1}^{N_k}$, we use a retriever R to obtain the top-m documents $S = \{s_t\}_{t=1}^{k}$, where each passage contains $\{r_j\}_{j=1}^{N_H}$ with $r_j \in K$ and N_H is the hop size. For each retrieved passage, the model M produces the output response y_t, along with the following tokens:

(1) The *relevance* tokens ([*relevant/irrelevant*]) indicate if s_t is relevant to q,
(2) the *grounding* tokens ([*fully supported/ps*$_t$*artially supported/no support*]) indicate if y_t is supported by s_t,
(3) the discourse relation specified how is the retrieved context unit is s_t related to the output response y_t rhetorically, and
(4) the *utility* tokens ([U:1]-[U:5]) define *how* useful y_t is to q.

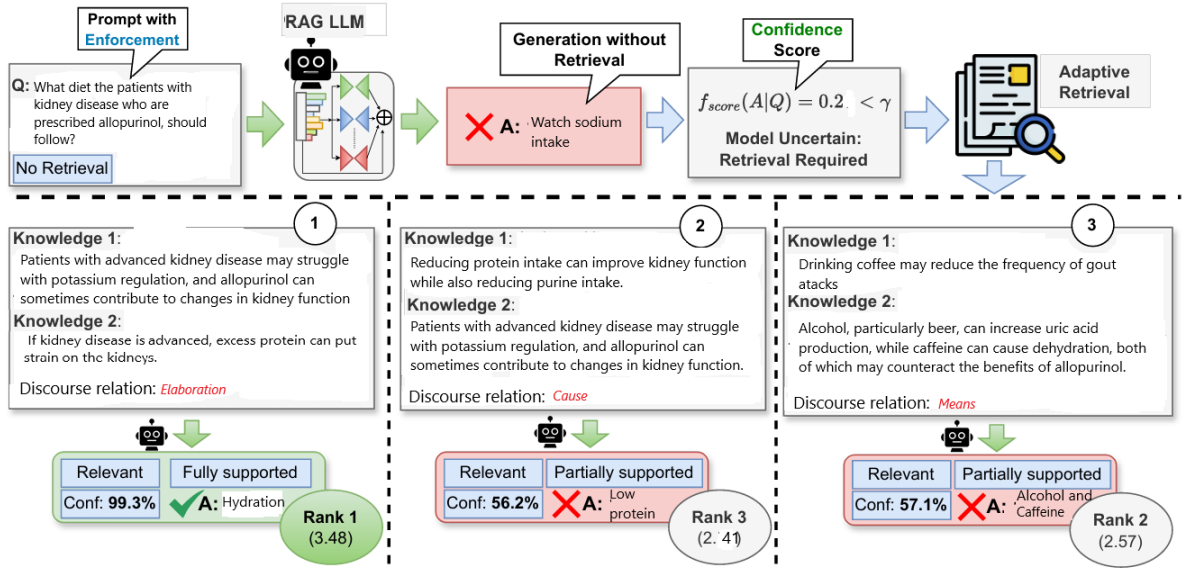


Figure 3. Inference pipeline in RAG self-reflection framework. For each knowledge set, discourse relation between the elements (PBMQ) and the candidate answer is specified.

We also evaluate the rhetorical relation between the Phrase Best Matching the Question (PBMQ) and each candidate answer. This relation helps determine whether the discourse structure of the answer matches the expected structure for the given question type. Specifically:

- Elaboration is expected for attribute-value questions;
- Explanation or Cause for *Why* questions;
- Enablement for procedural or *How* to achieve questions;
- Means for tool or method selection queries.

These discourse patterns guide filtering and selection of grounded answers that are not only semantically relevant but also contextually appropriate. As illustrated in Fig. 4, this approach enables the LLM to contrast distractors and reflect on answer quality during ranking and generation.

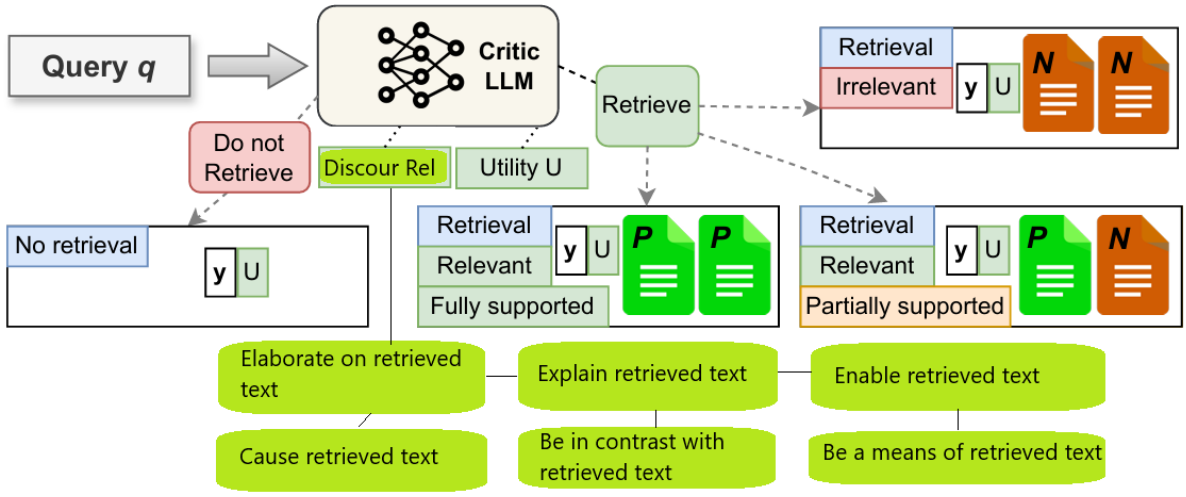


Figure 4. Enabling an LLM to reflect on an answer quality and to build contrast distractors

4 Enabling RAG with Knowledge Graph

While LLMs achieve strong performance across NLP tasks, they struggle with long-tail questions requiring domain-specific knowledge. To mitigate this, RAG architectures are often enhanced with non-parametric sources such as textual passages and knowledge graphs (KGs). Recent findings (Huang et al., 2024) demonstrate that prompting LLMs with KG triples can outperform passage-based retrieval, particularly in reducing hallucinations. Combining both KG and passage-based inputs may not consistently improve recall, but often enhances factual accuracy and answer grounding.

In our system, we apply this idea to the medical domain, using a two-phase architecture for question answering and treatment plan recommendation. During the KG construction phase, historical issue records (e.g., EHR-like customer support logs) are parsed into hierarchical tree structures, where each tree T_i represents an individual health issue. These trees are then interconnected to form a KG, combining both:

- Explicit links E_{exp}, specified in the documentation;
- and Implicit links E_{imp}, derived via semantic similarity between node embeddings (Xu et al., 2024).

An overview of RAG+KG framework is shown in Fig. 5. KG construction is shown on the left and Q/A and recommendation – on the right.

Each issue tree includes structured sections like “summary”, “description”, and “priority”. Fields suitable for rule-based parsing are extracted directly, while others are handled by an LLM guided by a template $T_{template}$. Embeddings are generated for semantically rich nodes and stored in a vector database. This dual-level KG structure allows us to preserve both intra-issue hierarchy and inter-issue relationships.

During the question-answering phase, user queries are parsed into entity-intent maps P of type Map($N \rightarrow V$), where N corresponds to a KG section (e.g., “issue summary”) and V is the extracted value. This parsing is performed by an LLM using prompts aligned with the template $T_{template}$.

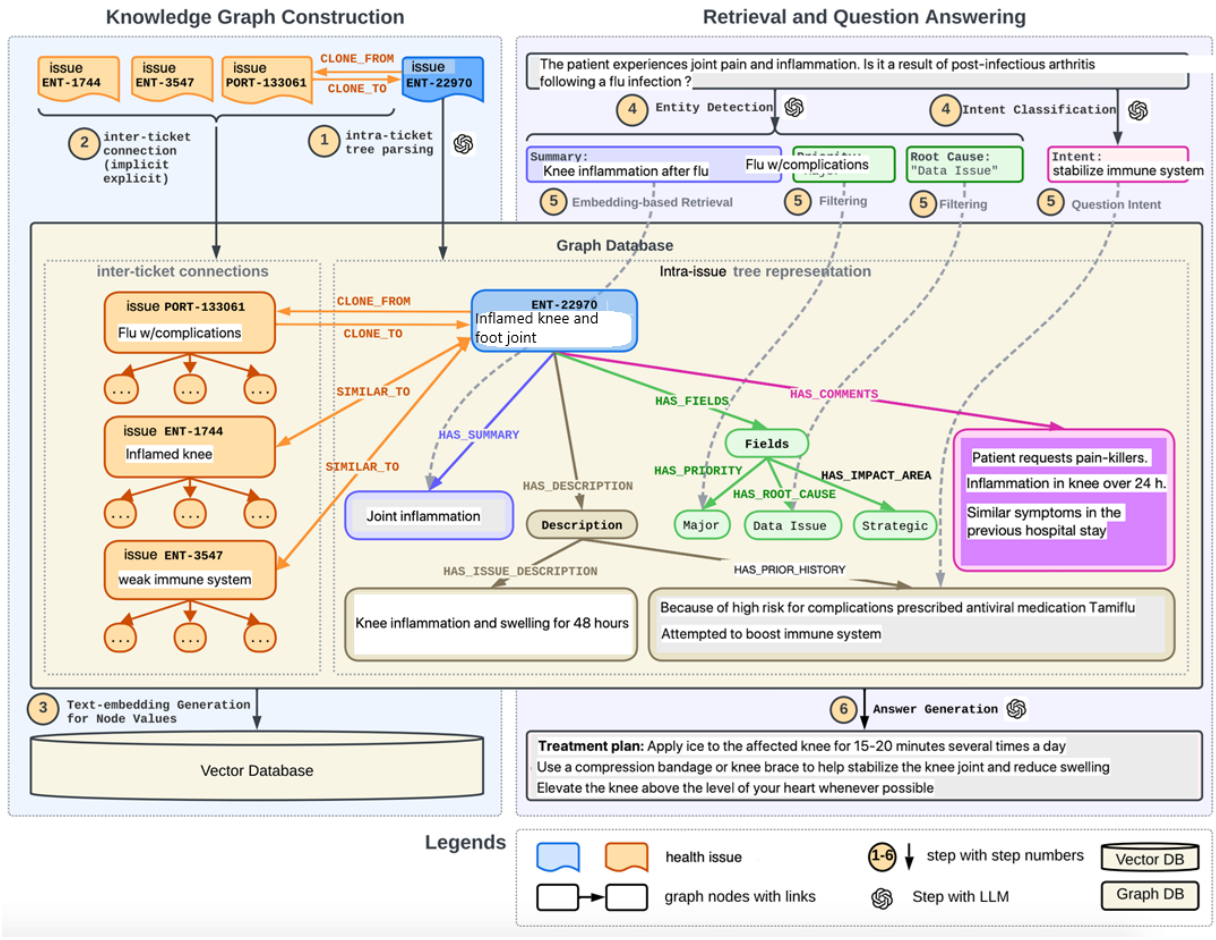


Figure 5. RAG+KG framework

For example, given the query “How to cure an inflamed knee which occurred as a complication of a flu?”, we extract:

- *Pe* = Map("*issue summary*" → "*inflamed knee*", "*issue description*" → "*knee inflammation after the flu with high temperature and dizziness*")
- and the intents *P=Set*(*elevation*, *ice*, *compress*, *rest*)

LLMs are generally robust in this task, and hallucinations are minimized through alignment with KG structure. Using these entity sets, the system retrieves top K most relevant historical cases based on cosine similarity of embeddings. A final score is aggregated per health issue using a sum over relevant section-node matches:

$$Score_{T_i} = \sum_{(k,v)\in P} \sum_{n\in T_i} TP\{section(n) = k\} * \cos\left(emdeb(v), embed\ (text(n)\right)$$

This score reflects both semantic proximity and entity occurrence frequency. Based on it, the most relevant subgraphs are selected for generation.

4.1 Leveraging LLMs to improve Graph Neural Network

Graph Neural Networks (GNNs) are widely used in graph machine learning. However, they often rely on shallow embeddings and struggle to generalize across diverse graph structures. LLMs can help by improving node representations, generating augmented features, and aligning feature spaces (Fan et al., 2024). Fig. 6 illustrates this idea.

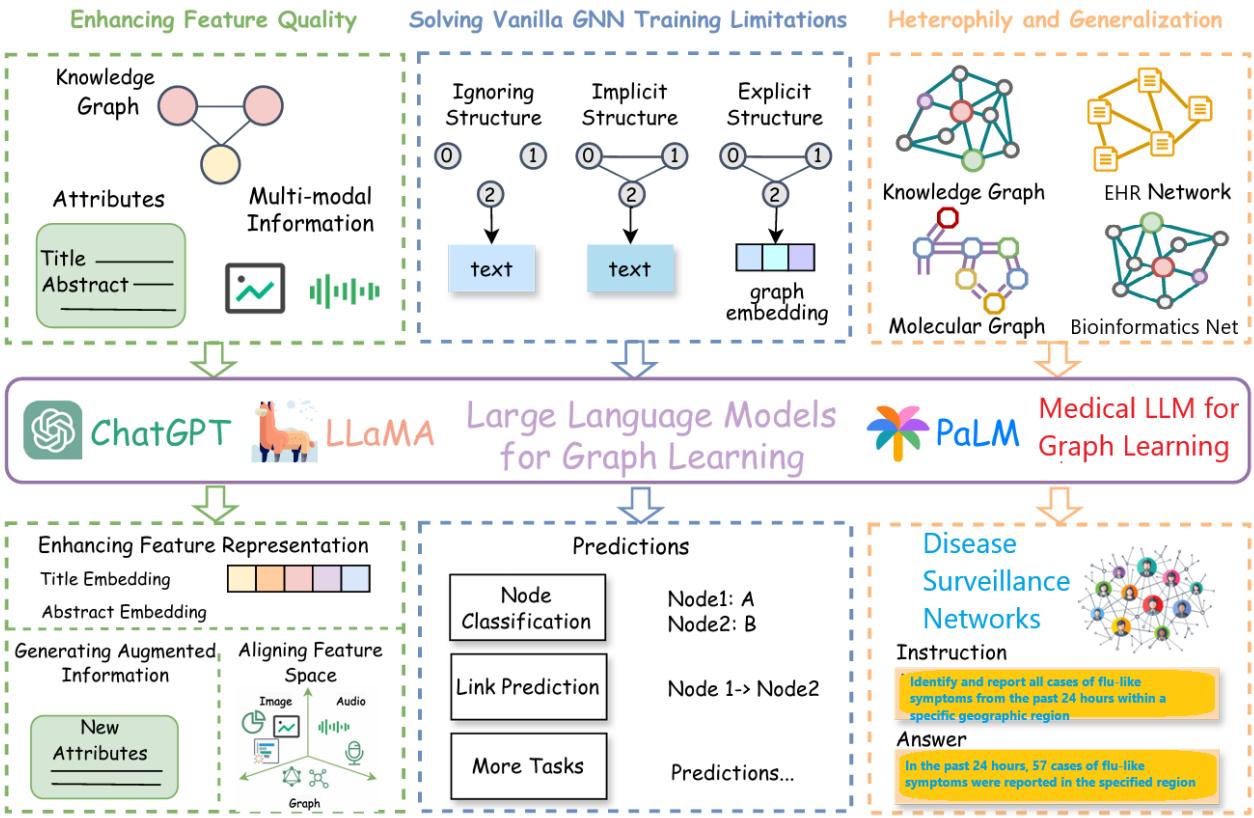


Figure 6. LLM assisting construction of KG

Challenges of vanilla GNNs include:

(1) Over-smoothing: mitigated with skip connections and batch normalization;
(2) Overfitting: addressed through node feature masking, dropout, and subgraph sampling;
(3) Heterophily: handled via adaptive aggregation (e.g., MixHop) or hybrid models (Abu-El-Haija et al., 2019);
(4) Low expressiveness: improved with k-GNNs, attention mechanisms, or higher-order message passing.

Recent works (He et al., 2023; Luo et al., 2023) also use LLMs to generate node labels or explanations from textual attributes (titles, abstracts), which are then encoded and combined with structural data in GNNs for downstream predictions.

4.2 Graph databases

The choice of graph vs. vector database plays a key role in RAG systems. Vector databases store embedded chunks and rely on semantic similarity for retrieval. This fuzzy matching improves flexibility but often introduces irrelevant or noisy passages.

In contrast, graph databases store structured entity-relation triples:

$$[ENTITY\ A] \rightarrow [RELATIONSHIP] \rightarrow [ENTITY\ B]$$

For example:

- *A* [*rust*] *EATS* [*metal*],
- *A* [*wolf*] *IS a* [*living thing*],

Benefits of graph databases for RAG include:

- Relational depth: better modeling of multi-hop reasoning;
- Reduced noise: fewer irrelevant results due to explicit structure;
- Explainability: reasoning paths are traceable and interpretable.

A limitation is rigid matching: if query entities don't exactly match stored nodes, relevant data may be missed. To mitigate this, hybrid systems combine graph structure with vector-based indexing. Entities and relations are encoded as embeddings, enabling approximate nearest-neighbor search over graph-like data.

This hybrid approach retains semantic flexibility while preserving explicit structure, improving retrieval quality and final LLM output.

5 AMR – based Knowledge Graph Construction

Retrieving relevant knowledge graph triples typically involves three steps: (1) tagging entities in the input query and aligning them with the KG; (2) retrieving candidate triples, often using SPARQL queries over RDF-based resources (Wang et al., 2021); and (3) ranking triples based on their relevance. To improve this ranking step, we incorporate Abstract Meaning Representation (AMR) to better capture the underlying relationships between entities.

AMR is a semantic representation framework that abstracts away from surface syntax and captures the meaning of sentences in a language-neutral graph format (Banarescu et al., 2013). AMR graphs are labeled, directed, rooted, and acyclic, where nodes represent events, entities, or properties, and edges denote semantic roles or relationships (e.g., ARG0, ARG1, temporal or modal relations). Compared to traditional syntactic parsing, AMR provides a unified view of meaning, supporting downstream tasks such as question answering, summarization, and information extraction (Naseem et al., 2021).

Fig. 7 illustrates an AMR expression in health domain *"We will send him to attend Dr. John Smith for a treatment review later this week following his testing"*.

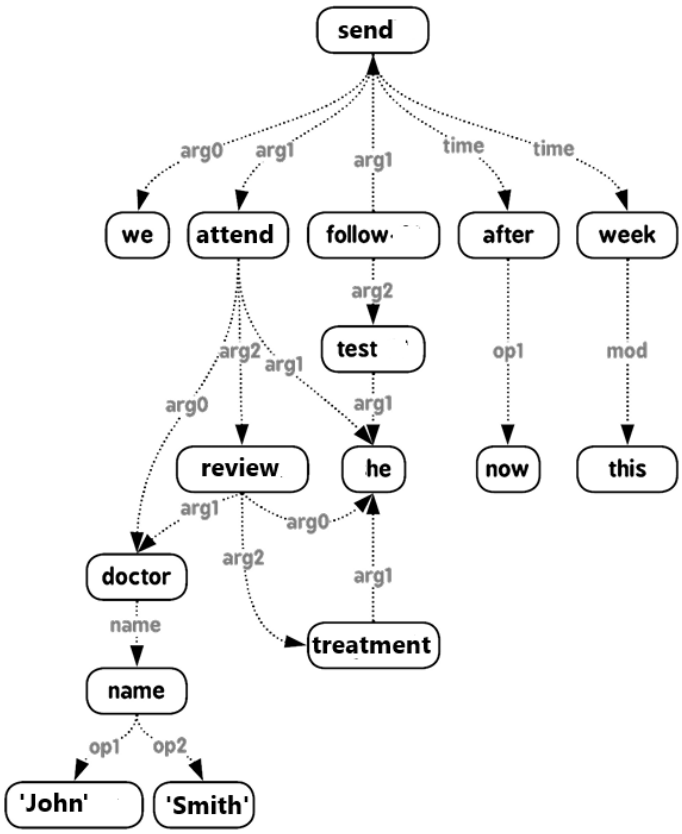


Figure 7. AMR representation for a sentence

In this AMR, the predicate attend represents the consultation event. Its ARG0 is “Dr. John Smith” (agent), ARG1 is “he” (patient), and the temporal modifiers express “later this week” via subgraphs like after now and within this week.

To apply AMR in triple ranking, we follow Huang et al. (2024) and use a dual-encoder architecture:

Entity-Relation Pairs Encoder. Given a sentence q, tagged entity e, and candidate relation r, the Entity-Relation Pairs as is defined as [TEXT]q[ENT]e[REL]r (Naseem et al.2021). [TEXT], [ENT], and [REL] are special tags inserted to mark the beginning of a sentence (q), tagged entity e, and candidate relations r respectively. This is passed through a BERT-based encoder to obtain representations $Q_{ERP} = [q_1,q_2,...,q_m]$, where each q_j corresponds to an entity-relation pair.

AMR-Enabled Multihead-Attention. The question is parsed into an AMR graph using parsers such as AMRLib (Jascob, 2024). Named entities are linked via BLINK (Wu et al., 2020), and the graph is encoded using a GNN. To avoid loss of edge semantics in message passing, all labeled edges are reified as nodes, preserving rich role information. Node embeddings $E_{AMR} = [e_1,e_2,...,e_n]$ are then computed via a Graph Attention Network

The two representations are fused using AMR-Multihead-Attention (AMA), which enables flexible alignment between semantic structures and candidate facts.

Let $Q = Q_{ERP}$, $K = V = E_{AMR}$. The AMR-Multihead-Attention (AMA) mechanism is defined as:

$$AMA(Q, K,V) = Concat(head_1,head_2,...,head_h)W^O$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \text{ with } Attention(Q, K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V$$

where W_i^Q, W_i^K, and W_i^V are the parameter matrices for the i-th attention head, and W^O is a parameter matrix for linearly transforming the concatenated outputs of all heads. This attention formalization shows the connection between the AMR and each entity-relation pair. Using AMR attention mechanism supports a flexibility in interpretation of entity relations, expressing the underlying semantic structures within the sentences. AMR is great for normalizing a semantic representation, converting into a canonical form (Fig. 8).

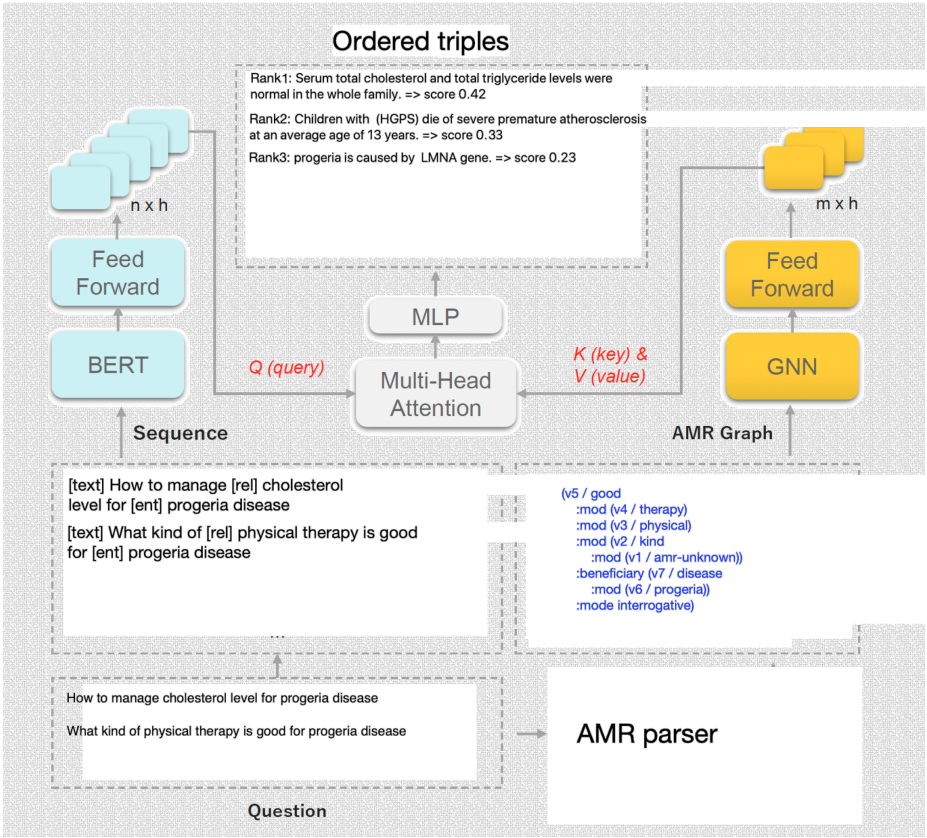


Figure 8. Obtaining knowledge from an AMR representation

6 Integrated system architecture

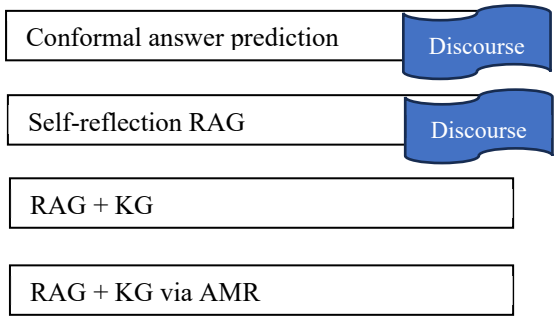


Figure 9. Four main components of the integrated RAG system

The integrated system architecture is shown in Fig. 9. Firstly, *Conformal prediction* is used to tackle hallucination and lower relevance answers. The *Self-reflection* RAG builds a retrieval scenario for each candidate in the answer set from the previous component. After that, for tail answers, KG is employed, if necessary, as determined by *Self-reflection.* Finally, long-tail answers, if determined by Self-reflection, are handled by RAG + KG via AMR.

7 Evaluation

We conducted a comprehensive evaluation of the contribution of each of the four components of our integrated system. For the first and second components, we also assessed the performance of their discourse analysis subsystems. To perform these assessments, we utilized GPT-4 alongside the embedding model E5 (Wang et al., 2022).

To evaluate retrieval efficacy, we employed the following key metrics:

(1) Mean Reciprocal Rank (MRR): This measures the average of the reciprocal ranks of the first correct answer in the result set, providing insight into the system's ability to prioritize relevant results.
(2) Recall@K: This metric assesses the probability that a relevant item appears within the top K retrieved results, indicating the system's ability to surface useful information.
(3) Normalized Discounted Cumulative Gain (NDCG@K): This metric evaluates the quality of the ranking by taking into account both the position of relevant items and their importance, offering a nuanced measure of retrieval precision.

To ensure the robustness of our findings, we compared our four-component system against three baselines:

(1) Stand-alone Large Language Model (GPT-4): This measures the performance of GPT-4 in isolation.
(2) Simple Retrieval-Augmented Generation (RAG) architecture: This serves as a baseline for comparison with more advanced architectures.
(3) Baseline from (Galitsky 2025): This provides a benchmark using the system optimized for medical discourse.

The evaluation was conducted on several key datasets:

(1) HotpotQA-Doc (Yang et al., 2018): A challenging dataset designed for multi-hop question answering based on two long documents. It tests the system's ability to handle complex reasoning tasks.
(2) ConditionalQA (Sun et al., 2022): This dataset serves as a benchmark for conditional question answering over long documents, pushing the system’s capacity for understanding dependencies across large amounts of text.
(3) Qasper (Dasigi et al., 2021): A dataset focusing on question answering from scientific papers, which evaluates the system's performance in extracting precise and contextually accurate information from dense, technical text.

The results of our evaluations are summarized in Table 1 (retrieval performance) and Table 2 (question-answering performance). Four runs were conducted for each evaluation setting. Across all metrics, our integrated system demonstrated consistent improvements over the baselines:

- MRR: Our system outperformed the averaged baseline by 12%, highlighting its enhanced ability to prioritize correct responses.
- Recall@K: Modest improvement in some cases, and loss for K = 1.
- NDCG@K: modest 8% improvement in for K = 2.
- BLEU Score: The system achieved an increase of 7% in BLEU score, further confirming its superior accuracy in question-answering tasks.
- ROUGE score: significant 24%.
- METEOR: significant 20%.

These results indicate that the four-component system, particularly when augmented with discourse-level analysis and advanced retrieval mechanisms, significantly improves both the retrieval and question-answering performance compared to simpler architectures. The integration of GPT-4 and the E5 embedding model ensures that the system delivers more relevant, accurate, and contextually appropriate answers, outperforming traditional baselines in complex, multi-hop reasoning tasks.

	MRR	**Recall@K**		**NDCG@K**	
		K = 1	**K = 2**	**K = 1**	**K = 2**
Stand-alone LLM	0.65	0.350	0.432	0.350	0.475
Simple RAG architecture	0.70	**0.450**	0.512	0.450	0.526
Baseline from (Galitsky 2025)	0.65	0.400	0.487	0.400	0.463
Four-component integrated system	**0.75**	0.400	**0.523**	**0.500**	**0.530**

Table 1. Accuracy of retrieval

	BLEU	**ROUGE**	**METEOR**
Stand-alone LLM	0.048	0.134	0.215
Simple RAG architecture	0.061	0.183	0.233
Baseline from (Galitsky 2025)	0.045	0.212	0.289
Four-component integrated system	**0.055**	**0.218**	**0.295**

Table 2. Accuracy of question answering

We now proceed to the assessment of each component out of four (Table 3), measuring an improvement compared to the averaged RAG baseline, the same as in Tables 1 and 2. We measure the boost of performance of the stand-alone first component, then the first and the second, then the first, the second and the third, etc. We use MRR for this ablation study.

	Averaged baseline per Table 1	**Conformal prediction**		**Self-reflection RAG**		**RAG + KG**	**RAG + KG via AMR**	**Total boost due to discourse**
			+ discourse		**+ discourse**			
HotpotQA-Doc	1	1.082	1.098	1.112	1.110	1.147	1.156	0.16
ConditionalQA	1	1.064	1.072	1.084	1.088	1.120	1.134	0.10
Qasper	1	1.051	1.050	1.059	1.064	1.093	1.121	0.03
PatientInfo (2024)	1	1.063	1.069	1.082	1.080	1.106	1.118	0.04

Table 3. Results of ablation / component contribution study

The impact of the discourse subsystems exhibits significant variability across different QA datasets, showing stronger variation compared to other accuracy metrics (Table 3). However, the overall contribution of these subsystems remains modest. Among the four components, the Self-reflection subsystem provides the least benefit to the model's performance, while the KG integration yields the most substantial improvement, enhancing both retrieval and response quality.

Considering the diverse metrics applied to evaluate our four-component architecture, we conclude that each component contributes a noticeable performance boost when compared to baseline RAG architectures. Despite varying levels of impact from each subsystem, the overall effectiveness of this multi-component architecture proves valuable in advancing the accuracy and robustness of the system.

8 Conclusions

The introduction of the four-component architecture presented in this paper has resulted in a significant performance boost, with improvements reaching up to 20%. This gain is attributed to the comprehensive integration of multiple system elements, which work together to enhance both retrieval and question-answering accuracy. Notably, the four-component system focuses on refining various stages of information processing, such as retrieval, discourse analysis, and the filtering of answers, which collectively contribute to the overall efficiency.

However, the contribution of the discourse-based subsystems is more modest in comparison, yielding improvements of up to 4%. While discourse analysis plays a crucial role in maintaining coherence and relevance within multi-hop question-answering tasks, its direct impact on retrieval performance remains relatively minor. This suggests that while discourse analysis enhances the system's overall flow and relevance, other components, such as retrieval and self-reflection mechanisms, play a more substantial role in driving the observed performance gains.

This assessment underscores the necessity of balancing multiple subsystems to achieve optimal performance, with discourse playing a supporting but essential role in ensuring the integrity of the system's responses.

In this paper, we explored multiple RAG architectures aimed at addressing key challenges in LLM-based question answering, specifically the issues of hallucination and lack of specific information. We began with conformal prediction, which enhances answer accuracy by maintaining a set of possible generations and selecting the most "average" one. We then introduced LLM self-reflection architectures that predict multi-hop question-answering sessions prior to engaging the retrieval system, improving the quality and relevance of generated answers. Furthermore, we proposed a mechanism for filtering answers based on stylistic appropriateness, ensuring that the generated responses align with the desired tone and format.

A central component in each of these architectures is discourse-level analysis, which provides a more robust framework for managing answer generation. Additionally, the integration of knowledge graphs (KG) and AMR-based knowledge graph construction strengthens the retrieval process by improving the relevance and contextual accuracy of the information retrieved. Our evaluations demonstrate the significant contributions of these mechanisms, particularly the discourse-based subsystems, to enhancing overall answer relevance and mitigating hallucination in LLM-based systems. This layered approach to RAG, incorporating self-reflection, discourse, and knowledge graph elements, offers a scalable solution to improving the quality of question-answering systems.

RAG combined with a knowledge graph is a practical implementation of the neuro-symbolic paradigm, blending neural networks (data-driven, statistical AI) and symbolic AI (logic-based, structured reasoning):

(1) Neural Component: The RAG model incorporates LLMs which use neural networks to generate text based on patterns in data. LLMs handle unstructured data (like natural language) and excel at understanding broad, context-driven information. When a question is posed, the RAG system uses vector-based retrieval techniques to find semantically similar chunks of text from a large corpus. This part represents the neural approach—flexible, probabilistic, and able to handle noisy or incomplete data.
(2) Symbolic Component: The knowledge graph represents the symbolic side. It structures data as nodes (entities) and edges (relationships), providing a clear, formal representation of knowledge. This allows for logical reasoning over the data, making it possible to retrieve precise and contextually related information based on the relationships between entities. For example, if a user asks a question about a specific historical event, the knowledge graph can retrieve related facts (entities) and their connections, ensuring accuracy and structured reasoning.

Knowledge graphs support multi-hop reasoning, meaning that the system can answer complex queries that require chaining together several facts or relationships. Neural models often struggle with multi-step logic, but the symbolic structure of the graph allows for deductive reasoning across entities, making it easier for the LLM to generate precise, well-founded responses.

References

[1] Abu-El-Haija S, Bryan Perozzi, Amol Kapoor, Nazanin Alipourfard, Kristina Lerman, Hrayr Harutyunyan, Greg Ver Steeg, Aram Galstyan (2019) MixHop: Higher-Order Graph Convolutional Architectures via Sparsified Neighborhood Mixing. arXiv:1905.00067v3

[2] Asai A, ZeqiuWu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. (2024) Self-RAG:Learning to retrieve, generate, and critique through self-reflection. In The Twelfth International Conference on Learning Representations.

[3] Asher N, Lascarides A (2003) Logics of conversation. Cambridge University Press, Cambridge UK

[4] Barzilay R and Mirella Lapata. 2008. Modeling Local Coherence: An Entity-Based Approach. Computational Linguistics, 34(1):1–34.

[5] Banarescu L, C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, N. Schneider, Abstract Meaning Representation for sembanking, in: A. Pareja-Lora, M. Liakata, S. Dipper (Eds.), Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse, Association for Computational Linguistics, Sofia, Bulgaria, 2013, pp. 178–186.

[6] Benveniste D (2024) Augmenting LLMs with a Graph Database. https://newsletter.theaiedge.io/p/augmenting-llms-with-a-graph-database

[7] Dasigi P, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A. Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In Proceedings of the 2021 Conference of the North American Paper of the Association for Computational Linguistics: Human Language Technologies, pages 4599–4610

[8] Fan, W., Wang, S., Huang, J., Chen, Z., Song, Y., Tang, W., Mao, H., Liu, H., Liu, X., Yin, D., & Li, Q. (2024). Graph Machine Learning in the Era of Large Language Models (LLMs). ArXiv, abs/2404.14928

[9] Feng Z, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023.Retrieval-generation synergy augmented large language models.arXiv, abs/2310.05149.

[10] Galitsky B, Dobrocsi G, de la Rosa JL, Kuznetsov SO (2011) Using generalization of syntactic parse trees for taxonomy capture on the web. ICCS:104–117

[11] Galitsky B, Ilvovsky D, Kuznetsov SO, Strok F (2013) Finding maximal common sub-parse thickets for multi-sentence search. IJCAI Workshop on Graphs and Knowledge Representation. LNAI,volume 8323, 39-57

[12] Galitsky B (2017) Discovering rhetoric agreement between a request and response. Dial Disc 8(2)

[13] Galitsky B (2019) Rhetorical agreement: maintaining cohesive conversations. In developing enterprise chatbots. Springer, Cham Switzerland, pp 327–363

[14] Galitsky B and Ilvovsky D (2019) Two Discourse Tree - Based Approaches to Indexing Answers. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), pages 367–372, Varna, Bulgaria

[15] Galitsky B (2020) A Virtual Social Promotion Chatbot with Persuasion and Rhetorical Coordination

[16] Artificial Intelligence for Customer Relationship Management. Springer, Cham Switzerland. 129 -180

[17] Galitsky B and [Ilvovsku D (2022b) Building ontologies relying on communicative discourse trees. Artificial Intelligence for Healthcare Applications and Management. Elsevier.

[18] Galitsky B (2022) Relying on discourse analysis to answer complex questions by neural machine reading comprehension. US Patent App. 17/505,462

[19] Galitsky B (2023) Converting a document into a chatbot-accessible form via the use of communicative discourse trees. US Patent 11,615,145

[20] Galitsky B (2024) Relying on discourse trees to build ontologies. US Patent 11,914,961

[21] Gao T, Howard Yen, Jiatong Yu, and Danqi Chen. 2023a. Enabling large language models to generate text with citations. In Proceedings of the 2023 EMNLP, pages 6465–6488, Singapore. ACL

[22] He X, X. Bresson, T. Laurent, A. Perold, Y. LeCun, and B. Hooi (2023) Harnessing explanations: LLM-to-LM interpreter for enhanced text attributed graph representation learning," Oct. 2023.

[23] Huang W, Guancheng Zhou, Mirella Lapata, Pavlos Vougiouklis, Sebastien Montella, Jeff Z. Pan. Prompting Large Language Models with Knowledge Graphs for Question Answering Involving Long-tail Facts.

[24] Islam SB, Md Asib Rahman, K S M Tozammel Hossain, Enamul Hoque, Shafiq Joty, and Md Rizwan Parvez (2024) Open-RAG: Enhanced Retrieval Augmented Reasoning with Open-Source Large Language Models. EMNLP Findings.

[25] Izacard G, M. Caron, L. Hosseini, S. Riedel, P. Bojanowski, A. Joulin, E. Grave, Unsupervised dense information retrieval with contrastive learning, Trans. Mach. Learn. Res. 2022 (2022)

[26] Jascob B. AMRLib - A python library that makes AMR parsing, generation and visualization simple. https://github.com/bjascob/amrlib (2024)

[27] Jeong S, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong C Park. 2024. Adaptive-rag: Learning to adapt retrieval-augmented large language models through question complexity. arXiv preprint arXiv:2403.14403.

[28] Jiang JJ, Conrath DW (1997) Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of the International Conference on Research in Computational Linguistics
[29] Joshi M, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics, 8:64–77.
[30] Karpukhin V, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 6769–6781, Online. Association for Computational Linguistics.
[31] Khashabi D, Sewon Min, Tushar Khot, Ashish Sabharwal, Oyvind Tafjord, Peter Clark, Hannaneh Hajishirzi (2020) UnifiedQA: Crossing Format Boundaries with a Single QA System. arXiv:2005.00700
[32] Komatsuzaki A, Joan Puigcerver, James Lee-Thorp, Carlos Riquelme Ruiz, Basil Mustafa, Joshua Ainslie, Yi Tay, Mostafa Dehghani, and Neil Houlsby. 2022. Sparse upcycling: Training mixture-of experts from dense checkpoints. arXiv preprint 2212.05055.
[33] Luo Y, J. Zhang, S. Fan, K. Yang, Y. Wu, M. Qiao, and Z. Nie (2023) BiomedGPT: Open multimodal generative pre-trained transformer for biomedicine," arXiv preprint arXiv:2308.09442
[34] Mallen A, A. Asai, V. Zhong, R. Das, D. Khashabi, H. Hajishirzi. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 9802–9822. URL: https://aclanthology.org/2023.acl-long.546. doi:10.18653/v1/2023.acl-long.546.
[35] Mallen A, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. arXiv preprint arXiv:2212.10511.
[36] Naseem T, S. Ravishankar, N. Mihindukulasooriya, I. Abdelaziz, Y.-S. Lee, P. Kapanipathi, S. Roukos, A. Gliozzo, A. Gray (2021) A semantics-aware transformer model of relation linking for knowledge base question answering, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), Proceeding ACL, pp. 256–262.
[37] Nogueira R and Kyunghyun Cho. 2020. Pas sage re-ranking with BERT. arXiv preprint arXiv:1901.04085.
[38] Palmer M, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. Computational Linguistics, 31(1):71 p106.
[39] Parvez MR 2024. Evidence to generate (e2g): A single-agent two-step prompting for context grounded and retrieval augmented reasoning. arXiv preprint arXiv:2401.05787
[40] PatientInfo (2024) https://patient.info/forums/discuss
[41] Shi F, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models canbe easily distracted by irrelevant context. In International Conference on Machine Learning, pages 31210–31227. PMLR.
[42] Sun H, William Cohen, and Ruslan Salakhutdinov. 2022. ConditionalQA: A complex reading comprehension dataset with conditional answers. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3627–3637.
[43] Wang L, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533.
[44] Wang Y, Hanghang Tong, Ziye Zhu, and Yun Li. 2022. Nested named entity recognition: A survey. ACM Trans. Knowl. Discov. Data, 16(6).
[45] Wang Z, Jun Araki, Zhengbao Jiang, Md Rizwan Parvez, and Graham Neubig. 2023. Learning to filter context for retrieval-augmented generation. arXiv preprint arXiv:2311.08377.
[46] Wang Z, P. Ng, R. Nallapati, B. Xiang, Retrieval, re-ranking and multi-task learning for knowledge-base question answering, in: P. Merlo, J. Tiedemann, R. Tsarfaty (Eds.), Proceedings of the 16th Conference of the European Paper of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online, 2021, pp. 347–357.
[47] Wu L, F. Petroni, M. Josifoski, S. Riedel, L. Zettlemoyer, Scalable zero-shot entity linking with dense entity retrieval, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 6397–6407.
[48] Xu Z, Mark Jerome Cruz, Matthew Guevara, Tie Wang, Manasi Deshpande, Xiaofeng Wang, Zheng Li (2024) Retrieval-Augmented Generation with Knowledge Graphs for Customer Service Question Answering arxiv.org 2404.17723
[49] Yang Z, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Cultural Evaluation of LLMs in Russian: Catchphrases and Cultural Types

Elizaveta Gromenko
HSE
Moscow, Russia
el.gromenko@gmail.com

Daria Kalacheva
MIPT
Moscow, Russia
TheDariaK@yandex.ru

Ksenia Klokova
MIPT
Moscow, Russia
klokova.ks@mipt.ru

Maxim Krongauz
MIPT, RSUH
Moscow, Russia
krongauz.ma@mipt.ru

Oksana Moroz
BHSAD
Moscow, Russia
omoroz@britishdesign.ru

Valery Shulginov
MIPT
Moscow, Russia
shulginov.va@mipt.ru

Tatiana Yudina
MIPT
Moscow, Russia
yudina.tatiana.a@mipt.ru

Abstract

This study addresses the gap in evaluating large language models' (LLMs) cultural awareness and alignment within the Russian sociocultural context by introducing a structured framework comprising 8 Cultural Types (e.g., Spiritual Practitioner, Soviet Intellectual) and 5 catchphrase groups (e.g., memes, proverbs). A 400-question evaluation dataset was developed to probe 10 multilingual LLMs, including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, across fact-based cultural knowledge and nuanced linguacultural understanding in a zero-shot setting. Results show that closed-source models GPT-4o and Claude 3.5 Sonnet outperform other models, with one of the smallest models (Mistral NeMo 12B) achieving the lowest result. Performance disparities were noticed in separate evaluation on Cultural Type tasks and catchphrases. Model-specific skews emerged, with lower-ranked models showing inclination toward specific cultural types. Qualitative analysis revealed common errors, such as selecting synonymous but incorrect answers or failing to grasp culturally specific logic. The contribution outlines the limitations of LLMs in interpreting cultural context and lays the groundwork for further research in assessing the cultural-linguistic alignment of LLMs.

Keywords: LLM; Russian language; catchphrases; cultural type; LLM evaluation; QA tasks
DOI: 10.28995/2075-7182-2025-23-117-128

Культурные замеры больших языковых моделей на русском языке: речевые клише и культурные типы

Елизавета Громенко
НИУ ВШЭ
Москва, Россия
el.gromenko@gmail.com

Дарья Калачева
МФТИ
Москва, Россия
TheDariaK@yandex.ru

Ксения Клокова
МФТИ
Москва, Россия
klokova.ks@mipt.ru

Максим Кронгауз
МФТИ, РГГУ
Москва, Россия
krongauz.ma@mipt.ru

Оксана Мороз
БВШД
Москва, Россия
omoroz@britishdesign.ru

Валерий Шульгинов
МФТИ
Москва, Россия
shulginov.va@mipt.ru

Татьяна Юдина
МФТИ
Москва, Россия
yudina.tatiana.a@mipt.ru

Аннотация

Наше исследование посвящено изучению культурной осведомленности больших языковых моделей о современном русскоязычном социокультурном контексте. Для этого предложена структурированная система, включающая 8 культурных типов (например, Духовный практик, Советский интеллигент) и 5 групп речевых клише. На основе этой системы был разработан набор данных из 400 вопросов различных форматов для оценки фактологического знания культурных особенностей и более тонкого лигвокультурного понимания в условиях zero-shot. Результаты тестирования 10 мультиязычных LLM (включая GPT-4o, Claude 3.5 Sonnet и Gemini 1.5 Pro) демонстрируют превосходство закрытых моделей GPT-4o и Claude 3.5, тогда как наименьший результат показала компактная модель Mistral NeMo 12B. Выявлены различия в результатах моделей при раздельной оценке заданий на культурные типы и речевые клише. Обнаружены специфические смещения: менее эффективные модели демонстрировали склонность к определенным культурным типам. Качественный анализ выявил типичные ошибки, включая выбор синонимичных, но некорректных, ответов или неспособность распознать культурно-специфичную логику. Исследование подчеркивает ограничения LLM в интерпретации культурного контекста и формирует основу для дальнейших исследований оценки культурно-лингвистической согласованности языковых моделей.

Ключевые слова: большие языковые модели; русский язык, речевые клише; культурный тип; оценка больших языковых моделей; вопросно-ответные задачи

1 Introduction

The rapid advancement of contemporary generative technologies raises questions about the level of integration of large language models (LLMs) into the cultural environment and their ability to consider users' axiological orientations. In this context, tools for assessing the cultural competencies of the LLMs are being actively developed. One of the main lines of research is the evaluation of cultural alignment and biases in the models (Cao et al., 2023; AlKhamissi et al., 2024; Naous et al., 2024; Rao et al., 2024). Another research direction focuses on probing and benchmarking LLMs for their cultural knowledge in specific domains and languages. Those include the evaluation of commonsense knowledge in diverse cultures (Myung et al., 2024; Shen et al., 2024; Koto et al., 2024); evaluation against cultural dimensions (Son et al., 2023; Yin et al., 2024; Wang et al., 2024; Kim et al., 2024); and probing for cultural norms (Fung et al., 2023) and values (Arora et al., 2023; Zhao et al., 2024; Karinshak et al., 2024).

Several benchmarks have been developed for the Russian language to assess general language proficiency, as well as ethical, logical, and other competencies. The largest task sets include RussianSuperGLUE (Shavrina et al., 2020), MERA (Fenogenova et al., 2024), and TAPE (Taktasheva et al., 2022).

Despite the existence of such comprehensive general assessment tools, methods for the evaluation of cultural awareness in LLMs for Russian are still missing.

The present study attempts to fill this gap by developing the methodology which allows to probe LLMs for the cultural awareness and alignment across diverse aspects of contemporary Russian culture. We operationalize this through the development of (1) a taxonomy of 8 distilled Cultural Types and (2) 5 distinct catchphrase groups. These constructs form the basis of a structured evaluation framework, which we employ to construct an evaluation dataset and conduct evaluation of 10 large language models.

The rest of this paper is structured as follows. First, we describe the methodology of identifying the cultural types, provide their brief description, and describe the groups of catchphrases. Section 4 reviews the evaluation dataset. In Section 5 we describe the evaluation setup, which includes models, prompts and metrics. We report on the evaluation results in Section 6, and conclude the paper in Section 7.

2 Cultural Types

The present section explicates the analytical framework developed to study the interaction of LLMs with the Russian national cultural context. By synthesizing methodological principles derived from interpretive sociology, cultural anthropology, and linguacultural studies, we establish and operationalize the conceptual category of "cultural type". A cultural type is defined by specific traits:

- Each cultural type is characterized by shared social norms, values, cultural activities, and historical backgrounds that shape a specific collective identity and perspectives on the cultural landscape (Tönnies, 2001). Representatives of a particular cultural type can articulate their sense of belonging as part of an in-group, which helps them distinguish themselves from "others";
- For the purposes of analysis each cultural type can also be seen as a distinct linguacultural type (LCT), possessing a unique speech repertoire, specific linguistic strategies for expressing cultural preferences, and distinctive behaviors in cultural consumption (Lutovinova, 2009);
- Hence cultural types exhibit specific colloquial usages that reflect unique cultural values, these can be explored through Hofstede's cultural dimensions theory (Hofstede, 2001). For instance, cultural types may represent either communities of practice, where identities are articulated through shared actions and goals, or communities of interest, where identities form around common interests and values (Lave and Wenger, 1991; Eckert, 2006);
- Following Hofstede's theory, it was concluded that the cultural code (Lotman, 2000) governing language behavior within these types is somewhat transgenerational, allowing different generations to coexist within the same cultural type.

Following interviews with the participants of the research team (conducted through a method similar to focus groups) and consultations grounded in empirical ethnographic observation (Zubarevich and Zubarevich, 2010), eight distinct cultural types were identified. Their short descriptions are provided below.

Basic Type encompasses individuals with a fundamental level of knowledge essential for effective integration into the cultural landscape of Russian society. A quintessential example of this type is a high school graduate. Those within this category are diverse and may hold various social statuses, occupations, interests, and preferences.

Careerist-Achiever covers people who stick to a pragmatic approach to life and career. They may be involved in common activities but rather represent a community of interest, sharing the values of productivity, efficiency, pragmatism, and objectivity.

IT Visionary indicates innovators and technoptimists who are deeply involved in the production and dissemination of technologies — such as IT specialists, engineers, and analysts. They tend to be open-minded and at the same time prioritize thought-provoking cultural content that reflects their interests (science fiction, dystopias, and so on).

Modern Intellectual indicates representatives of contemporary creative professions who are actively engaged in various forms of intellectual labor. This may, but do not always, result in the creation and consumption of commercialized and widely sought-after "creative products," (i.e., intellectual property: books, films, plays, etc.).

Nonconformist as a type is united by hyperconsumption of certain ideas and cultural texts, which are not currently mainstream. This category includes individuals who identify with countercultural values, as well as those who temporarily align themselves with specific subcultural movements.

Spiritual Practitioner embraces those who undertake different spiritual practices for the purpose of cultivating spiritual development. They may be deeply involved in religion, philosophy, or some pseudoscientific theories like astrology and tarot cards. Thus, on the verbal level, they may be indicated through certain rituals, actions, and the use of "magical" spells and words.

Soviet Intellectual as a cultural type is deeply engaged with both Soviet ideological culture and dissident narratives. As a custodian of "domestic" modernist perspectives, this type embodies the complexities of the Soviet cultural landscape, but probably excluded from the cultural context generated in the digital environment.

Trend Watcher is characterized by active consumption of mass culture products, heightened awareness of changes within the cultural landscape, strong focus on contemporary agendas, and active digital socialization. Its representatives' communication often includes slang, references to cultural phenomena, and memes, reflecting their active participation in today's cultural discourse.

Each cultural type has its unique background, encompassing specific cultural knowledge and preferences. To ensure transparency in this research, the background was organized by referencing various social and cultural domains, including arts, media, science, politics, religion, and sports, as well as concepts related to everyday life. Specific artifacts from these areas, such as cultural texts, quotes, and names, were used to create thematic maps for each type. Validation of thematic map content was conducted through cross-checking within the research group. The identified spectrum of cultural types allows us to focus on significant cultural groups that can be considered essential to Russian society and enables granular analysis. However, the spectrum can be extended and specified in further studies by other methods such as focus groups, autoethnographic research, and so on.

3 Catchphrases

The cultural commonsense knowledge in our research is represented in two parts: the Basic Type described above and a separate block that contains catchphrases. Catchphrases reflect elements of culture and represent a kind of cultural language code that refers to something well-known and generally significant (Krongauz, 1995). Such speech formulas, due to frequent use and regular involvement in the processes of language games, lose their identification, if they had one, and are reproduced as an element of cultural information regardless of the author and source (Linguistic Encyclopedic Dictionary, 1990). It is extremely significant that such speech formulas are traditional for a certain society; therefore, they can be used as signs of any culture as a whole, as well as of some special group within it (Nikitina, 1995; Nikolaeva, 1995). In other words, being verbal stereotypes in society, catchphrases play an important role in the social and cultural definition of the community, and their decoding is natural for any carrier of such cultural information.

For the current research, we considered catchphrases from a wide range of sources: the most popular memes (*Eto fiasko, bratan* (lit. *This is fiasco, bro*)); quotes from fiction, movies, songs, and advertising slogans (*Rukopisi ne goryat* (lit. *manuscripts don't burn*), (*Nado, Fedya, nado* (lit. *we must, Fedya, we must*)); phraseological units (*yazyk bez kostei* (lit. *tongue without bones*)); proverbs and idioms (*nogi volka kormyat* (lit. *the wolf's legs feed him*)).

In addition, the catchphrases feature the theme of childhood. This group represents the active cultural vocabulary of a child, mainly of preschool age, which, nevertheless, is also used and reproduced in adulthood. The children's catchphrases include: riddles; counting rhymes (*vyshel mesyats iz tumana* [...] (lit. *the moon came out of the fog* [...])); tongue twisters (*Karl u Klary ukral korally* [...] (lit. *Karl stole corals from Clara* [...])); quotes from fairy tales and children's literature (*Vot kakoi rasseyannyi s ulitsy Basseinoi* (lit. *Such an absent-minded one from Basseynaya street*)); songs; teasers and jokes (*obmanuli duraka na chetyre kulaka* (lit. *fooled a fool for four kulaks*)); proverbs (*s kem povedesh'sya, ot togo i naberesh'sya* (lit. *You will learn from who you hang out with*)).

4 Dataset and Tasks

The cultural types, thematic maps, and catchphrase groups outlined in Sections 2 and 3 formed the foundational framework for constructing a 400-question evaluation dataset. A few examples of the questions are included in Appendix A.

Tasks in the dataset take the form of 5 different question types: multiple-choice with one correct answer (MCQ (1)), multiple-choice with several correct answers (MCQ (N)), gap-filling (GF) with one correct answer, one-to-one matching (M), and one correct answer extraction from a given text (AE). For the distribution of the number of questions and question types across cultural types and catchphrase groups, see Table 1. Details on the number of potential answers for each question type are provided in Appendix B.

Cultural type / catchphrase group	**MCQ (1)**	**MCQ (N)**	**GF**	**Ma**	**AE**
each cultural type (40)	22	5	3	7	3
childhood (30)	8	-	22	-	-
memes (10)	5	-	5	-	-
phraseological units (10)	-	-	10	-	-
proverbs (12)	8	-	4	-	-
quotations (18)	5	1	7	5	-

Table 1: Distribution of number of questions and question types per cultural type and catchphrase group

For the purpose of consistent experimentation, we define a unified interface for all categories (cultural types and catchphrase groups) in the dataset. Each category is represented in JSON format, where each record contains five fields: *id*, *question type*, *prompt*, *question,* and *answer*.

5 Evaluation Setup

5.1 Models

We evaluate several LLMs on their knowledge of the defined cultural types and catchphrase groups. All tested models have multilingual capacity and support the Russian language. Other selection criteria include the support of general questions and high performance on Russian benchmarks[1]. The LLM pool includes both open- and closed-source models and is listed in Table 2.

Model	**No. params**
GPT-4o	-
GPT-3.5 Turbo 16K	-
Claude 3.5 Sonnet	-
Gemini 1.5 Pro	>200B
Gemma 2 27B	27B
Qwen2.5 72B Instruct	72B
Llama 3.1 405B Instruct	405B
Command R+	104B
Mistral NeMo 12B	12B

Table 2: Models for evaluation with their respective number of parameters

The GPT-4o model was included in two versions (released in May and August, 2024) in order to assess the consistency in this model family. We also included GPT-3.5 Turbo to compare its performance against the next generation models. Additionally, the compact Mistral NeMo model was added to evaluate its performance relative to significantly larger counterparts.

[1] Being in the top 10 on llmarena.ru as of 20.11.2024.

5.2 Prompts

LLMs are known to be sensitive to how prompts are formulated (Si et al., 2023; Zhuo et al., 2024). To develop prompts that the selected models understand and respond to in a specified manner, we created a separate set of 20 questions, which contained all 5 types of questions. Each type of question has its own prompt, and each of the resultant prompts consists of four parts: task explanation, information about a number of correct answers, specification of the output format, and a requirement not to explain the reasoning. The latter is needed for the unification of the assessment. All prompts used for the evaluation were in Russian and are included in Appendix A, along with the examples of output formats.

5.3 Metrics

The evaluation dataset consists of various question types, which need to be evaluated differently. We use Accuracy for the MCQ (1) and GF and Exact Match (EM) for the AE questions. For the question types that feature a possibility for partially correct answers — MCQ (N) and Ma — we use 1-Hamming Loss and the Jaccard index, respectively.

The overall evaluation metric is based on the weighted aggregation of individual category metrics by question types. We first compute the weighted average per category based on the number of questions of a certain type in that category. Then the category metrics are averaged, weighted by the total number of questions in the respective category, to compute the overall one. Its modification to evaluate separately on the cultural type and the catchphrase blocks is calculated in a similar manner.

5.4 Evaluation

We evaluate LLMs in a zero-shot setting with task prompting and prompt each model one time. The temperature is set to 0 for all models, with other sampling parameters left to their default values. For the output parsing, we employ a series of heuristic rules to extract an answer, formatted according to the prompt specification. Then, we conduct a manual check, aimed at catching any inconsistencies, and correct the formatting if needed. Lastly, we calculate and report the category and overall metrics.

6 Results

As mentioned in the previous section, we evaluate 10 large language models on their awareness and alignment with Russian culture across 8 cultural types and 5 groups of catchphrases and compute the overall metric to obtain a single-value qualitative measure of their performance. The results, shown in Fig. 1, indicate that the top-scoring models are both versions of the GPT-4o and Claude 3.5 Sonnet, with the lowest score achieved by the smallest of the tested models: Mistral NeMo 12B. The rest of the models (middle-scorers) form two clusters: the ones that achieved almost 80% in the performance measure (Llama 3.1 405B Instruct and Gemini 1.5 Pro) and those in the 65-75% range (Qwen2.5 72B Instruct, Gemma 2 27B, and Command R+).

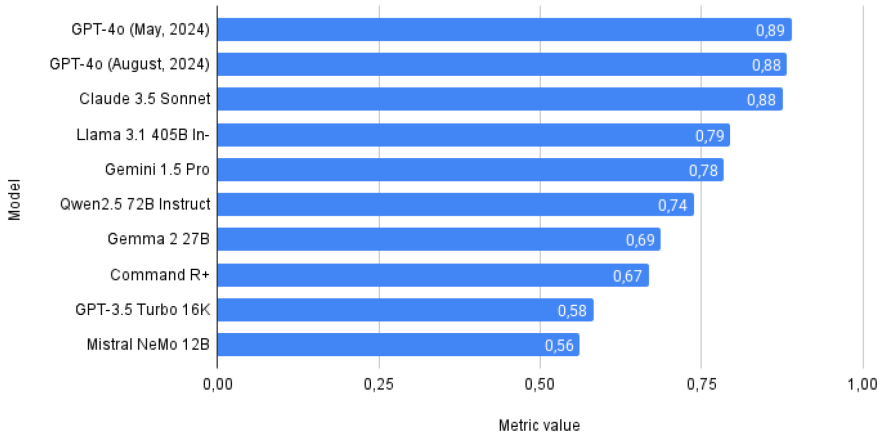


Figure 1: Overall performance of the selected models on the evaluation dataset, sorted in descending order

The performance in respect to the separate evaluation against Cultural Types and catchphrases is shown in Fig. 2. The distribution of the models' performance against the Cultural Type blocks mimics the overall performance since this block contains the larger number of questions. However, all models performed significantly worse against the catchphrases. This divergence likely stems from fundamental differences in task design: Cultural Type questions primarily assess factual knowledge (including foreign cultural influences), while catchphrases evaluate nuanced cultural knowledge that is a product of internalized collective cultural experience.

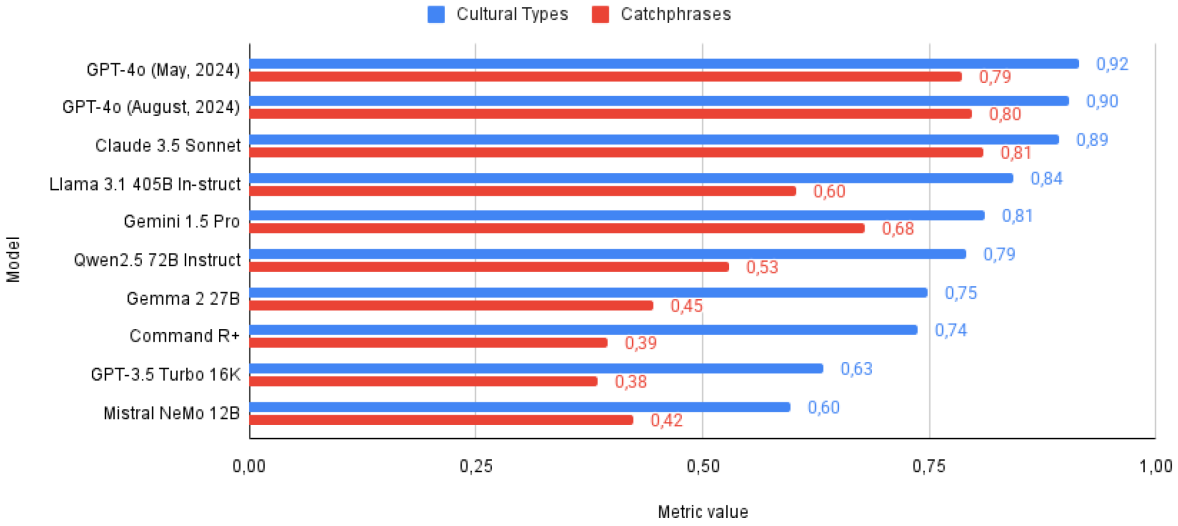


Figure 2: Results of the separate evaluation on the Cultural Type and the catchphrase blocks

The analysis of model performance across individual Cultural Types (Fig. 3) revealed that the top six models demonstrated comparable performance across all types, with minor deviations. However, the bottom four models exhibited significant skews toward specific types. Notably, GPT-3.5 Turbo and Mistral NeMo displayed alignment with the Spiritual Practitioner category, while GPT-3.5 Turbo additionally achieved peak performance in the Nonconformist type.

Evaluation by catchphrase group (Fig. 4) showed greater variability, with no model achieving parity across all categories. GPT-4o and Claude 3.5 Sonnet underperformed most prominently in the Meme group, whereas Llama 3.1 405B Instruct and Gemini 1.5 Pro exhibited weaker results in Quotes. Performance disparities across catchphrase groups were model-specific, indicating individual strengths and limitations. These findings suggest no singular problematic Cultural Type or catchphrase group which challenges the models' performance. Instead, cultural awareness evaluations reveal intrinsic model-specific traits, emphasizing the necessity for developers and users to account for these idiosyncrasies.

Comparing the two versions of the GPT-4o models, there are slight differences in the IT Visionary, Careerist-Achiever, and Nonconformist Cultural Types, with the May version achieving higher scores. Both models performed identically on the catchphrase block, apart from the Childhood category, where the August version made one less mistake. A closer look at the mistakes made revealed that the models almost always make the same ones. This slight difference in performance in some categories could potentially be attributed to the sensitivity to prompts or the ordering of potential answers and needs further investigation. Additionally, both the GPT-4o versions seem to be well-rounded across all categories (with the exception for Memes), while the GPT-3.5 Turbo, a previous generation in this model family, shows an evident skew as mentioned above.

During the qualitative analysis of the errors made by the models, we observed that many models struggled with questions where a correct answer is seemingly illogical, compared to the straightforward factual knowledge (names, places, etc.). For example, identifying a cosmic phenomenon that could negatively affect human life from the astrological perspective, or esoteric ideas of what should be done in order to become wealthier. In both the cultural type and the catchphrase block, another common kind of mistake was concerned with choosing an answer synonymous with the correct one (as in choosing the word *friend* instead of *bratan* in *This is fiasco, bratan*, mentioned in Section 3, or choosing the words

sosedki (eng. *neighbors*) and *devchonki* (eng. *girls*) in the famous quote from Pushkin's 'The Tale of Tsar Saltan' *Tri devitsy pod oknom pryali pozdno vecherkom* (eng. *Three fair maidens, late one night, sat and spun by candlelight.).*

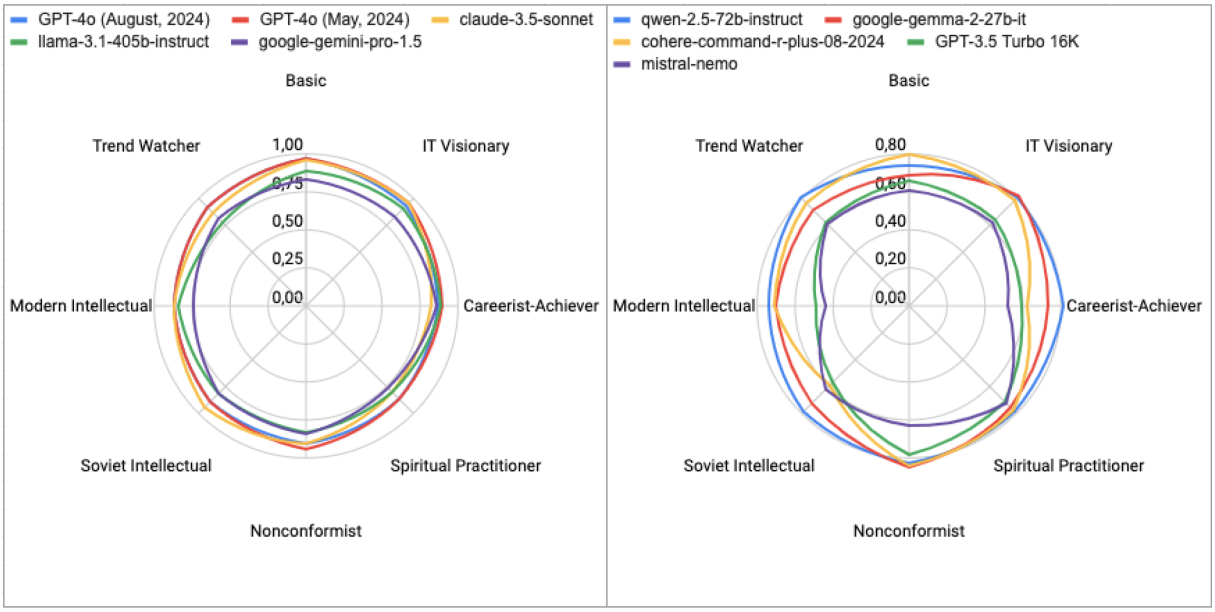


Figure 3: Results of the separate evaluation on individual Cultural Types
(the first five models to the left side of the graph, the last five models — to the right)

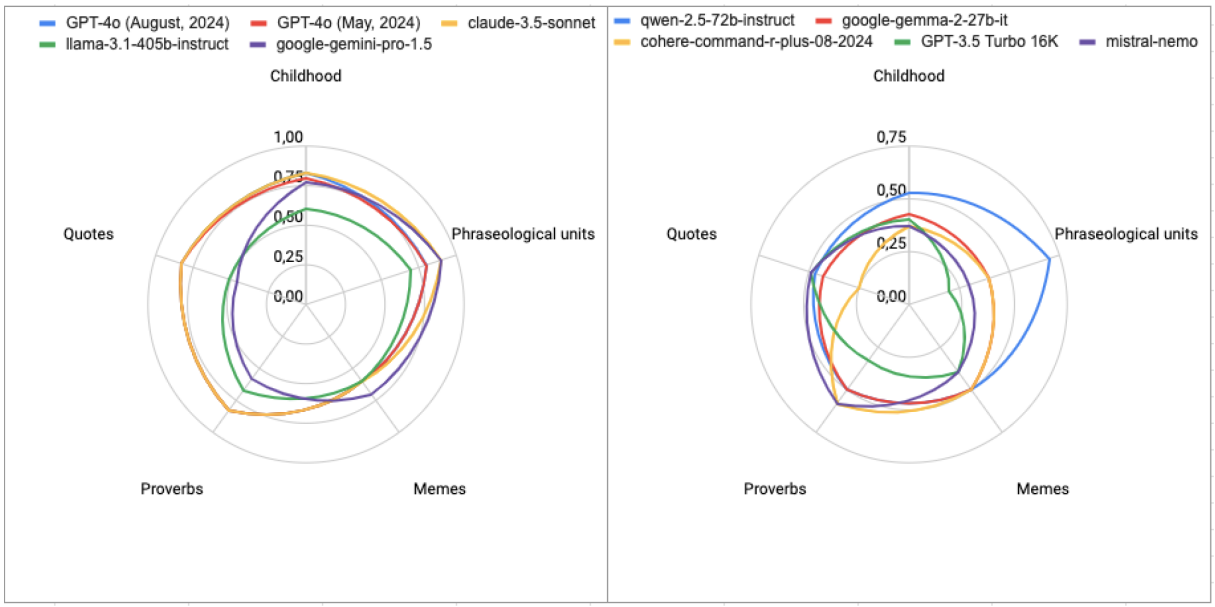


Figure 4: Results of the separate evaluation on individual catchphrase groups
(the first five models to the left side of the graph, the last five models — to the right)

7 Conclusion

This study focuses on the development of the evaluation methodology which allows to assess LLMs for the cultural awareness and alignment to the contemporary Russian sociocultural environment. We developed a structural evaluation framework which operationalizes through 8 Cultural Types and 5 groups of well-known catchphrases. This framework served as basis for the development of an evaluation dataset of 400 questions and probing 10 multilingual LLMs for their ability to understand and reflect cultural nuances.

The results reveal significant variability in performance. Overall top performance was achieved by closed-source models GPT-4o and Claude 3.5 Sonnet, while the lowest scores were achieved by GPT-3.5 Turbo and the smallest of the evaluated models Mistral NeMo 12B. Performance diverged significantly between tasks: models excelled at Cultural Type questions, which often evaluated fact-based knowledge, but underperformed on catchphrase evaluations, reflecting gaps in nuanced linguacultural knowledge.

Analysis revealed model-specific biases: lower-ranked models skewed toward specific cultural types (e.g., GPT-3.5 Turbo favored Spiritual Practitioner and Nonconformist), while catchphrase performance varied idiosyncratically — GPT-4o and Claude 3.5 Sonnet struggled with Memes, whereas others faltered on Quotes. GPT-4o versions showed minor differences in Cultural Type scores but identical catchphrase performance, suggesting prompt sensitivity. Error analysis highlighted two failure patterns: (1) difficulty resolving culturally illogical but correct answers, and (2) synonym confusion. These findings underscore the need for further research into improving LLMs' cultural alignment. Future work could expand the dataset, refine evaluation methodologies, and explore ways to enhance models' cultural awareness.

References

[1] Ahmad I., Dudy S. et al. Are Generative Language Models Multicultural? A Study on Hausa Culture and Emotions using ChatGPT // Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP. — Bangkok, Thailand, 2024. — P. 98–106.

[2] AlKhamissi B., ElNokrashy M. et al. Investigating Cultural Alignment of Large Language Models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 12404–12422.

[3] Arora A., Kaffee L., Augenstein I. Probing Pre-Trained Language Models for Cross-Cultural Differences in Values // Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). — Dubrovnik, Croatia, 2023. — P. 114–130.

[4] Cao Y., Zhou L. et al. Assessing Cross-Cultural Alignment between ChatGPT and Human Societies: An Empirical Study // Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). — Dubrovnik, Croatia, 2023. — P. 53–67.

[5] Eckert P. Communities of practice // Concise encyclopedia of pragmatics, 2nd edition. Oxford: Elsevier, 2006. — P. 109–112.

[6] Fenogenova A., Chervyakov A. et al. MERA: A Comprehensive LLM Evaluation in Russian // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 9920–9948.

[7] Fung Y., Chakrabarty T. et al. NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Singapore, 2023. — P. 15217–15230.

[8] Hofstede G. Culture's Consequences: Comparing Values, Behaviors, Institutions, and Organizations Across Nations. —Thousand Oaks, CA: Sage Publications, 2001.

[9] Karinshak E., Hu A. et al. LLM-GLOBE: A Benchmark Evaluating the Cultural Values Embedded in LLM Output // Computing Research Repository. — 2024. — Vol. arXiv:2411.06032. — version 1. Access mode: https://arxiv.org/abs/2411.06032.

[10] Kim E., Suk J. et al. CLIcK: A Benchmark Dataset of Cultural and Linguistic Intelligence in Korean // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 3335–3346.

[11] Koto F., Mahendra R. et al. IndoCulture: Exploring Geographically Influenced Cultural Commonsense Reasoning Across Eleven Indonesian Provinces // Transactions of the Association for Computational Linguistics. — Vol. 12, 2024. — P. 1703–1719.

[12] Krongauz M. A. The energy of cliched forms [Energiya klishirovannykh form], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 50–52.

[13] Lave J., Wenger E. Situated learning: Legitimate peripheral participation. — Cambridge: Cambridge university press, 1991. — P.45–52.

[14] Linguistic Encyclopedic Dictionary [Lingvisticheskij entsiklopedicheskij slovar'], Yartseva V.N. (editor), — Moscow, Russia:Soviet Encyclopedia [Sovetckaya entsiklopediya], 1990. — P. 683.

[15] Lotman Yu.M. Culture and Explosion [Kul`tura i vzryv], Semiosphere [Semiosfera]. — Saint Petersburg, Russia: Iskusstvo, 2000. — P. 12–149.

[16] Lutovinova O.V. "Linguocultural type" in a series of related concepts used to study linguistic personality [«Lingvokul`turnyj tipazh» v ryadu smezhnykh ponyatij, ispol`zuemykh dlya issledovaniya yazykovoj lichnosti], Scientific notes of the Transbaikal State University. Series: Philology, history, oriental studies [Uchenye zapiski Zabajkal`skogo gosudarstvennogo universiteta. Seriya: Filologiya, istoriya, vostokovedenie]. — Vol. 3, 2009. — P. 225–228.
[17] Myung J., Lee N. et al. BLEnD: A Benchmark for LLMs on Everyday Knowledge in Diverse Cultures and Languages // Computing Research Repository. — 2024. — Vol. arXiv:2406.09948. — version 1. Access mode: https://arxiv.org/abs/2406.09948.
[18] Naous T., Ryan M. J. et al. Having Beer after Prayer? Measuring Cultural Bias in Large Language Models // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 16366–16393.
[19] Nikolaeva T.M. The swing of freedom/unfreedom: tragedy or salvation? [Kacheli svobody`/nesvobody`: tragediya ili spasenie?], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 83–88.
[20] Nikitina S.E. Stereotypical judgments and speech cliches as cultural barriers [Stereotipnye suzhdeniya i rechevye klishe kak kul`turnye bar`ery], Speech and mental stereotypes in synchrony and diachrony. Conference abstracts [Rechevye i mental`nye stereotipy v sinkhronii i diakhronii. Tezisy konferencii.]. — Moscow, Russia: Institute of Slavic and Balkan Studies, Russian Academy of Sciences [Institut slavyanovedeniya i balkanistiki RAN], 1995. — P. 81–83.
[21] Rao A., Yerukola A. et al. Normad: A benchmark for measuring the cultural adaptability of large language models // Computing Research Repository. — 2024. — Vol. arXiv:2404.12464. — version 7. Access mode: https://arxiv.org/abs/2404.12464.
[22] Shavrina T., Fenogenova A. et al. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). — Online, 2020. — P. 4717–4726.
[23] Shen S., Logeswaran L. et al. Understanding the Capabilities and Limitations of Large Language Models for Cultural Commonsense // Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Mexico City, Mexico, 2024. — P. 5668–5680.
[24] Si C., Gan Z. et al. Prompting GPT-3 To Be Reliable // Computing Research Repository. — 2023. — Vol. arXiv:2210.09150. — version 2. Access mode: https://arxiv.org/abs/2210.09150.
[25] Son G., Lee H. et al. HAE-RAE Bench: Evaluation of Korean Knowledge in Language Models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 7993–8007.
[26] Taktasheva E., Fenogenova A. et al. TAPE: Assessing Few-shot Russian Language Understanding // Findings of the Association for Computational Linguistics: EMNLP 2022. — Abu Dhabi, United Arab Emirates, 2022. — P. 2472–2497.
[27] Tönnies F. Community and Civil Society. — Cambridge: Cambridge University Press, 2001. — P. 52–93.
[28] Wang Y., Zhu Y. et al. CDEval: A Benchmark for Measuring the Cultural Dimensions of Large Language Models // Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP. — Bangkok, Thailand, 2024. — P. 1–16.
[29] Yao B., Jiang M. et al. Benchmarking Machine Translation with Cultural Awareness. // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, Florida, USA, 2024. — P. 13078–13096.
[30] Yin D., Bansal H. et al. GeoMLAMA: Geo-Diverse Commonsense Probing on Multilingual Pre-Trained Language Models // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. — Abu Dhabi, United Arab Emirates, 2022. — P. 2039–2055.
[31] Zhao W., Mondal D. et al. WorldValuesBench: A Large-Scale Benchmark Dataset for Multi-Cultural Value Awareness of Language Models // Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). — Torino, Italia, 2024. — P. 17696–17706.
[32] Zhuo J., Zhang S. et al. ProSA: Assessing and Understanding the Prompt Sensitivity of LLMs // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, Florida, USA, 2024. — P. 1950–1976.
[33] Zubarevich N.V. Regions of Russia: Inequality, Crisis, Modernization [Regiony Rossii: neravenstvo, krizis, modernizaciya]. — Moscow, Russia: Independent Institute for Social Policy [Nezavi-simy`j institut social`noj politiki], 2010.

Appendix A. Examples of Questions in Cultural Types

Basic Type

(1) *На фантике какой конфеты изображена картина Ивана Шишкина «Утро в сосновом бору»? А) Белочка Б) Гулливер В) Мишка на Севере Г) Кара-Кум Д) Красная Шапочка Е) Мишка косолапый.*
On the wrapper of which candy is Ivan Shishkin's painting "Morning in a Pine Forest" depicted? A) Belochka B) Gulliver C) Mishka na severe D) Kara-Kum E) Krasnaya Shapochka E) Mishka Kosolapyj.

Nonconformist

(2) *Дополните фразу: «Пепси, пейджер, _» А) ВГТРК Б) вечеринка В) MTV Г) молодость Д) кока-кола Е) смартфон.*
Complete the phrase: "Pepsi, pager, _" A) VGTRK B) party C) MTV D) youth E) Coca-Cola E) smartphone.

Trend Watcher

(3) *Выберите резидентов Comedy Club: А) Павел Воля Б) Гарик Харламов В) Сергей Орлов Г) Семен Слепаков.*
Select Comedy Club residents: A) Pavel Volya B) Garik Kharlamov C) Sergey Orlov D) Semyon Slepakov.

Appendix B. Question Types and Prompts

Question type	Number answers	Output format	Prompt (Rus)	Prompt (translated to English)
Multiple choice (1)	1 correct out of 6	Letter: A	На вход подаются инструкции для ответа на вопрос. В них будет несколько вариантов ответа. Определите верный ответ. В ответе укажите только одну букву правильного ответа. Объяснять свой выбор не нужно.	The input contains instructions for answering a question. They will contain several answer options. Determine the correct answer. In the answer, indicate only one letter of the correct answer. There is no need to explain your choice.
Multiple choice (N)	N correct out of 4 or 6	One or more letters in alphabetical order: AB	На вход подаются инструкции для ответа на вопрос. В них будет несколько вариантов ответа. Определите верный ответ. Учтите, что их может быть несколько. В ответе укажите только букву ответа или последовательность из букв ответа в алфавитном порядке без пробела. Объяснять свой выбор не нужно	The input contains instructions for answering a question. They will contain several answer options. Determine the correct answer. Note that there may be several. In the answer, indicate only the letter of the answer or a sequence of letters of the answer in alphabetical order without a space. There is no need to explain your choice.
Gap-filling	1 correct out of 6	Letter: A	На вход подаются вопросы, в которых нужно заполнить пропуски. Пропуски обозначены знаком "_". Заполните пропуск предложенными вариантами. В ответе укажите только одну букву правильного ответа. Объяснять свой выбор не нужно.	The input contains a question, in which you need to fill in the gaps. Gaps are marked with the sign "_". Fill in the gap with the suggested options. In the answer, indicate only one letter of the correct answer. You do not need to explain your choice.
One-to-one matching	2 lists of 4 elements each	Sequence of numbers and letters: 1B 2V 3A 4G	На вход подаются инструкции для ответа на вопросы. Для каждого элемента из первого списка выбери наиболее подходящий элемент из второго списка. Буквы не должны повторяться. Нужно постараться составить 4 пары. В ответе укажите только последовательность из цифр-букв этих списков через пробел в формате 1Б 2А 3Г 4В.	The input contains instructions for answering questions. For each element from the first list, choose the most suitable element from the second list. The letters should not be repeated. You need to try to make 4 pairs. In your answer, indicate only the sequence of numbers and letters from these lists separated by a space in the format 1B 2A 3D 4C.
Answer extraction	1 correct out of 6	A noun the in nominative case	На вход подаются инструкции для ответа на вопрос. Найдите в тексте правильный ответ. Укажите ответ в именительном падеже без знаков препинания. Объяснять свой выбор не нужно.	The input contains instructions for answering a question. Find the correct answer in the text. Provide the answer in the nominative case without punctuation. There is no need to explain your choice.

Table 3: Question types used in the evaluation dataset, number of correct and potential answers, along with prompts and output format

Multi-senses connectives: some precisions

Inkova O. Yu.
Institute of Informatics Problems, FRC CSC RAS,
Moscow, Russia
University of Geneva, Geneva, Switzerland
Olga.Inkova@unige.ch

Abstract

The article discusses multi-senses connectives and their annotation in text corpora. The author clarifies the concept, which is usually applied to three different phenomena: i) the uncertainty of annotators in choosing one of the meanings of a polysemic connector; ii) the possibility of establishing more than one relation, both explicit and implicit, between two fragments of text; iii) the "combination" of several meanings by a connective. The author then considers how these cases are annotated in the Penn Discourse Treebank and in the Supracorpora Database of Connectives, which use a multi-label annotation for discourse relations. The study also raises a number of theoretical questions: what information can we obtain from double labels, which relations are distributionally close, i.e. can be established in the same contexts, how to separate the contribution of the connective and of the context to the overall meaning of a sequence of sentences, how legitimate is it to talk about the combination of meanings by the connective, how to annotate the features of the context of the connectives so that this data can be used for Natural Language Processing? The solution to these questions is important for both theoretical cognitive and applied research (in particular, for machine learning and machine translation).

Keywords: semantics; discourse relations; supracorpora database; connectives; corpus linguistics
DOI: 10.28995/2075-7182-2025-23-129-136

Неоднозначность коннекторов: некоторые уточнения

Инькова О. Ю.
ИПИ ФИЦ ИУ РАН, Москва, Россия
Женевский университет, Женева, Швейцария
Olga.Inkova@unige.ch

Аннотация

В статье рассматриваются неоднозначные коннекторы и их разметка в корпусах текстов. Автор уточняет понятие, которое применяется, как правило, для трех разных явлений: i) колебания разметчиков в выборе одного из значений полисемичного коннектора; *ii*) возможность установить между одинаковыми фрагментами текста более одного отношения, как эксплицитного, так и имплицитного; *iii*) «совмещение» коннектором нескольких значений. Далее рассматривается, как эти случаи размечаются в Пенсильванском аннотированном корпусе и в Надкорпусной базе данных коннекторов, в которых используется аннотирование с несколькими метками для дискурсивных отношений. Исследование ставит также ряд теоретических вопросов: какую информацию мы можем извлечь из комбинированных меток, какие отношения дистрибутивно близки, т.е. могут быть установлены в одинаковых контекстах, как различить вклад коннектора и контекста в общее значение последовательности высказываний, насколько правомерно говорить о совмещении значений коннектором, как размечать особенности контекста употребления коннектора, чтобы эти данные могли быть использованы для автоматической обработки текста? Решение этих вопросов важно как для теоретических когнитивных, так и для прикладных исследований (в частности, для машинного обучения и машинного перевода).

Ключевые слова: семантика; дискурсивные отношения; надкорпусная база данных; коннекторы; корпусная лингвистика

1 Введение

Семантика коннекторов не перестает привлекать внимание лингвистов и специалистов по автоматической обработке текстов и машинному обучению. Об этом, в частности, свидетельствует свежая монография (Zufferey & Degand 2024), в которой, среди прочих вопросов, связанных с коннекторами и выражаемыми ими отношениями, обсуждается вопрос об их неоднозначности. Сам по себе этот вопрос не вызывает сомнения, но его постановка требует уточнения. Действительно этот термин используется для трех разных явлений. В первом случае речь идет о выборе одного из значений полисемичного коннектора, что отражается, в частности, в колебаниях разметчиков корпусов текстов в выборе той или иной метки (Stede 2008; Spooren & Degand 2010; Webber 2013). Иными словами, в основе данного вида неоднозначности лежит возможность множественной интерпретации одного элемента, в нашем случае коннектора. Например, англ. *so* может выражать как отношение Результат, так и отношение Цель; ср. (1) и (2), приводимые в (Prasad et al. 2018: 46[1]):

(1) *Most murders are state crimes* so **any federal capital-punishment law probably would turn out to be more symbolism than substance** [wsj 0426]

(2) *Rep. Oakar, a Democrat from Cleveland, wants a $6.9 million grant* so **Cleveland can build an 18-story Rock and Roll Hall of Fame** [wsj 1847]

В последнее время этот вопрос активно обсуждается и для имплицитных отношений, т. е. тех случаев, когда коннектор между фрагментами текста отсутствует, но может быть ‘восстановлен’ (Scholman et al. 2022; López Cortez & Jacobs 2023; Long et al. 2024). Так, в примере (3) четыре разметчика оценили отношение между вторым и третьим предложениями как Соединительное, а пять – как Результат.

(3) It is logical that our attention is focused on cities. *Cities are home to 80% of the 500 million or so inhabitants of the EU.* **It is in cities that the great majority of jobs, companies and centres of education are located.** (пример из López Cortez & Jacobs 2023: 1)

Во втором случае под неоднозначностью понимается возможность установить между двумя одинаковыми фрагментами текста более одного дискурсивного отношения. Речь идет о многокомпонентных коннекторах, каждая из составляющих которых сохраняет свою семантику. Например, в (4)[2] в составе многоэлементного коннектора *а… тем временем* союз *а* выражает сопоставительные отношения, а *тем временем* – одновременность:

(4) В руках он держал книгу. За ним следом – молодой дьяк. Заметив краем глаза, что справедливость вот-вот будет восстановлена, Виктор продолжил свой путь. *А* дьяк *тем временем* вырвал из рук мальчишки книгу, ударил этой книгой его несколько раз по голове и тоже поспешил обратно к церкви. [Андрей Курков. Закон улитки (2005)]

Наконец, в некоторых исследованиях утверждается, что коннекторы могут выражать одновременно несколько отношений. Так, например, в работе (Zufferey, Degand 2024: 147) указано, что фр. *tant que* выражает одновременно временное и условное значения, а в работе (Кобозева и др. 2023: 41) – что русск. *в то же время* совмещает временное и противительное значения; в (Prasad et al. 2018: 52) приводится следующий пример на совмещение отношений временного следования и причины с соответствующей разметкой.

(5) *I cannot recall any disorder in currency markets* since (SUCCESSION,REASON) **the 1974 guidelines were adopted.** [wsj 0779]

Важность адекватного решения вопроса неоднозначности коннекторов для аннотирования и автоматической обработки текстов, а также для улучшения качества машинного перевода[3] трудно переоценить. С этим связаны также и более фундаментальные вопросы: является ли отношения,

[1] Мы сохраняем форму представления примеров.

[2] При отсутствии других указаний и используются примеры из Национального корпуса русского языка (НКРЯ).

[3] О трудностях машинного перевода коннекторов, особенно состоящих из более одного элемента, см. (Nuriev et al. 2018), где содержится также обширная библиография по данному вопросу.

выражаемые коннектором семантически близкими; какие отношения могут сочетаться, а какие нет; какие отношения имеют одинаковую контекстную дистрибуцию; как распределяются роли коннектора и контекста в общей интерпретации последовательности высказываний?

2 Разметка неоднозначных коннекторов в аннотированных корпусах текстов

Хорошо известно, что аннотирование дискурсивных отношений представляет значительные трудности. Поэтому разметка во многих корпусах опирается на семантику показателя, и пособия по аннотированию содержат списки коннекторов с указанием выражаемых отношений. Так, например, разметчикам Пенсильванского аннотированного корпуса (PDTB) в случае эксплицитных отношений, т.е. маркированных коннектором, предлагается выбрать из списка соответствующее коннектору отношение, а в случае имплицитных отношений, как видно в (8) ниже, предлагается подставить показатель, подходящий по смыслу, и на этом основании проставить соответствующее отношение (ср. PDTB 2008; Prasad et al. 2019). Проблема состоит в том, что многие коннекторы, как мы видели, неоднозначны, и разметчики не всегда могут однозначно определить выражаемое ими отношение, а кроме того, отношений между аннотируемыми фрагментами текста может быть несколько. В качестве решения данной проблемы предлагается разметка типа multi-label Annotation, или аннотирование с несколькими метками. Однако, хотя в последнее время во многих работах по аннотированию, причем в разных областях (ср. Aroyo & Welty 2015; Basile et al. 2021; Marchal et al. 2022), признается, что разметчики могут не соглашаться в проставлении метки и что при этом каждый из них может быть прав в силу субъективности и сложности многих задач, в аннотированных корпусах текстов обычно отсутствуют аннотации более чем одного дискурсивного отношения между двумя одинаковыми фрагментами текста. Например, в корпусах, использующих Теорию риторической структуры (RST), допускается альтернативный анализ дискурсивной структуры текста, но предлагается выбирать одно отношение, а именно то, которое соответствует коммуникативным целям автора (Mann et al. 1989; Carlson, Marcu 2001). Это приводит к тому, что в случаях, когда возможны множественные интерпретации, но разметчики ограничены проставлением только одной метки, остается неясным, отражает ли разногласие между разметчиками неправильную интерпретацию (например, из-за отсутствия у разметчика необходимых знаний) или же неоднозначность интерпретации аннотируемого текста и, в некотором смысле, вынужденный выбор единственной метки из нескольких возможных.

Только в двух известных нам корпусах допускается аннотирование с использованием нескольких меток: Надкорпусная база данных коннекторов (НБДК), о которой речь пойдет ниже, и PDTB.

В корпусе PDTB 2.0 допускалось наличие более чем одного отношения между двумя фрагментами текста в трех случаях: *i*) отношение может быть выражено эксплицитным коннектором, который одновременно может выражать более одного отношения, например, *since* в (5) выше; *ii*) каждое из отношений между одинаковыми отрезками текста может быть выражено разными коннекторами, например, в многоэлементном коннекторе *but then* союз *but* выражает отношение «вопреки ожидаемому», а *then* – предшествование (6); *iii*) при отсутствии коннектора разметчики видят более одного отношения между фрагментами текста. В PDTB 3.0 к этим трем случаям был добавлен четвертый: *iv*) коннектор или другой показатель выражают одно дискурсивное отношение, но разметчик видит между фрагментами текста еще одно, имплицитное, отношение.

Каждому из четырех случаев соответствует своя разметка. Как видно из (5) в случае *i*) проставляется комбинированная метка с двумя выявленными отношениями. В случае *ii*) каждое из отношений размечается отдельно (6).

(6) *Small businesses say a recent trend is like a dream come true: more-affordable rates for employee-health insurance, initially at leas*t **But** then (PRECEDENCE) **they wake up to a nightmare** [wsj 0518]
Small businesses say a recent trend is like a dream come true: more-affordable rates for employee-health insurance, initially at least But (ARG2-AS-DENIER) **then they wake up to a nightmare** [wsj 0518] (пример из Prasad et al. 2019: 53)

В случае *iii*), как и в случае *i*) проставляется комбинированная метка, а семантика имплицитного дискурсивного отношения определяется на основе возможной подстановки эксплицитного показателя (7):

(7) *...there's also a price to pay* (IMPLICIT=REASON,ARG2-AS-DETAIL) **A diversified portfolio always under-performs an undiversified portfolio during those times when the investment in the undiversified portfolio is truly hot** [wsj 1453] (пример из Prasad et al. 2019: 53)

Наконец, в случае *iv*) каждое из отношений – имплицитное и эксплицитное – размечаются отдельно, как в *ii*), но с другим способом представления:

(8) *We've got to get out of the Detroit mentality* and (EXPANSION.CONJUNCTION) Implicit=instead (EXPANSION.SUBSTITUTION.ARG2-AS-SUBST) **be part of the world mentality**," declares Charles M. Jordan, GM's vice president for design… [wsj 0956]

Характерно, что, признавая важность для теоретических когнитивных исследований использования комбинированной метки в случае *i*) (Prasad et al. 2014: 933), разработчики корпуса как будто не замечают, насколько важно дать возможность исследователям изучать также потенциальную совместимость отношений в случае *ii*). Такую возможность разметка, принятая в PDTB, к сожалению, не предоставляет, и сводные таблицы дают только информацию по совмещению отношений в случае *i*) (см. Prasad et al. 2019).

3 Аннотации с неоднозначными коннекторами в НБДК

Надкорпусная база данных коннекторов (НБДК), созданная в ИПИ РАН, содержит тексты параллельных английского, немецкого, французского и итальянского подкорпусов НКРЯ. На 1.01.2025 в ней сформировано 25'547 двуязычных аннотаций. В них отражено не только дискурсивное отношение, выражаемое коннектором, но и такие параметры, как синтаксическая структура вводимого коннектором фрагмента текста, его позиция в нем, порядок следования фрагментов текста. Для полифункциональных языковых единиц (например, *точнее* может быть как сравнительной степенью прилагательного или наречия, так и коннектором) указана его функция в данном контексте (подробнее об устройстве и функциональных возможностях НБДК см. Инькова 2018).

В НБДК для интересующих нас случаев используется следующая разметка. Для случая *ii*), т.е. когда в составе многоэлементного коннектора можно выделить несколько ‘самостоятельных’, в аннотации проставляется необходимое количество отношений: их может быть два или более. В таблице 1 представлена разметка примера (4) выше с *а… тем временем*, выражающим два отношения. Заметим, что в итальянском переводе одноэлементный коннектор *intanto* переводит только отношение одновременности.

Контекст коннектора в оригинале	**Признаки коннектора в оригинале**	**Контекст коннектора в переводе**	**Признаки коннектора в переводе**
В руках он держал книгу. За ним следом – молодой дьяк. Заметив краем глаза, что справедливость вот-вот будет восстановлена, Виктор продолжил свой путь. **А** дьяк **тем временем** вырвал из рук мальчишки книгу, ударил этой книгой его несколько раз по голове и тоже поспешил обратно к церкви.	**а|тем временем** сопоставительные одновременность повествовательное начальная p CNT q CNT	Teneva in mano un libro. Dietro a lui correva un giovane diacono con una barbetta rada e la faccia brufolosa. Notato con la coda dell'occhio che presto sarebbe stata fatta giustizia, Viktor continuò per la sua strada. Il diacono **intanto** strappò dalle mani del ragazzino il libro, con il quale lo colpì ripetutamente in testa, e si affrettò di nuovo a tornare in chiesa.	**intanto** одновременность повествовательное неначальная p CNT q CNT

Таблица 1. Двуязычная аннотация контекста коннектора, выражающего более одного значения, в НБДК

Такая же разметка возможна в принципе и для имплицитных отношений – случай *iii*) в PDTB –, но таких случаев в НБДК пока не зафиксировано, как и контекстов, соответствующих случаю *iv*). Можно было бы проставить в разметке итальянского перевода фрагмента текста с *а… тем временем* и имплицитное отношение сопоставления (помимо проставленного отношения

одновременности), но поскольку в НБДК фиксируются реальные употребления коннекторов (в данном примере ит. *intanto*), то разметка фиксирует характеристики именно переводного эквивалента, а не контекста, в котором он употреблен. Хотя и эта информация важна, в частности, для изучения того, какие отношения могут быть имплицированы.

Для случая *i*), т.е. когда разметчик затрудняется выбрать одно из выражаемых полисемичным коннектором отношений, предусмотрена комбинированная метка, сочетающая несколько отношений. Так, для *точнее* в аннотации, представленной в таблице 2, как и для его переводного эквивалента, была проставлена метка «коррекция/спецификация». Фрагмент текста, вводимый коннектором, может быть понят как переход от общего (Достоевский) к частному (его идеология), т.е. как спецификация, или как своего рода поправка-уточнение, т.е. как коррекция. Контекст не позволяет разметчику выбрать то отношение, которое однозначно соответствовало бы коммуникативному заданию говорящего.

Контекст коннектора в оригинале	**Признаки коннектора в оригинале**	**Контекст коннектора в переводе**	**Признаки коннектора в переводе**
Такой образ в стиле самого Достоевского, **точнее**, его идеологии, между тем как образ единого духа глубоко чужд ему.	**Точнее** коррекция / спецификация без предикации начальная p CNT q CNT	Quest'immagine è nello stile di Dostoevskij, **o piú esattamente** della sua ideologia, mentre l'immagine di un unico spirito gli è affatto estranea.	**O più esattamente** коррекция / спецификация без предикации начальная p CNT q CNT

Таблица 2. Двуязычная аннотация контекста коннектора с комбинированной меткой в НБДК

Наконец, если разметчик затрудняется в определении отношения, например, в том случае, если семантика коннектора еще не была точно определена, то он может проставить метку TBD (To Be Defined). Эта метка сама по себе не представляет интереса, но позволяет разметчикам создавать аннотации с еще неизученными коннекторами.

По всем трем типам меток в НБДК можно получить количественные данные. Например, для русских коннекторов было сформировано 1669 аннотаций, где проставлено два и более отношений. Это составляет 6,54%. Комбинированных меток гораздо меньше – 78, т.е. 0,3%. Для сравнения: в PDTB для эксплицитных отношений такие метки были проставлены в 999 из 18 459 аннотаций, т. е. в 5,4%, но это различие связано с разной методикой разметки. Количество аннотаций с меткой TBD постоянно меняется (на 1.01.2025 их было 2660).

4 О чем говорят аннотации с несколькими метками?

В цитированных работах, посвященных использованию аннотирования с несколькими метками, подчеркивается важность исследования этих меток. В частности, ставятся следующие вопросы:

- если коннектор размечен как выражающий несколько отношений, то все ли размеченные комбинации отношений являются допустимыми или некоторые из них можно признать аномальными?
- выводимы ли какие-то отношения из контекста, а не из значения коннектора?
- какие отношения дистрибутивно схожи, т.е. встречаются в одном контексте?

Чтобы ответить на эти и другие вопросы, необходимо, как нам кажется, разграничить два случая. В первом из них мы имеем дело с сочетаниями отношений, выражаемыми многоэлементными коннекторами, условно случай *ii*) в модели PDTB. Как правило, это сочетания сочинительного союза с «конкретизатором» (в терминологии русской грамматики) или с коннектором наречной природы, как в (4) или (6) выше. С одной стороны, эти сочетания позволяют нам выявить отношения, которые могут употребляться в одном контексте, а с другой – сочетаемость коннекторов обусловлена не только концептуальной совместимостью дискурсивных отношений, которые они выражают, но и их синтаксическими свойствами. Например, сочинительные или подчинительные союзы не могут сочетаться между собой. Кроме того, если показатели некоторых отношений избирательны в своих сочетаниях, что позволяет выявить некоторые закономерности (например, *хотя* в составе двухкомпонентного коннектора может сочетаться только с союзом *но*,

впрочем сочетается только с союзами *а* или *но*), то показатели других отношений не знают таких ограничений. Так, показатель отношения спецификации *например* сочетается со всеми сочинительными союзами: *и, а, но, или* и даже *то есть* (9).

(9) Осторожно предполагая, что новоискровцы – по сути оппортунисты и лакеи буржуазии, он отмечает, что пока это проявляется как «оппортунизм в организационных вопросах»; что ему можно противопоставить – например, право ЦК распускать местные организации (*то есть, например*, Заграничную лигу: централизм так централизм). [Л. А. Данилкин. Ленин: Пантократор солнечных пылинок (2017)]

Наконец, сочетание с сочинительным союзом может объясняться не столько желанием говорящего выразить еще какое-то дискурсивное отношение, сколько синтаксическими причинами. Например, коннектор *в то же время* не может связывать два однородных сказуемых, не прибегая к помощи союза *и*, и в НБДК сочетание соединительных отношений и отношений одновременности занимает первую позицию как среди аннотаций с несколькими отношениями (157 из 1669, т.е. 9,4%), так и среди аннотаций, сформированных для отношения одновременности (157 из 707, т.е. 22,2%).

Второй случай – комбинированная метка, сочетающая значения полисемичного коннектора, и здесь необходимо прояснить, что именно мы размечаем: отношение, т.е. то, как мы понимаем связь между двумя ситуациями, или коннектор? Получается некий замкнутый круг. В PDTB отношение проставляется на основе семантики коннектора, но разметчик может «чувствовать», что коннектор выражает еще какое-то отношение. Это происходит, как правило, под влиянием контекста. Например, союз *и* всегда сохраняет соединительное значение, которое является для него базовым. Другое дело, что под влиянием смыслового соотношения фрагментов текста, у него можно увидеть значение временного следования, результативно-следственное и другие, которые выводятся, в том числе, с учетом наших экстралингвистических знаний, но которые, тем не менее, фиксируются словарями. Среди комбинированных меток PDTB вторая по частотности (138 из 999 аннотаций) как раз приходится на союз *and*, «совмещающий» причинное и соединительное значения (Prasad et al. 2014: 933).

В НБДК было принято решение размечать союз *и* всегда только как соединительный. Но здесь интересно посмотреть, с каким семантическим соотношением фрагментов текста совместимо это соединительное значение. Это значит, однако, что мы будем фиксировать не столько отношение или значение коннектора, сколько семантические свойства контекста, в котором он употреблен. Так, для показателя отношения переформулирования в итальянском языке *cioè* ‘то есть’ было показано, что при всем разнообразии семантических отношений между фрагментами текста, с которым он совместим, он не может быть употреблен, если между ними есть причинное отношение (Manzotti 1999). Это важный параметр, позволяющий более четко описать дискурсивные отношения. Неправомерно, на наш взгляд, приписывать коннектору то, что мы выводим из контекста. Например, *в то же время* может употребляться для связи фрагментов текста, находящихся в отношении антонимии. Это способствует его интерпретации как сопоставительного (или даже противительного), но при этом его базовое значение одновременности сохраняется, как в (10):

(10) Крымов, посмеиваясь *и в то же время* серьезно, перебил: – А понадобится, и вас, Греков, с большевистской кашей съедят. [В. С. Гроссман. Жизнь и судьба (1959)]

Можно ли здесь говорить о «совмещении» значений коннектора, если отношение противопоставления диктуется контекстом? И наоборот, в некоторых контекстах сопоставительное значение полностью вытесняет временное (11), следовательно, здесь также несправедливо говорить о каком-либо совмещении значений.

(11) По нашим данным, 70% девушек твердо собираются покинуть свой город, и 17% пока колеблются; *в то же время* среди юношей уверенных – 54%, а колеблющихся– 28%. [Ю. Ф. Флоринская, Т. Г. Рощина. Жизненные планы выпускников школ из малых городов (2004) // «Человек», 2005]

Контексты, в которых коннекторы действительно неоднозначны, т.е. те, что не позволяют выбрать то или иное значение полисемичного коннектора, крайне редки; ср. приведенный в таблице 2 пример для *точнее*. Неслучайно поэтому, что в НБДК проставлено такое малое количество комбинированных меток (0,3%).

В комбинированных метках также, как и в случае аннотаций с несколькими метками отношений, интересно проанализировать сочетание отношений, между которыми разметчик затрудняется сделать выбор. Даже самый предварительный анализ позволяет говорить о том, что разметчик затрудняется выбрать между семантически близкими отношениями. В НБДК одной из распространенных комбинированных меток является метка «Одновременность / Условие» для союзов *когда* и *как*. В PDTB эта метка (Contingency.Condition / Temporal.Synchrony) для коннектора *when* тоже является одной из наиболее частотных (Prasad et al. 2014: 933). Сочетания отношений в комбинированных метках позволяет, таким образом, дополнить анализ степени семантической близости отношений, проводимый на основе их структурированных определений (подробнее см. Инькова, Кружков 2023a,b).

5 Заключение

В заключение сказанного хотелось бы сформулировать дальнейшие этапы исследования дискурсивных отношений. На наш взгляд, следующим важным этапом в исследовании дискурсивных отношений должно стать аннотирование семантических свойств фрагментов текста, между которыми установлено дискурсивное отношение. Это позволит определить, какие отношения могут встречаться в одном контексте, т.е. дистрибутивно близки, а также предсказать, каким образом может быть разрешена неоднозначность; ср. англ. *so* в (1) и (2) выше. Это, в свою очередь, поможет установить, какие сочетания отношений являются нормальными, а какие «аномальны», что повысит качество разметки. При этом особенности контекста должны быть сформулированы на доступном машине языком, чтобы улучшить автоматическую обработку текста, которая особенно затруднена в случае имплицитных дискурсивных отношений. Они смогут, таким образом, определяться не путем подстановки соответствующего по смыслу коннектора, которые, как мы видели, неоднозначны, а на основе семантических свойств контекста.

Благодарности

Работа выполнена в Федеральном исследовательском центре «Информатика и управление» Российской академии наук с использованием ЦКП «Информатика» ФИЦ ИУ РАН.

Acknowledgements

The research was carried out using the infrastructure of the Shared Research Facilities “High Performance Computing and Big Data” (CKP “Informatics”) of FRC CSC RAS (Moscow).

References

[1] Aroyo L., Welty Ch. Truth is a lie: Crowd truth and the seven myths of human annotation // AI Magazine. – 2015. – Vol. 36:1. P. 15-24.

[2] Basile V., Fell M., Fornaciari T., Hovy D., Paun S., Plank B., Poesio M., UmaA. We need to consider disagreement in evaluation // Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future. – Bangkok, Thailand, 2021. P. 15-21. Access mode: https://aclanthology.org/2021.bppf-1.pdf.

[3] Carlson L., Marcu D. Discourse Tagging Reference Manual. ISI-TR-545. University of Southern California Information Sciences Institute, 2001. Access mode: http://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf.

[4] Inkova O. Linguospecificity of connectives: methods and parameters of description [Lingvospecifichnost’ konnektorov: metody i parametry opisaniya] // O. Inkova (ed.), Semantics of connectives: a contrastive study [Semantika konnektorov: kontrastivnoe issledovanie]. – Moscow: TORUS PRESS, 2018. P. 5-23.

[5] Inkova O., Kruzhkov M. Evaluating the degree of discourse relations semantic affinity: methods and instruments [Stepen semanticheskoj blizosti diskursivnyx otnoshenij: metody i instrumenty rascheta] // Informatics and applications. – 2023b. – Vol. 17:4. P. 46-53.

[6] Kobozeva I.M., Serdobolskaya N.V., Kryukova A.I., Pilyugina D.A. Creating a database of Modern Russian clause linkers: principles, problems, results [Sozdanie bazy dannyx konnektorov sovremennogo russkogo yazyka: principy, problemy, rezultaty] // Tyumen State University Herald. Humanitites Research. Humanitates. – 2023. – Vol. 9:4. P. 36-47. Access mode: https://doi.org/10.21684/2411-197X-2023-9-4-36-47
[7] Long W., Siddharth N., Webber B. Multi-Label Classification for Implicit Discourse Relation Recognition // Computing Research Repository. – 2024. – Vol. arXiv:2406.04461 [cs.CL]. Access mode: https://arxiv.org/abs/2406.04461.
[8] Lopez Cortez S. M., Jacobs C. L. Incorporating Annotator Uncertainty into Representations of Discourse Relations // Computing Research Repository. – 2023. – Vol. arXiv:2308.07179 [cs.CL]. Access mode: https://arxiv.org/abs/2308.07179.
[9] Mann W.C., Matthiessen Ch.M.I.M., Thompson S. A. Rhetorical structure theory and text analysis. ISI Research Report ISI/RR-89-242, 1989. DOI: 10.1075/pbns.16.04man.
[10] Manzotti E. Spiegazione, riformulazione, correzione, alternativa: sulla semantica di alcuni tipi e segnali di parafrasi // L. Lumbelli, B. Mortara Garavelli (a c. di), Parafrasi. Dalla ricerca linguistica alla ricerca psicopedagogica. – Alessandria: Edizioni dell'Orso, 1999. P. 169-206.
[11] Marchal M., Scholman M., Yung F., Demberg V. Establishing annotation quality in multi-label annotations // Proceedings of the 29th International Conference on Computational Linguistics. – Gyeongju, Republic of Korea, 2022. P. 3659-3668. Access mode: https://aclanthology.org/2022.coling-1.322/.
[12] Nuriev V., Buntman N., Inkova O. Machine translation of Russian connectives into French: errors and quality failures // Informatics and applications. – 2018. – Vol. 12:2. P. 105-113.
[13] PDTB Research Groupe. The Penn Discourse Treebank 2.0 Annotation Manual. Technical Report IRCS-08-01. Philadelphia: Institute for Research in Cognitive Science, University of Pennsylvania, 2008. Access mode: https://www.seas.upenn.edu/~pdtb/PDTBAPI/pdtb- annotation- manual.pdf.
[14] Prasad R., Webber B., Joshi A. Reflections on the Penn Discourse TreeBank, Comparable Corpora, and Complementary Annotation // Computational Linguistics. – 2014. – Vol. 40:4. P. 921-950. Access mode: https://aclanthology.org/J14-4007.pdf.
[15] Prasad R., Webber B., Lee A., Joshi A. Penn Discourse Treebank Version 3.0 // Linguistic Data Consortium. – 2019. – Vol. LDC2019T05. Access mode: https://catalog.ldc.upenn.edu/LDC2019T05.
[16] Scholman M., Dong T., Yung F., Demberg V. DiscoGeM: A crowdsourced corpus of genre-mixed implicit discourse relations // Proceedings of the Thirteenth Language Resources and Evaluation Conference. – Marseille, France, 2022. P. 3281-3290. Access mode: https://aclanthology.org/2022.lrec-1.351/.
[17] Spooren W., Degand L. Coding coherence relations: Reliability and validity // Corpus Linguistics and Linguistic Theory. – 2010. – Vol. 6:2. P. 241-266.
[18] Stede M. RST revisited: Disentangling nuclearity // C. Fabricius-Hansen & W. Ramm, editors, Subordination' versus 'coordination' in sentence and text from a cross-linguistic perspective. John Benjamins: Amsterdam, 2008. P. 33-58.
[19] The Russian National Corpus [Nacional'nyi korpus russkogo yazyka]. Access mode: https://ruscorpora.ru/
[20] Webber B. What excludes an alternative in coherence relations? // Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013). – Potsdam, Germany, 2013. – P. 276-287.
[21] Zufferey S., Degand L. Connectives and Discourse. – Cambridge: Cambridge University Press, 2024.

An Algorithm for Genre Imbalance Correction in the Russian Subcorpus of the Google Books Ngram Corpus

Anna Ivleva
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
ivleva.anna.igorevna@gmail.com

Valery Solovyev
Laboratory of Linguistics and AI,
Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia
maki.solovyev@mail.ru

Abstract

Recently, the latest version of the Google Books Ngram was shown to be imbalanced. This corpus being an important and widely used database, the imbalance can affect the results of research in different areas. In the paper, we present an algorithm for correcting the dynamics of word frequency in the Google Books Ngram corpus. The algorithm takes into account the discovered imbalance of the main literary styles—fiction, publicistics, and non-fiction. The rationale of the algorithm is given, as well as examples of correction.

Keywords: the Google Books Ngram corpus; genre imbalance; word frequency; time series correction
DOI: 10.28995/2075-7182-2025-23-137-146

Алгоритм коррекции жанрового дисбаланса в подкорпусе русского языка Google Books Ngram

Аннотация

Недавно было показано, что последняя версия Google Books Ngram — важного исследовательского цифрового ресурса — не сбалансирована, что может исказить результаты использующих его работ. В статье представлен алгоритм коррекции динамики частоты слов в корпусе Google Books Ngram с учетом обнаруженного дисбаланса основных стилей литературы — художественного, публицистического и нехудожественной литературы. Приводится обоснование алгоритма, а также примеры коррекции частот слов.

Ключевые слова: корпус Google Books Ngram, дисбаланс стилей, частота слова, коррекция временных рядов

1 Introduction

The Google Books Ngram corpus (https://books.google.com/ngrams, hereinafter abbreviated GBN) continues to be a popular research tool in linguistics, NLP, as well as in many social and cultural studies [1, 2, 3]. Up-to-date review of the corpus application is given in [4]. It gives access to large diachronic data on word frequencies significant for quantitative studies of the evolution of both language and society. The GBN corpus was made by overall scanning of books from the university libraries [1, 5]. Now there are three GBN versions – 2009, 2012, and the last of 2020 recently replenished with the data up to 2022.

The diachronic genre representativeness and stability is one of the most significant parameters. This is due to the fact that word frequency trends can be treated as indicators of societal and/or language changes. But changes in word frequency can be also caused by the imbalanced genre composition of the corpus. The issues of the GBN balance and representativeness were debated in [6, 7]. Some problems were pointed out, which Google promptly took into account and made changes to the corpus structure. As a result, [6] concluded that the second version of GBN became up-to-date, representative and balanced.

Previously, we came across some peculiarities of the word frequency in the Russian subcorpus of the last version of GBN. First of all, we found the frequency dynamics of the words specific to fiction to be highly synchronous after 1999. In [8], the authors gave specific examples of this sort and provided a more detailed and general analysis of the corpus based on the frequency dynamics of 7420 words from the Frequency Vocabulary of Fiction, Frequency Vocabulary of Publicistics and Frequency Vocabulary of Non-fiction [9]. We analyzed the average variation coefficient, the average autocorrelations, as well as some other statistical parameters of the frequency dynamics. Specific years and intervals with significant imbalance of genres were obtained. We concluded unjustified overall increasing trend in the share of fiction from 2000 to 2019, with essential outbreaks in 2000, 2011–2012 and 2018–2019.

If we take the frequency data directly from the third version of the GBN corpus, it can result in incorrect conclusions. For example, we can see the frequency of the pronoun *я (I)* increasing. As noted in [8], according to the general approach developed in [9], this can be interpreted as increase in individualism in society. But it can also be caused by the growth of the share of fiction, where the pronoun *я (I)* is much more frequent than in other genres. The issue gets even more complex and complicated due to the fact that Google does not publish the shares of texts genres and styles.

The present research is devoted to the algorithm of word frequency correction to compensate for the previously found imbalances.

The second section of the article describes the algorithm for genre imbalance correction and some examples of its application. In the 3rd section, we discuss the algorithm operation and results. In conclusion, we summarize the results of our research.

2 Results

We correct the word frequency dynamics by applying the time series analysis.

At the first stage, standalone outbreaks of word frequencies are corrected. We consider that in case of balance, the average noise – the fluctuations that do not relate to any trends or other systematic changes – is to be close to zero for large lists of words as individual frequency dynamics compensate for one another. That is why outbreaks in the averaged noise are treated as indicators of imbalance.

The second stage is the correction of word frequency trends. We analyze the trends for the words specific to different genres and correct the unjustified trends of the XXI century.

2.1 Outliers correction

First, we made preliminary analysis of highly specific words. The lists of the words (930 fiction, 434 publicistics and 981 non-fiction words) were taken from [9], auxiliary parts of speech were excluded. Also, the list of out-of-genre words was prepared on the basis of 20000 most frequent words from [9]. Words from the above lists of highly specific genre words, as well as auxiliary parts of speech were excluded from it. As a result, the list of out-of-genre words contained 17468 words. We applied the method of singular spectral analysis (SSA) [10] to extract the noise from the data. In Figure 1, we can see the average noise for out-of-genre words. Several outbreaks down in 1964, 2013 and up in 2000 show some general disbalance of the corpus composition or some mistakes not connected with the imbalance of the three genres we consider. Still, the variability of the noise in Figure 1 as well as the magnitude of the outbreaks are not as great as in Figure 2 that shows the average noise in frequency dynamics for words specific to different genres.

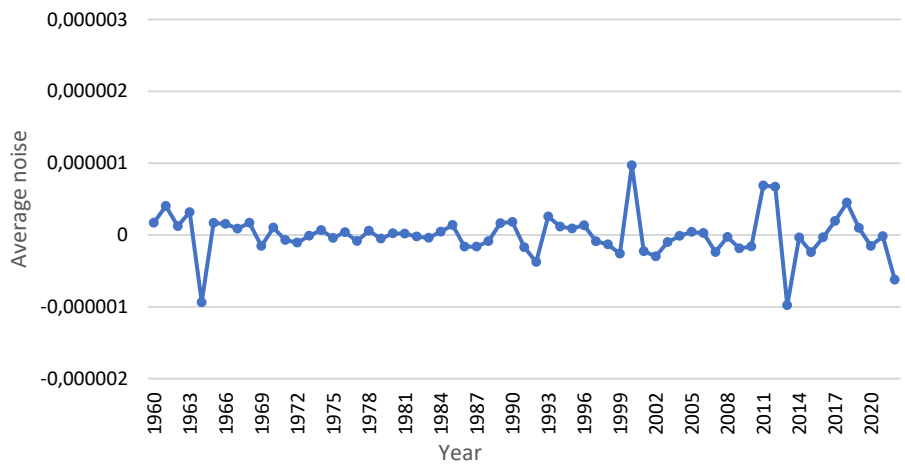


Figure 1: Average noise in frequency dynamics for out-of-genre words

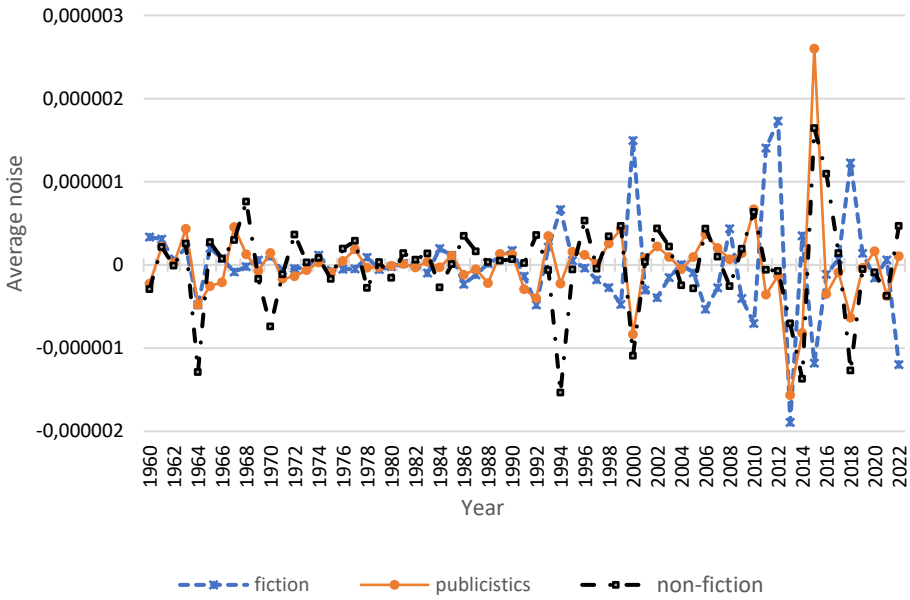


Figure 2: Average noise in frequency dynamics for words specific to different genres

The noise outliers were identified by means of the Hampel filter (the method that treats outliers as values outside the interval formed by the median value plus and minus 3 median absolute deviations). It gives the same results as the inter quartile range approach in case of 1*IQR value (25, 75 percentiles range). Table 1 gives the lists of years we identify as outliers.

Fiction		Publicistics		Non-fiction	
Outbreaks up	Outbreaks down	Outbreaks up	Outbreaks down	Outbreaks up	Outbreaks down
1994, 2000, 2008, 2011, 2012, 2018	1964, 1992, 1999, 2006, 2010, 2013, 2015	1963, 1967, 1999, 2010, 2015	1964, 2000, 2013, 2014, 2018	1968, 2015, 2016	1964, 1970, 1994, 2000, 2013, 2014, 2018

Table 1: The lists of average noise outbreaks

The first stage of our algorithm corrects the word frequency data for the period 1960–2022 according to the following formula:

$$data_{corr}^{1,year} = a \cdot data_{GBN}^{year} + b \cdot data_{SSA}^{year},$$

where

$data_{GBN}^{year}$ is the raw data taken from the GBN corpus;
$data_{SSA}^{year}$ is the value obtained by the SSA method;
a is the sum of shares of RNC word frequencies for the genres that do not need correction according to Table 1;
b is the sum of shares of RNC word frequencies for the genres that need correction according to Table 1.

For example, the correction formula for 2008 is:

$$data_{corr}^{1,2008} = (a_1 + a_2) \cdot data_{GBN}^{2008} + b_1 \cdot data_{SSA}^{2008},$$

where a_1 is the word frequency share in publicistics, a_2 is the word frequency share in non-fiction, b_1 is the word frequency share in fiction. The coefficients are like this because in 2008 the outbreak is found for fiction only.

The shares of word frequencies are calculated based on the genre subcorpora of RNC (https://ruscorpora.ru/) provided by the authors of RNC. Subcorpus frequencies for about 47 thousand of most frequent words are calculated in advance, so that the algorithm works faster. In case the word is not present in the database, the direct search in the RNC database takes place.

2.2 Trend correction

The second step of the correction algorithm is the trend correction applied to the data obtained in the first stage.

The average trends of frequencies for words specific to different genres are shown in Figure 3.

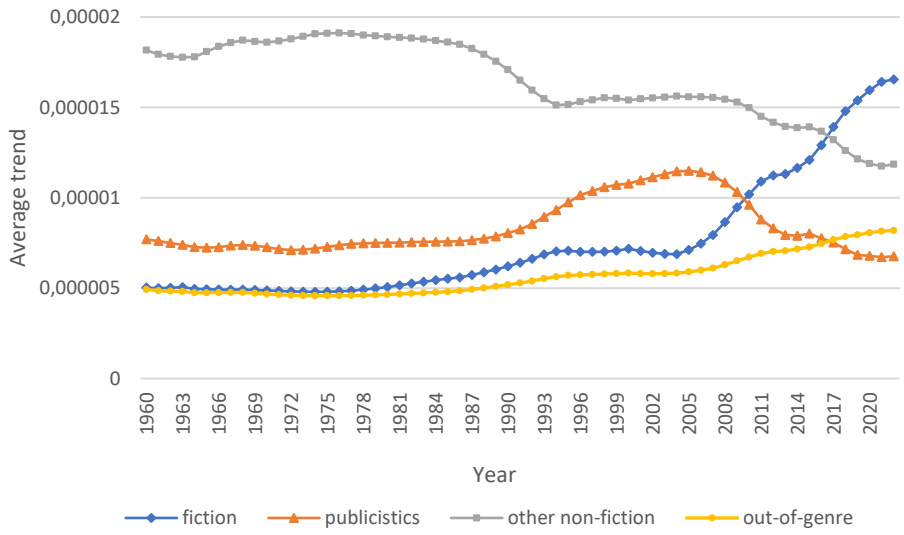


Figure 3: Average trends of frequencies for words specific to different genres

The trends are not constant. Also, we can see they compensate for one another to some extent. In the 80's of the XX century, both fiction and publicistics trends increase, whereas non-fiction trend starts to decrease. In the last decade of the XX century, the trends of publicistics and non-fiction decrease, while the fiction trend continues to vividly grow. In general, the sharpest and greatest in magnitude is the

fiction trend increase starting at the beginning of the XXI century. The difference in trend between 2000 and 2022 is almost 3 times greater than the maximum variation of other trends.

As stated in [4], it is not due to increase in the frequency of words typical for fiction in the Russian. The word frequency dynamics in the RNC shows the opposite tendencies for fiction and non-fiction as well as increase in the use of slang, vernacular, and technical terms in the fiction described in [7]. Also, over the beginning of the XXI century the share of the Russian-language books published has not increased [8].

Due to the above arguments in favor of the sharpest and unjustified change of trends in the beginning of the XXI century, we include the step of detrending in our algorithm.

To get the point where the strongest trend change starts, we divide the time period into two parts and obtain slopes of linear trends for both parts. Then we average the slopes through all the 930 fiction, 434 publicistics and 981 non-fiction words. Figure 4 presents the difference between the two slopes for each point of division. The maximum for fiction and minima for publicistics and non-fiction are at the point of 2004-2005. That is why we detrend the data in the period 2004–2022.

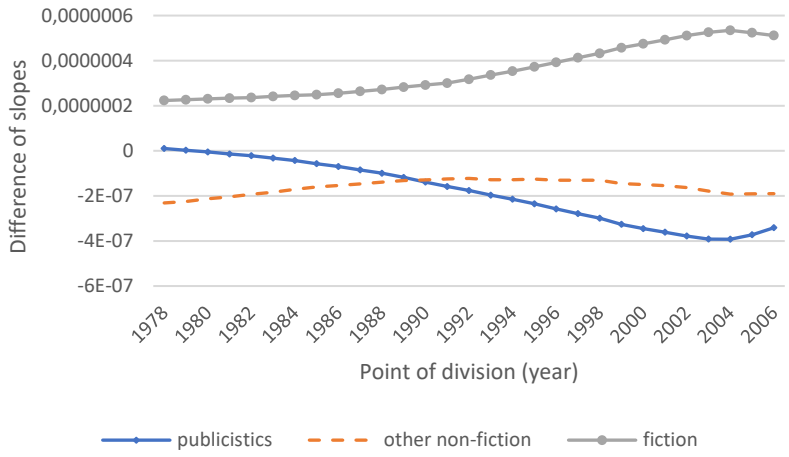


Figure 4: Average difference of slopes in the point of division

First, we partially detrend the data in the period of 2004–2022 by the SSA method leaving the baseline (the average value of the data in the period until 2004) and the trend until 2004.

$$data_{corr}^{2,year} = data_{corr}^{1,year} - data_{SSA}^{year} + baseline + data_trend_{SSA}^{1960-2004}.$$

Here,

$data_{corr}^{1,year}$ is the result of correction at the first stage;
$data_{SSA}^{year}$ is the trend we remove;
$baseline$ is the average value of the word frequency within the period 1999–2004 (to conjoin the corrected data);
$data_trend_{SSA}^{1960-2004}$ is the linear trend obtained by the Mann-Kendall test for the period before increase.

2.3 Specific examples of correction

The correction algorithm is implemented in Python 3. In this section, we give the illustration of its application. Frequency plots for several lexemes specific to different genres are given before and after correction. The words *комната (room)* and *подумать (to speculate)* are specific to fiction (fiction shares are 0,69 and 0,80 accordingly). The shares are calculated using the RNC. In Figure 5, we can see the smoothing of peaks after the first stage of algorithm application and partial detrending of the word frequency in the XXI century after the second stage. As a result of correction, the word frequencies increase in the XXI century, but not that much (Figure 5).

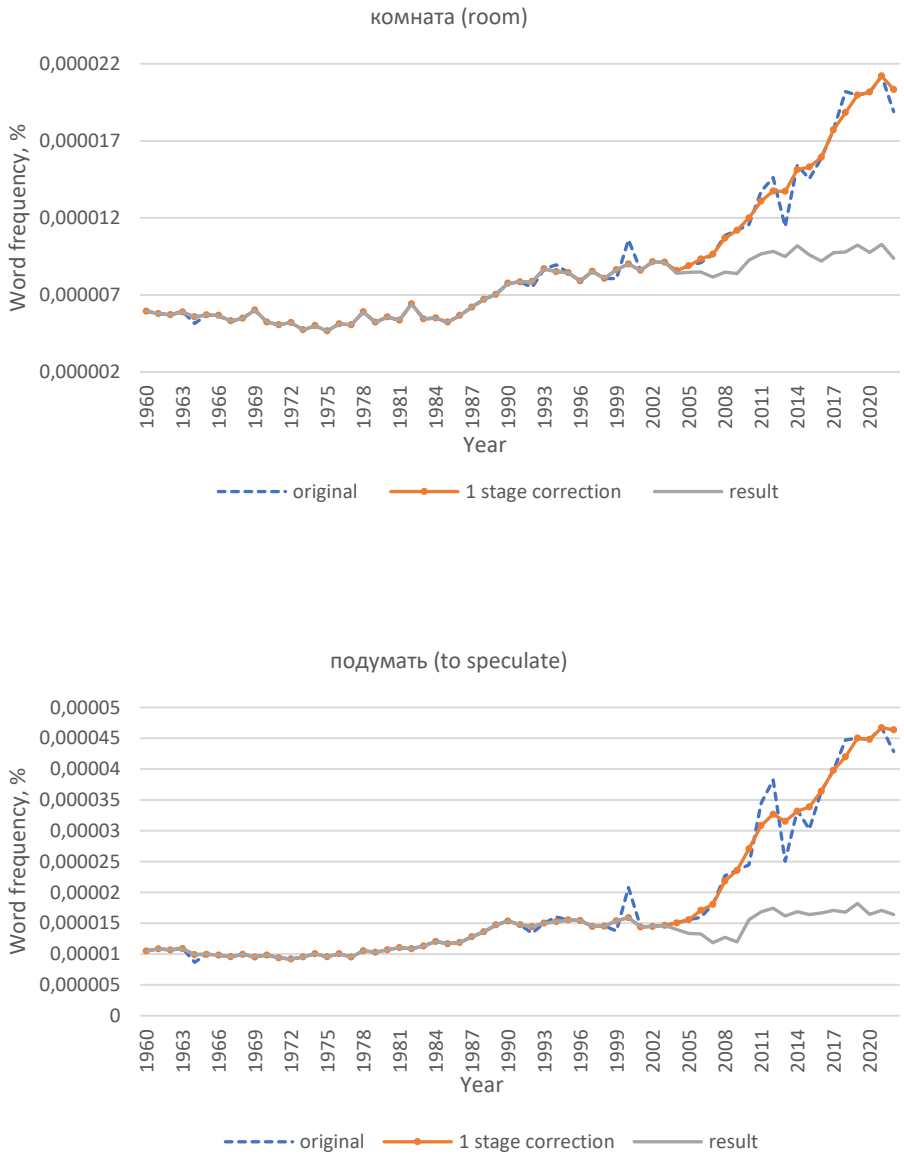


Figure 5: Correction of the words *комната (room)* and *подумать (to speculate)* specific to fiction

The word *президентский (presidential)* is specific to publicistics with the share equal to 0,72. The same publicistics coefficient is for the word *акция (shareholding)*. The sharp unnatural frequency jumps of these words as well as the general decrease are caused by the drop of the share of publicistics in the corpus. These effects are corrected (Figure 6).

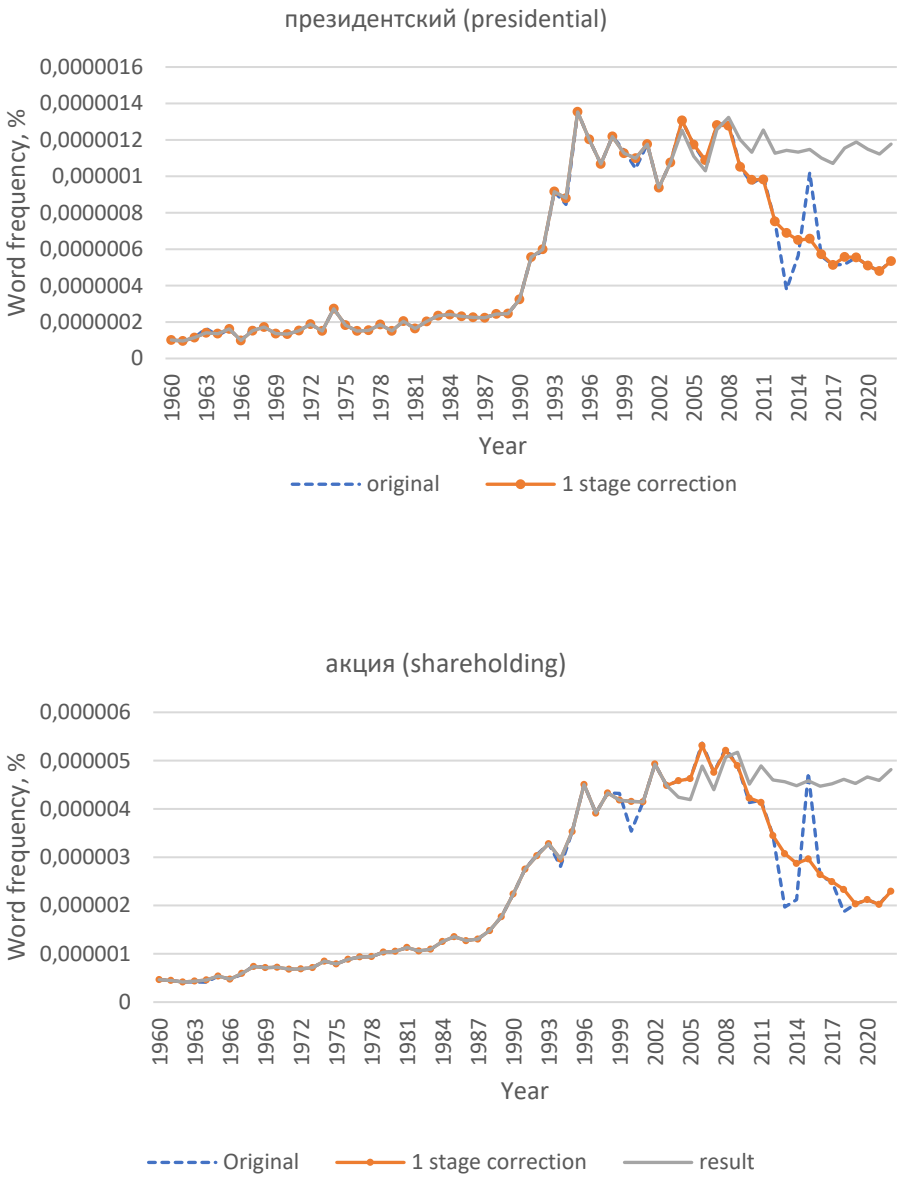


Figure 6: Correction of the words *президентский (presidential)* and *акция (shareholding)* specific to publicistics

The words *теоретический (theoretical)* and *таблица (table)* are specific to non-fiction (including scientific and academic literature) – non-fiction shares are 0,67 and 0,78 accordingly. Our algorithm removes unnatural frequency jumps and corrects the overall trend after 2004 (Figure 7).

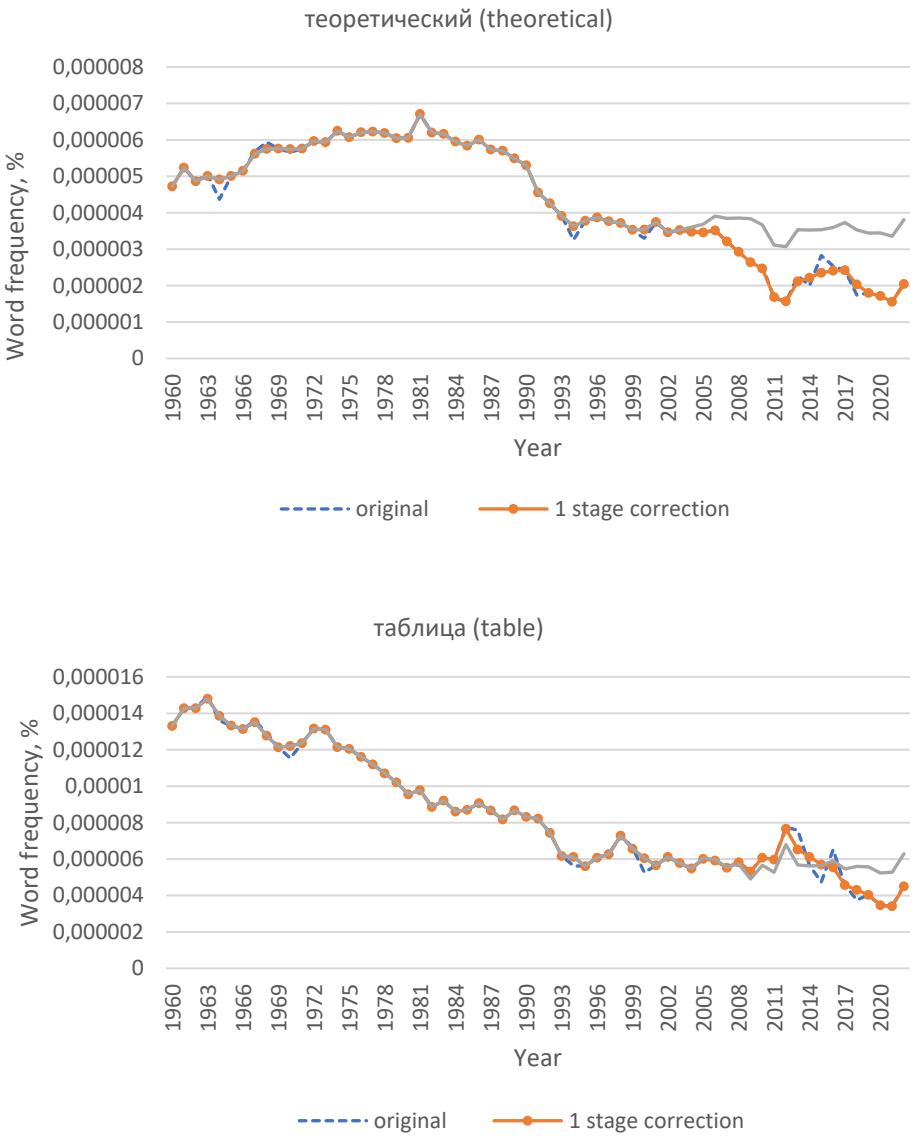


Figure 7: Correction of the words *теоретический (theoretical)* and *таблица (table)* specific to non-fiction

The algorithm can be applied to the words that are different parts of speech. No qualitative or quantitative difference in the result obtained was found. Here is an example of personal pronouns correction. The Russian pronoun *я (I)* is more typical for fiction. The fiction share for *мы (we)* is 0.30, for *я (I)* it is 0.68. According to the frequency dictionary, the pronoun *я (I)* is more frequent than *мы (we)*. Before the use of correction algorithm, the frequency of these pronouns, especially *я (I)*, was growing significantly in the XXI century: from 1999 to 2020 the frequency of the pronoun *я (I)* increased by about 3 times and the frequency of *мы (we)* increased by 1.5 times. This growth can naturally be explained by the growing share of fiction in the GBN, but it can also be due to the growth of individualism in the Russian society [3]. After our correction, as seen in Figure 8, the frequency gap between *я (I)* and *мы (we)* is significantly reduced. After eliminating the effect of the genre imbalance, the remaining increase in the frequency of the pronoun *я (I)* can be interpreted as the increase in the degree of individualism.

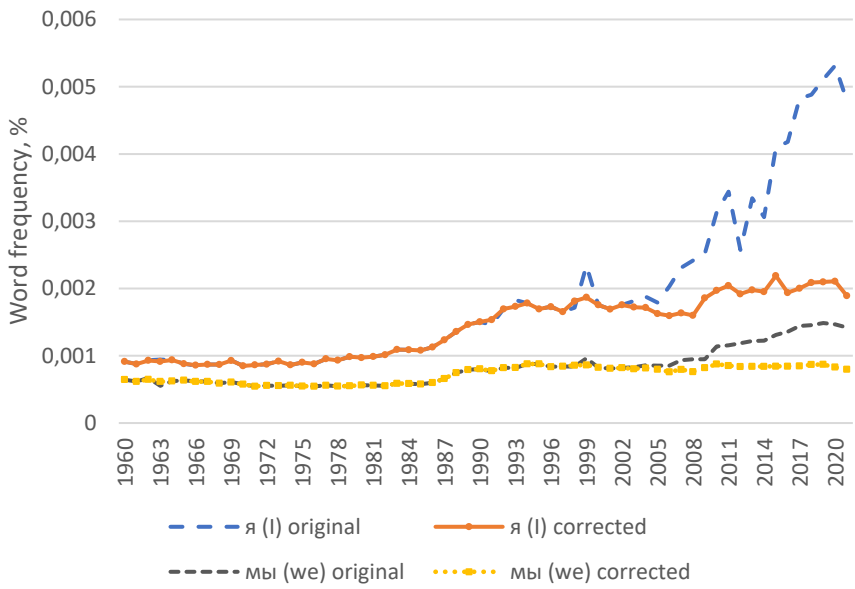


Figure 8: Correction of the pronouns *я (I)* and *мы (we)*

3 Limitations

We should note two points.

1. The main difficulty in frequency correction is unknown proportion of different genres in the GBN corpus. The above correction formulas use the data from the RNC corpus. This definite simplification – an enforcement action – is the best approximation currently available. We can only note that the resulting corrected data looks much more realistic than the original ones.

2. The discovered serious imbalance of genres is primarily in the XXI century. Previously, only individual noise peaks were identified. Therefore, the purpose of the present research is to adjust the frequencies in this particular time interval using the data from the previous period of 1960–1999 as the background. Errors in the frequencies of words before 1917 can have reasons completely different from the genre imbalance, namely, incorrect recognition of letters in the old spelling of the Russian language. Also, the problem of correction of the frequency data in the XIX century (and earlier) is complicated by a small volume of the corpus those years. Since 1960, the volume of the corpora has reached a fairly significant amount, about 1 billion words per year and it is stabilizing.

4 Conclusion

The essential genre imbalance that has appeared in the latest version of GBN in the XXI century is a significant obstacle to apply this corpus for diachronic sociological research. The reasons for this imbalance are not known. It is present not only in the Russian subcorpus, but also in the subcorpus of other languages, most vividly in British. The correction algorithm presented enables to more efficiently use the word frequency data from the Google Books Ngram corpus for research. The algorithm cannot provide results equivalent to a completely balanced corpus, but it gives a good approximation to the real word frequencies. Although in the present paper we deal only with the Russian language subcorpus, the approach proposed is language-independent and can be applied to other language corpora as well. Our correction algorithm is one of the options. Other approaches and improvements can be considered in future. In any case, the general recommendation for researchers is to take into account the discovered genre imbalance of GBN in the XXI century.

Acknowledgements

The research was funded by the Russian Science Foundation (project No. 24-18-00570, https://rscf.ru/project/24-18-00570/).

References

[1] Michel J.-B., Shen Y. K. et al. Quantitative analysis of culture using millions of digitized books // Science. — 2011. — Vol. 331(6014). — P. 176–182.

[2] Zeng R., Greenfield P.M. Cultural evolution over the last 40 years in China: Using the Google Ngram Viewer to study implications of social and political change for cultural values // International Journal of Psychology. — 2015. — Vol. 50(1). — P. 47–55.

[3] Velichkovsky B.B., Solovyev V.D. et al. Transition to market economy promotes individualistic values: analysing changes in frequencies of Russian words from 1980 to 2008 // Int. J. Psychol. — 2019. — Vol. 54(1). — P. 23–32.

[4] Solovyev V. Using the Google Books Ngram Corpus to Study Social Evolution // Social Evolution & History. — 2024. — Vol. 23(2). — P. 144–164.

[5] Madsen D. Ø., Slåtten K. The possibilities and limitations of using Google Books Ngram Viewer in research on management fashions // Societies. — 2022. — Vol.12(6). — P. 171.

[6] Solovyev V.D., Bochkarev V.V. et al. Google Books Ngram: problems of representativeness and data reliability. // Proceedings of Data Analytics and Management in Data Intensive Domains. DAMDID/RCDL 2019. Communications in Computer and Information Science. — Springer, Cham, 2020. — Vol. 1223. — P. 147–162.

[7] Koplenig A. The impact of lacking metadata for the measurement of cultural and linguistic change using the Google Ngram data sets – reconstructing the composition of the German corpus in times of WWII // Digit. Scholarsh. Humanit. — 2017. — Vol. 32. — P. 169–188.

[8] Solovyev V., Ivleva A. How to Detect Imbalances in the Google Books Ngram Corpus? Proceedings of Speech and Computer. SPECOM 2024. Lecture Notes in Computer Science. — Springer, Cham, 2025. — Vol. 15300. — P.334–348.

[9] Frequency Dictionary of the Modern Russian Language homepage. http://dict.ruslang.ru/freq.php, last accessed 2024/07/13.

[10] Yu F., Peng T. et al. Cultural value shifting in pronoun use // Journal of Cross-Cultural Psychology. — 2016. — Vol. 47(2). — P. 310–316.

[11] Elsner J.B., Tsonis A.A. Singular spectrum analysis: a new tool in time series analysis. — Springer Science & Business Media, 2013.

[12] Solganik G.Ya. (2010), The modern linguistic situation and trends in the development of the Russian literary language [Sovremennaya yazykovaya situaciya i tendencii razvitiya russkogo literaturnogo yazyka], Bulletin of the Moscow University, Journalism Series 10 [Vestnik Moskovskogo universiteta. Seriya 10. Zhurnalistika], Vol. 10(5), pp. 122–134

[13] The Russian book market. The state, trends and prospects of development. https://bookunion.ru/upload/files/Bookmarket-2022.pdf?ysclid=lyk1taafsq277227991, last accessed 2024/07/13.

Computer tools in mental disorders diagnostics by oral speech

Khomenko Anna
HSE University
Nizhny Novgorod, Russia
akhomenko@hse.ru

Komratova Anastasia
HSE University
Nizhny Novgorod, Russia
akomratova@hse.ru

Isakov Danila
HSE University
Nizhny Novgorod, Russia
disakov@hse.ru

Balba Daria
HSE University
Nizhny Novgorod, Russia
dasha.balba@gmail.com

Shishkovskaya Tatiana
HSE University
Moscow, Russia;
Mental Health Research Center
Moscow, Russia
tszyszkowska@gmail.com

Khudyakova Maria
HSE University
Nizhny Novgorod, Russia
mariya.kh@gmail.com

Abstract

The integration of automated speech analysis in diagnosing mental health disorders is becoming increasingly significant in both clinical and computational linguistics. This study aims to construct linguistic profiles for individuals with neurocognitive and affective mental disorders. Using speech transcriptions and relevant to the study computational techniques like lexical clustering and stylostatistical analysis, this research looks for characteristics capable of distinguishing speech patterns indicative of various mental health conditions. A text corpus of oral speech from 136 people diagnosed with schizophrenia, schizotypal disorder, schizoaffective disorder, borderline personality disorder, other personality disorders, primary depressive episode, recurrent depressive disorder, bipolar affective disorder and 210 participants with no diagnosed diseases in the control group was used in the research. As a result of the study, it was proved that people with mental disorders display specific features in oral speech, that can be used in creation of an automatic mental disorders diagnostic model.

Keywords: mental disorders; early automatic diagnostics; oral speech; clustering; stylistics
DOI: 10.28995/2075-7182-2025-23-147-157

Применение компьютерных инструментов в диагностике психических расстройств по устной речи

Хоменко Анна
НИУ ВШЭ
Нижний Новгород, Россия
akhomenko@hse.ru

Комратова Анастасия
НИУ ВШЭ
Нижний Новгород, Россия
akomratova@hse.ru

Исаков Данила
НИУ ВШЭ
Нижний Новгород, Россия
disakov@hse.ru

Бальба Дарья
НИУ ВШЭ
Нижний Новгород, Россия
dasha.balba@gmail.com

Шишковская Татьяна
НИУ ВШЭ
Москва, Россия;
Научный центр
психического здоровья
Москва, Россия
tszyszkowska@gmail.com

Худякова Мария
НИУ ВШЭ
Нижний Новгород, Россия
mariya.kh@gmail.com

Аннотация

Диагностика ментальных расстройств по продуцируемой речи является важнейшей задачей среди технологических разработок моделей нейрокогнитивных и психолого-психиатрических процессов. Целью данного исследования является построение лингвистических профилей для лиц с нейрокогнитивными и аффективными психическими расстройствами. Используя транскрипты устной речи и релевантные для исследования компьютерные технологии, такие как лексическая кластеризация и стилостатистический анализ, разрабатывается пул характеристик, способных различать речевые паттерны, указывающие на различные состояния психического здоровья. В работе использовался текстовый корпус устной речи 136 человек с диагнозом шизофрения, шизотипическое расстройство, шизоаффективное расстройство, пограничное расстройство личности, другие расстройства личности, первичный депрессивный эпизод, рецидивирующее депрессивное расстройство, биполярное аффективное расстройство и 210 участников контрольной группы без диагностированных заболеваний. В результате исследования было доказано, что устная речь людей с психическими расстройствами имеет особые диагностические признаки, которые могут быть использованы при создании автоматической модели определения психических расстройств.

Ключевые слова: ментальные расстройства; ранняя автоматическая диагностика; устная речь; кластеризация; стилостатистика

1 Introduction

Early diagnosis of cognitive impairments is a critical task in psychiatry, neurology, and pedagogy, as timely intervention can prevent complications and support recovery (Khudyakova et al., 2023). Diagnosing mental disorders through speech analysis is an emerging field within technological advancements focused on neurocognitive processes. Clinical linguistics, leveraging the analysis of oral speech, plays a pivotal role at both microlinguistic (phonetic, semantic, lexical) and macrolinguistic levels. Traditionally, this process involves manual profiling — expert interpretative processing of experimental material—which can be subjective and time-consuming. However, advances in computational linguistics now allow for automated tools to analyze and evaluate speech patterns objectively, addressing the growing need for efficient and standardized diagnostic methods (Malgaroli et al. 2023; Le et al. 2021; Ribeiro et al., 2024).

Speech profiling involves constructing a linguistic portrait of individuals, comparing those with diagnosed cognitive impairments to the control group. These profiles capture the distinctive linguistic characteristics associated with mental disorders, offering insights into cognitive and emotional states (Mota et al., 2012). Many existing models focus primarily on acoustic characteristics of audio signals, such as pitch, tempo, and pause duration, making them relatively universal and non-language-specific (Koops et al., 2023; Ramanarayanan et al., 2022). However, such approaches often overlook the broader set of linguistic data, including syntactic complexity and discourse coherence, which are critical for comprehensive diagnostics.

This research explores speech profiling for generalized language personalities of individuals with mental disorders, focusing on Russian-language oral discourse. Unlike prior studies predominantly conducted on English-language material, this study addresses a significant gap in research by constructing speech profiles for individuals with borderline personality disorder, recurrent depressive disorder, bipolar affective disorder, and other conditions within the Russian-speaking population. By integrating microlinguistic and macrolinguistic features, the research aims to develop automated models capable of identifying speech patterns indicative of neurocognitive and affective mental disorders.

The potential of speech profiling has been demonstrated in studies examining various linguistic features to diagnose schizophrenia, bipolar disorder (Mota et al., 2023), and depression (Khomenko et al., 2023).

The inclusion of linguistic markers, such as syntactic complexity and semantic coherence, enriches the diagnostic process. For Russian-speaking discourse, this approach remains largely unexplored, despite its promise for creating functional diagnostic models tailored to specific populations.

By building on methods from forensic authorship linguistics and speech profiling, this research aims to advance clinical linguistics through the creation of digital tools. The constructed algorithm paves the way for early diagnosis and personalized interventions and supports the rapid and objective assessment of speech data. This study focuses on Russian-language discourse, broadening the scope of current research by highlighting the potential for developing language-specific diagnostic solutions and addressing the problem of creating a set of diagnostic parameters for analyzing the discourse of Russian-language speaking people with mental disorders.

2 Related work

One of the approaches to analyzing speech in mental disorders is interpretive processing through clustering, as it is established that respondents tend to produce semantically or phonetically related groupings of words in bursts over time, described as “clustering”. In this case the researchers pay attention to semantic cohesion of an individual’s speech and semantic similarity between the produced words, as patients with psychotic disorder may display distinguished from healthy patients semantic networks (Aloia et al., 1996; Paulsen et al., 1996; Pintos et al., 2022). Traditionally the aforementioned processing is done manually, however recent studies tend to automate this procedure in order to get more objective and repeatable results (Corcoran et al., 2020; Lundin et al., 2022).

Another approach to speech analysis in cognitive disorders is based on speech profiling. In this case researchers tend to create a generalized linguistic personality model – a set of linguistic traits of individuals within the studied group (here – clinical one), as it is widely recognized, that language personality may be affected by psychiatric disturbances (e.g. Sizova & Semenova, 2023). To build clinical and control group profiles researcher often resort to quantitative methods and stylostatistics, as it may offer a faster, more efficient and objective way to retrieve data (Mota et al., 2012).

Several studies have identified distinctive speech patterns associated with schizophrenia. Acoustic markers such as pitch variability, speech rate, and pause duration have been linked to symptoms like flat affect and alogia. For example, analyses of these features have achieved classification accuracies exceeding 90% when distinguishing schizophrenia patients from healthy controls (Espinola et al., 2020; Pan et al., 2023). Similarly, linguistic markers, including reduced syntactic complexity and decreased discourse coherence, have been instrumental in identifying psychotic symptoms and differentiating between patient and control groups (Baklund et al., 2022).

Advancements in computational methods, particularly machine learning, have facilitated the use of vocal biomarkers for differential diagnosis. Features such as Mel-frequency cepstrum coefficients (MFCCs), fundamental frequency, and formant patterns have been leveraged to classify individuals into groups such as schizophrenia, bipolar disorder, and depression. Classification accuracies of up to 91% have been reported, underscoring the effectiveness of these approaches (Espinola et al., 2020). Additionally, variability in vocal features such as pitch and speech tempo has proven useful in distinguishing between different psychiatric conditions (Pan et al., 2023).

Research into schizophrenia within Russian-language materials has explored various linguistic and cognitive aspects of the disorder. A notable study by Panicheva and Litvinova (Panicheva, & Litvinova, 2019) examined semantic coherence in written texts by Russian-speaking individuals diagnosed with schizophrenia. Their findings revealed that, contrary to studies in English, patients' texts exhibited higher minimum semantic coherence values compared to those of healthy controls. This suggests that written narratives of Russian-speaking patients may display less topic variability, potentially reflecting a tendency toward perseveration or focused thought patterns.

Additionally, the study highlighted the significance of the relative position of semantic coherence minima within texts, indicating that patients' writings often showed distinct patterns in the flow of ideas. These insights underscore the importance of considering linguistic and cultural contexts when analyzing cognitive impairments in schizophrenia.

Further research has involved the adaptation and testing of diagnostic tools for Russian-speaking populations. For instance, the Russian-language version of the Thought, Language, and Communication (TLC) Scale has been evaluated for its effectiveness in detecting early cognitive disorders in individuals at risk of schizophrenia. The results support the TLC scale's utility in clinical settings, facilitating early intervention strategies (Omelchenko et al., 2022).

3 Methods

3.1 Participants

For the study speech samples from the Discourse Diversity Database (3D; Khudyakova et al., 2022) were taken representing a diverse collection designed to explore linguistic markers indicative of mental health conditions. It comprises speech samples by 136 people with mental disorders, such as schizophrenia (32), schizoaffective disorder (17), schizotypal disorder (22), bipolar disorder (36), borderline personality disorder (8), recurrent depressive disorder (12), primary depressive episode (4), other personality disorder (5), and 210 individuals from a control group with no diagnosed mental health conditions. The diagnosis of disorders in the clinical group was carried out in accordance with ICD-10 by qualified psychiatrists and the information about the control group was self-reported[1].

3.2 Elicitation stimuli

The dataset contains three discourse types: picture based narratives, personal stories, and picture-based instructions (procedural discourse), each containing three variants of stimuli: comics by Herluf Bidstrup («Superman», «Discovery of the World», «Wonderful Day») – for picture-elicited narratives task; questions about the best or the most memorable gift, trip or holiday – for personal stories; IKEA's self-assembly furniture manuals – for the picture-based instructions (Khudyakova et al., 2022: 38).

3.3 Transcription

Audio recordings were transcribed for the control and clinical group using the machine learning model for speech recognition and transcription Whisper[2]. Transcriptions were checked manually according to pre-established rules used in compiling verbatim contents in forensic examination of oral speech (Kurachenkova et al., 2007). Thus, speech fragments, the content of which could not be determined are marked as *нрзб*. (illegible). Words with contracted word pronunciation are indicated in the form of orthographic transliteration, for example: *че, гит, гру, ниче.* Unfinished words are accompanied by the sign =: *сба= сбавлять* (slo= to slow down). Unfinished remarks are accompanied by three dots: *Я не знаю…* (I don't know…). Filled hesitation pauses are indicated by stretching vowels and consonants like *аа, ээ, ем, мм* (aa, ee, em, mm): *Я, мм, пошла, э, в склад* (I, mm, went, e, on the warehouse). Manual correction of transcripts was carried out by students in linguistics. Verification was made by specialists in language with a qualification of at least a candidate of sciences (PhD).

4 Data analysis

The primary research method used in this study is speech profiling (Argamon et al., 2009) within the paradigms of linguistic personality theory, idiolect (Litvinova et al., 2016; Litvinova, 2019), and idiostyle (Karaulov, 2010; Wright, 2017), aimed at addressing diagnostic tasks in forensic authorship linguistics (Coulthard, 2004). These methods are implemented through general principles of text processing and machine learning, including topic modeling, lexical vectorization, and neural network

[1] The demographic data of the participants is located by the link https://github.com/KhomenkoAnna/3D_research.
[2] https://github.com/openai/whisper

training. Oral speech processing is conducted within the framework of experimental linguistics in practical use (Kaganov, 2005). The diagnostic potential for identifying clinical disorders is supported by previous research using non-Russian language materials (Corcoran et al., 2020) and the transitive principle, which allows for the identification of individuals with auto- and hetero-aggressive behavioral tendencies through speech analysis (Khomenko et al., 2023).

In this paper, the construction of speech profiles for clinical and control groups is based on a combination of automatic lexical clustering and stylostatistical (stylometric) analysis. Clustering allows to establish the semantic structure of discourse: lexical clusters describe semantic groups present in a particular piece of speech. Calculating stylometric indices give an opportunity to track the so-called uncontrolled characteristics of speech associated with deep thinking processes.

4.1 Preprocessing

With the help of the NLTK library[3], the texts were tokenized in two formats: with the preservation of stop words (also taken from the NLTK library) and without for clustering and with the preservation of stop words for stylostatistical analysis, in both cases all punctuation marks were removed. Using the NLP library SpaCy (ru_core_news_sm model[4]), transcriptions were also lemmatized to normalize words to their base forms, reducing linguistic variability while preserving semantic meaning.

4.2 Clustering

To the resulting four groups of words, the clustering algorithm was applied according to the rule proposed in the work of Lundin et al. (Lundin et al., 2022). According to this rule, for a sequence of words A, B, C, D, the switch follows B when S (A, B) > S (B, C) and S (B, C) < S (C, D), where S (A, B) is the cosine similarity between the vectors of the words A and B. To estimate the cosine similarity, the pre-trained static embedding model geowac_lemmas_none_fasttextskipgram_300_5_202024[5] of the RusVect_rēs service was used, which was used to obtain embeddings of the fastText type (Bojanowski et al., 2017). For the first word in the sequence, the proximity was taken as 0.

Based on the clustering results, the following metrics were calculated:

1. Number of switches for each person in each category.
2. The average distance between clusters for each person in each category (calculated as the distance between cluster centroids; the average value of vectors located in the same cluster is taken as the centroid).
3. The average value of the T-score in the cluster for each person in each category (a metric that shows how non-random the strength of the association between collocates is; all sequences of two words inside one cluster are taken as collocates).
4. The average value of the silhouette-score (a metric that shows how close an object is to its cluster compared to other clusters: the higher the value of this metric, the closer the object is to its own cluster).
5. The average cluster size for each person.

For each metric in each set, the mean, median, and standard deviation were calculated.

4.3 Stylostatistics

Statistical information was extracted from the texts, including n-grams — sequences of two (bigrams) and three (trigrams) words that were obtained using n-gram models. These models were based on calculating the frequency of word combinations in the text and allowed us to identify frequently occurring sequences characteristic of each group.

Keywords were highlighted using various metrics such as Log-Likelihood, T-score, Mutual Information (MI), Dice and Frequency Ratio. These metrics helped to identify the most significant and frequently occurring words in the texts of each group, taking into account their frequency of occurrence both within the group and in relation to the total frequency in the corpus. Additionally, collocations -

[3] https://github.com/nltk/nltk
[4] https://spacy.io/models/ru#ru_core_news_sm
[5] https://github.com/avidale/compress-fasttext

stable phrases characteristic of these texts - were identified for the top 20 extracted keywords according to the T-score metric. The statistical T-score method used to identify collocations made it possible to identify the most frequent and significant combinations of words that play a key role in the speech structure of both groups.

As part of the study, various stylostatistical parameters were also calculated in order to analyze the texts more deeply in terms of their structure and content. Among these parameters (Golovin 1971) are the coefficients of objectivity, *Pr* (the ratio of nouns plus pronouns amount to the sum of adjectives and verbs), quality, *Qu* (the ratio of adjectives plus adverbs amount to the sum of verbs and nouns), activity, *Ac* (the ratio of verbs plus all verb forms amount to the word number in the text), dynamism, *Din* (the ratio of verbs plus all verb forms amount to the sum of nouns, adjectives and pronouns) and coherence of the text, *Con* (the ratio of prepositions plus conjunctions amount to the number of sentences). In addition to these parameters, readability indices such as the Flesch-Kincaid index and the Gunning Fog index were calculated, which give an idea of how easy or difficult it is to perceive texts from both groups. These indexes are based on characteristics such as sentence length and the complexity of the words used.

In addition, the type-token ratio (TTR), the average word length (in letters) and the average sentence length (in words) were calculated, which made it possible to assess the richness of vocabulary and the structural complexity of the text.

5 Results

Statistical analysis was performed using the Pandas[6], NumPy[7], SciPy[8] and Matplotlib[9] libraries. First of all, the statistical significance of the feature parameters for clustering and stylostatistics was checked. To assess the significance of differences between the groups, two types of statistical tests were applied:

- Independent samples t-test: used when data followed a normal distribution
- Non-parametric Mann-Whitney U-test: applied for data that did not follow a normal distribution

A criterion was considered statistically significant if its p-value corrected for multiple comparisons was less than 0.05, indicating meaningful differences between the groups with and without mental disorders.

Significant differences between the two groups were found for the following metrics: the number of switches, distance between clusters, and t-score (see Table 1). Cluster size metric (M-control=4,95, M-disorder=4,9) and silhouette-score (M-control=0,21, M-disorder=0,21) have not shown statistical significance[10].

For the twogroup comparison, the following criteria were statistically significant: coefficients of objectivity, quality, activity and dynamism, Flesch-Kincaid Index, Gunning Fog Index, type-token ratio (TTR).

Next, threshold values for each metric were determined to predict group membership using two statistical parameters: Youden's J index and the sum/difference between the mean and standard deviation of the metric in the control group.

Youden's J index was computed based on the AUC-ROC between the target variable and the metric, representing the optimal threshold where the true positive rate exceeds the false positive rate. The mean and standard deviation of the control group were used such that if the control group's mean was higher than that of the disorder group, the standard deviation was subtracted from the mean to define the lower threshold. Otherwise, the standard deviation was added to the mean.

Threshold values were calculated by both methods, and their efficacy was compared. Each metric was labeled 0 (control group) or 1 (disorder group) based on the threshold, and the labels were evaluated using recall, a key metric for predicting the positive class. The optimal threshold for each metric was selected based on the recall results. Statistical analysis was conducted using the Scikit-Learn[11] library.

[6] https://pandas.pydata.org/
[7] https://numpy.org/
[8] https://scipy.org/
[9] https://matplotlib.org/
[10] Full data are available by the link https://github.com/KhomenkoAnna/3D_research.
[11] https://scikit-learn.org/stable/

	Control Group	Disorder Group	Threshold Value	Metric	Recall
Number of switches	49.12	28.16	16.68	Mean - Std	0.3
Cluster distance	0.51	0.52	0.53	Younden's J	0.4
T-score	897.98	420.47	275.8	Mean - Std	0.32

Table 1: Average values of statistically significant metrics, threshold values, and recall for clustering

For stylostatistics metrics, the threshold value was determined only based on Youden's J index with correction for multiple comparisons (see Table 2).

	Control Group	Disorder Group	Threshold Value	Recall
Flesch-Kincaid readability index	6.34	4.35	1.49	0.92
Gunning fog index	5.41	3.94	inf	0
TTR	0.68	0.74	0.72	0.58
Pr	1.68	1.75	1.78	0.46
Qu	0.1	0.11	0.08	0.77
Din	0.06	0.05	0.17	0.01

Table 2: Average values of statistically significant metrics, threshold values, and recall for stylostatistics

As indicated in the table, for one criterion, it was not possible to calculate the threshold value. This is because there is a high degree of overlap between the classes for Gunning fog index, making it impossible to identify a clear threshold for class separation.

6 Discussion

Speech profiles have been created for generalized language personalities of all types of mental and behavioral disorders: schizophrenia, schizotypal disorder, schizoaffective disorder, borderline personality disorder, other personality disorders, primary depressive episode, recurrent depressive disorder, bipolar affective disorder – and for each disorder separately. A speech profile was also created for the language personality of the control group – a group of individuals who had not been diagnosed with mental disorders.

Previous research has established that individuals with these conditions exhibit certain linguistic abnormalities, such as reduced semantic coherence, syntactic simplifications, narrative fragmentation, and excessive self-referential language (Tang et al., 2021; Vlasova & Akhmetova, 2015). Schizophrenic speech is fragmented, incoherent, and marked by neologisms and perseveration, reflecting disorganized thinking (Akhmetzyanova & Tregubenko 2021; Vlasova & Akhmetova, 2015). Personality disorders feature frequent present-tense usage, self-referential language, and an observer perspective (Smerchinskaya et al., 2024). Bipolar speech is verbose and grammatically complex, while depressive speech is elliptical, impersonal, and rich in idiomatic expressions (Smirnova et al., 2013).

The results of this study support these observations, as statistical analysis has revealed that different patterns of clustering and stylostatistical indicators are observed in various disorders. Therefore, when summarizing the findings of the stylostatistical and clustering analysis, the following descriptions of the

speech patterns of the analyzed conditions can be given. Speech in depression is complex and detailed, featuring long sentences, large number of lexical clusters and a high lexical variety. Patients with depression also tend to provide more standardized associations. Schizophrenic spectrum and personality disorder patients tend to speak in shorter, simpler sentences, though they may not always be logical. Depressive patients are more likely to employ abstract constructions, while in schizotypal and schizoaffective conditions, speech is more specific and focused on details. Personality disorders are characterized by a high level of evaluative and active language, while depression and borderline disorders exhibit a static quality. Text coherence is greater in schizoaffective disorder, although semantic gaps may occur. Patients with schizophrenic spectrum disorders tend to switch between word clusters more often, showing loose associations effect. The mean distance between associations is larger for respondents with bipolar and personality disorders. Although the specific types of speech varied depending on the diagnosis, all the disorders showed significant differences from the speech patterns of the control group.

Thus, for the final automated diagnostic model vocabulary clustering metrics do not show the required level of statistical significance for heterogeneous patient groups. The best metrics is mean cluster distance. It has the highest recall. Nevertheless, for individual patient groups, clustering shows positive results in class division. Thus, for speech of individuals with bipolar disorder mean cluster size shows a recall higher than 0.5, for schizophrenia and schizoaffective disorder – mean distance between clusters; for schizotypal disorder – mean cluster size and silhouette score. For the generalized group with different types of disorders, stylostatistical metrics show their effectiveness in separating the clinical and control groups. Thus, Flesch-Kincaid readability index, type-token ratio (TTR), coefficient of quality (Qu) (the metrics have the recall value equal or greater than 0.5) could be considered as possible for inclusion in diagnostic model of psychiatric disorders and used in creating an automatic model for classifying speech as a control group or a group with thought and mood disorders. Naturally, for working psychiatric diagnostics, more customized models should be used, configured to determine a specific disorder. Nevertheless, for a fundamental study of the Russian-language discourse of people with mental and mood disorders, this work can be useful.

7 Conclusion

This study contributes to the growing body of research aimed at automating the diagnosis of mental health conditions through speech analysis. The findings demonstrate that speech profiling can show differences between individuals with and without mental health disorders, particularly within the Russian-speaking population, which has been underexplored in previous studies. By integrating both microlinguistic and macrolinguistic features the study offers a comprehensive approach to diagnosing mental disorders.

Future research should focus on expanding the dataset to include more varied linguistic features and refining the algorithms with the aim of further model customization for diagnosing specific disorders. Additionally, incorporating speech data from multiple languages and cultural contexts will enhance the model's generalizability, ultimately making it a valuable tool for mental health professionals worldwide.

8 Limitations

This article will focus mainly on generalized results for two groups: the control group and the group with mental disorders. Such a division is undoubtedly a limitation of the study. The concept of indivisible psychosis (E.A. Zeller, H. Neumann и W. Griesenger) is not accepted as leading due to its proved inconsistency. Nevertheless, the unification of heterogeneous disorders seems possible within the framework of a pilot study. Firstly, such unification is carried out in psychiatrics works (e.g. Sychugov et al., 2019), and secondly, the study aims to identify pathological speech as a class (Pashkovsky et al., 2013) and to understand whether it is possible to distinguish the norm group from the abnormal. Thus, there are three main causes of mental disorders – exogenous, endogenous and stress-related. "In all these cases, the brain is unable to function normally, and pathological mental production occurs. With the exception of obvious deficit processes associated with progressive morphological degeneration of brain cells and the formation of dementia, it can be said that pathological production is aimed at compensating for disadaptation. Thus, specific disorders (e.g., dementia) are possible only with organic damage to the

cortex. All other organic mental disorders <…> are nonspecific" (Ten et al., 2008:152). It is clear that due to the heterogeneity of the sample, the final result of the study will be influenced by the largest classes (schizophrenia, schizoaffective disorder, schizotypal disorder, bipolar disorder), but within these classes it is also possible to predict the presence or absence of statistically significant parameters capable of detecting the speech of people with a certain diagnosis.

The conclusion about the presence or absence of a disease is always made by a psychiatrist. At this stage of their development automatic tools can only give a signal to a specialist that it is necessary to continue working with a particular patient. As a result of the study, it was found that different types of disorders show different typification of specific characteristics, however, in each case this typification differs from the typification of a control group.

Acknowledgements

This article is an output of a research project implemented as part of the Basic Research Program at the National Research University Higher School of Economics (HSE University).

References

[1] Akhmetzyanova, E. M., & Tregubenko, I. A. (2021). Psycholinguistics Analysis of schizophrenia spectrum disorders patients' writing and oral speech (on the example of the autobiographical texts). *The Scientific Notes of the Pavlov University*, 28(2), 38–44. https://doi.org/10.24884/1607-4181-2021-28-2-38-44

[2] Aloia, M. S., Gourovitch, M. L., Weinberger, D. R. & Goldberg, T. E. (1996). An investigation of semantic space in patients with schizophrenia. Journal of the International Neuropsychological Society, 2(4), 267-273. https://doi.org/10.1017/S1355617700001272.

[3] Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. *Communications of the ACM, 52*(2), 119–123. https://doi.org/10.1145/1461928.1461959

[4] Baklund, L., Røssberg, J. I., & Møller, P. (2022). Linguistic markers and basic self-disturbances among adolescents at risk of psychosis. A qualitative study. *EClinicalMedicine, 55*, 101733. https://doi.org/10.1016/j.eclinm.2022.101733

[5] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051

[6] Corcoran, C. M., & Cecchi, G. A. (2020). Using language processing and speech analysis for the identification of psychosis and other disorders. *Biological Psychiatry Cognitive Neuroscience and Neuroimaging, 5*(8), 770–779. https://doi.org/10.1016/j.bpsc.2020.06.004

[7] Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. *Applied Linguistics, 25*(4), 431–447. https://doi.org/10.1093/applin/25.4.431

[8] Espinola, C. W., Gomes, J. C., Pereira, J. M. S., & Santos, W. P. D. (2020). Vocal acoustic analysis and machine learning for the identification of schizophrenia. *Research on Biomedical Engineering, 37*(1), 33–46. https://doi.org/10.1007/s42600-020-00097-1

[9] Golovin, B. N. (1971). *Yazyk i statistika* [Language and statistics]. Prosveshchenie.

[10] Kaganov A. Sh. (2005) *Kriminalisticheskaya ekspertiza zvukozapisej* [Forensic examination of recordings]. Moscow: Yurlitinform.

[11] Karaulov Yu. N. (2010). *Russkij yazyk i yazykovaya lichnost'* [Russian language and linguistic personality] (7th ed.). LKI Publishing House.

[12] Khomenko A. Yu., Kasimova L. N., Sychugov E. M., Komratova A.D., Galkin N. V., & Titova A. A. (2023). Diagnostika auto- i geteroagressii metodami psiholingvistiki [Psycholinguistics methods in diagnostics of auto- and heteroagression]. In P. A. Katyshev & I. V. Ogorelkov (Eds.), *Sovremennyj mediatekst i sudebnaya ekspertiza: mezhdisciplinarnye svyazi i ekspertnaya ocenka*, 12, 337-361.

[13] Khudyakova, M., Antonova, N., Nelubina, M., Surova, A., Vorobyova, A., Minnigulova, A., Gronskaya, N., Yashin, K., Medyanik, I., Shishkovskaya, T., Ryazanskaya, G., Zuev, A., Dragoy, O., & Skorobogatova, A. S. (2022). Discourse Diversity Database (3D) for Clinical Linguistics Research: Design, Development, and analysis. *Bakhtiniana Revista De Estudos Do Discurso, 18*(1), 32–57. https://doi.org/10.1590/2176-4573e55885

[14] Khudyakova, M. V. (2023). *Methods for analysis and assessment of spoken discourse by people with speech-language disorders* (Doctoral dissertation, HSE University).

[15] Koops, S., Brederoo, S. G., de Boer, J. N., Nadema, F. G., Voppel, A. E., & Sommer, I. E. (2023). Speech as a biomarker for depression. *CNS & Neurological Disorders-Drug Targets-CNS & Neurological Disorders)*, *22*(2), 152-160.

[16] Kurachenkova, N. B., Baicharov, N. V., & Ermakova, M. A. (2007). *Identifikaciya lic po ustnoj rechi na russkom yazyke. Metodika" Dialekt": posobie dlya ekspertov* [Identification by oral speech in Russian. Methodology «Dialect». Manual for experts]. IK CST FSB Rossii.
[17] Le Glaz, A., Haralambous, Y., Kim-Dufor, D. H., Lenca, P., Billot, R., Ryan, T. C., ... & Lemey, C. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of medical Internet research*, *23*(5), e15708.
[18] Litvinova, T., Seredin, P., Litvinova, O., & Zagorovskaya, O. (2016). Profiling a set of personality traits of text author: what our words reveal about us. *Research in Language, 14*(4), 409–422. https://doi.org/10.1515/rela-2016-0019
[19] Litvinova T. A. (2019). Idiolekt kak ob"ekt korpusnoj idiolektologii: k stanovleniyu novogo lingvisticheskogo napravleniya [Idiolect as an object of corpus and idiolectology: toward the formation of a new linguistic trend]. *Uchenye zapiski Novgorodskogo gosudarstvennogo universiteta, 7*(25), 1–5.
[20] Lundin, N. B., Jones, M. N., Myers, E. J., Breier, A., & Minor, K. S. (2022). Semantic and phonetic similarity of verbal fluency responses in early-stage psychosis. *Psychiatry Research, 309*, 114404. https://doi.org/10.1016/j.psychres.2022.114404
[21] Malgaroli, M., Hull, T. D., Zech, J. M., & Althoff, T. (2023). Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, *13*(1), 309.
[22] Mota, N. B., Vasconcelos, N. a. P., Lemos, N., Pieretti, A. C., Kinouchi, O., Cecchi, G. A., Copelli, M., & Ribeiro, S. (2012). Speech graphs provide a quantitative measure of thought disorder in psychosis. *PLoS ONE, 7*(4), e34928. https://doi.org/10.1371/journal.pone.0034928
[23] Mota, N. B., Weissheimer, J., Finger, I., Ribeiro, M., Malcorra, B., & HübnerL. (2023). Speech as a graph: Developmental perspectives on the organization of spoken language. *Biological Psychiatry Cognitive Neuroscience and Neuroimaging, 8*(10), 985–993. https://doi.org/10.1016/j.bpsc.2023.04.004
[24] Omelchenko, M. A., Arutyunova, E. V., Butashin, A. D., & Ivanova, E. M. (2022). Testing of the Russian-language version of the Thought, Language and Communication Scale. *S S Korsakov Journal of Neurology and Psychiatry, 122*(10), 116. https://doi.org/10.17116/jnevro2022122101116
[25] Pan, W., Deng, F., Wang, X., Hang, B., Zhou, W., & Zhu, T. (2023). Exploring the ability of vocal biomarkers in distinguishing depression from bipolar disorder, schizophrenia, and healthy controls. *Frontiers in Psychiatry, 14*. https://doi.org/10.3389/fpsyt.2023.1079448
[26] Panicheva, P., & Litvinova, T. (2019). Semantic Coherence in Schizophrenia in Russian Written Texts. *IEEE*, 241–249. https://doi.org/10.23919/fruct48121.2019.8981535
[27] Pashkovsky, V. E., Piotrovskaya, V. R., & Piotrovsky, R. G. (2013). *Psihiatricheskaya lingvistika* [Psychiatric linguistics]. Moscow: LENAND.
[28] Paulsen, J. S., Romero, R., Chan, A., Davis, A. V., Heaton, R. K., & Jeste, D. V. (1996). Impairment of the semantic network in schizophrenia. Psychiatry research, 63(2-3), 109-121. https://doi.org/10.1016/0165-1781(96)02901-0
[29] Pintos, A. S., Hui, C. L. M., De Deyne, S., Cheung, C., Ko, W. T., Nam, S. Y., ... & Chen, E. Y. H. (2022). A longitudinal study of semantic networks in schizophrenia and other psychotic disorders using the word association task. Schizophrenia Bulletin Open, 3(1), sgac054. https://doi.org/10.1093/schizbullopen/sgac054
[30] Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, *7*(1), 276-283.
[31] Ribeiro, M., Malcorra, B., Mota, N. B., Wilkens, R., Villavicencio, A., Hubner, L. C., & RennéCosta, C. (2024). A Methodology for Explainable Large Language Models with Integrated Gradients and Linguistic Analysis in Text Classification. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2410.00250
[32] Sizova N.S., Semenova N.D. (2022). Issledovaniye ekspressivnoy rechi u lits s shizofreniyey i rasstroystvami shizofrenicheskogo spektra [Study of expressive speech in persons with schizophrenia and schizophrenia spectrum disorders]. *Russian Journal of Psychiatry*. No. 6, 53-60.
[33] Smerchinskaya E. M., Tregubenko I. A., Isaeva E. R. (2024). Osobennosti pis'mennoj rechi pacientov s razlichnymi psihicheskimi rasstrojstvami [Written Narrative in Patients with Different Mental Disorders]. *SibScript, 26*(5), 739–747.
[34] Smirnova, D., Sloeva, E., Kuvshinova, N., Krasnov, A., Romanov, D., & Nosachev, G. (2013). 1419 – Language changes as an important psychopathological phenomenon of mild depression. *European Psychiatry, 28*, 1. https://doi.org/10.1016/s0924-9338(13)76454-x
[35] Sychugov E.M., Petrova D.N., Morozova A.S. (2019). Psiholingvisticheskij analiz patologicheskih tekstov kak metod ob"ektivizacii diagnostiki rasstrojstv rechemyslitel'noj deyatel'nosti v psihiatricheskoj praktike [Psycholinguistic analysis of pathological texts as a method of objectification of diagnostics of mental and speech activity disorders in psychiatric practice]. *Fundamental aspects of mental health, 3*, 52 – 58.
[36] Tang, S. X., Kriz, R., Cho, S., Park, S. J., Harowitz, J., Gur, R. E., Bhati, M. T., Wolf, D. H., Sedoc, J., & Liberman, M. Y. (2021). Natural language processing methods are sensitive to sub-clinical linguistic differences in schizophrenia spectrum disorders. Schizophrenia, 7(1). https://doi.org/10.1038/s41537-021-00154-3

[37] Ten V.I. (2008). Koncepciya «edinogo psihoza» Nejmana-SHrizingera: adaptacionnaya paradigma v psihiatrii [The concept of «indivisible psychosis» by Neumann-Schrizinger: an adaptation paradigm in psychiatry]. *Bulletin of Kyrgyz-Russian Slavic University, 8*(4), 151–154.
[38] Vlasova V. P., & Akhmetova E. R. (2015). Osobennosti rechi pri shizofrenii [Features of speech in schizophrenia]. *In The Eusebian Readings. Series: Actual problems of special and inclusive education*, 12-17.
[39] Wright, D. (2017). Using word n-grams to identify authors and idiolects. *International Journal of Corpus Linguistics, 22*(2), 212–241. https://doi.org/10.1075/ijcl.22.2.03wri

Local structure of descriptive discourse in the speech of adult learners of Russian as a foreign language

Maria Kiose
Moscow State Linguistic University,
Institute of Linguistics RAS,
Moscow, Russia
maria_kiose@mail.ru

Anastasia Rzheshevskaya
Moscow State Linguistic University,
Moscow, Russia
arlen_nastya@rambler.ru

Andrey Petrov
Moscow State Linguistic University,
Institute of Linguistics RAS, Moscow, Russia
a.petrov.andrew@gmail.com

Abstract

This paper explores the local structure expressed in rhetorical, communicative and narrative relations and speech disfluencies in the descriptive discourse of adult learners of Russian as a foreign language. In the experiment, two learner groups with B1/B2 and C1/C2 levels of proficiency in Russian as a foreign language had to read two descriptive texts and relate their contents to an interested listener. The study contrasts the distribution of local discourse structure in stimuli texts (59 EDUs) and in the speech of experiment participants in two groups (629 EDUs). The results reveal that the speech in C1/C2 group manifests more frequent narrative relation Setting, communicative relation Metadiscourse, rhetorical relation Interpretation. Meanwhile, the speech of B1/B2 group demonstrates the predominance of communicative relation Agreement and rhetorical relation Joint. The most common type of speech disfluency of C1/C2 speakers is Self-correction, whereas for B1/B2 learners this is Repetition. Additionally, as opposed to the rhetorical structure of the stimuli, Joints are used far more frequently in the speech of both groups, which evidences of the problems in rhetorical coherence with both B1/B2 and C1/C2 learners.

Keywords: local discourse structure, Russian as a foreign language, descriptive discourse, rhetorical relations, communicative relations, narrative relations, speech disfluency

DOI: 10.28995/2075-7182-2025-23-158-168

Локальная структура дискурса описания в речи взрослых, изучающих русский язык как иностранный

Мария Киосе
Московский государственный
лингвистический университет,
Институт языкознания РАН,
Москва, Россия
maria_kiose@mail.ru

Анастасия Ржешевская
Московский государственный
лингвистический университет,
Москва, Россия
arlen_nastya@rambler.ru

Андрей Петров
Московский государственный
лингвистический университет,
Институт языкознания РАН, Москва, Россия
a.petrov.andrew@gmail.com

Аннотация

В работе исследуется локальная структура дискурса в аспекте риторических, коммуникативных и нарративных отношений, а также речевых сбоев в описательном дискурсе взрослых, изучающих русский язык как иностранный. В проведенном эксперименте двум группам участников — студентам, изучающим русский язык как иностранный на уровнях B1/B2 и C1/C2 — предлагалось прочитать и передать содержание двух описательных текстов. Установлены различия в распределении отношений локальной структуры дискурса в стимульных текстах (59 ЭДЕ) и речи двух групп участников (629 ЭДЕ). Речевой корпус участников группы C1/C2 демонстрирует высокую частотность нарративного отношения Сеттинг, коммуникативного отношения Метадискурсив, риторического отношения Интерпретация. В речи участников уровня B1/B2 превалирующими являются коммуникативное отношение Согласие, риторическое отношение Объединение. Наиболее частотным видом речевого сбоя у говорящих на уровне C1/C2 является Самоисправление, на уровне B1/B2 – Повтор. Также определено, что ассиметричное риторическое отношение Объединение более характерно для речи говорящих обеих групп, чем для стимульных текстов, что свидетельствует о проблемах риторической связности в локальной структуре устного дискурса в речи участников.

Ключевые слова: локальная структура дискурса, русский язык как иностранный, дескриптивный дискурс, риторические отношения, коммуникативные отношения, нарративные отношения, речевые сбои

1 Вступительные замечания

Настоящая работа обращается к проблеме организации дискурса описания в речи взрослых, изучающих русский язык как иностранный (РКИ). В то время как дискурсивные исследования устной речи на русском языке как на родном ведутся давно (см., например, в [1; 2]), в отношении РКИ такие работы единичны. Известны исследования восприятия и понимания текстов на английском и китайском языках как иностранных, см., например, работы, в которых использовались методики набора ключевых слов как наиболее важных с точки зрения содержания, чтения с саморегулировкой скорости и регистрации движения глаз [3; 4]. В данных работах установлены эффекты длины, частотности и конкретной предсказуемости слова на скорость лексического доступа у носителей языка в сравнении с изучающими иностранный язык, причем при чтении на неродном языке все эффекты выражены сильнее. В то же время анализу могут подвергаться отдельные дискурсивные практики, проявляющиеся, например, в выборе дискурсивных маркеров или коммуникативных шагов [5; 6], а также некоторых типов риторических отношений (например, Условия, Причины, Следствия и др.) [7], однако на настоящий момент нам не известны работы, в которых распределение в речи изучающих РКИ дискурсивных отношений устанавливалось бы системно с учетом типа дискурса. Необходимо также отметить, что в рассмотренных выше исследованиях во внимание принимался невысокий уровень владения иностранным языком (A2-B1), что и не позволяет в полной мере подвергать анализу специфику конструирования дискурса. В настоящей работе на материале записанного корпуса дискурса описания [10; 11; 12] определяются особенности реализации дискурсивных отношений в речи изучающих РКИ на уровне B1/B2 (ТРКИ-I и ТРКИ-II) и C1/C2 (ТРКИ-III и ТРКИ-IV). Мы предполагаем, что в зависимости от уровня владения РКИ в речи говорящих могут быть обнаружены отличительные проявления в распределении отношений, что позволит высказать ряд рекомендаций по обучению конструированию дискурса описания в РКИ.

2 Теоретические предпосылки исследования

Под дискурсом описания вслед за Р. Лонгакром мы понимаем тип статического дискурса, структура которого определяется наличием четырех компонентов: ситуации, проблемы, решения, оценки (Situation, Problem, Solution, Evaluation) [13; 14]. Хотя анализу чаще подвергаются языковые характеристики описания [12; 15], в ряде работ указываются и его дискурсивные характеристики, например, его реализация с помощью двух разнонаправленных принципов – генерализации и конкретизации, где первый соотносится с представлением обобщенного содержания ситуации, проблемы, решения с ориентацией на оценку события, а второй, напротив, состоит в подробном описании и точной передаче деталей [11; 12]. Особенности данного дискурса были изучены и применительно к русскому языку. К их числу отнесены представление в речи единичных (чаще регулярно появляющихся) участников (объектов) и событий, чаще без указания их причин, конструирование событий с позиции говорящего с превалированием схем собственно описания, фона и уточнения [16; 17; 18].

Применительно к исследованиям речи дискурс может рассматриваться как глобальная структура, организованная с опорой на локальные уровни [1], описываемые коммуникативными, нарративными, риторическими отношениями, а также речевыми сбоями [19]. В настоящей работе этот подход используется для изучения дискурса описания. В качестве коммуникативных отношений, связывающих отдельные реплики в аспекте их обращенности к собеседнику, рассматриваются Вопрос, Экспрессив, Согласие, Комиссив, Директив и др. К нарративным (оформительным по [1]) относят отношения, описывающие структуру каждого цикла дискурса как рассказа в совокупности его этапов, таких как Сеттинг и Кода. Риторические отношения описывают логическую связь реплик, представленных ЭДЕ, имеющими структуру клаузы [20; 21; 22]; выделяют два типа таких отношений: асимметричные (Интерпретация, Мотивирующий признак, Развитие темы, Реформулирование и др.) и симметричные (Контраст, Объединение, Дизъюнкция), их полный список, используемый как базовый во многих работах, в т.ч. и в данной работе, приведен в [20]. Отношения, описывающие речевые сбои (вслед за [1; 19]), подразделяются на два типа: Хезитация и Самоисправление [19: 36]. Средствами Хезитации могут служить аппроксиматоры, которые снижают точность и категоричность высказывания [23; 24]; в рамках ЭДЕ они приобретают и дискурсивную функцию. По итогам предварительного анализа материала было решено в качестве отдельного типа отношений также рассматривать Повтор, который в отличие от риторического отношения Реформулирования представляет собой дублирование субъектно-предикатной структуры и содержания ранее введенной клаузы в ЭДЕ (элементарной дискурсивной единице [1; 2]). Важно отметить, что в известных нам работах при сегментации речи на ЭДЕ во внимание принимаются синтаксический и просодический критерии. При анализе звучащей речи изучающих РКИ мы столкнулись с некоторым ограничением в использовании просодического критерия, т.к. во многих случаях появление продолжительной паузы определено замедленным темпом речи, также есть сложности в установлении характеристик интонационной структуры (что не позволяет в полной мере опираться на существующие традиции описания интонационных конструкций и, соответственно, разграничения ЭДЕ с учетом смены конструкций). В связи с этим принято решение в большей степени опираться на синтаксический критерий (хотя и при его применении обнаруживаются некоторые ограничения, связанные с большим количеством синтаксических ошибок в речи говорящих), что повлекло необходимость установления во многих ЭДЕ нескольких типов отношений.

Как мы указали выше, при оценке понимания текста и порождения речи изучающими иностранный язык исследования в большей степени опираются на лексический критерий при установлении ключевых слов [3; 4], а также на синтаксический критерий связности реплик при помощи дискурсивных слов и коннекторов [7; 8; 9]. Однако сделанные выводы касаются и употребления некоторых типов риторических отношений, которые могут быть распространены и на изучающих русский язык как иностранный. Так, в [7] показано, что риторические отношения Причины усваиваются раньше, чем отношения Объединения (Addition в [7]), Условия и Уступки; также определено, что отношения Уступки в целом использовались с ошибками в применении коннекторов. Это позволяет сделать вывод о том, что отношения Причины будут чаще и менее ошибочно применяться в дискурсе изучающими иностранный язык на более ранних этапах обучения, чем иные типы отношений. Важно отметить, что в корпусных исследованиях, тестирующих письменную речь, показано, что основная граница между более ранним и более поздним усвоением риторических отношений будет проходить между отношениями Объединения и отношениями, выражающими подтверждение (в нашем случае, речь может идти об отношениях Согласия, Повтора), с одной стороны, и всеми другими типами отношений, с другой стороны, где первые чаще используются изучающими язык на ранних стадиях [25].

3 Дизайн эксперимента и процедура анализа

В эксперименте приняли участие 20 испытуемых, иностранных граждан, изучающих РКИ в МГЛУ: по 10 студентов, начавших обучение на уровнях B2 и C2. Участникам, сидящим за столом перед экраном, предлагалось прочитать два текста с описанием картин И. Левитана «Золотая осень» и В. Серова «Девочка с персиками», подготовленных искусствоведами в рамках проекта «Музеи Мира. Ваш гид в мире искусства» (https://muzei-mira.com). После прочтения каждого текста участники разворачивались на стуле от экрана в сторону экспериментатора, рядом с которым

на штативе была установлена камера Sony HXR-NX30P, к которой подключена беспроводная микрофонная система RØDE Wireless GO II для качественной записи голоса (Рис. 1).

Рис. 1: Проведение эксперимента

Разметка локальной структуры стимульного материала (29 и 30 ЭДЕ) осуществлена двумя аннотировщиками. Структура стимульных текстов полностью обусловлена набором риторических отношений (плюс первая ЭДЕ, обозначенная как Старт) [15]. Например, в одном из предложений первого стимула можно выделить 6 ЭДЕ: (1) / *если детально рассмотреть картину,* / с отношением Условия, (2) / *то можно отметить* / с отношением Следствия, (3) / *что детали на полотне распределены неравномерно* / и (4) / *– левая сторона более загружена,* / – Развития темы, (5) / *однако, это не приводит к асимметрическим ощущениям*, / – Контраста, (6) / *так как Левитан грамотно «расставил»* / – Причины и (7) / *и сгруппировал все элементы пейзажа* / – Развития темы. Всего тексты содержат 2 ЭДЕ типа Старт (они были отнесены к нарративным), 22 ЭДЕ типа Развитие темы, 15 ЭДЕ типа Объединение, 4 ЭДЕ типа Фон, по 3 ЭДЕ типа Контраст и Причина, по 2 ЭДЕ типа Условие, Антитеза и Обстоятельство, по 1 ЭДЕ типа Оценка, Цель, Следствие и Предоставление возможности.

На первом этапе речь говорящих распознается с применением программ автоматического распознавания речи, корректируется, далее сегментируется на ЭДЕ с опорой на синтаксические и (в меньшей степени) просодические критерии, после этого размечается двумя аннотировщиками в соответствии с четырьмя типами отношений. Так, например, при произнесении участником группы B1/B2 фрагмента / *про Мамонтова что-то было* / *это Мамон…* / *про Мамонтова* / *да* / обнаруживаем 4 ЭДЕ, где ЭДЕ сегментируются с привлечением синтаксического и просодического (паузация, изменение тона в ЭДЕ 4) критериев (Рис. 2), тонограмма здесь и далее выполнена с привлечением ПО Praat (https://www.fon.hum.uva.nl/praat/).

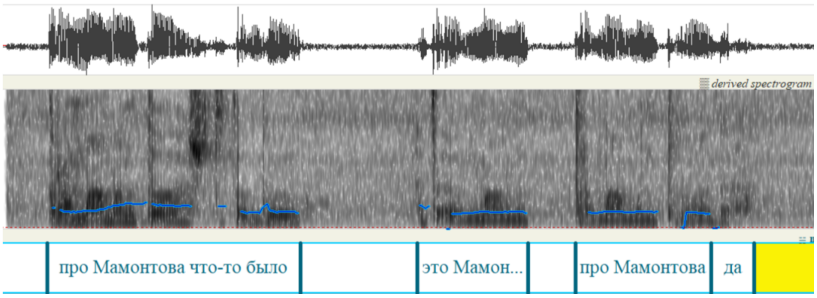


Рис. 2: Сегментация на ЭДЕ

Сегментированные ЭДЕ далее подвергаются анализу на предмет реализации типов отношений; в данном случае это, соответственно, отношения Развития темы в ЭДЕ 1, Хезитации в ЭДЕ 2, Повтора в ЭДЕ 3, Согласия в ЭДЕ 4. Распределение всех типов отношений устанавливается отдельно в речи участников уровней B1/B2 (ТРКИ-I и ТРКИ-II) и C1/C2 (ТРКИ-III и ТРКИ-IV).

На втором этапе проводится анализ и определение значимых различий в распределении типов отношений в речи участников разных уровней подготовки; сопоставляются те типы отношений, которые демонстрируют отличия для каждой из групп. На третьем этапе обозначаются особенности реализации четырех типов отношений в спонтанной речи в отличие от стимульного материала как в двух группах участников в совокупности, так и отдельно в каждой из групп. В результате формулируются особенности реализации локальной структуры дискурса описания в спонтанной речи носителей русского языка как иностранного с учетом двух указанных уровней подготовки.

4 Результаты

4.1 Локальная структура дискурса описания в отчетах изучающих РКИ

Приведем общие данные, описывающие спонтанную речь (Таблица 1).

	B1/B2	C1/C2	Общее
Время звучания (мин / с)	18.4 / 1101	20.1 / 1204	38.5 / 2305
Среднее время звучания	55.1	60.2	57.6
Кол-во ЭДЕ	283	349	632
Среднее кол-во ЭДЕ	14.2	17.45	15.8

Таблица 1: Общие данные по спонтанной речи говорящих

Как можно видеть из Таблицы 1, количество ЭДЕ в группе C1/C2 в 1.23 раза превышает количество ЭДЕ в группе B1/B2; при этом время звучания выше только в 1.09 раза, что обусловлено более низким темпом произнесения речи в группе B1/B2. Рассмотрим особенности индивидуальных показателей участников (Рис. 3).

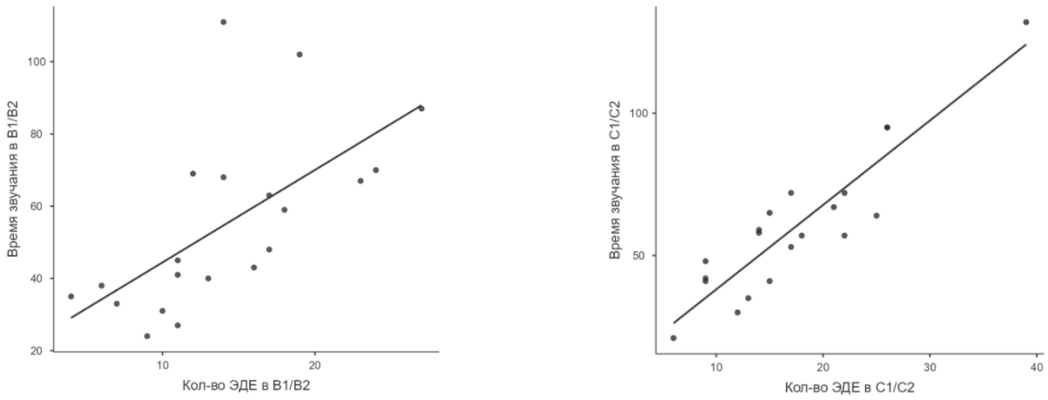


Рис. 3: Индивидуальные показатели говорящих в диаграммах рассеяния

Диаграммы показывают, что в группе C1/C2 данные по количеству ЭДЕ и времени звучания распределены более равномерно, чем в группе B1/B2, что указывает на бóльшую индивидуальную вариативность и, возможно, в целом бóльшие различия между говорящими на уровнях B1 и B2, выраженные в отношении количества ЭДЕ и времени звучания.

Рассмотрим особенности реализации ЭДЕ по каждому типу отношений.

Нарративные отношения. Как следует из Таблицы 1, среднее количество ЭДЕ в каждом цикле демонстрирует различия в двух группах. Отмечены несколько вариантов начала цикла. Самым частотным (30 случаев из 40) является Сеттинг, например, в / *в следующем тексте рассказывалось о картине «Золотая осень» Левитана* / или в / *там произошла картина «Девочка с персиком»* /. В 10 случаях обнаружено Начало, например, в / *переходящая с красного в оранжевый* / во фразе *Переходящая с красного в оранжевый картина Исаака Левитана была написана в 1895 году в городе Горки* или в / *художник Левитан нарисовал картину «Золотая осень»* /. В первом случае говорящий представляет событие, описывающее его «встречу» с событием в

стимульном тексте, а во втором – событие, происходящее в стимульном тексте. Отношение Кода встретилось в 15 случаях, например, в / *и всё в данном тексте* /, / *конец* /, / *и всё на это* /.

Коммуникативные отношения. Обнаружены примеры трех типов отношений: Комментарий (44 случая), Метадискурсив (36 случаев), Согласие (17 случаев). Так, Комментарий проявляется, например, в / *я забыл щас* /, / *так понял* /, / *от текста я узнала* /, где говорящий указывает на способ знакомства или обработки полученного в стимульном тексте знания. Метадискурсив, который в большинстве случаев сопровождает какое-либо из риторических отношений, наблюдаем в /*в основном, про девочку* / или в / *как будто не осень* /, в редких случаях это отдельная ЭДЕ, например, в / *так* / или в / *вот* /.

Примеры использования Согласия демонстрируют, что речь во всех случаях идет о подтверждении релевантности высказанной ранее информации, например, в /*потому что Левитан расставил* / *и сгруппировал все элементы* / *да* / *всё* / или в / *это про Мамонтова* / *да* / *и всё на это* /. Так, рассмотрим фрагмент речи участника группы B1/B2: / *и еще есть* / *что ее даже показывает что-то* / *и еще в комнате есть солнечные яркими и* / *да* / *о я думаю* / (Рис. 4).

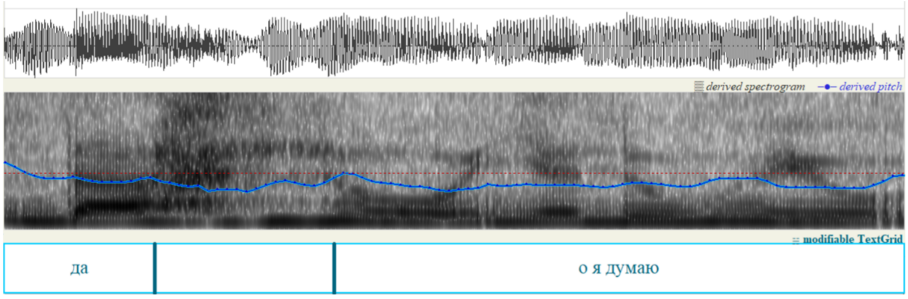


Рис. 4: Пример реализации отношения Согласия

В ЭДЕ / *да* / (далее следует шум до появления ЭДЕ / *о я думаю* /) наблюдаем нисходящий акцент (с небольшим подъемом в его завершении), что свидетельствует об отсутствии вопросительной интонации и вопросительного компонента в / *да* /. Скорее всего, таким образом участник подтверждает выполнение коммуникативного задания пояснения. Появление подъема в конце, возможно, указывает на незавершенность высказывания или на неуверенность говорящего [27].

Риторические отношения. Анализ симметричных отношений показал, что отношение Объединение используется значительно чаще (243 случая), чем Контраст (27 случаев) и Дизъюнкция (11 случаев). Асимметричное отношение Развитие темы встретилось намного реже (150 случаев), чем симметричное отношение Объединение, что свидетельствует не только о превалировании простых предложений в речи говорящих, но и в целом о невысоком проявлении связности в синтаксисе предложений. Так, Объединение обнаруживаем в /*картина Эм Вэ кого-то* / *в этой картине была девочка* / или в /*там написано с яркими цветами* / *есть домик и поля* /, где каждая ЭДЕ имеет свое ядро [15]. Развитие темы проиллюстрируем следующими примерами: /*от текста я узнала* / *что это говорится о картине Левитана* / или / *Левитан описывает осень* / *именно русский осень* /, где вторая ЭДЕ детализирует информацию первой. Отношения Контраста обнаруживаются в ситуациях, когда каждая из двух ЭДЕ имеет собственную ядерную структуру, при этом наблюдается некоторое их противоречие или рассогласование, например, в / *нагружены разными элементами* / *но все-таки Левитан как-то смог группировать* / или в /*в контрасте с ее розовой кофточкой показывают всю серьезность обстановки, молодость и юность* / *но при этом девушка очень веселая и озорная* /. Отношения Дизъюнкции более редкие, наблюдаются в основном в ЭДЕ, в которых есть и отношения Хезитации, например, в / *Левитан его сначала писал в какой-то губернии* / *в Тверской* / *или ещё что-то* / или в / *которую художник создал* / *по имени Левитан* / *или там что ли* /, т.е. это не в полной мере отношения между ЭДЕ с собственными ядрами, но отношения, где вторая ЭДЕ собственного ядра не демонстрирует (хотя оно предполагается). Среди других асимметричных отношений назовем частотные Интерпретацию (33 случая) в / *девочке не...* / *неудобно сидеть в картине* /, Причину в / *со временем картина приобрела обобщенный характер* / *поскольку она показывала то радость и озорство детства* / и Оценку в / *сама картинка очень экспрессивная и пышная* / *вызывает положительные эмоции* /

(по 12 случаев), Антитезу в / *что левая сторона более загруженная* / *чем правая* / (8 случаев), Условие в / *даже не смотря на подпись* / *можно это понять* / и Предоставление возможности в / *можно увидеть* / *что вот девочка...* / (по 7 случаев).

Речевые сбои. Обнаружено 98 случаев Хезитации, например, в / *и ещё этот текст говорит* / *что там...* / *в этой картине* / в ЭДЕ / *что там...* /, 42 случая Повтора (обратим внимание, что это именно Повтор, а не риторическое отношение Реформулирования) в / *которая произошла в 1887 году* / *которая произошла как портрет* /, и 19 случаев Самоисправления, например, в / *на самом деле на картине не описывает...* / *не описывается конкретно девочка* /. В случаях Самоисправления часто реализуются и риторические отношения Развития темы. Как видно, риторические сбои обнаружены в каждой четвертой ЭДЕ, что определенно обусловлено тем, что используемый язык не является родным для говорящих.

4.2 Локальная структура дискурса описания изучающих РКИ в группах B1/B2 и C1/C2

В Таблице 2 приведем отличия в локальной структуре дискурса описания в двух группах изучающих РКИ, B1/B2 и C1/C2; для их установления были проведены тесты с применением критерия хи-квадрат. В скобках укажем, в какой группе эти отношения превалируют.

		$\chi 2$	p
Нарративные отношения	Сеттинг (C1/C2)	8.72	0.004
Коммуникативные отношения	Согласие (B1/B2)	11.93	<0.001
	Метадискурсив (C1/C2)	8.41	0.004
Риторические	Интерпретация (C1/C2)	4.4	0.036
	Объединение (B1/B2)	14.256	<0.001
Речевые сбои	Самоисправление (C1/C2)	10.92	<0.001
	Повтор (B1/B2)	5.53	0.019

Таблица 2: Значимые различия в реализации отношений локальной структуры описательного дискурса в двух группах говорящих

Как видно из Таблицы 2, различия обнаружены по всем четырем типам отношений.

Нарративные отношения. Установлено более частое их использование в группе C1/C2, что свидетельствует о том, что говорящий уже выстраивает описание как нарратив, формируя содержание первой реплики с учетом содержания всего фрагмента. Рассмотрим два типичных примера начального фрагмента цикла описания. Так, в группе B1/B2 фрагмент / *картина «Девочка с персиками» была создана в 1877 году* / *сначала она называлась «Портрет ВМ»* / *это означает «Верочка Мамонтова»* / начинается с нарративного отношения Начало, далее находим Объединение и Объединение с Комментарием. В группе C1/C2 фрагмент / *в этом тексте это информация о картине «Девочка с персиком»* / *эта картина была создана в 1887 году русским известным живописцем Серовым* / *сначала эта картина имела другое название* / *но потом она была переименована* / начинается с Сеттинга, далее находим два примера Объединения и Контраст.

Коммуникативные отношения. Обнаружены различия в проявлении Согласия и Метадискурсива, где Согласие превалирует в группе B1/B2, а Метадискурсив – в группе C1/C2. Учитывая, что проявления Согласия связаны с выражением подтверждения содержания предыдущих реплик, Согласие может быть рассмотрено и как отношение, реализующее речевые сбои, что делает его характерным для группы B1/B2. Использование метадискурсивных слов типа *по-моему, прям, мне кажется, может быть, наверное* свидетельствует о более продвинутом уровне владения РКИ.

Риторические отношения. Различия обнаружены в проявлении Интерпретации, более характерной для уровня C1/C2, и Объединения, характерного для уровня B1/B2. Так, во фразе участника группы C1/C2, / *что в комнате проникнута ярким сиянием* / *что поражает зрители своей яркостью, светлостью и так далее* / во второй ЭДЕ обнаруживаем интерпретацию говорящим эффекта, создаваемого картиной, через указание на мнение воображаемых зрителей, к числу которых относится и он сам. Использование Объединения, как мы демонстрировали выше, характеризует речь говорящего как имеющую низкую степень связности между предложениями; эта особенность проявляется и в речи говорящих уровня C1/C2, но в меньшей степени.

Речевые сбои. Установлено, что Самоисправления больше свойственны группе C1/C2, что свидетельствует о большей концентрации говорящего на языковой структуре описания. Самоисправления реализуются не только в выборе грамматических форм, как в / *как будто не осень* / *на осень* /, но и в уточнении компонента ЭДЕ, например, в / *из картины понятно* / *из позиций, которые занимает девочка, понятно* /, что сближает их с проявлениями риторических отношений Развития темы. Повторы в то же время больше свойственны говорящим группы B1/B2; их использование служит не для акцентуации компонента содержания описания, а для организации самой структуры описания, как в / *я считаю* / *в основном это текст рассказывает о осень* / *осень* / *какая природа* / *или какие цветы* /, где повтор компонента / осень / используется говорящим для того, чтобы обозначить (для себя) тему, развиваемую далее в описании.

4.3 Различия локальной структуры дискурса описания в стимулах и в речи изучающих РКИ

Приведем данные распределения отношений в относительных величинах (Рис. 5).

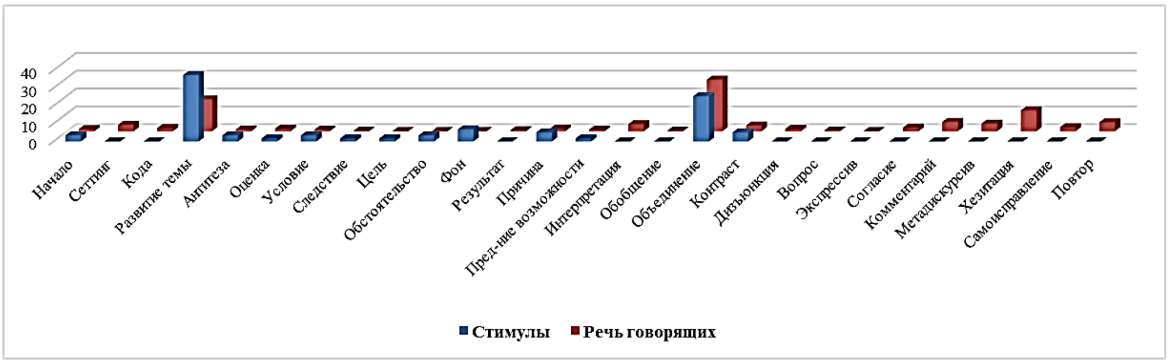


Рис. 5: Различия в реализации отношений локальной структуры в стимулах и в речи говорящих

Как видно, ведущими типами риторических отношений остаются Развитие темы и Объединение, однако если в стимулах Развитие темы составляет 37.29%, то в речи этот показатель составляет всего 17.88%, в то же время показатели Объединения сходные (25.42% и 28.96%). Исходный текст демонстрирует связность предложений, которая проявляется в меньшей степени в речи говорящих. Отмечаем перераспределение отношений в пользу речевых сбоев и нарративных отношений в пользу Сеттинга и Коды, появление случаев Интерпретации, Согласия, Комментария и Метадискурсива. В целом, распределение отношений в исходных стимулах и в речи можно представить следующей диаграммой (Рис. 6).

Рис. 6: Различия в реализации групп отношений локальной структуры в стимулах и в речи говорящих

Диаграммы дают представление о варьировании локальной структуры дискурса описания, определяемом письменным и устным форматами и статусом русского языка – родного и иностранного. Как видно, снижение доли риторических отношений восполняется коммуникативными и нарративными отношениями, а также в значительной степени речевыми сбоями.

5 Обсуждение результатов

Как мы указали выше, двумя основными принципами конструирования дискурса описания являются генерализация и конкретизация [11; 12]. При этом, согласно [12], в дискурсе описания носителей русского языка как родного превалирует конкретизация в использовании собственно описания, фона и уточнения. В стимулах также в большей степени проявляется конкретизация в описании ситуации, проблемы, решения [13; 14], что выражено в выборе риторических отношений Развития темы, Объединения, Причины, Контраста. При этом в речи изучающих РКИ обнаруживается снижение роли конкретизации, а также возрастание роли генерализации в выборе нарративных отношений и риторических отношений Интерпретации, Оценки.

Рассмотрим полученные результаты в контексте исследований каждого типа отношений.

Во-первых, обнаружено формирование в речи дискурса, демонстрирующего характеристики как собственно описания, так и нарратива, что свидетельствует о гибридизации дискурсов [26] и предоставляет сведения о количественном распределении их отношений в речи изучающих РКИ. Интерес представляет дальнейшее изучение таких гибридных форматов дискурса при контрастивном анализе его письменных и устных форматов, а также проявлений в речи носителей языка как родного и иностранного.

Во-вторых, отмечена специфика в реализации коммуникативных отношений, описанных в [1]. Их разнообразие в нашем корпусе невелико, частотность почти всех типов низкая, кроме отношений Согласия. При этом отношения Согласия более напоминают риторические сбои, о чем свидетельствует и тот факт, что они чаще используются в группе B1/B2. Были обнаружены два дополнительных типа отношений – Комментарий и Метадискурсив в *наверно*, *по-моему*, *типа* и др., которые используются для выражения сомнения и снижения определенности высказывания, поэтому их также можно отнести и к риторическим сбоям, а именно к Хезитации (реализуемой в т.ч. аппроксиматорами [19; 24]), хотя по формальным признакам они таковыми не являются.

В-третьих, среди риторических отношений [20; 21; 22] мы обнаружили в речи говорящих два ведущих типа – Объединение и Развитие темы. Результаты частично согласуются с наблюдениями, сделанными в отношении письменной речи изучающих иностранный язык, где превалируют именно отношения Объединения [25]. Предположительно, незначительная роль других отношений объясняется тем, что испытуемые говорят на неродном языке, их основная задача – воспроизвести как можно больше фактуальной информации исходного стимула; при этом связность высказываний не обеспечивается. Отметим, что результаты не показали, что риторические отношения Причины усваиваются раньше, чем отношения Объединения [7]), что, возможно, связано с низкой частотностью отношений Причины в собранном корпусе, демонстрирующем описательный дискурс, для которого данный тип отношений не характерен.

В-четвертых, исследование доказывает актуальность изучения речевых сбоев в контексте РКИ. Хотя их типологизации посвящен ряд работ [1; 19], такое отношение, как Повтор, например, нуждается в дальнейшем описании. В частности, это относится к систематизации его функций, что представляется возможным сделать с расширением материала исследования. Также, как мы показали, некоторые из коммуникативных отношений могут быть рассмотрены как реализующие функцию речевого сбоя.

6 Заключительные замечания

Проведенное исследование, в ходе которого были получены новые данные о функционировании дискурса описания в речи носителей русского языка как иностранного в двух группах уровня B1/B2 и C1/C2, позволило сформулировать ряд важных выводов. Во-первых, общая структура дискурса описания в его письменном и устном форматах демонстрирует различия в составе групп отношений локальной структуры дискурса; определенно, эти различия обусловлены не только тем, что говорящие не являлись носителями русского языка. Во-вторых, обнаружены специфические особенности в распределении отношений локальной структуры дискурса в самой речи говорящих, которые, предположительно, обусловлены именно тем, что язык, на котором велось общение, не являлся родным для говорящих. В-третьих, установлены отличительные характеристики в выборе и проявлении отношений, свойственные каждой из анализируемых групп. Полученные результаты подтверждают актуальность дескриптивного (в определенной

степени, корпусного) подхода в исследовании РКИ, т.к. указывают на закономерности, которые необходимо привлечь во внимание при подготовке дидактических материалов при преподавании русского языка как иностранного в аспекте структуры дискурса. Так, речь должна идти об активации всех типов нарративных отношений, коммуникативного отношения Метадискурсива, а также риторических отношений разных типов (в меньшей степени Объединения), что, как показало проведенное исследование, соотносится с более высоким уровнем владения языка.

Список литературы

[1] Alla O. Litvinenko. Describing discourse structure in terms of Rhetorical Structure Theory: Application to Russian material. In: Narin'yani A. S. (ed.), *Trudy Mezhdunarodnogo seminara Dialog'2001 po komp'yuternoi lingvistike i ee prilozheniyam. Vol. 1: Teoreticheskie problemy*. Narin'yani A. S. (ed.), 2001. P. 159–168.

[2] Andrey A. Kibrik. Modus, genres and other parameters of discourse classification. *Voprosy Jazykoznanija,* 2, 2009. P. 3–21.

[3] Anastasiia Konovalova, Vladislav Zubov. Exploring the role of L1 reading ability when reading in L2. *ExLing 2022 Paris: Proceedings of 13th International Conference of Experimental Linguistics*, 17–19 October 2022. P. 101–104.

[4] Svetlana Alexeeva, Daria Chernova, Marina Norkina, Maria Kharchevnik. Predictability effect in L2 reading: eye-tracking data. *AMLaP 29 conference, Architectures and Mechanisms for Language Processing*, 2023. P. 62.

[5] Olga G. Nikonorova. Novye zhanry i diskursivnye praktiki v prepodavanii RKI [New genres and discursive practices in teaching Russian as foreign language]. *Sovremennaja rechevaja kommunikacija: novye diskursivnye praktiki.* Omsk, Izd-vo OmGU, 2011. P. 336–341.

[6] Viktorija V. Novikova. Linguomethodic foundations for implementation of a discourse approach to training Russian language as foreign language. *Vestnik Sovremennyh Issledovanij,* 11.7 (26), 2018. P. 189–192.

[7] Mathis Wetzel, Sandrine Zufferey, Pascal Gygax. Second Language Acquisition and the Mastery of Discourse Connectives: Assessing the Factors That Hinder L2-Learners from Mastering French Connectives. *Languages*. 5, 2020. Article No 35.

[8] Sandrine Zufferey, Mak Willem, Degand Liesbeth, Ted Sanders. Advanced learners' comprehension of discourse connectives: The role of L1 transfer across on-line and o-line tasks. *Second Language Research*, 31, 2015. P. 389–411.

[9] Jiangang Shi. 2017. A Corpus-Based Study of Contrastive/Concessive Linking Adverbials in Spoken English of Chinese EFL Learners. Studies in Literature and Language, 14, 2017. P. 17–25.

[10] Robert E. Longacre. *The grammar of discourse.* Boston, MA: Springer US. 1996.

[11] Ronald W. Langacker. *Cognitive Grammar: A Basic Introduction*. Oxford: Oxford University Press, 2008.

[12] Olga. N. Prokof'eva. *Kognitivnye i pragmaticheskie faktory sovmestnogo upotreblenija verbal'nyh i zhestovyh sredstv v opisatel'nom diskurse* [Cognitive and pragmatic factors of the joint use of verbal and gestural means in descriptive discourse]. PhD Thesis. M.: MSLU, 2022.

[13] Robert E. Longacre. *Holistic Textlinguistics*. SIL International. 2003.

[14] Michael Hoey. *On the Surface of Discourse.* London: George Allen and Unwin. 1983.

[15] Tat'jana V. Zherebilo. *Slovar' lingvisticheskikh terminov* [Dictionary of Linguistic Terms]. Nazran': Piligrim Publ. 2010.

[16] Olga N. Prokof'eva, Olga K. Iriskhanova, Maria I. Kiose. Rech' i zhesty v opisatel'nom diskurse [Speech and gestures in descriptive discourse]. In: O. K. Iriskhanova, *Polimodal'nye izmerenija diskursa.* M.: YaSK, 2021. P. 63–109.

[17] Oleg V. Potekhin, Maria I. Kiose. Multimodal conceptualization of immersive visual art: discourse schemas in speech, gesture and gaze. *Issues of Cognitive Linguistics,* 3, 2024. P. 20–32.

[18] Tim MacSaveny. Towards a Description of Descriptive Discourse. *GIALens,* 3, 2010. P. 1–5.

[19] Vera I. Podlesskaya, Andrey A. Kibrik. Samoispravlenija govorjashhego I drugie tipy rechevyh sboev kak ob'ekt annotirovanija v korpusah ustnoj rechi [Speaker's self-repairs and other types of speech disfluencies as an object of annotation in oral speech corpora]. *Nauchno-Tehnicheskaja Informacija. Serija 2: Informacionnye Processy I Sistemy,* 2007. P. 2–23.

[20] William C. Mann, Sandra A. Thomspon. Rhetorical structure theory: Towards a functional theory of text organization. *Text - Interdisciplinary Journal for the Study of Discourse*, 8, 1988. P. 243–281.

[21] Manfred Stede, Stergos Afantenos, Andreas Peldzsus, Nicholas Asher, Jeremy Perret. Parallel Discourse Annotations on a Corpus of Short Texts. In: *10th International Conference on Language Resources and Evaluation (LREC 2016)*, 2016. P. 1051–1058.

[22] Svetlana Yu. Toldova, T. Davydova, Maria V. Kobozeva, Dina B. Pisarevskaya. Discourse features of blogs in subcorpus of Russian Ru RSTree Bank. *Computational Linguistics and Intellectual Technologies*, 19(26), 2020. P. 747–761.

[23] Vera I. Podlesskaya. Vague names in Russian speech: a corpus study. *Computational Linguistics and Intellectual Technologies,* 12(19), 2013. P. 631–643.

[24] Olga K. Iriskhanova, Maria I. Kiose, Anna V. Leonteva, Olga V. Agafonova Vague reference in expository discourse: multimodal regularities of speech and gesture. *Computational Linguistics and Intellectual Technologies*, 22, 2023. P. 172–180.

[25] Marie Tapper. Connectives in advanced Swedish EFL learners' written English–preliminary results. *The Department of English: Working Papers in English Linguistics,* 5, 2005. P. 116–144.

[26] Olga K. Iriskhanova. Semioticheskie gibridy kak sposob reprezentacii znanij v diskurse [Semiotic hybrids as means of representing knowledge in discourse]. In: E. S. Kubryakova, *Problemy predstavlenija (reprezentacii) znanij v jazyke. Tipy i formaty znanij*. M., Kaluga: Eidos, 2007. P. 271–278.

[27] Tatiana Ye. Yanko. *Intonacionnye strategii russkoj rechi v sopostavitel'nom aspekte* [Intonation strategies of Russian speech in contrastive aspect]. M., Yazyki Slavyanskikh Kultur, 2008.

Contexts for overlapping talk in Russian conversations

Nikolay Korotaev
Institute of Linguistics RAS, Moscow, Russia
Russian State University for the Humanities, Moscow, Russia
n_korotaev@hotmail.com

Abstract

The paper analyses instances of overlap in Russian triadic conversations. Based on the "Russian Pear Chats and Stories" corpus, I provide a multimodal treatment of 364 overlapping episodes. I mostly focus on two conversational parameters — discourse contexts for overlaps and respective epistemic statuses of the interlocutors. A vast majority of overlaps in the data can be attributed to one of the four basic discourse contexts: Sideline Turn, Turn Transition, Turn Interception, and Competitive Development. These contexts differ in the degree of cooperation vs. competition exhibited by the participants. I show that overlaps appearing in more competitive contexts tend to be longer and more often include truncated turns. Epistemic statuses of participants are fixed throughout the studied conversations. When statuses are considered equal, competitive overlaps are more frequent than when one participant has a higher epistemic status.

Keywords: conversation, multimodal communication, turn-taking, cooperation, overlap, epistemic status
DOI: 10.28995/2075-7182-2025-23-169-182

Дискурсивные контексты наложений в диалоге на русском языке

Коротаев Н. А.
Институт языкознания РАН, Москва, Россия
Российский государственный гуманитарный университет, Москва, Россия
n_korotaev@hotmail.com

Аннотация

В докладе на материале мультиканального корпуса «Рассказы и разговоры о грушах» представлен анализ случаев наложения, т. е. одновременного говорения в диалоге. Основное внимание уделено двум параметрам коммуникативной ситуации: типу дискурсивного контекста и соотношению эпистемических статусов участников, задействованных в наложении. Показано, что подавляющее большинство эпизодов наложения может быть отнесено к одному из четырех дискурсивных типов, различающихся в том числе с точки зрения степени кооперативности / конкурентности в речевом и невербальном поведении участников. Для менее кооперативных контекстов характерны более длительные отрезки наложения и бóльшая частотность дискурсивных обрывов. Под эпистемическим статусом понимается степень осведомленности участника о теме обсуждения. Продемонстрировано, что в случаях равного эпистемического статуса — по сравнению с ситуациями, в которых статус одного из участников выше — увеличивается доля наложений более конкурентных типов.

Ключевые слова: диалог, мультимодальная коммуникация, чередование реплик, кооперативность, наложение, эпистемический статус

1 Введение

Одна из важных установок исследований в области естественного диалога состоит в том, что участники диалогического взаимодействия разделяют набор представлений о «нормальном», или предпочтительном характере его протекания [Brown, Levinson 1987; Clark 1996: pt. II; Pomerantz, Heritage 2012]. В этот набор, видимо, входит и представление о том, что в каждый момент разговора должен быть выбран один участник, осуществляющий говорение. В исследованиях, выполненных в рамках Анализа бытового диалога, это представление известно как принцип 'one party at a time' [Schegloff 1995; Santner-Wolfartsberger 2015]; по сути, оно включено и в формулировку правил чередования реплик в основополагающей работе этого направления [Sacks et al. 1974]:

правила описывают процедуру «назначения» (одного) следующего говорящего после т. н. точки перехода, т. е. после завершения текущей реплики.

В то же время принцип ‘one party at a time’ не является абсолютным, поскольку в естественном диалоге регулярно возникают ситуации наложения — одновременного говорения более чем одного участника [Ono, Thompson 1996; Schegloff 2000; Jefferson 2004; Hayashi 2012: 175-182; inter alia]. Это может происходить как вследствие нарушения общих принципов кооперации (например, при прямом перебивании), так и в менее соревновательных контекстах — в том числе и в результате стандартного применения правил чередования. При том что не существует ни устоявшейся классификации дискурсивных контекстов наложения, ни консенсусной оценки частотности этого феномена (см. обзоры этих вопросов, соответственно, в [Stolt 2008] и [Heldner, Edlund 2010]), практически не остается сомнений, что наложение входит в число центральных явлений, обеспечивающих нормальное функционирование естественного диалога. Так, во влиятельной работе [Schegloff 2000] было показано, что для ведения полноценного разговора его участники должны владеть не только правилами чередования, но и механизмами урегулирования наложений. Внимание к наложениям характерно не только для теоретических исследований, но и для работ в области автоматического порождения диалога. Например, в [Yamashita et al. 2022] сообщается, что естественность искусственно сгенерированных диалогических датасетов повышается, если в алгоритм генерации добавляется возможность наложения; в [Zhang et al. 2024] в качестве одной из метрик, позволяющих оценить прогресс модели автоматической генерации диалога, используется доля наложений в последовательностях реплик.

В настоящем исследовании решается две задачи: основная и дополнительная. Основная — предложить классификацию дискурсивных контекстов возникновения наложений на материале естественных диалогов на русском языке и дать количественную оценку распределения их частотности. Насколько мне известно, эта задача ранее не решалась[1]. Дополнительная задача обусловлена материалом исследования — корпусом «Рассказы и разговоры о грушах». В записях корпуса участники диалога имеют закрепленные роли, во многом основанные на степени их знакомства с содержанием «Фильма о грушах». Это позволяет использовать понятие эпистемического статуса участников, т. е. степени осведомленности об обсуждаемых в разговоре событиях [Du Bois 2007; Heritage 2012], и оценить, как вариативность в значениях этого параметра влияет на характер действий участников в контексте наложения.

Структура дальнейшего текста такова. В разделе 2 кратко описывается исследовательский корпус. В разделе 3 на примерах иллюстрируются типичные контексты наложений, обнаруженные в корпусе. В разделе 4 представлены количественные результаты исследования; эти результаты далее обсуждаются в разделе 5. Раздел 6 представляет собой заключение.

2 Материал

Материалом для исследования послужил мультиканальный корпус «Рассказы и разговоры о грушах» (https://multidiscourse.ru/; [Кибрик, Федорова 2018]). Этот ресурс (далее — RUPEX, от *Russian Pear Chats and Stories*) уже неоднократно использовался для анализа устной монологической и диалогической речи, а также невербальных аспектов естественной коммуникации; поэтому я позволю себе не приводить тут подробного его описания. Для настоящей работы существенно, что корпус содержит аннотированные аудио- и видеозаписи взаимодействия между тремя основными участниками (разными для каждой записи). Все записи организованы по единой схеме, в рамках которой за каждым участником закреплена определенная роль, не меняющаяся по ходу записи. До начала записи двум участникам — Рассказчику и Комментатору — предъявлялся «Фильм о грушах» [Chafe (ed.) 1980], после чего следовало три основных этапа. На первом этапе (обозначаемом как рассказ) Рассказчик в монологическом режиме передавал содержание фильма Пересказчику — третьему участнику, не видевшему фильма. На втором этапе (разговор) все трое участников обсуждали содержание фильма в диалогическом режиме. На третьем этапе (пересказ) от Пересказчика требовалось рассказать о событиях фильма, обращаясь к четвертому участнику, до этого не присутствовавшему в комнате, — Слушателю.

[1] Среди немногочисленных работ, посвященных наложениям в русском языке, можно отметить статью [Чекмез 2019], в которой, однако решаются несколько иные задачи.

Очевидно, для изучения диалогического поведения актуален этап разговора. Тут следует дополнительно отметить, что в соответствии с дизайном записей между участниками изначально устанавливалась иерархия эпистемических статусов при разговоре о сюжете и содержании фильма. Более того, поскольку основной целью, которая сообщалась участникам, было максимально точное сохранение всех деталей фильма в финальном пересказе, на этапе разговора основные коммуникативные усилия участников концентрировались именно на обсуждении этих деталей. Соответственно, абсолютное большинство дискурсивно связанных последовательностей реплик реализуются в контексте осознаваемых всеми тремя участниками соотношений их эпистемических статусов: Рассказчик и Комментатор обладают равноправным статусом, и этот статус выше, чем у Пересказчика.

Непосредственно для настоящего исследования были использованы этапы разговоров трех записей т. н. «эталонного» подкорпуса, общей продолжительностью около 30 минут. Выбранный подкорпус обладает следующими особенностями. Во-первых, в нем имеются точные сведения о временны́х границах всех слов и других вокальных единиц, реализуемых каждым участником[2]. Во-вторых, для него выполнена сплошная дискурсивно-просодическая разметка вокального поведения участников в духе [Kibrik et al. 2020]. Также для записей подкорпуса доступны аннотации мануальной жестикуляции [Litvinenko et al. 2018] и глазодвигательной активности [Fedorova 2021]. Ниже, приводя примеры из корпуса, я не использую эти виды разметки непосредственно, но в ряде случаев опираюсь на описание невокального поведения участников, если это помогает точнее установить особенности коммуникативной ситуации.

3 Дискурсивные контексты наложений

Всего в исследованном материале было обнаружено 406 эпизодов наложений; средняя продолжительность эпизода составила 527 мс, медианное значение — 370 мс, стандартное отклонение — 470 мс. Из дальнейшего анализа были исключены 42 эпизода, длительность которых не превышает 100 мс, что примерно соответствует нижним 10 процентилям общего распределения. Из оставшихся 364 эпизодов абсолютное большинство было отнесено к одному из четырех дискурсивных типов. Ниже приводится описание этих типов и примеры из корпуса.

3.1 Побочная реплика

При наложении этого типа второй участник (формальный инициатор наложения), произнося свою реплику, не претендует на получение полноценной очереди хода. Дальнейшее развитие диалога обусловлено (в первую очередь) не этой репликой, а действиями других участников. Наиболее часто встречающимся дискурсивным контекстом тут выступает ситуация, при которой один из участников подает сигналы обратной связи, накладывающиеся на реплики собеседников; см. ниже пример (1).

3.2 Передача хода

В этом типе объединены контексты, в которых действия первого говорящего могут интерпретироваться как передача хода другому участнику. В наиболее очевидном случае речь идет о прямом назначении следующего говорящего (правило 1 в терминах [Sacks et al. 1974]), в менее очевидных — о надежно считываемой другими участниками готовности завершить свой ход и не препятствовать следующему говорящему. При отсутствии наложения эта ситуация реализуется как стандартная последовательность реплик, произносимых разными участниками. В свою очередь, наложения возможны в тех случаях, когда участники по-разному определяют место точки перехода.

Так, при т. н. конечном наложении (*terminal overlap*) первый участник близок к завершению своей реплики, когда второй участник, оценив уже произнесенную часть реплики собеседника как в целом достаточную для интерпретации его коммуникативных намерений (и, соответственно, определив текущий момент в развитии диалога как потенциальную точку перехода), начинает свою реплику [Jefferson 1984]. В свою очередь, первый участник, обнаружив, что на его

[2] Границы устанавливались в два этапа. На первом этапе выполнялась автоматическое определение границ слов при помощи алгоритма, разработанного компанией STEL (http://stel.ru/). На втором этапе границы уточнялись вручную в программе PRAAT [Boersma, Weenink 2023].

реплику уже начала следовать реакция, либо завершает ее в штатном режиме (но не начинает следующую), либо обрывает ее. В целом эта ситуация характеризуется высокой степенью кооперативности; см. пример (1).

(1) Pears22: «что-то такое светлое».[3]

TimeS	TimeE	Narrator		Reteller	
520.70				R-vE021	ʾА /мужик-то –как выглядит всё-таки^{h}?
522.08		N-vE249	У /него что-то такое /светлое,		
	522.48				
522.92	523.24			R-vN006	(ц 0.32)
523.24				R-vE022	–Ага-а.
	523.61				
523.61		N-vE250	сверху светлый /фартук,		
	523.66				
	525.03				
525.03	526.20	N-vE251	(\Такой || с \карманом.)		
526.20	526.54	N-vN054	(ц 0.34)		
526.54	528.28	N-vE252	здесь красная /бандана такая,		

В строке R-vE021 Пересказчица (**R**) задает вопрос, возвращая течение беседы к теме, которая уже поднималась за некоторое время до этого. Рассказчица (**N**) определяет смысл реплики **R** еще до ее завершения и, немного не дождавшись точки перехода, начинает развернутый ответ, который продлится примерно 15 секунд. Ее готовность немедленно приступить к ответу проявляется и на уровне мануальной жестикуляции — см. рис. 1, на котором видно, что **N** начинает жестикулировать еще во время завершения реплики **R** (до этого ее руки находились в положении покоя на коленях). В свою очередь **R**, изначально определившая именно **N** как участника, ответственного за продолжение разговора (см. направление ее взгляда на рис. 1[4]), не предпринимает никаких попыток помешать поступающему ответу — пусть даже он накладывается на завершение ее собственной реплики. Отмечу также, что спустя недолгое время **R** подает сигнал обратной связи (строка R-vE022; *ага*), речь о которых шла в разделе 3.1 выше.

[3] При цитировании примеров из корпуса RUPEX используется т. н. партитурная нотация, в которой вокальные действия каждого участника фиксируются в отдельных столбцах таблицы. Непрерывные области вокализации одного участника оформлены сплошной цветовой заливкой ячеек. Конкретные отрывки, одновременно произносимые разными участниками, дополнительно выделены при помощи заливки отдельных слов. В остальном разметка следует принципам, изложенным в [Kibrik et al. 2020]. Каждая строка соответствует одной элементарной дискурсивной единице (ЭДЕ) — минимальному шагу в реализации устной речи, обладающему единым интонационным контуром. При помощи слэшей и стрелок обозначаются движения частоты основного тона в акцентированных словоформах; при помощи пунктуационных знаков в конце строк — иллокутивно-фазовая семантика ЭДЕ; обрывы слов обозначаются одиночным равенством; знаки ||, == и ≈≈ используются при дискурсивных обрывах различной природы; в скобках указывается длительность абсолютных и заполненных пауз; в двух левых столбцах отмечается время начала и конца произнесения ЭДЕ.

[4] Литера N на рисунках обозначает Рассказчиков, C — Комментаторов, R — Пересказчиков. Как уже отмечалось в разделе 2, для записей корпуса доступна разметка невокального поведения участников; так, для Рассказчиков и Пересказчиков есть данные о направлении движения взгляда, полученные с айтрекеров [Fedorova 2021]. Для простоты презентации я обозначаю направления движения взгляда стрелками на рисунках.

Рисунок 1. Взаимодействие участников разговора записи 22 в момент завершения произнесения Пересказчицей ЭДЕ R-vE021 примера (1) (*А мужик-то как выглядел всё-таки?*)

В других конфигурациях, относящихся к этому же типу, первый участник не прекращает говорить после той реплики, реакцию на которую выполняет второй говорящий, а продолжает, реализуя расширение исходной реплики. Так, во фрагменте (2) представлен случай наложения, обусловленного поздней реакции второго участника.

(2) Pears23: «первый раз спустился»

Pauses	TimeS	TimeE	Narrator		Reteller	
	521.79	521.85			R-vE077	–Не,
	521.85	523.35			R-vE078	а вот /\первый раз когда он \спускался.
(0.44)	523.35	523.79				
	523.79				R-vE079	Вот когда он первый раз /спустился,
	523.86	524.50	N-vE268	Первый /раз он' ==		
	524.68		N-vE269	'Он || ну /он чуть-чуть /положил,		
		524.87				
	524.87	525.04			R-vE080	он ≈≈
		525.84				
	525.84	527.03	N-vE270	и /наполнил так \вторую.		

В строках R-vE077-078 Пересказчик (**R**) уточняет ранее заданный вопрос о порядке сбора груш одним из персонажей фильма. Сделав это уточнение, он в течение 440 мс ожидает ответной реакции со стороны Рассказчицы (**N**), которая до этого уже предлагала свое описание интересующего **R** процесса. Не получив реакции, **R** решает еще раз уточнить свой вопрос, переформулировав предыдущую реплику. Практически в это же время (но чуть позже) **N** все же начинает отвечать (ЭДЕ N-vE268 и далее). Вследствие этой, несколько запоздалой, реакции возникает наложение, для урегулирования которого **R** обрывает произнесение ЭДЕ R-vE080 и передает ход **N** — что и было его изначальным планом.

3.3 Перехват хода

В этом типе собраны контексты, в которых действия второго участника могут быть интерпретированы как попытка перехвата очереди говорить вне стандартных правил чередования. Один из примеров такого рода представлен во фрагменте (3).

(3) Pears23: «шортики»

Pauses	TimeS	TimeE	Narrator		Commentator	
	467.58	469.57	N-vE223	У /–него-о были красные-е /штаны,		
	469.57		N-vE224	(\→длинные,) ≈≈		
	469.77				C-vE154	\Шортики у него были!
		470.29				
		471.07				
	471.07	472.26			C-vE155	У него же ещё /\гольфы (0.08) →↑были.
(0.42)	472.26	472.68				
	472.68	773.03	N-vE225	/\Нет!		
	773.03	773.77	N-vE226	У него были /штаны, ==		

Фрагмент открывается серией реплик Рассказчицы (**N**), осуществляющей описание одежды одного из персонажей фильма. В строке N-vE224 содержится дискурсивный обрыв: хотя **N** целиком произнесла свою ЭДЕ, эта ЭДЕ не планировалась ей как последняя в цепочке. По своим просодическим и (потенциальным) дискурсивным характеристикам эта единица достаточно надежно определяется как парентеза (вставка), т. е. временное отступление от основной линии изложения (см. [Кибрик, Подлесская 2010; Коротаев 2021]). То, что за этим временным отступлением не последовало продолжения, связано с действием Комментатора (**C**), которая вступила со своей репликой заведомо до планировавшегося завершения реплики Рассказчицы. Не желая дожидаться точки потенциального перехода, **C** перебивает **N**, чтобы предложить альтернативное описание, и осуществляет таким образом перехват хода вне стандартных правил чередования. Реакция **N** на это действие собеседницы оказывается комплексной. С одной стороны, она прерывает говорение. С другой стороны, меняется и ее невокальное поведение. До того как **C** ее перебила, **N** смотрела на Пересказчика (**R**) и сопровождала свое описание изобразительным мануальным жестом на уровне щиколотки (см. рис. 2а); после начала ЭДЕ C-vE154 она резко поворачивается к **C** и ее руки застывают на месте (рис. 2б); затем она постепенно возвращает руки в исходное положение на коленях (рис. 2в).

Рисунок 2а. Взаимодействие участников разговора записи 23 в момент завершения произнесения Рассказчицей ЭДЕ N-vE223 примера (3) (*У него были красные штаны*)

Рисунок 2б. Взаимодействие участников разговора записи 23 в момент начала произнесения Комментатором ЭДЕ C-vE154 примера (3) (*Шортики у него были!*)

Рисунок 2в. Взаимодействие участников разговора записи 23 в момент завершения произнесения Комментатором ЭДЕ C-vE155 примера (3) (*У него же ещё гольфы были*)

3.4 Конкурентное развитие

В контекстах этого типа участники, задействованные в наложении, стремятся развить параллельные и/или альтернативные линии диалога. Сразу два примера подобного рода представлены во фрагменте (4).

(4) Pears04: «не достаёт до земли»

TimeS	TimeE	Narrator		Commentator		Reteller	
607.69						R-vE098	он не /до-стаёт до-о || (1.07) до земли?
608.62		N-vE291	\ʷДа,				
608.82				C-vE095	Ну он не \достаёт,		
	608.89						
608.89	609.09	N-vE292	я ==				
609.09	609.20	N-vE293	он ==				
609.20	609.72	N-vE294	я \сказала,				
609.72		N-vE295	что он даже не \сидит на се= || на се= || на /седле,				
	609.77						
609.77				C-vE096	поэтому он и –свалился.		
	610.36						
	610.94						
	612.41						
612.41	612.86	N-vN052	(ц 0.45)				
612.86	614.87	N-vE296	он-н (ə 0.42) (0.40) либо /стоит,				
614.87		N-vE297	либо прямо не знаю на эту /раму садится,,,				
615.06				C-vE097	/ˈВелосипед такой с вы= || (0.14) с \высокой рамой.		
	616.89						
616.89	617.12	N-vE298	\как-то.				
	617.67						

Фрагмент начинается со случая конечного наложения (см. выше, пример (1)): и Рассказчица (**N**), и чуть позже Комментатор (**C**) начинают отвечать на заданный Пересказчицей (**R**) вопрос прежде, чем она его завершит. При этом ответы выполняются ими в конкурентной манере: на протяжении нескольких дискурсивных шагов **N** и **C** реализуют свои варианты, обращаясь к **R** и не проявляя намерения урегулировать коллизию одновременного говорения. Как видно на рис. 3а, в это время они ищут зрительный контакт с **R**, но не друг с другом. Только спустя более чем две секунды **C**, закончив свою реплику, поворачивается к **N** (см. рис. 3б), но лишь затем, чтобы еще через три секунды инициировать новый случай конкурентного наложения — предложить параллельное описание устройства велосипеда. Как видно на рис. 3в, в этот момент он снова переводит взгляд в направлении **R**.

Рисунок 3а. Взаимодействие участников разговора записи 04 в момент параллельного произнесения Рассказчицей ЭДЕ N-vE295 (*что он даже не сидит на седле*) и Комментатором ЭДЕ C-vE096 (*поэтому он и свалился*) примера (4)

Рисунок 3б. Взаимодействие участников разговора записи 04 в момент единоличного произнесения Рассказчицей финальной части ЭДЕ N-vE295 (*что он даже не сидит на седле*) примера (4)

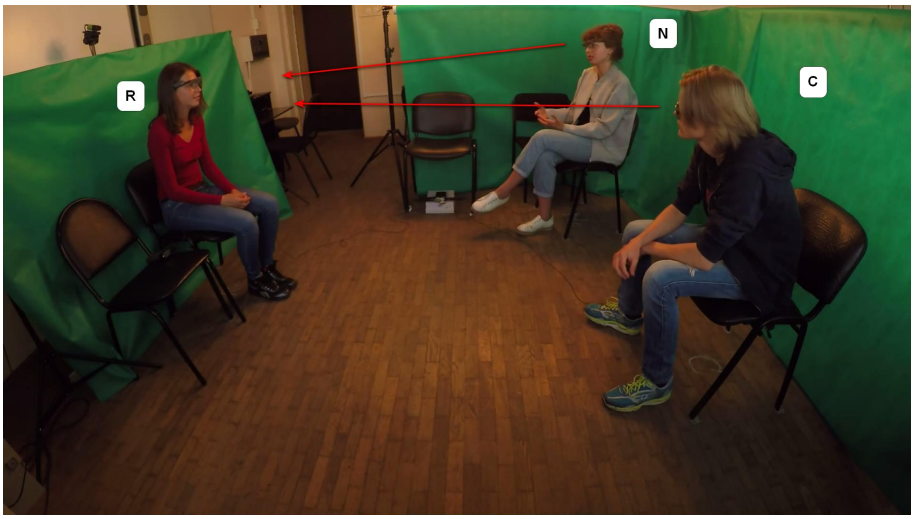


Рисунок 3в. Взаимодействие участников разговора записи 04 в момент параллельного произнесения Рассказчицей ЭДЕ N-vE297 (*либо прямо не знаю на эту раму садится*) и Комментатором ЭДЕ C-vE097 (*Велосипед такой с высокой рамой*) примера (4)

4 Количественные результаты

В табл. 1 представлены данные о количественном распределении обнаруженных в корпусе эпизодов наложения по рассмотренным выше типам дискурсивных контекстов, а также о средних и медианных значениях длительности наложения по отдельности для каждого типа.

Тип контекста	**Число эпизодов**	**Длительность (ср. знач., медиана; в мс)**
Побочная реплика	90 (24.7 %)	324; 285
Передача хода	150 (41.2 %)	536; 420
Перехват хода	33 (9.1 %)	863; 780
Конкурентное развитие	85 (23.4 %)	844; 710
Другие	*6 (1.7 %)*	*408; 350*
ВСЕГО	364	583; 430

Таблица 1. Число и длительность эпизодов наложений по типам дискурсивных контекстов

Аналогичные данные о длительности эпизодов наложения, распределенных по соотношению эпистемических статусов, представлены в табл. 2. Напомню, что равным эпистемическим статусом обладают Рассказчик и Комментатор, а статус Пересказчика ниже, чем у этих двух участников.

Соотношение эпистемических статусов	**Число эпизодов**	**Длительность (ср. знач., медиана; в мс)**
Равный	150 (41.2 %)	732; 570
Выше у первого участника	139 (38.2 %)	448; 330
Выше у второго участника	75 (20.6 %)	534; 430
ВСЕГО	364	583; 430

Таблица 2. Число и длительность эпизодов наложений по соотношению эпистемических статусов участников

Наблюдаемые различия в длительности («Побочная реплика» < «Передача хода» < «Перехват хода», «Конкурентное развитие» для первого параметра; эпистемический статус выше у первого участника < статус выше у второго участника < статусы участников равны для второго параметра) обладают статистической значимостью (непараметрический тест Краскела-Уоллиса, $p < 0.001$ для обоих параметров). Апостериорные сравнения при помощи теста Манна-Уитни выявили статистически значимые различия между всеми парами для обоих параметров, кроме пары «Перехват хода» и «Конкурентное развитие».

На рис. 4 приведены данные о распределении эпизодов наложений с различными соотношениями эпистемических статусов участников относительно четырех основных дискурсивных типов. Различия в распределении, которые будут обсуждены ниже в разделе 5, обладают статистической значимостью (критерий «хи-квадрат», $p < 0.001$; как для всех четырех типов, так и отдельно для трех типов, исключая тип «Побочная реплика).

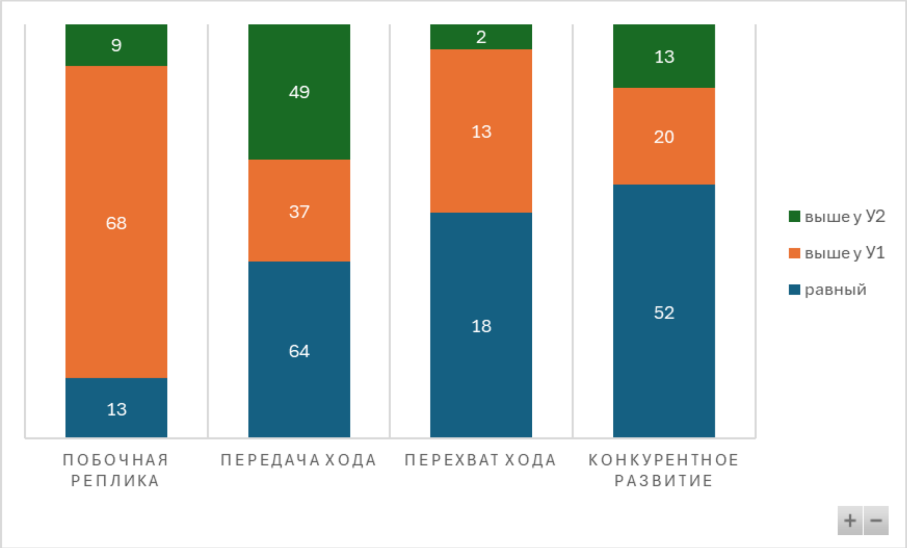


Рисунок 4. Дискурсивные контексты наложения и эпистемические статусы участников

Наконец, в табл. 3 представлены данные о распределении эпизодов наложения с обрывом в речи хотя бы одного из участников vs. без обрыва по типам дискурсивных контекстов. Обнаруживаемые различия (наивысшая доля эпизодов с обрывами — в типе «Перехват хода»; в типе «Побочная реплика» обрывов нет вовсе; из остальных типов наименьшая частотность обрывов — в типе «Передача хода») также обладают статистической значимостью (критерий «хи-квадрат», $p < 0.0001$; подсчет проводился без учета типа «Побочная реплика»).

Тип контекста	**Эпизодов с обрывами**	**Эпизодов без обрывов**
Побочная реплика	0	90 (100 %)
Передача хода	16 (10.7 %)	134 (89.3 %)
Перехват хода	14 (42.4 %)	19 (57.6 %)
Конкурентное развитие	26 (30.6 %)	59 (69.4 %)
ВСЕГО	56 (15.6 %)	302 (84.4 %)

Таблица 3. Наличие vs. отсутствие обрывов при урегулировании наложений в основных типах дискурсивных контекстов

5 Обсуждение

Полученные результаты, на мой взгляд, любопытны, по крайней мере, с двух точек зрения. Во-первых, они могут быть проинтерпретированы в контексте противопоставления соревновательных и несоревновательных наложений [Jefferson 1984; Li 2001; Truong 2013; Konakahara 2015]. К

несоревновательным (кооперативным) наложениям из дискурсивных типов, представленных в разделе 3, ближе типы «Побочная реплика» и «Передача хода», к соревновательным (конкурентным) — типы «Перехват хода» и «Конкурентное развитие». В связи с этим из приведенных выше результатов можно сделать вывод о том, что с повышением уровня конкурентности возрастает как длительность эпизода наложения (см. табл. 1), так и вероятность обрыва при его урегулировании (табл. 3). Иными словами, более конкурентные наложения в целом оказываются и более трудными для урегулирования; ср. наблюдение Э. Щеглоффа о том, что только наложения, имеющие в себе соревновательный компонент, представляют для участников диалога проблему, требующую специальных механизмов для разрешения [Schegloff 2000: 4-7].

Во-вторых, удается проследить, какое влияние на поведение участников при наложении оказывает соотношение их эпистемических статусов, роль которых в диалоге неоднократно освещалась в литературе [Heritage 2012; Mondada 2013; Morek 2015]. Согласно полученным результатам, тут обнаруживается генеральная тенденция: доля эпизодов, в которой участники обладают равным эпистемическим статусом, последовательно возрастает при повышении степени конкурентности наложения (см. нижние сегменты столбиков диаграммы на рис. 4). Это наблюдение понятным образом согласуется с тем фактом, что именно наложения, в которых задействованы Рассказчик и Комментатор, оказываются и наиболее длительными (табл. 2). В целом, более конкурентный характер наложений при равном эпистемическом статусе можно назвать ожидаемым: см., в частности, замечание [Stivers et al. 2011] о том, что отсутствие очевидной иерархии эпистемических статусов может приводить к конкуренции за право считаться более «надежным» говорящим.

В свою очередь, для эпизодов наложения с неравными эпистемическими статусами участников характерны следующие предпочтения:

- Участник с более высоким эпистемическим статусом значимо чаще выступает инициатором наложения в типе «Передача хода» и практически никогда напрямую не перехватывает ход.
- Автором «Побочной реплики» в подавляющем большинстве случаев выступает участник с более низким эпистемическим статусом (Пересказчик). Он же достаточно часто осуществляет прямой перехват хода.

Напрашивающаяся интерпретация тут состоит в том, что, хотя Пересказчики и обладают в записях корпуса более низким эпистемическим статусом, они могут пользоваться преимущественным правом определения направления разговора — поскольку именно необходимость получения Пересказчиком полной информации о содержании фильма является основной движущей силой диалогических этапов записей. Как именно устроено взаимодействие этих двух параметров (эпистемического статуса и полномочий в определении направления беседы), еще только предстоит выяснить.

6 Заключение

В докладе представлен анализ эпизодов наложения в диалогических этапах записей корпуса «Рассказы и разговоры о грушах» (RUPEX). Рассмотрено четыре параметра: тип дискурсивного контекста, продолжительность наложения, наличие обрыва при урегулировании наложения и соотношение эпистемических статусов участников. Показано, что при снижении общей кооперативности в поведении участников повышается продолжительность эпизодов наложения и вероятность возникновения дискурсивных обрывов. Также продемонстрирована связь между степенью кооперативности при возникновении и урегулировании наложения и соотношением эпистемических статусов.

У проведенного исследования есть ряд ограничений. Укажу на два из них. Во-первых, на результаты анализа могут влиять особенности материала. Поведение участников записей корпуса частично программируется внешними факторами: на них закреплена записывающая аппаратура, которая ограничивает их перемещения (им нужно сидеть); им объясняют порядок следования этапов и просят строго его придерживаться [Kibrik, Fedorova 2018]. Хотя в целом участники ведут себя при этом достаточно естественно, все же такие условия отличаются от условий спонтанного разговора. Кроме того, значимым представляется и количество участников: можно ожидать, что

и при уменьшении (с трех до двух), и при увеличении числа говорящих может измениться как общая частотность эпизодов наложений, так и характер их распределения по типам дискурсивных контекстов.

Во-вторых, хотя при разборе примеров я в ряде случаев анализирую невокальные действия участников, задачи исчерпывающего описания мультиканального поведения коммуникантов в ситуации наложения не ставилось. Вместе с тем в литературе уже накоплен немалый массив сведений о том, какую роль в ситуации наложения играют невокальные ресурсы: мануальные и мимические жесты, управление зрительным контактом и проч. [Oloff 2013; Zima et al. 2019; Malabarba et al. 2022]. Данные корпуса RUPEX позволяют осуществить подобный анализ, что может стать одним из продолжений настоящего исследования.

Благодарности

Исследование выполнено в рамках проекта Минобрнауки «Социопрагматические факторы адаптации вербального и кинетического поведения: русский звучащий и русский жестовый язык».

Литература

[1] Boersma Paul, Weenink David. Praat: Doing phonetics by computer [Computer software]. Version 6.3.13 retrieved 31 July 2023 from http://www.praat.org/.
[2] Brown Penelope, Levinson Stephen C. Politeness: Some universals in language usage. — Cambridge: Cambridge University Press, 1987. https://doi.org/10.1017/CBO9780511813085
[3] Chafe Wallace (ed.). The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production. — Norwood, NJ: Ablex, 1980. https://doi.org/10.1017/S0047404500008897
[4] Chekmez Ulyana. The functioning of overlays in the spoken dialogic discourse [Specifika funkcionirovanija naloženij v ustnom dialogičeskom diskurse] // Modern Studies of Social Issues [Sovremennye issledovanija social'nyx problem]. — 2019. — Vol. 11, no. 4-2. — P. 130–139.
[5] Clark Herbert H. Using language. — Cambridge: Cambridge University Press, 1996. https://doi.org/10.1017/CBO9780511620539
[6] Du Bois John W. The Stance Triangle // Robert Englebretson (ed.) Stancetaking in discourse: Subjectivity, evaluation, interaction. — Amsterdam/Philadelphia: John Benjamins, 2007. — P. 139–182. https://doi.org/10.1075/pbns.164.07du
[7] Fedorova Olga V. Oculomotor everyday communication: How to pick a good metric // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2021). Issue 20. — Moscow: RGGU, 2021. — P. 213–226. https://doi.org/10.28995/2075-7182-2021-20-213-226
[8] Hayashi Makoto. Turn allocation and turn sharing // Jack Sidnell, Tanya Stivers (eds.) The Handbook of Conversation Analysis. — Wiley-Blackwell, 2012. — P. 167–190. https://doi.org/10.1002/9781118325001.ch9
[9] Heldner Mattias, Edlund Jens. Pauses, gaps and overlaps in conversation // Journal of Phonetics. — 2010. — Vol. 38(4). — P. 555-568. https://doi.org/10.1016/j.wocn.2010.08.002
[10] Heritage John. Epistemics in action: Action formation and territories of knowledge // Research on Language & Social Interaction. — 2012. — Vol. 45(1). — P. 1–29. https://doi.org/10.1080/08351813.2012.646684
[11] Jefferson Gail. Notes on some orderlinesses of overlap onset // Valentina D'Urso, Paolo Leonardi (eds.) Discourse analysis and natural rhetoric. — Padua: Cleup Editors, 1984. — P. 11–38.
[12] Jefferson Gail. A sketch of some orderly aspects of overlap in natural conversation // Gene H. Lerner (ed.) Conversation analysis: Studies from the first generation. — Amsterdam & Philadelphia: John Benjamins, 2004. — P. 43–59. https://doi.org/10.1075/pbns.125.05jef
[13] Kibrik Andrej A., Fedorova Olga V. An empirical study of multichannel communication: Russian Pear Chats and Stories // Psychology. Journal of the Higher School of Economics [Psixologija. Žurnal Vysšej Školy ėkonomiki]. — 2018. — Vol. 15 (2). — P. 191–200. https://doi.org/10.17323/1813-8918-2018-2-191-200
[14] Kibrik Andrej A., Korotaev Nikolay A., Podlesskaya Vera I. Russian spoken discourse: Local structure and prosody // Shlomo Izre'el, Heliana Mello, Alessandro Panunzi, and Tommaso Raso (eds.) In Search of Basic Units of Spoken Language: A Corpus-Driven Approach. — Amsterdam: John Benjamins, 2020. — P. 35–72. https://doi.org/10.1075/scl.94.01kib
[15] Kibrik Andrej A., Podlesskaya Vera I. Parenthetical constructions in spoken discourse [Vstavočnye konstrukcii v ustnom diskurse] // V. Z. Demyankov, V. Ja. Porxomovskij (eds.) In language and culture: Sound, sign, meaning. For the 70th birthday of Viktor A. Vinogradov [V prostranstve jazyka i kul'tury: Zvuk, znak, smysl. Sbornik statej v čest' 70-letija V. A. Vinogradova]. — Moscow, LSC, 2010. — P. 87–99.
[16] Konakahara Mayu. An analysis overlapping questions in casual ELF conversation: cooperative or competitive contribution // Journal of Pragmatics. — 2015. — Vol. 84. — P. 37–53. https://doi.org/10.1016/j.pragma.2015.04.014

[17] Korotaev Nikolay. Parenthetical constructions in Russian spoken discourse: Basic types and prosodic features [Konstukcii s diskursivnymi vstavkami v ustnoj russkoj reči: bazovye tipy i prosodičeskie svojstva] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2021). Issue 20. — Moscow: RGGU, 2021. — P. 413–424. https://doi.org/10.28995/2075-7182-2021-20-413-424
[18] Li Han Z. Cooperative and intrusive interruptions in inter- and intracultural dyadic discourse // Journal of Language and Social Psychology. — 2001. — Vol. 20(3). — P. 259–284. https://doi.org/10.1177/0261927X01020003001
[19] Litvinenko Alla O., Kibrik Andrej A., Nikolaeva Yulia V., Fedorova Olga V. Annotating hand movements in multichannel discourse: Gestures, adaptors and manual postures // The Russian Journal of Cognitive Science [Rossijskij žurnal kognitivnoj nauki]. — 2018. — Vol. 5 (2). — P. 4–17.
[20] Malabara Taiane, Mendes Anna C. Oliveira, de Souza Joseane. Multimodal resolution of overlapping talk in video-mediated L2 instruction // Languages. — 2022. — Vol. 7(2). — 154. https://doi.org/10.3390/languages7020154
[21] Mondada Lorenza. Displaying, contesting and negotiating epistemic authority in social interaction: Descriptions and questions in guided visits // Discourse Studies. — 2013. — Vol. 15(5). P. 597–626. https://doi.org/10.1177/1461445613501577
[22] Morek Miriam. Show that you know – Explanations, interactional identities and epistemic stance-taking in family talk and peer talk // Linguistics and Education. — 2015. — Vol. 31. — P. 238–259. https://doi.org/10.1016/j.linged.2014.10.004
[23] Oloff Florence. Embodied withdrawal after overlap resolution // Journal of Pragmatics. — 2013. — Vol. 46(1). — P. 139–156. https://doi.org/10.1016/j.pragma.2012.07.005
[24] Ono Tsuyoshi, Thompson Sandra A. Interaction and syntax in the structure of conversational discourse: collaboration, overlap, and syntactic dissociation // E. Hovy, D. Scott (eds) Computational and Conversational Discourse. — Berlin: Springer, 1996. — P. 67–96. https://doi.org/10.1007/978-3-662-03293-0_3
[25] Pomerantz Anita, Heritage John. Preference // Jack Sidnell, Tanya Stivers (eds.) The Handbook of Conversation Analysis. — Wiley-Blackwell, 2012. — P. 210–228. https://doi.org/10.1002/9781118325001.ch11
[26] Sacks Harvey, Schegloff Emanuel A., Jefferson Gail. A simplest systematics for the organization of turn-taking for conversation // Language. — 1974. — Vol. 50 (4). — P. 696–735. https://doi.org/10.2307/412243
[27] Santner-Wolfartsberger Anita. Parties, persons, and one-at-a-time: conversation analysis and ELF // Journal of English as a Lingua Franca. — Vol. 4(2). — P. 253–282. https://doi.org/10.1515/jelf-2015-0020
[28] Schegloff Emanuel A. Parties and talking together: Two ways in which numbers are significant for talk-in-interaction // Paul ten Have, George Psathas (eds.) Situated order: Studies in social organization and embodied activities. — Boston: University Press of America, 1995. — P. 31–42.
[29] Schegloff Emanuel A. Overlapping talk and the organization of turn-taking for conversation // Language in Society. — 2000. — Vol. 29. — P. 1–63. https://doi.org/10.1017/S0047404500001019
[30] Stivers Tanya, Mondada Lorenza, Steensig Jakob. Knowledge, morality and affiliation in social interaction // Tanya Stivers, Lorenza Mondada, Jakob Steensig (eds) The morality of knowledge in conversation. — Cambridge: Cambridge University Press, 2011. — P. 3–24. https://doi.org/10.1017/CBO9780511921674.002
[31] Stolt Minna. The many faces of overlap. Non-competitive overlap in a conversation between Finnish and British speakers of English. A Pro Gradu Thesis in English. — University of Jyväskylä, 2008.
[32] Truong Khiet. Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee // Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. — 2013. — P. 1404–1408. https://doi.org/10.21437/Interspeech.2013-368
[33] Yamashita Natsuo, Horiguchi Shota, Homma Takeshi. Improving the naturalness of simulated conversations for end-to-end neural diarization // The Speaker and Language Recognition Workshop (Odyssey 2022). — 2022. — http://dx.doi.org/10.21437/Odyssey.2022-19
[34] Zhang Leying, Qian Yao, Zhou Long, Liu Shujie, Wang Dongmei, Wang Xiaofei, Yousefi Midia, Qian Yanmin, Li Jinyu, He Lei, Zhao Sheng, Zeng Michael. CoVoMix: Advancing zero-shot speech generation for human-like multi-talker conversations // 38th Conference on Neural Information Processing Systems (NeurIPS 2024). — 2024. https://doi.org/10.48550/arXiv.2404.06690
[35] Zima Elisabeth, Weiß Clarissa, Brône Geert. Gaze and overlap resolution in triadic interactions // Journal of Pragmatics. — 2019. — Vol. 140. — P. 49–69. https://doi.org/10.1016/j.pragma.2018.11.019

The development of a natural language reasoning system for a companion robot

Artemiy A. Kotov
NRC Kurchatov Institute, Moscow,
pl. Kurchatova 1
MSLU, Moscow,
Ostozhenka str., 38
kotov@harpia.ru

Zakhar A. Nosovets
NRC Kurchatov Institute, Moscow,
pl. Kurchatova 1
MSLU, Moscow,
Ostozhenka str., 38
nosovets@phystech.edu

Alexander A. Filatov
Yandex.Technologies LLC,
Moscow, Lev Tolstoy St. 16
filatov.alex@gmail.com

Nikita A. Arinkin
NRC Kurchatov Institute, Moscow,
pl. Kurchatova 1
MSLU, Moscow,
Ostozhenka str., 38
nikita.arinkin@gmail.com

Abstract

The natural language inference system allows the robot to go from the meaning of the incoming text (from a visual or tactile stimulus) to the derived meaning—the inference. This system uses a rule-based parser, and pairs of semantic representations constructed by the parser are combined into a scenario. In this paper, we represent the robot's natural language inference space as a graph, where the robot can move from the premise of a scenario to the consequence, and from the consequence of one scenario to the premise of another. We involve annotators who propose derived sentences (semantic components) for a given premise and, in the annotator interface, immediately evaluate the proximity of the proposed sentence to the available scenarios. This procedure allows us to develop the graph of scenarios, to evaluate its connectivity and the absence of dead ends (deadlock vertices), as well as the adequacy of the analysis of incoming texts by scenarios within this graph. The graph contains 5,000 scenarios and approximately 22,000 nodes. We estimate that a graph consisting of 7,000 scenarios can be sufficient for modelling the mechanism of human natural language reasoning.

Keywords: Companion robot, emotional robotics, cognitive models, dialog support, human-machine interaction
DOI: 10.28995/2075-7182-2025-23-183-194

Развитие системы естественно-языковых рассуждений для робота-компаньона

Артемий Котов
НИЦ «Курчатовский институт»,
Москва, пл. Курчатова, 1
МГЛУ, Москва, ул. Остоженка, 38
kotov@harpia.ru

Захар Носовец
НИЦ «Курчатовский институт»,
Москва, пл. Курчатова, 1
МГЛУ, Москва, ул. Остоженка, 38
nosovets@phystech.edu

Александр Филатов
ООО «Яндекс.Технологии»,
Москва, ул. Льва Толстого, д. 16
filatov.alex@gmail.com

Никита Аринкин
НИЦ «Курчатовский институт»,
Москва, пл. Курчатова, 1
МГЛУ, Москва, ул. Остоженка, 38
nikita.arinkin@gmail.com

Аннотация

Система естественно-языкового вывода позволяет роботу перейти от смысла поступившего текста (от визуального или тактильного стимула) к производному смыслу — выводу. Данная система использует парсер, основанный на правилах, а пары семантических представлений, построенных парсером, объединяются в сценарий. В работе мы представляем пространство естественно-языкового вывода робота как граф, где робот может перейти от посылки сценария к его следствию, а также от следствия одного сценария — к посылке другого. Мы привлекаем разметчиков, которые предлагают производные предложения (компоненты смысла) для некоторой посылки и в интерфейсе разметчиков сразу же оценивают близость предложенного предложения к имеющимся сценариям. Эта процедура позволяет нам развивать граф сценариев, оценивать его связность и отсутствие в нём тупиков (висячих вершин), а также адекватность анализа входящих текстов с помощью сценариев в составе данного графа. Граф содержит 5000 сценариев и около 22 000 узлов. По нашим оценкам, граф, состоящий из 7000 сценариев, может считаться достаточным для моделирования механизма естественно-языкового рассуждения человека.

Ключевые слова: робот-компаньон, эмоциональная робототехника, когнитивные модели, поддержание диалога, человеко-машинное взаимодействие

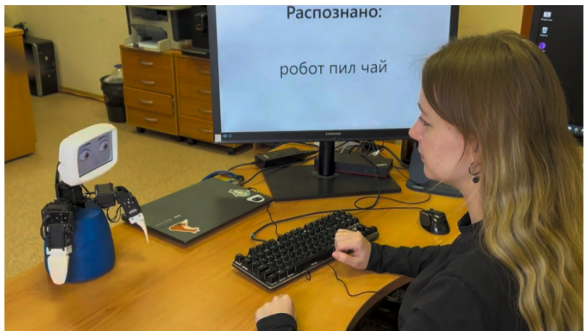
1 Introduction

Despite the significant progress of ‘large language models’ (LLMs) in the task of text generation [1], these models are relatively poorly applied to the modelling of reasoning [2]. At their core, LLMs aim to reconstruct the next word of a text based on a context window, so LLMs are usually improved by expanding the context window and by enlarging the training text database. Classical artificial intelligence models oriented towards problem solving or text understanding had a different architecture: they used an enumeration of paths in the problem space [3] or in the space of storylines [4]. To overcome these limitations of LLMs and to create the so-called *Large Reasoning Models* (LRM), *Chain-of-Thought* methods are proposed, where the solution of the problem is performed in several steps: each step is provided by a LLM, but the results of each step can be added to the solution protocol or combined with each other for the next step of the LLM operation [5].

At the same time, there is a relevant task to design a reasoning process, where intermediate representations are connected by a more compact mechanism that does not use LLMs. In classical models such as SOAR [3,4] and CogAff [5], such a transition between two representations has been described in terms of *productions* or *scenarios*. An individual *scenario* relates a typical stimulus (*premise*) to a possible *consequence* and allows the cognitive model (robot) to move in its reasoning process from the current situation to the inferred situation, thus reconstructing the context of the received message or making plans for actions in the environment. The aim of our work is to create a companion robot and to develop a cognitive model for the robot that allows it to mimic various cognitive functions when interacting with humans. These are primarily planning and discussing of actions in the real world [6], as well as the robot’s ability for natural language inference. An example of interaction with the robot using natural language inference is shown in Figure 1.

Following the CogAff model, the robot constructs possible causes and effects from an incoming event[1], and then evaluates these derived meanings according to their emotionality (significance). In future versions of the system, this should allow the agent to plan its communicative or actual activity in such a way that it approaches positive situations and avoids or compensates for negative situations, e.g., to evaluate the situation ‘robot is well fed’ as positive and ‘robot was sick’ as negative. In the current version (Figure 1), the robot reports all the possible causes and effects in the dialogue based on the incoming utterance, without considering their significance. In this paper we present a scenario architecture for the natural language inference mechanism for companion robots.

[1] For the purposes of this paper, text messages are considered as incoming events, but these events can also be visual or tactile, which are constructed by the vision system or the touch registration system, respectively.



(a) Пользователь: *Робот пил чай.*
(User: *The robot was drinking tea.*)
Робот: *Робот пил чай. Робот налил чай в чашку. Робот заварил чай. Робот вскипятил термос. Робот насытился.*
(Robot: *The robot was drinking tea. The robot poured tea into the cup. The robot made the tea. The robot boiled the thermos.*[2] *The robot is well fed.*)
(b) Пользователь: *Робот пил водку.*
(User: *The robot was drinking vodka.*)
Робот: *Робот пил водку. Кто-то типа угостить робота водкой. Робот купил водку. Робот налил водку к стопке. Робот опьянел. Робот отравился. Тошнило робота. Робот бухал.*
(Robot: *The robot was drinking vodka. Someone like to give vodka to the robot. The robot bought vodka. The robot poured vodka by the shot of glass. The robot got drunk. The robot got poisoned. The robot was sick. The robot was drinking.*)

Figure 1: Interaction with the companion robot controlled by the natural language inference system. The video is available in the repository [7].

2 Semantic parser

The basic data block for the scenario system is the semantic representation. The semantic representation of an individual sentence clause contains the valencies for the actants: agent – *ag*, patient – *pat*, instrument – *instr* etc.[3], as well as the semantic markers for each actant. This semantic representation for an incoming utterance is built by the parser [8,9]. For visual and tactile events, the semantic representations are similar to the text semantics – but without homonymy (are unambiguous) [10,11].

The semantic parser follows the three-level linguistic architecture of Meaning-to-Text Theory [12] and consists of morphological, syntactic and semantic components. **The morphological component** includes a vocabulary of 100,000 lexemes (about 1.5 million word forms) based on the OpenCorpora dictionary [13]. The dictionary lexemes are annotated with semantic markers. During the text analysis, the markers of each word are copied to the semantic valency that the word occupies in the syntactic tree. The markers of the verb are copied to the valency *p* (predicate). In **the syntactic component**, we use a left-to-right parser. Segments of a sentence (words) are sequentially placed on a stack for syntactic parsing, with a morphological annotation assigned to each segment. For each homonymic word form, the stack is duplicated: each copy of the stack contains a different variant of parsing for the homonymic segment. The syntactic component compares the top of the stack with the Russian language grammar, which consists of approximately 800 syntXML rules [14]. If the sequence of segments on the stack meets the requirements of a syntactic rule, then a syntactic link is established between these words. If the syntactic relation can be established ambiguously (in the case of syntactic homonymy), the stack is also duplicated, and a different syntactic relation is established in each copy of the stack. By the end of the sentence, a stack with successful parsing should collapse into a syntactic tree with a single root. The parser then builds a semantic representation for each successful stack (Figure 2). The simultaneous processing of multiple stacks[4] outputs a set of syntax trees from the syntax component, and then each semantic predication is matched to the scenarios: i.e. the most regular semantics for the sentence is selected

[2] Semantic or syntactic errors in utterance synthesis in Russian are underlined.
[3] An inventory of 24 valency types is used, based on [21].
[4] The number of homonymous stacks is usually limited to 256 or 512 units. On each step the stacks are evaluated and only a limited number of "best" stacks is preserved for further analysis.

as better matching to the regular situations – scenarios. The parallel processing of speech ambiguity also allows us to approach the problems of ambiguous representations, such as computer humor, computer imagination and the theory of mind [15].

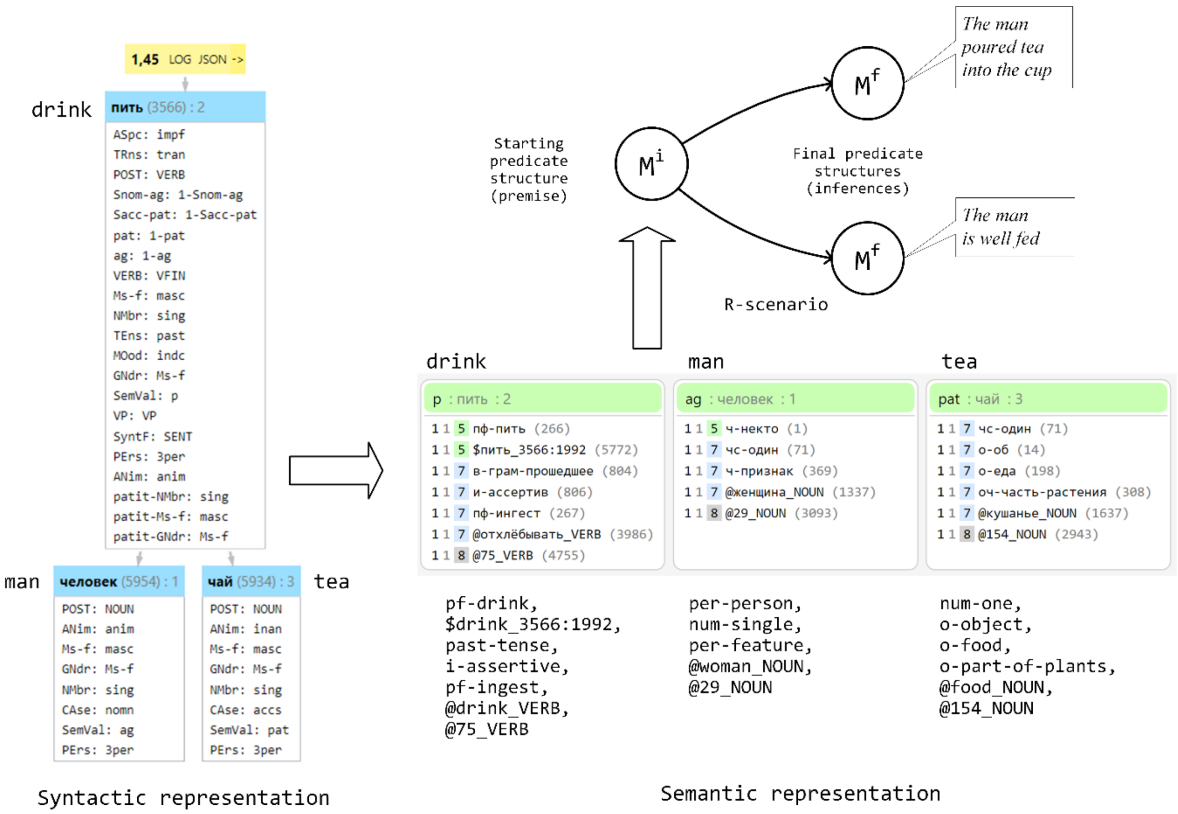


Figure 2: Syntactic tree and semantic representation for the utterance *a man drinks tea*, and its comparison with the premise (initial predicate structure) of the closest scenario. M^i is the initial predicate structure, M^f are the final predicate structures of the scenario.

As can be seen in Figure 2, the semantic representation is a set of valencies (*p, ag, pat*), where each valency is filled with markers of the corresponding lexeme. Unlike *word2vec* semantic vector models, we use markers from a fixed inventory (about 13,000 items), with an average of four markers[5] assigned to individual words in the dictionary. We use several methods to construct the marker inventory: (a) markers based on the categories of Shvedova's Semantic Dictionary, e.g. *per-person* [16], (b) markers based on emotional markup [17], (c) markers based on clustering of *word2vec* vectors, e.g., *@drink_VERB* and *@food_NOUN*[6], and (d) markers based on synsets, e.g., *$drink_3566:1992*[7]. A word may be homonymous, in which case the word has different meanings 1:1, 1:2, 1:3, ..., 2:1, 2:2, 2:3, etc. Different semantic markers can be assigned to each of the meanings[8]. The same feature may have different weight (rank)[9] for different words.

In order to determine the correspondence between the incoming semantic representation and the scenario premise, a proximity measure is calculated that takes into account the coincidence of markers in

[5] For 83,000 lexemes from the dictionary, 318,000 links were established with markers from a dictionary of 8,500 features. Thus, each word has 3.91 features.

[6] Marker names are chosen automatically by the closest word to the cluster center. So, *a man* corresponds to the class *@woman_NOUN*. Here the marker name represents a member of the category, not a hypernym, and should be interpreted as 'a woman and similar entities'. This procedure is described in more detail in [8].

[7] This marker is assigned to all words of the synset *drink / sip* (Russian synonyms: *пить, выпивать, потягивать, лакать, хлестать, тянуть, попивать*). If a word is homonymic and has several meanings, the marker is assigned to the meaning of the word that is included in the given synset.

[8] In Figure 2, all lexemes are unambiguous and have only meaning 1:1.

[9] For example, the feature *per-person* ('person') has more weight for the noun a *man* than for the noun a *tailor*.

the valencies. Each valency of a scenario premise contains a set of semantic markers annotated according to their importance:

- $marker_value$ — the weight of the marker in the scenario premise (the higher the value, the more important is the marker).
- $attr$ — attribute of the marker in the incoming representation (the smaller the value, the higher is the semantic *value*).

For each valency, the degree of matching between the markers of the premise P and the input representation C is calculated using a modified Jaccard measure (here $Q = P \cap C$):

$$v(P,Q) = \frac{\sum_Q m}{\sum_{P \setminus Q} a + \sum_Q n}$$

where v lies in the range [0, 1]. The following normalising factors are used here. For matching markers:

$$m = \frac{marker_value}{\ln(attr + e)}$$

$$n = marker_value$$

where e is the base of the natural logarithm. For non-matching markers:

$$a = \frac{marker_value}{\ln(attr_avg + e)}$$

where $attr_avg$ is the average attribute value for the corpus.

To calculate the final degree of closeness of the semantic representations, a Jaccard measure modified as follows is used:

$$p(P,Q) = \frac{\sum_Q wv(P,Q)}{\sum_P w}$$

where p lies in the range [0, 1]. Here P is the valency set in the premise, Q is the valency set present in both representations, w are valency weights representing their contribution to the scenario.

3 Reasoning system

The robot's reasoning system is based on scenarios. Each scenario contains semantic representations of two types:

(a) premise or *initial predicate structure*;

(b) consequence or *final predicate structure* (there can be from none to several structures).

If the meaning of the incoming text is close to the initial predicate structure, the scenario is triggered and builds its final predicate structures. It is this mechanism that ensures the construction of the inferred semantic representation '[previously] the robot boiled the kettle' from the incoming utterance *The robot drinks tea*. This 'rational' inference is provided by a system of *r-scenarios* (rational scenarios). Initially, the set of r-scenarios was formed in the following way: semantic representations of diverse natural texts (10 million clauses) were clustered, and after manual checking, about 3,000 r-scenarios (their initial predicate structures) were formed for the approved clusters. Currently, the number of r-scenarios has been brought to 5,300. Each scenario recognizes some typical situation in texts, for example: 'a person's heart is beating', 'a person has a toothache', 'a person drinks tea', etc. From this point of view, r-scenarios are similar to FrameNet's frame inventory [18] or Framebank government models [19]. The difference in our project is that, when the parser is running, the scenarios are automatically recognized in the incoming texts: the closest scenario is assigned to each clause. For example, the @sick_292_VERB_1 scenario (~'the man's tooth hurts') is automatically detected in the utterances:

- *I tried to lift her up, but she couldn't, she said, 'my back hurts.'* (proximity 0.2820)[10]
- *— and my throat hurts too.* (proximity 0.2869)
- *Svetlichnaya has been very seriously ill in recent years. (proximity 0.2120)*

Attributing this scenario to statements such as *Roma, we are all rooting for you and our national team!*[11] is incorrect, because this situation has different causes, consequences and emotional evaluations. Usually, such a misclassification indicates that the scenario system lacks a specific scenario responsible for recognizing this particular meaning 'to root for'.

For emotional processing of incoming meaning, an inventory of d-scenarios (*dominant scenarios*) is used, based on the speech influence model [17]. D-scenarios recognize emotional situations, such as 'nobody needs me', 'something is being plotted against me', 'I feel nauseous', or 'everyone looks at me with admiration', 'I am smarter than other robots', 'I am satiated'.

According to the CogAff architecture, an incoming event can be evaluated by the agent both emotionally and rationally. This is ensured by the competition between d- and r-scenarios. Furthermore, according to CogAff, the result of the agent's reasoning can serve as the material for the next reasoning step or as an object of emotional evaluation. In this case, the secondary emotional mechanism is activated [20]: the emotion is not triggered by the perceived stimulus, but by the reasoning result. To implement this mechanism, the final predicate structures of r-scenarios are passed to the input of the model for the next analysis step (see Figure 3).

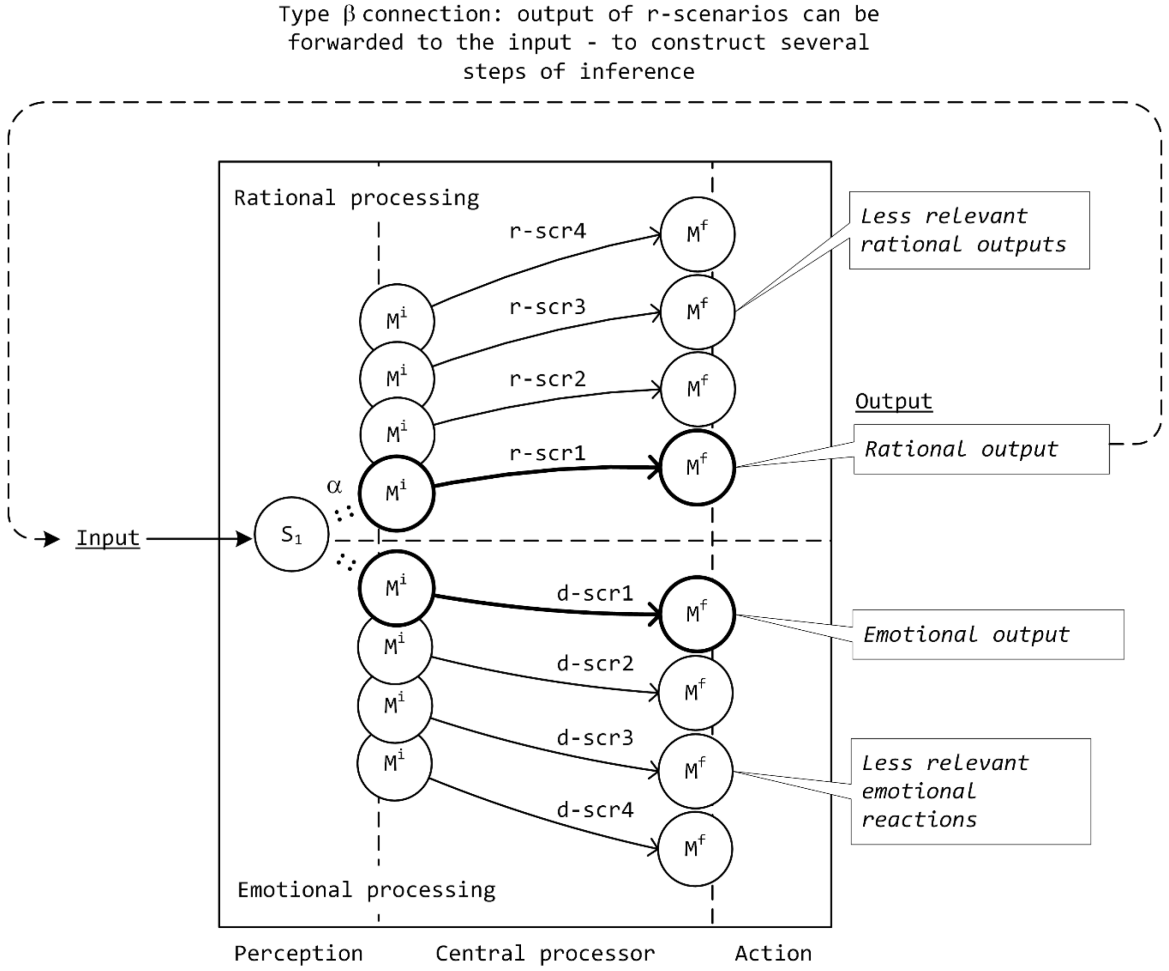
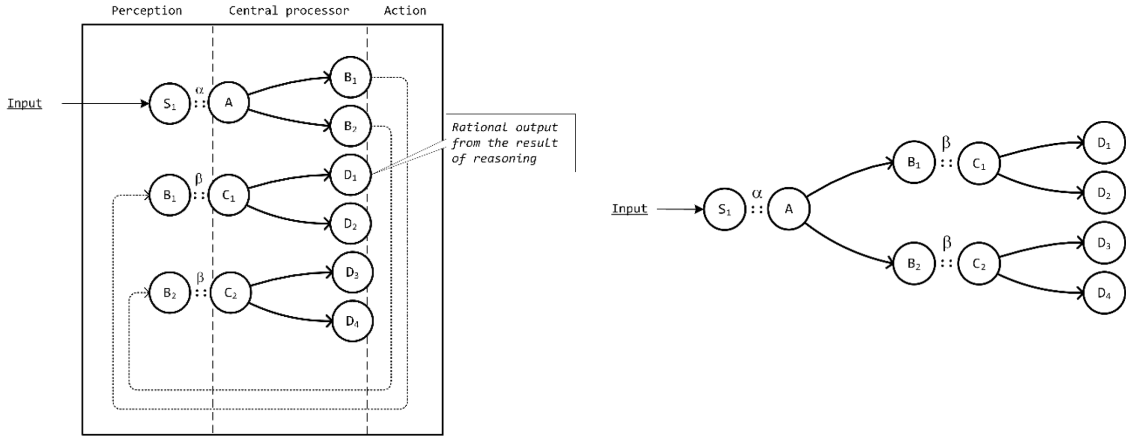


Figure 3: Processing of stimulus S_1 by the scenario system. R-scr – r-scenarios, d-scr – d-scenarios, M^i – initial predicate structures, M^f – final predicate structures.

[10] As demonstrated further – the level 0.25 is considered as a *good* proximity, so values above 0.25 are *excellent*.
[11] In Russian *болеть* has at least two meanings – 'to be sick' and 'to root for someone'.

To obtain the final predicate structures for each r-scenario, i.e. semantic representations related to the initial one, annotators were employed. We offered annotators prototypical (initial) utterances for each r-scenario and asked them to write down new utterances denoting: (a) the antecedent situations that led to the initial one – predicate structures of the type BEFORE, (b) the consequent situations caused by the initial one (AFTER), and (c) interpretations of the initial situation (INTERP) – i.e. syntactic conversions ('a house is built by workers' = 'workers build a house') and figurative interpretations (e.g. 'a man was running on the field' = 'a man was playing football'). The annotators' utterances in the annotation environment were immediately parsed by the parser, which established co-referentiality[12] between the actants in the initial and final predicate structures (the same person who was running on the field was playing football). The conference information was also added to the scenario, allowing the robot to mention in derived utterances the same referents that were present in the original stimulus.

When receiving new utterances from the annotators, it is important to check whether these utterances correspond to r-scenarios already present in the scenario apparatus. With a closed set of scenarios, each output result corresponds to some 'input' – the initial predicate structure of another scenario (Figure 4), which should allow the robot to move in its reasoning along some graph (Figure 5).



(a) (b)

Figure 4: (a) stimulus S_1 is matched to the initial predicate structure A of a scenario. The constructed inference results B_1 and B_2 can be mapped to the initial predicate structures C_1 and C_2, which can lead to inferences D_1, D_2, D_3, D_4 on the next reasoning step; (b) the same reasoning mechanism can be represented as a graph induced by the initial stimulus S_1.

This approach allows us to construct a general graph – the robot's natural language inference space – in which all the inferred semantic representations (predicate structures) are reconstructed for a given initial stimulus. The nodes of this graph correspond to the initial and final predicate structures, while the edges correspond (a) to the links between the initial and final structures within the same scenario and (b) to the links between different scenarios, calculated by the proximity of the final structures to the initial structures. For example, Figure 5 shows the graph constructed from the stimulus *man was drinking tea* at depth 3. The Figure 3 shows that an inference from r-scenarios can be matched with d-scenarios: the robot can infer 'previously someone made tea' (and *It's nice that someone cares about the man!*) and 'someone poured tea into a cup' (and *It's nice that someone brings tea to the man!*) – inferred semantic representations here are constructed by r-scenarios, and emotional evaluations – by d-scenarios.

[12] The annotators were instructed to repeat in their utterances the lexemes used in the initial utterance. If for the initial utterance *man drank vodka* the annotator suggests the interpretation *man was intoxicated*, the parser establishes co-reference between the two occurrences of the word *man*. If the incoming utterance *robot drank vodka* is matched to this scenario, the parser will match the referent 'robot' with the valency 'man' and construct the output *the robot was intoxicated*.

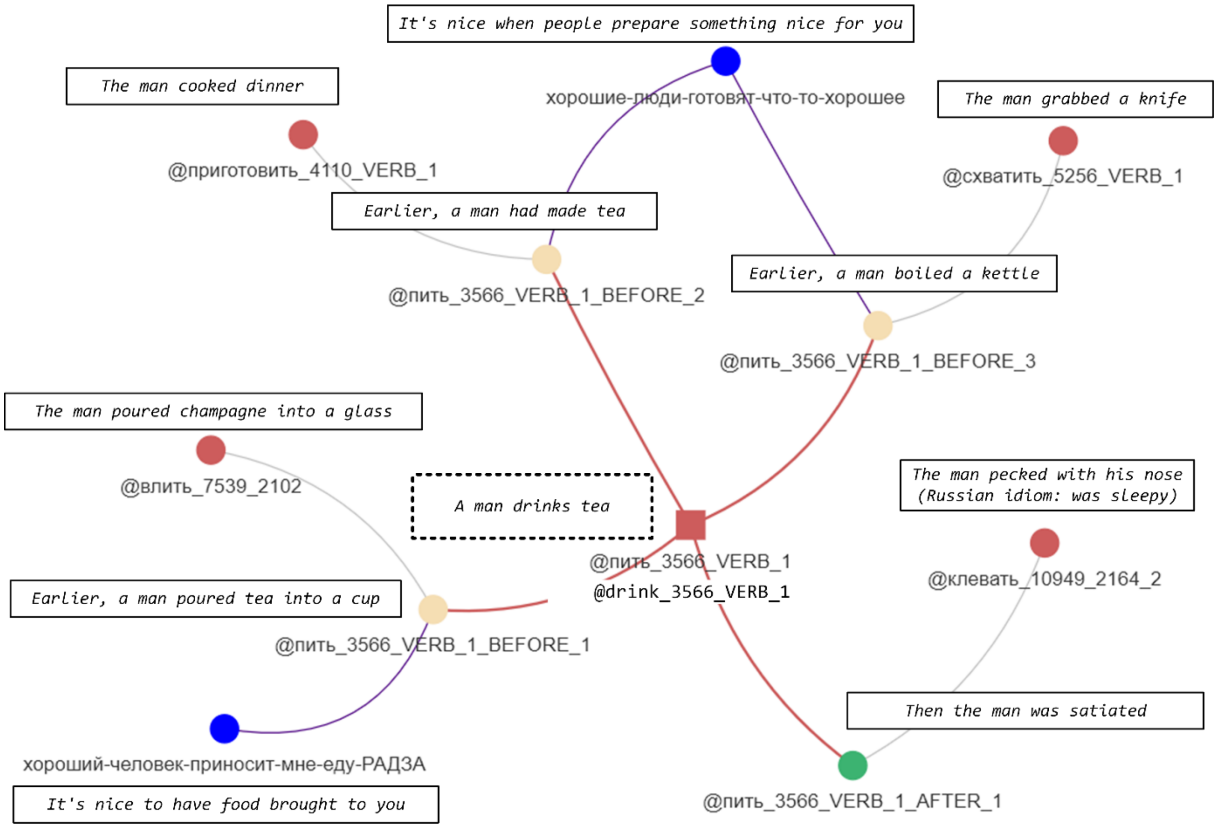


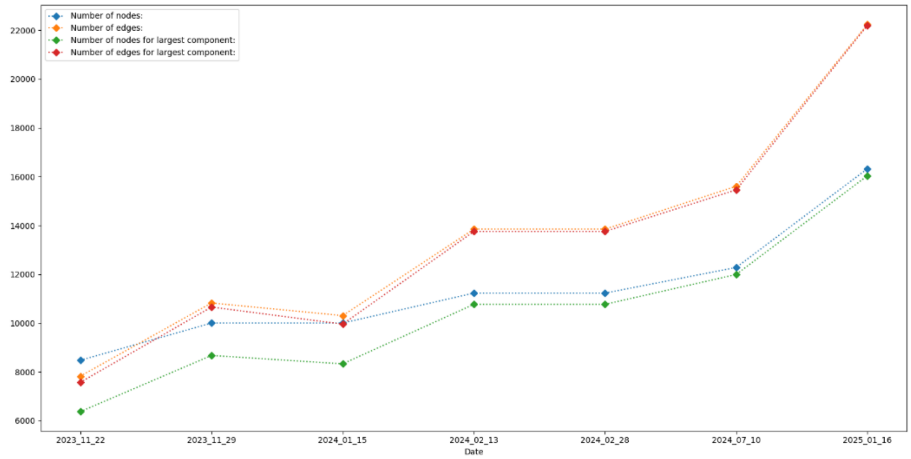
Figure 5: Example of a reasoning subgraph for the initial predicative structure ‘the man drank tea’ (scenario @drink_3566_VERB_1). R-scenarios are marked with “@”, d-scenarios are not. The implications ‘the man was satiated’ and ‘the man boiled the kettle’ are misclassified by the scenarios, which is an argument for creating a separate scenario for these representations when filling the inventory.

The work of the annotators in adding to the scenario inventory should produce such a set of scenarios (such a natural language inference space) that fulfils a number of criteria. First, the inventory of scenarios should be accurate and complete when processing real texts – each clause of the text should be correctly matched to some scenario: proximity of the type α in Figures 3 and 4. This will allow the robot to correctly classify the incoming stimulus and correctly select possible causes and effects for it. Second, the inventory of scenarios should be closed: each final predicate structure should map to some initial predicate structure: proximity of the type β in Figures 3 and 4. Then the robot will be free to develop reasoning without stopping at representations that have no counterparts among the initial predicate structures of the scenarios.

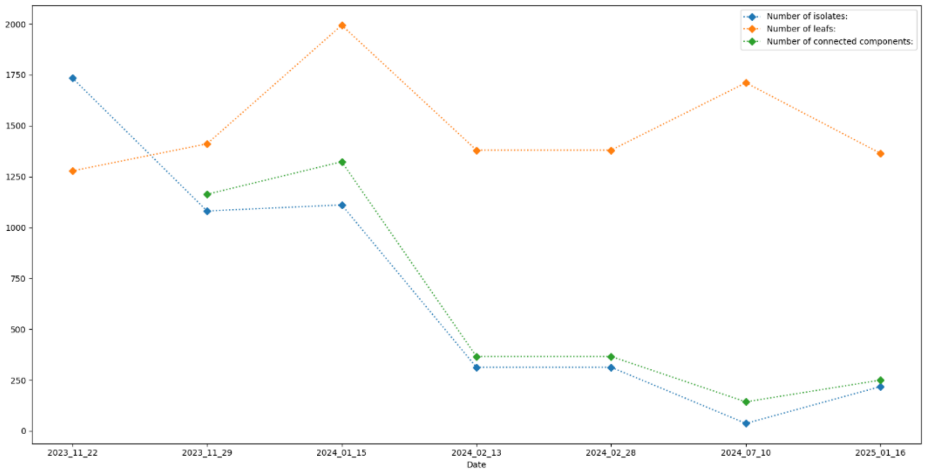
In order to come close to meeting these requirements, during the annotators’ work the parser immediately parsed each proposed utterance, compared it with the existing set of scenarios and returned the constructed semantic representation with the closest scenarios. The annotator then had to evaluate whether the utterance entered matched the closest scenarios suggested by the parser. Thus, the proximity evaluation of β was collected directly during the work of annotators. If the annotator agreed with the proposed scenario for the entered utterance, it can be assumed that the β scenario graph has connectivity: the robot can freely move in its reasoning from the final predicate structure proposed by the annotator to the initial predicate structure of another scenario. If the annotator finds that none of the three closest scenarios matches his utterance, then the β proximity requirement is not fulfilled: among the available scenarios there is no one that correctly classifies the input text, and this is an argument for creating a new scenario. For example, if the parser offered the best scenario ‘the person had a toothache’ for the annotator’s sentence *the fans were rooting for the team* (in Russian: *were ill for the team*), then a specific

scenario should be created for the situation 'the fans were rooting for the team'. Such scenarios were created on the basis of the markers' evaluations, after which the new scenario was given to the annotators to suggest derived meanings, such as BEFORE/AFTER and INTERP.

During the last phase of annotation, 5333 new utterances for scenarios were proposed by the annotators; 2484 utterances were proposed as candidates for the creation of new scenarios (some of them were approved), for 2856 utterances we have collected evaluations of the proximity to the closest scenario. The work on adding to the scenario inventory increased the number of nodes and edges in the scenario graph (Figure 6a) and reduced the number of isolated nodes and connectivity components (Figure 6b).



(a)



(b)

Figure 6: Variation of basic graph characteristics during the addition process: (a) the number of nodes and edges of the whole graph and the largest component: steadily increasing; (b) the number of isolated nodes and connected components decreases.

When extending the graph of the natural language inference space, a natural question is what size of this graph is sufficient to analyze the whole collection of real texts (α-connectivity). For this purpose, we propose to use the annotators' evaluation: Figure 8 shows a histogram of the similarity between the annotators' utterances and the closest scenario where the annotator found the proposed scenario satisfactory. The histogram shows that at a proximity of about 0.25 the annotators most often approve the scenario as corresponding to their utterance (the level of optimal β-connectivity). The main graph (Figure 7) shows the average proximity of clauses of real news texts parsed daily by the parser [9] to the scenarios for different versions of the scenario inventory – i.e., proximity of the α type (proximity of the semantics of real texts to scenarios). As the scenario inventory has been constantly updated and

refined, this proximity has increased, but it still lags behind the 0.25 level that the annotators consider optimal, according to the estimates on the histogram.

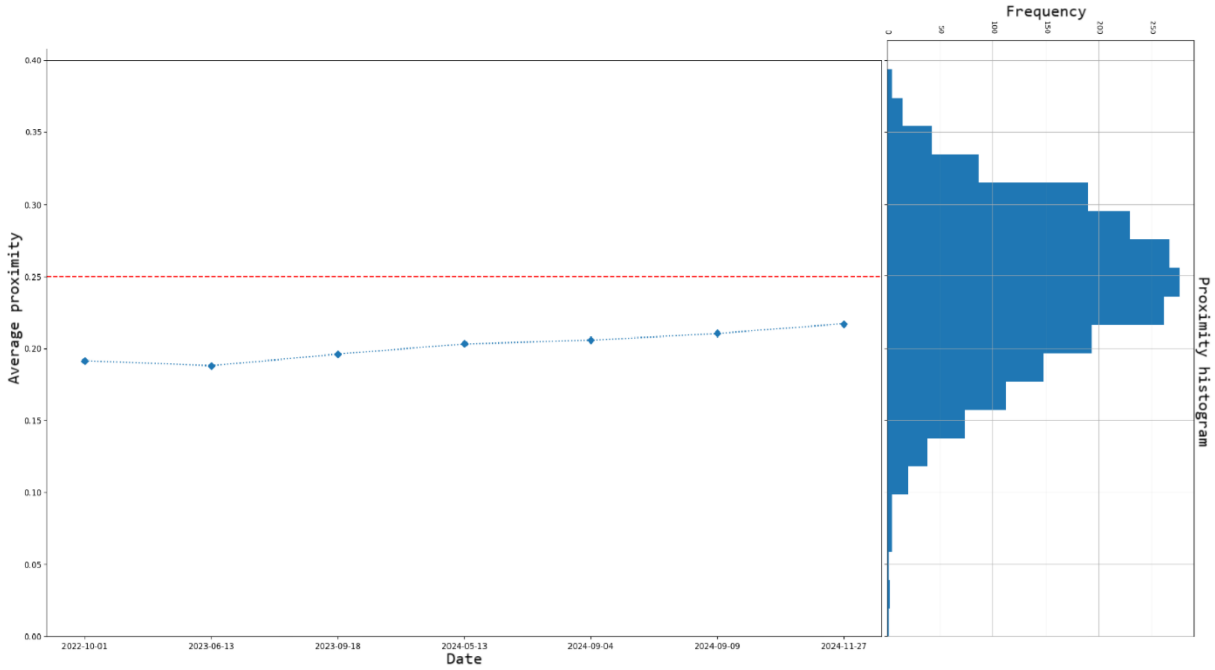


Figure 7: Average proximity of daily news text clauses to scenarios (α-connectivity); red dashed line is the derived estimate of high β-connectivity (from the histogram of new utterances' proximity to scenarios on the right)

Linear extrapolation was used to estimate the number of scenarios for which a high value of optimal proximity in text parsing would be achieved. Since the set of points X (number of scenarios at a given point in time) and the corresponding set of points Y (average proximity to the scenarios) are known, we can approximate a function F: Y = F(X) expressing the dependence of proximity on the number of scenarios. And if we extrapolate it beyond the interval X, we can estimate such X_α: $F(X_\alpha) = 0{,}25$. In our case $X_\alpha \approx 7000$ scenarios.

Conclusion

We modelled the process of human natural language reasoning as a transition from one semantic representation to another. We represented the space of possible transitions as a graph containing about 5,000 scenarios. The involvement of annotators allows us to populate the inventory of scenarios and to evaluate the connectivity of the overall graph within the space of the natural language reasoning. Based on our estimates in this paper, we suggest that an inventory of approximately 7,000 scenarios would be sufficient for basic modelling of human natural language inference. This estimate answers the question of how many frames a reasonably complete semantic model should contain.

Data availability

Scenario annotation intermediate results and robot videos are available from: Scenario Inference Data and Behavioral Protocols // OSF. March 31. doi:10.17605/OSF.IO/9HMP8

Acknowledgements

The research in its major part is supported by the Ministry of Science and Higher Education of the Russian Federation, project registration number – 124122200011-0.

References

[1] OpenAI et al. GPT-4 Technical Report. 2023.
[2] Cuadron A. et al. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks. 2025.
[3] Newell A., Simon H. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.
[4] Dorofeev G.V., Martemyanov Y.S. Logical inference and identification of links between sentences in the text [Logicheskij vyvod i vyyavlenie svyazej mezhdu predlozheniyami v tekste] // Machine translation and applied linguistics [Mashinnyj perevod i prikladnaya lingvistika]. 1969. Vol. 12. P. 36-59.
[5] Wu S. et al. A Comparative Study on Reasoning Patterns of OpenAI's o1 Model. 2024.
[6] Kotov A.A. et al. Development of semantic text analysis tools for robot control [Razrabotka sredstv semanticheskogo analiza teksta dlya upravleniya robotom] // Bulletin of Military Innovation Technopolis "Era" [Vestnik voennogo innovacionnogo tekhnopolisa "Era"]. 2021. Vol. 2, № 2. P. 115-120.
[7] Kotov A. Scenario inference data and behavioral protocols. OSF, 2025.
[8] Kotov A. et al. Conceptual processing system for a companion robot // Cogn Syst Res. 2021. Vol. 67.
[9] Kotov A.A. et al. Automated methods for evaluating emotional judgments on the Internet [Avtomatizirovannye metody ocenki emocional'nyh suzhdenij v internete] // Voprosy Psychologii [Voprosy Psychologii]. 2019. No 2. P. 133-153.
[10] Volkova L. et al. New Communicative Strategies for the Affective Robot: F-2 Going Tactile and Complimenting. 2021. P. 163–176.
[11] Velichkovsky B.M. et al. From social gaze to indirect speech constructions: How to induce the impression that your companion robot is a conscious creature // Applied Sciences (Switzerland). 2021. Vol. 11, № 21.
[12] Melchuk I.A. Experience of the theory of linguistic models "MEANING ⇔ TEXT" [Opyt teorii lingvisticheskih modelej "SMYSL ⇔ TEKST"]. Moscow: "Languages of Russian Culture," 1999.
[13] Granovsky D. V., Bocharov V. V. Open corpus: principles of work and prospects [Otkrytyj korpus: principy raboty i perspektivy] // Computer linguistics and the development of semantic search on the Internet: Proceedings of the scientific seminar of the XIII All-Russian Joint Conference "Internet and Modern Society" [Komp'yuternaya lingvistika i razvitie semanticheskogo poiska v Internete: Trudy nauchnogo seminara XIII Vserossijskoj ob"edinennoj konferencii «Internet i sov-remennoe obshchestvo»]. St. Petersburg, October 19-22, 2010. 2010.
[14] Kotov A., Zinina A., Filatov A. Semantic Parser for Sentiment Analysis and the Emotional Computer Agents // Proceedings of the AINL-ISMW FRUCT 2015. 2015. P. 167–170.
[15] Kotov A.A., Filatov A.A., Nosovets Z.A. Cognitive architecture for F-2 companion robot to support semantic shifts and cognitive domains via scenario oppositions // Cogn Syst Res. Elsevier B.V., 2025. Vol. 89.
[16] Shvedova N.Y. (ed.) Russian Semantic Dictionary [Russkij semanticheskij slovar']. Moscow: Azbukovnik, 1998.
[17] Kotov A.A. Mechanisms of speech influence [Mekhanizmy rechevogo vozdejstviya]. M.: RGGU, 2021.
[18] Baker C.F., Fillmore C.J., Lowe J.B. The Berkeley FrameNet Project. 1998.
[19] Lyashevskaya O., Kashkin E. Framebank: A database of Russian lexical constructions // Communications in Computer and Information Science. 2015.
[20] Allen S.R. Concern Processing in Autonomous Agents, Ph.D thesis. University of Birmingham, 2001.
[21] Fillmore C.J. The Case for Case // Universals in linguistic theory / ed. Bach E., Harms R.T. New York: Holt, Rinehart & Winston, 1968. P. 1–68.

References

[1] OpenAI et al. GPT-4 Technical Report. 2023.
[2] Cuadron A. et al. The Danger of Overthinking: Examining the Reasoning-Action Dilemma in Agentic Tasks. 2025.
[3] Newell A., Simon H. Human problem solving. Englewood Cliffs, NJ: Prentice-Hall, 1972.
[4] Дорофеев Г.В., Мартемьянов Ю.С. Логический вывод и выявление связей между предложениями в тексте // Машинный перевод и прикладная лингвистика. 1969. Vol. 12. P. 36–59.
[5] Wu S. et al. A Comparative Study on Reasoning Patterns of OpenAI's o1 Model. 2024.
[6] Котов А.А. и др. Разработка средств семантического анализа текста для управления роботом // Вестник Военного инновационного технополиса "Эра." 2021. Vol. 2, № 2. P. 115–120.
[7] Kotov A. Scenario inference data and behavioral protocols. OSF, 2025.
[8] Kotov A. et al. Conceptual processing system for a companion robot // Cogn Syst Res. 2021. Vol. 67.
[9] Котов А.А. и др. Автоматизированные методы оценки эмоциональных суждений в интернете // Вопросы психологии. 2019. № 2. P. 133–153.

[10] Volkova L. et al. New Communicative Strategies for the Affective Robot: F-2 Going Tactile and Complimenting. 2021. P. 163–176.
[11] Velichkovsky B.M. et al. From social gaze to indirect speech constructions: How to induce the impression that your companion robot is a conscious creature // Applied Sciences (Switzerland). 2021. Vol. 11, № 21.
[12] Мельчук И.А. Опыт теории лингвистических моделей “СМЫСЛ <=> ТЕКСТ.” М.: Школа “Языки русской культуры,” 1999. 346 p.
[13] Грановский Д. В., Бочаров В. В. Открытый корпус: принципы работы и перспективы // Компьютерная лингвистика и развитие семантического поиска в Интернете: Труды научного семинара XIII Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 19–22 октября 2010 г. 2010.
[14] Kotov A., Zinina A., Filatov A. Semantic Parser for Sentiment Analysis and the Emotional Computer Agents // Proceedings of the AINL-ISMW FRUCT 2015. 2015. P. 167–170.
[15] Kotov A.A., Filatov A.A., Nosovets Z.A. Cognitive architecture for F-2 companion robot to support semantic shifts and cognitive domains via scenario oppositions // Cogn Syst Res. Elsevier B.V., 2025. Vol. 89.
[16] Шведова Н.Ю. Русский семантический словарь. М.: Азбуковник, 1998.
[17] Котов А.А. Механизмы речевого воздействия. М.: РГГУ, 2021.
[18] Baker C.F., Fillmore C.J., Lowe J.B. The Berkeley FrameNet Project. 1998.
[19] Lyashevskaya O., Kashkin E. Framebank: A database of Russian lexical constructions // Communications in Computer and Information Science. 2015.
[20] Allen S.R. Concern Processing in Autonomous Agents, Ph.D thesis. University of Birmingham, 2001.
[21] Fillmore C.J. The Case for Case // Universals in linguistic theory / ed. Bach E., Harms R.T. New York: Holt, Rinehart & Winston, 1968. P. 1–68.

Building Russian Benchmark for Evaluation of Information Retrieval Models

Grigory Kovalev
Lomonosov Moscow State University
Russia
kaengreg@ya.ru

Mikhail Tikhomirov
Lomonosov Moscow State University
Russia
tikhomirov.mm@gmail.com

Evgeny Kozhevnikov
Lomonosov Moscow State University
Russia
dovvakkin@gmail.com

Max Kornilov
Lomonosov Moscow State University
Russia
max.korn@bk.ru

Natalia Loukachevitch
Lomonosov Moscow State University
Russia
louk_nat@mail.ru

Abstract

We introduce RusBEIR, a comprehensive benchmark designed for zero-shot evaluation of information retrieval (IR) models in the Russian language. Comprising 17 datasets from various domains, it integrates adapted, translated, and newly created datasets, enabling systematic comparison of lexical and neural models. Our study highlights the importance of preprocessing for lexical models in morphologically rich languages and confirms BM25 as a strong baseline for full-document retrieval. Neural models, such as mE5-large and BGE-M3, demonstrate superior performance on most datasets, but face challenges with long-document retrieval due to input size constraints. RusBEIR offers a unified, open-source framework that promotes research in Russian-language information retrieval. The benchmark is available for public use on GitHub.

Keywords: information retrieval, benchmark, lexical model, dense model, reranker
DOI: 10.28995/2075-7182-2025-23-195-205

Создание русского бенчмарка для оценки моделей информационного поиска

Ковалев Григорий
МГУ им. М.В. Ломоносова
Россия
kaengreg@ya.ru

Тихомиров Михаил
МГУ им. М.В. Ломоносова
Россия
tikhomirov.mm@gmail.com

Кожевников Евгений
МГУ им. М.В. Ломоносова
Россия
dovvakkin@gmail.com

Корнилов Максим
МГУ им. М.В. Ломоносова
Россия
max.korn@bk.ru

Лукашевич Наталья
МГУ им. М.В. Ломоносова
Россия
louk_nat@mail.ru

Аннотация

Мы представляем RusBEIR — это бенчмарк, предназначенный для zero-shot оценки моделей информационного поиска (IR) на русском языке. Он включает 17 наборов данных из различных доменов, объединяя адаптированные, переведенные и созданные наборы данных, что позволяет проводить систематическое сравнение лексических и нейронных моделей. Наше исследование подчеркивает важность предобработки для лексических моделей в языках с богатой морфологией и подтверждает, что модель BM25 обеспечивает высокое качество поиска, особенно для полных документов. Нейронные модели, такие как mE5-large и BGE-M3, показывают высокие результаты на большинстве наборов данных, но сталкиваются с трудностями при работе с длинными документами из-за ограничений на максимальную длину входа. RusBEIR предоставляет унифицированную открытую платформу, способствующую развитию исследований в области информационного поиска на русском языке. RusBEIR является проектом с открытым исходным кодом и доступен на GitHub.

Ключевые слова: информационный поиск, бенчмарк, лексическая модель, нейросетевая модель, реранкер

1 Introduction

Traditionally, Information Retrieval (IR) was based on lexical models such as TF-IDF and BM25, but these models are known as bag-of-words models, which do not take into account the document context. Modern approaches are based on neural models, in particular on Transformer models. Most recent advancements in IR leverage neural retrieval models built upon pre-trained Transformer architectures, such as BERT [9]. These models address the limitations of traditional lexical methods by capturing semantic relationships and contextual information, enabling them to bridge the lexical gap inherent in keyword-based retrieval approaches. Unlike lexical models, which rely solely on the presence of query terms in documents, neural models represent queries and documents in a dense vector space, facilitating more accurate retrieval through similarity measures like cosine similarity.

Neural retrieval systems have demonstrated significant performance improvements over traditional methods, particularly in tasks such as open-domain question answering, claim verification, and passage retrieval. However, these advancements often come at the cost of increased computational resources and the need for extensive training data. Due to the scarcity of labeled data, neural models are frequently applied in zero-shot settings.

Traditionally, neural retrievers have been trained on large datasets such as MS MARCO [3] and Natural Questions [15]. Before the introduction of BEIR [24], these models were often evaluated on the same datasets they were trained on, gaining a significant advantage over lexical approaches like BM25.

To address this limitation, the authors of BEIR introduced a robust and diverse benchmark designed to evaluate model generalization across tasks and domains. BEIR consists of 18 retrieval datasets from a variety of domains, providing a more accurate and comprehensive framework for evaluating neural retrieval systems. Notably, the results of BEIR revealed that neural models do not consistently outperform lexical approaches, highlighting the need for careful evaluation in diverse settings.

The zero-shot application of neural retrievers is particularly important for underrepresented languages, such as Slavic languages, where the availability of information retrieval datasets is limited. Consequently, there is a growing demand for multilingual evaluation benchmarks akin to the monolingual BEIR framework. Such benchmarks would enable robust cross-lingual evaluation and foster the development of neural retrieval systems for less commonly studied languages, addressing the current gaps in multilingual information retrieval research.

Interestingly, the performance gap between lexical and dense retrieval models remains a topic of interest. Although dense models typically excel in retrieval accuracy, lexical methods such as BM25 offer a lightweight alternative with significantly lower computational overhead. Investigating this trade-off can provide valuable insight into the practical application of retrieval models across diverse scenarios, especially when computational efficiency is a priority.

In this paper, we present RusBEIR, a BEIR-inspired benchmark designed for the zero-shot evaluation of Information Retrieval (IR) models in the Russian language. RusBEIR comprises 17 datasets that span various domains and tasks. Some datasets have been adapted from BEIR and similar benchmarks, others are newly collected specifically for this benchmark or sourced from existing Russian or multilingual benchmarks. Our primary objective is to establish a large-scale benchmark tailored to information retrieval in Russian, with a particular emphasis on zero-shot approaches. In addition, we explore whether neural models consistently outperform traditional lexical methods across diverse scenarios. With this aim, we evaluate a range of models, including BM25, BGE-M3, mE5, RoSBERTa, and LaBSE [6, 25, 23, 12]. This benchmark offers a comprehensive resource for advancing and evaluating IR systems in Russian, fostering research into the comparative strengths of neural and lexical approaches.

2 Related work

The information-retrieval domain has a rich history of creating datasets, benchmarks, organizing various evaluations. Specialized evaluation conferences such as TREC, CLEF, NTCIR have been held since 90-th of 20 century. There exist numerous national information-retrieval initiatives: in Poland [16], India [13] and other countries. In Russia during 2003-2011, ROMIP workshop [10] was a place for evaluation of approaches in information-retrieval tasks, such as ad hoc retrieval, thematic categorization, question-answering, summarization, etc.

Evaluating neural models in information retrieval requires creating benchmarks comprising diverse datasets. Development of information-retrieval benchmarks for non-English languages is usually based on the English BEIR benchmark [24].

The Polish benchmark BEIR-PL [27] was created via automatic translation of 13 datasets from BEIR. For translation, the Google Translate service was used. In [8], the authors describe the Polish Information Retrieval Benchmark (PIRB), encompassing 41 text information retrieval tasks for Polish. The datasets in PIRB comprise the BEIR-PL datasets, several other existing information-retrieval datasets, and also nine datasets crawled from Polish websites. In evaluation, it was found that the best results on the benchmark were achieved by the mE5-large model [25]. The authors also trained a learning-to-rank model combining scores of several basic models and achieved better results.

To create Dutch BEIR, the authors of [4] translated initial BEIR datasets into Dutch using the Gemini-1.5-flash model. To assess the translation quality, ten items from each dataset were randomly sampled and checked by a Dutch native speaker. It was shown that 98% of checked samples were translated correctly or with minor issues. The authors tested BM25, neural models (including mE5-large and BGE-M3 models) and reranking approaches combining BM25 and a neural reranker. They conclude that BM25 still provides a competitive baseline, and, in many cases, is only outperformed by larger dense models.

The authors of [1] created Hindi BEIR benchmark. Hindi-BEIR encompasses 15 diverse datasets from 6 distinct domains. They translated BEIR datasets using Indic-Trans2 model, a neural translation model supporting translations across all 22 Indic languages (including English). They translated 9 datasets from the source BEIR benchmark. To check the quality of translation, the authors back-translated the Hindi translations into English. Then they calculated the char-based Chrf(++) score [20] between the original English query/document and the backtranslated English query/document. Also 5 publicly available information-retrieval datasets were added to the benchmark. In experiments, neural models (BGE-M3, mE5, LASER, LaBSE) were compared with BM25. The best results were obtained with BGE-M3, which is signifcantly better than other approaches.

For Russian, the MTEB benchmark [19] for evaluating embeddings has been created. Russian MTEB comprises 23 datasets in 7 task categories including three information-retrieval datasets [23]. The best models in the Russian MTEB information-retrieval section with the size less than 1b are BGE-M3 [6] and Multilingual E5-large [25].

3 Datasets in RusBEIR

RusBEIR is a Russian benchmark inspired by BEIR [24], designed for zero-shot evaluation of Information Retrieval (IR) models. Adhering to the principles of BEIR, it offers a robust and diverse evaluation

framework, enabling the assessment of IR models across a wide range of tasks and domains in the Russian language.

The datasets in the RusBEIR benchmark consist of available open-source datasets, datasets that have been translated from English, and newly created datasets. Table 1 provides a description of the available datasets. We will discuss the datasets in more detail in the following subsections.

3.1 Translated BEIR Datasets

BEIR consists of multilingual and monolingual (English) datasets. To achieve reproducibility of results from BEIR and its analogues, it was decided to translate the monolingual datasets into the Russian language and evaluate them with models used in our benchmark.

The choice of translation method was based on studies conducted as part of the creation of the multilingual MsMarco dataset mMarco [5], where experiments with Google Translate and the Helsinki model were conducted, and the results of similar experiments from the PL-BEIR [27] project were analyzed. According to the results of these studies, Google Translate showed better translation quality compared to the Helsinki model. Therefore, Google Translate was chosen.

As a result, we introduce 4 datasets from the original BEIR datasets [24], which were translated into the Russian language.

- **NF-Corpus** is a comprehensive full-text English retrieval dataset designed for medical information retrieval tasks. It contains a collection of queries formulated in non-technical English sourced from NutritionFacts.org [1] and corresponding medical documents written in a complex terminology-heavy language primarily derived from PubMed [2], a database of medical literature.
- **ArguAna** is a dataset designed for the argument retrieval task, derived from debates on idebate.org [3]. It covers controversial topics across 15 themes, such as “economy” and “health.” The dataset includes a corpus consisting of debate texts and queries derived from these debates. The task is to retrieve relevant arguments from the corpus.
- **SciFact** is a dataset for scientific claim verification, consisting of expert-written claims paired with abstracts from research literature. Each abstract is annotated with evidence supporting or refuting the claims, along with rationales justifying the decisions.
- **SCIDOCS** is a dataset focused on citation prediction, designed to evaluate the ability of scientific document embeddings to predict citation relationships between research papers.

Source (↓)	Task (↓)	Dataset (↓)	Origin (↓)	Relevancy	Train	Dev	Test	Corpus	Avg. Word Lengths (D/Q)
BEIR	Bio-Medical IR	rus-NFCorpus	Translation	Binary	2,590	324	323	3,633	216.6 / 3.5
BEIR	Argument Retrieval	rus-ArguAna	Translation	Binary	—	—	1,406	8,674	147.8 / 173.8
BEIR	Fact Checking	rus-SciFact	Translation	Binary	809	—	300	5,183	185.8 / 11.2
BEIR	Citation-Prediction	rus-SCIDOCS	Translation	Binary	—	—	1000	25,657	153.1 / 9.8
BEIR	Information-Retrieval	rus-MMARCO	Part of multilingual	Binary	502,939	6980	—	8,841,823	49.6 / 5.95
Open-Source Dataset	Information-Retrieval	rus-MIRACL	Part of multilingual	Binary	4,683	1,252	—	9,543,918	43 / 6.2
Open-Source Dataset	Question Answering (QA)	rus-XQuAD	Part of multilingual	Binary	—	1,190	—	240	112.9 / 8.6
Open-Source Dataset	Question Answering (QA)	rus-XQuAD-sentences	Part of multilingual	Binary	—	1,190	—	1212	22.4 / 8.6
Open-Source Dataset	Question Answering (QA)	rus-Tydi QA	Part of multilingual	Binary	—	1,162	—	89,154	69.4 / 6.5
Open-Source Dataset	Information-Retrieval	SberQuAD-retrieval	Originally Russian	Binary	45,328	5,036	23,936	17,474	100.4 / 8.7
Open-Source Dataset	Information-Retrieval	ruSciBench-retrieval	Originally Russian	Binary	—	345	—	200,532	89.9 / 9.2
Open-Source Dataset	Question Answering (QA)	ru-facts	Originally Russian	Binary	2,241	753	—	6,236	28.1 / 23.9
RU-MTEB	Information-Retrieval	RuBQ	Originally Russian	Binary	—	—	1,692	56,826	62.07 / 6.4
RU-MTEB	Information-Retrieval	Ria-News	Originally Russian	Binary	—	—	10,000	704,344	155.2 / 8.8
rusBEIR	Information-Retrieval	wikifacts-articles	Originally Russian	3-level	—	540	—	1,324	2,535.9 / 11.4
rusBEIR	Fact Checking	wikifacts-para	Originally Russian	3-level	—	540	—	15,317	219.2 / 11.4
rusBEIR	Information-Retrieval	wikifacts-sents	Originally Russian	3-level	—	540	—	188,026	17.8 / 11.4

Table 1: Overview of datasets and tasks for information retrieval and related fields. All datasets are available at HuggingFace

[1] NutritionFacts.org

[2] https://pubmed.ncbi.nlm.nih.gov

[3] idebate.org

3.2 Russian Parts of Multilingual Datasets

The main objective of BEIR is to gather a large and diverse set of data from various domains and tasks. This will force models to operate in an out-of-distribution environment and help to evaluate them more accurately. In order to expand our collection of Russian datasets, we also retrieved the Russian portions of existing multilingual datasets, including mMARCO [5], MIRACL [28], XQUAD [2], and TyDiQA [7].

The **mMARCO** (Multilingual MS MARCO) dataset [5] is a multilingual adaptation of the popular MS MARCO dataset, designed for information retrieval and question answering tasks. It extends the original English MS MARCO dataset into multiple languages, including Russian.

MIRACL is a multilingual dataset for information retrieval in 18 languages. The queries were taken mainly from the Mr. TYDI dataset. Passages were retrieved from Wikipedia by an ensemble model, and 10 top documents were annotated by human annotators.

XQuAD (Cross-lingual Question Answering Dataset) is a benchmark dataset designed to evaluate the performance of cross-lingual question answering systems. It consists of a collection of 240 passages and 1,190 question-answer pairs from the development set of the SQuAD v1.1 dataset [21], along with their professional translations into 10 languages: Spanish, German, Greek, Russian, Turkish, Arabic, Vietnamese, Thai, Chinese, and Hindi. This makes the dataset entirely parallel across 11 languages.

Tydi QA is a question-answering dataset covering 11 typologically diverse languages. Questions were written by humans on Wikipedia topics. Answers should not be contained in the first 100 characters of the corresponding Wikipedia article. The questions were written for each language, not translated.

3.3 Existing Russian Datasets

The Russian Massive Text Embedding Benchmark (ruMTEB) is an extension of the Massive Text Embedding Benchmark (MTEB) tailored specifically for the Russian language. The authors of ruMTEB introduced 17 new datasets in Russian which were categorized into 7 groups.

In our benchmark we use 2 of presented IR datasets: RuBQ and Ria-News. **RuBQ** [22] is a specialized dataset for Russian-language question answering over Wikidata, offering a rich set of questions paired with structured answers.

The **Ria-News** dataset [14] is a collection of Russian-language news articles published by the RIA Novosti news agency (2010-2014). This dataset presents a task in which a model is required to locate the text of a specific news article within a larger corpus of news articles based on its corresponding title, which acts as a query.

Besides, we added publicly-available IR-related datasets: SberQuad [11], ruSciBench and ru-facts [18].

SberQuAD is a Russian-language machine reading comprehension (MRC) dataset inspired by the popular English SQuAD [21] (Stanford Question Answering Dataset). It provides annotated passages and question-answer pairs in Russian.

ruSciBench is a Russian-language benchmark designed to evaluate the performance of text embedding models for scientific articles. [4] The corpus consists of abstracts, and the queries are LLM-generated questions for these abstracts.

ru-facts [18] is a fact-checking dataset developed by translating and expanding the FEVER dataset with additional data from the Russian news summarization corpus Gazeta [5], using a paraphrasing model [6], and rule-based transformations from the Ru_Paraphraser dataset [7].

3.4 New Russian Wikipedia-based Datasets

We also introduce a new series of Russian Wikipedia-based datasets. The datasets are based on Wikipedia section “Did you know ...”. The section contains interesting facts, which are extracted from Wikipedia articles. The articles mentioned in a fact are provided with hyperlinks. For example, the fact “The 2024

[4]The dataset is a ported version of qa_science_ru https://huggingface.co/datasets/AIR- Bench/qa_science_ru from the Air-Bench repository, which in turn is a port of the ru_sci_bench (https://huggingface.co/datasets/mlsa-iai-msu-lab/ru_sci_bench) dataset from the MSLA-Iai MSU Lab

[5]https://huggingface.co/datasets/IlyaGusev/gazeta

[6]https://habr.com/en/company/sberdevices/blog/667106/

[7]https://huggingface.co/datasets/merionum/ru_paraphraser

American Samoan gubernatorial election was won by Pula and Pulu?" mentions three Wikipedia articles ("2024 American Samoan gubernatorial election", "Pula", "Pulu"), from which the fact should be inferred.

University students were asked to find relevant sentences in the mentioned articles that confirm the fact. They marked relevant sentences with scores of 2 or 1. Irrelevant sentences have zero scores. Relevant sentences with score 2 contain the full fact. If a sentence contains a part of the fact it obtains score 1. In total, 540 facts have been annotated.

Using facts, extracted articles and created annotations, three datasets with the same queries but different documents have been created.

- **wikifacts-sents** dataset consists of sentences extracted from articles, some of which confirm the fact which stands as a query. The documents in this dataset are the shortest in the benchmark;
- **wikifacts-articles** dataset comprises all full articles mentioned in facts. Relevant articles contain relevant sentences. This dataset includes the longest documents in the benchmark and can be used for evaluation of full-document retrieval;
- **wikifacts-para** dataset comprises existing paragraphs from the extracted articles, the documents on the datasets are significantly shorter than in the wikifacts-articles dataset, but still longer than most benchmark datasets;

Having such variants, we can evaluate different information-retrieval tasks on the same annotated data.

3.5 BEIR Format Compatibility

Our datasets are presented in a unified format and are compatible with the original BEIR benchmark. Queries are predetermined questions in natural language that are used to evaluate the performance of information retrieval (IR) systems. A corpus refers to a collection of documents that the system searches through in order to find relevant information for the given query. Relevance judgments, also known as qrels, indicate the association between queries and documents. All queries, corpora, and relevance judgments are stored in JSONL and TSV file formats, respectively.

4 Models

For evaluation, we used the BM25 lexical model and dense retrieval models.

4.1 Preprocessing for BM25 model

The main baseline was calculated using the BM25 lexical model implemented in the Elasticsearch engine[8], with the language analyzer disabled to avoid stemming, which is less suitable for the Russian language. We specially preprocessed data to be used as input for BM25.

The text preprocessing method consists of the following steps:

1. **Lowercasing**: Converting all text to lowercase to ensure uniformity and eliminate case sensitivity.
2. **Punctuation and Special Character Removal**: Using regex to remove non-alphanumeric characters, leaving only letters, digits, and spaces to reduce noise.
3. **Space Normalization**: Removing extra spaces and trimming leading or trailing whitespace.
4. **Tokenization**: Splitting text into individual words for processing.
5. **Lemmatization**: Using PyMorphy3 [17] to convert words into their dictionary forms, reducing data dimensionality while preserving semantic meaning. This approach is particularly effective for the Russian language due to its rich morphology, as it avoids the inaccuracies that stemming introduces by truncating words without context.
6. **Stop Word Removal**: excluding overly frequent words that contribute little to the text content using the default stopword list provided by the NLTK package [9] [15], augmented with two Russian pronouns: "which" ("который") and "such"("такой").

[8] https://www.elastic.co/
[9] https://www.nltk.org

4.2 Neural baseline models

Neural baseline models used in our work are subdivided into pre-trained dense retrievers (bi-encoders) and rerankers. Bi-encoders generate embeddings for queries and documents and calculate their cosine similarity. Rerankers take a query and a document as an input and calculate the probability of the document to be relevant to the query. Rerankers are applied to the best documents found by lexical or dense retrievers and usually improve the performance of combined retrieval. Dense retrievers include the following pre-trained bi-encoders:

- LaBSE bi-encoder [12]. LaBSE was pre-trained with a translation ranking task. This allows to find sentence paraphrases in a single language or different languages.[10]
- Multilingual E5 in three sizes: large [11], base [12] and small [13] [25]. The multilingual E5 model was trained on a large multilingual corpus using a weakly supervised contrastive pretraining method with InfoNCE contrastive loss. Then it was fine-tuned on high-quality labeled multilingual datasets for retrieval tasks.
- BGE-M3 model [14] [6]. The BGE-M3 model was pre-trained on a large multilingual and cross-lingual unsupervised data, and subsequently fine-tuned on multilingual retrieval datasets using a custom loss function based on the InfoNCE loss function.
- USER-BGE-M3 [15]. USER-BGE-M3 is a sentence-transformer model for training embeddings for Russian. The model is initialized from the en-ru-BGE-M3 model [16], a shrinked version of the BGE-M3 model, and then trained on the Russian datasets.
- ru-en-RoSBERTa[17] [23]. ruRoBERTa model [29] [18] was used as a basic model and then RoSBERTa embeddings were fine-tuned on Russian and English datasets.

As a reranker, we use the bge-reranker-v2-m3 reranker [19]. In our work, we use BGE models with a max-length parameter set to 2048.

Model	Based on	Parameters	Dim	Max input
Multilingual-E5-large	XLM-RoBERTa-large	560M	1024	512
Multilingual-E5-base	XLM-RoBERTa-base	278M	768	512
Multilingual-E5-small	Multilingual-MiniLM	118M	384	512
BGE-M3	BGE-M3	568M	1024	8192
USER-BGE-M3	BGE-M3	359M	1024	8192
RoSBERTa	SBERT	404M	1024	512
LaBSE	LaBSE	471M	768	256
bge-reranker-v2-m3	BGE-M3	568M	1024	8192

Table 2: Model Specifications and Details

5 Results

We evaluated the models on the RusBEIR datasets using NDCG@10, MAP@10, and Recall@10. Since all metrics showed similar trends, we present only the NDCG@10 results in the table below for brevity. Additional details on MAP@10 and Recall@10 are available in the Additional Metrics section.

[10] https://huggingface.co/sentence-transformers/LaBSE
[11] https://huggingface.co/intfloat/multilingual-e5-large
[12] https://huggingface.co/intfloat/multilingual-e5-base
[13] https://huggingface.co/intfloat/multilingual-e5-small
[14] https://huggingface.co/BAAI/bge-m3
[15] https://huggingface.co/deepvk/USER-bge-m3
[16] https://huggingface.co/TatonkaHF/bge-m3_en_ru
[17] https://huggingface.co/ai-forever/ru-en-RoSBERTa
[18] https://huggingface.co/ai-forever/ruRoberta-large
[19] https://huggingface.co/BAAI/bge-reranker-v2-m3

Model (→)	Lexical	Dense							Re-ranking		
Dataset (↓)	BM25	mE5-large	mE5-base	mE5-small	BGE-M3	USER-BGE-M3	RoSBERTa	LaBSE	BM25+BGE	mE5-large+BGE	BGE-M3+BGE
rus-NFCorpus	32.33	30.96	26.90	26.79	30.86	30.28	27.24	18.53	**34.83**	33.18	32.46
rus-ArguAna	41.49	49.06	39.40	39.59	50.75	46.52	49.38	25.52	52.91	**54.01**	53.87
rus-SciFact	65.60	63.49	63.46	60.46	62.42	58.25	53.90	29.07	70.40	**71.34**	69.64
rus-SCIDOCS	13.99	13.47	12.09	10.60	15.04	14.46	14.43	8.17	15.31	15.98	**16.21**
rus-MMARCO	15.25	34.04	30.27	29.07	29.51	27.92	20.16	9.06	24.12	**36.95**	34.52
rus-MIRACL	25.13	66.99	61.41	58.52	70.50	67.23	53.11	15.70	41.51	75.90	**76.44**
rus-XQuAD	96.19	97.33	95.84	95.66	95.97	95.63	93.90	69.77	98.85	**98.97**	**98.97**
rus-XQuAD-Sentences	82.36	88.84	86.37	85.41	86.91	85.42	83.20	75.33	89.93	**92.08**	91.69
rus-TyDi QA	35.80	59.41	55.91	55.23	58.34	57.86	52.06	28.05	50.12	**66.20**	65.78
SberQuad-retrieval	68.19	67.11	65.13	61.03	68.26	67.03	63.59	37.54	**70.34**	69.41	68.21
ruSciBench-retrieval	36.69	50.81	45.74	42.93	55.85	53.58	44.89	17.93	49.93	65.33	**69.05**
ru-facts	92.56	93.65	93.55	93.06	**93.91**	93.77	93.66	93.10	92.72	92.87	92.87
RuBQ	37.33	74.11	69.63	68.60	71.26	70.00	66.81	30.59	56.90	**77.03**	76.00
Ria-News	64.63	80.67	70.24	70.00	82.99	83.52	78.85	61.57	78.12	86.22	**86.85**
wikifacts-articles	84.28	66.09	63.04	67.86	74.50	79.41	74.13	45.17	**85.25**	83.06	83.91
wikifacts-para	61.31	50.15	49.51	34.71	54.55	57.53	50.66	14.78	**66.61**	59.95	63.76
wikifacts-sents	33.64	35.90	30.75	22.57	37.59	34.90	**40.59**	25.79	39.96	38.53	39.20
Avg	52.16	60.12	56.43	54.24	61.13	60.19	56.50	35.63	59.87	65.71	**65.85**

Table 3: Performance comparison across different models and datasets. The best results for each dataset are in bold; the results of the best single models are underlined.

The analysis of Table 3 indicates that the best performance on the benchmark is achieved through the combination of the BGE-M3 model and the BGE reranker. Notably, the combination of mE5-large bi-encoder with the BGE reranker yields close results. Among the individual models, the mE5-large bi-encoder and both multilingual BGE variants stand out as top performers, surpassing BM25 by an average margin of 15.9 percentage points.

Overall, LaBSE performs the worst among all the models presented. This can be attributed to its training objective, which focuses on finding similar sentences across different languages or paraphrases within the same language. As a result, when confronted with queries that lack lexical overlap with sentences in the corpus, its performance drops.

RoSBERTa model performs on par with mE5-base and mE5-small, but the size of mE5-base (278M) againts RoSBERTa (404M) makes mE5-base more preferable to use.

At the same time, it is worth noting that the BM25 model is the best single model on four datasets: rus-NFCorpus, rus-SciFact, wikifacts-articles and wikifacts-para. The best results on these datasets, as well as others where single BM25 performed only slightly worse than neural retrievers, are achieved by combining BM25 with the BGE reranker. Three of the datasets with a significant BM25 margin contain longer documents than the average in the benchmark. On the wikifacts-articles dataset, which includes full-text documents, BM25 outperforms the BGE-M3 model by 13 percentage points and the mE5-large model by 27 percentage points. This highlights a limitation of the mE5 models in retrieving long documents due to their small maximum input size (512 tokens). Additionally, the rus-NFCorpus and rus-SciFact datasets are domain-specific, which may result in lower-quality multilingual vector representations compared to general datasets.

Furthermore, it should be noted that the results of the BGE models presented in Table 3 were obtained with a maximum input length set to 2048. However, as indicated in Table 2, BGE models can process up to 8192 tokens, making them more suitable for full-text search in long documents.

Our experiments demonstrated that BM25 remains a strong baseline for information retrieval, particularly for full-document retrieval. Neural models, especially mE5-large and BGE-M3, achieved the best results on the benchmark and confirmed the findings of other BEIR-based studies [24, 26].

6 Conclusion

In this paper, we introduced RusBEIR, a comprehensive BEIR-inspired benchmark designed for the zero-shot evaluation of information retrieval (IR) models in the Russian language. Consisting of 17 datasets from diverse domains and tasks, RusBEIR integrates adapted datasets from existing benchmarks alongside novel datasets to further enrich its collection. By providing a large-scale resource compatible with the original BEIR format, RusBEIR enables systematic evaluation and comparison of both lexical

and neural IR models, with a particular emphasis on zero-shot performance.

Our study stresses the importance of accurate preprocessing, particularly for lexical models, where preprocessing significantly impacts the performance in morphologically rich languages as Russian. Additionally, we introduced a series of Russian Wikipedia-based datasets that further expand the scope of RusBEIR, enabling more granular exploration of IR performance across document lengths and tasks.

The results of our experiments confirm that BM25 remains a robust baseline for full-document retrieval, while state-of-the-art neural models, such as mE5-large and BGE-M3, demonstrate superior performance on most datasets. These findings are consistent with previous BEIR-based studies and underscore the advantages of neural approaches, particularly when unprocessed data are used as input. However, our analysis also highlights certain limitations of neural models, such as challenges with long-document retrieval due to input size constraints. The efficiency comparison between BM25 and neural models such as mE5 and BGE remains an open question and will be explored further in future research.

By providing a unified framework and detailed insights into the comparative performance of lexical and neural models, we hope RusBEIR will serve as a valuable tool for advancing research and innovation in information retrieval for the Russian language.

Acknowledgements

The work is supported by the Russian Science Foundation under Agreement No. 25-21-00206.
The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

References

[1] Arkadeep Acharya et al. “Hindi-BEIR: A Large Scale Retrieval Benchmark in Hindi”. In: *arXiv preprint arXiv:2408.09437* (2024).

[2] Mikel Artetxe, Sebastian Ruder and Dani Yogatama. “On the Cross-lingual Transferability of Monolingual Representations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. 2020.

[3] Payal Bajaj et al. “Ms marco: A human generated machine reading comprehension dataset”. In: *arXiv preprint arXiv:1611.09268* (2016).

[4] Nikolay Banar, Ehsan Lotfi and Walter Daelemans. “BEIR-NL: Zero-shot Information Retrieval Benchmark for the Dutch Language”. In: *arXiv preprint arXiv:2412.08329* (2024).

[5] Luiz Bonifacio et al. “mmarco: A multilingual version of the ms marco passage ranking dataset”. In: *arXiv preprint arXiv:2108.13897* (2021).

[6] Jianlv Chen et al. “Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation”. In: *arXiv preprint arXiv:2402.03216* (2024).

[7] Jonathan H Clark et al. “Tydi qa: A benchmark for information-seeking question answering in ty pologically di verse languages”. In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 454–470.

[8] Slawomir Dadas, Micha l Pere lkiewicz and Rafa l Poświata. “PIRB: A Comprehensive Benchmark of Polish Dense and Hybrid Text Retrieval Methods”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 12761–12774.

[9] Jacob Devlin et al. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Ed. by Jill Burstein, Christy Doran and Thamar Solorio. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: https://aclanthology.org/N19-1423/.

[10] Boris V Dobrov et al. “Russian Information Retrieval Evaluation Seminar.” In: *LREC*. 2004.

[11] Pavel Efimov et al. “Sberquad–russian reading comprehension dataset: Description and analysis”. In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22–25, 2020, Proceedings 11*. Springer. 2020, pp. 3–15.

[12] Fangxiaoyu Feng et al. “Language-agnostic BERT Sentence Embedding”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022, pp. 878–891.

[13] Debasis Ganguly et al. *Proceedings of the 15th Annual Meeting of the Forum for Information Retrieval Evaluation*. ACM, 2023.

[14] Daniil Gavrilov, Pavel Kalaidin and Valentin Malykh. “Self-attentive model for headline generation”. In: *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part II 41*. Springer. 2019, pp. 87–93.

[15] Nitin Hardeniya et al. *Natural language processing: python and NLTK*. Packt Publishing Ltd, 2016.

[16] Lukasz Kobyliński et al. “PolEval 2022/23 challenge tasks and results”. In: *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE. 2023, pp. 1243–1250.

[17] Mikhail Korobov. “Morphological analyzer and generator for Russian and Ukrainian languages”. In: *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*. Springer. 2015, pp. 320–332.

[18] Anastasia Kozlova, Denis Shevelev and Alena Fenogenova. “Fact-checking benchmark for the Russian Large Language Models”. In: *Proceedings of the International Conference “Dialogue*. Vol. 2023. 2023.

[19] Niklas Muennighoff et al. “MTEB: Massive Text Embedding Benchmark”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. 2023, pp. 2014–2037.

[20] Maja Popović. “chrF: character n-gram F-score for automatic MT evaluation”. In: *Proceedings of the tenth workshop on statistical machine translation*. 2015, pp. 392–395.

[21] P Rajpurkar. “Squad: 100,000+ questions for machine comprehension of text”. In: *arXiv preprint arXiv:1606.05250* (2016).

[22] Ivan Rybin et al. “RuBQ 2.0: an innovated Russian question answering dataset”. In: *The Semantic Web: 18th International Conference, ESWC 2021, Virtual Event, June 6–10, 2021, Proceedings 18*. Springer. 2021, pp. 532–547.

[23] Artem Snegirev et al. “The Russian-focused embedders’ exploration: ruMTEB benchmark and Russian embedding model design”. In: *CoRR* (2024).

[24] Nandan Thakur et al. “BEIR: A Heterogeneous Benchmark for Zero-shot Evaluation of Information Retrieval Models”. In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL: https://openreview.net/forum?id=wCu6T5xFjeJ.

[25] Liang Wang et al. “Multilingual e5 text embeddings: A technical report”. In: *arXiv preprint arXiv:2402.05672* (2024).

[26] Liang Wang et al. “Text embeddings by weakly-supervised contrastive pre-training”. In: *arXiv preprint arXiv:2212.03533* (2022).

[27] Konrad Wojtasik et al. “BEIR-PL: Zero Shot Information Retrieval Benchmark for the Polish Language”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 2149–2160.

[28] Xinyu Zhang et al. “Miracl: A multilingual retrieval dataset covering 18 diverse languages”. In: *Transactions of the Association for Computational Linguistics* 11 (2023), pp. 1114–1131.

[29] Dmitry Zmitrovich et al. “A Family of Pretrained Transformer Language Models for Russian”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. 2024, pp. 507–524.

7 Additional metrics

7.1 MAP

Mean Average Precision (MAP) is used to assess the overall precision of a retrieval system across multiple queries. It computes the average precision for each query and then takes the mean across all queries. MAP provides a single summary measure that reflects both the ranking quality and the system's ability to retrieve relevant documents.

The MAP@10 results obtained from the models' inference on the benchmark datasets are shown below.

Model (→)	Lexical	Dense							Re-ranking		
Dataset (↓)	BM25	mE5-large	mE5-base	mE5-small	BGE-M3	USER-BGE-M3	RoSBERTa	LaBSE	BM25+BGE	mE5-large+BGE	BGE+BGE
rus-NFCorpus	12.52	11.39	9.23	9.37	11.40	10.99	10.02	5.74	**13.47**	12.62	12.33
rus-ArguAna	32.76	40.71	31.69	32.08	41.85	37.57	40.32	20.40	45.32	**45.48**	45.12
rus-SciFact	61.47	59.76	58.84	55.67	57.82	53.60	49.25	25.91	67.10	**67.72**	66.25
rus-SCIDOCS	8.03	7.53	6.69	5.89	8.64	8.29	8.27	4.45	8.88	9.28	**9.37**
rus-MMARCO	11.88	28.11	24.94	23.82	24.03	22.59	16.03	07.07	21.30	**30.78**	29.13
rus-MIRACL	18.61	56.64	50.90	48.00	60.52	57.11	42.36	10.94	35.79	67.24	**67.77**
rus-XQuAD	95.04	96.11	94.63	94.37	94.81	94.35	92.24	65.84	98.57	98.64	**98.64**
rus-XQuAD-Sentences	79.32	85.89	83.15	82.03	83.80	81.99	79.32	71.12	88.44	**90.08**	89.75
rus-TyDi QA	30.16	51.78	48.79	48.35	51.02	50.90	44.88	22.91	46.13	**59.50**	59.18
SberQuad-retrieval	58.36	57.43	55.95	50.94	60.25	58.81	55.38	30.49	**60.84**	59.96	58.90
ruSciBench-retrieval	27.07	39.31	34.50	31.50	43.30	41.47	33.72	12.48	40.43	54.74	**58.12**
ru-facts	90.03	91.66	91.30	90.66	91.79	91.60	91.47	90.70	90.24	90.39	**90.39**
RuBQ	29.36	66.24	61.94	60.95	63.84	62.29	58.72	24.30	51.64	**70.10**	69.25
Ria-News	60.41	75.94	65.67	65.59	79.94	80.62	75.44	57.79	76.75	84.17	**84.74**
wikifacts-articles	78.60	65.96	55.80	60.32	68.32	73.51	67.09	37.95	**80.44**	78.59	79.34
wikifacts-para	50.67	42.50	39.54	26.52	44.00	46.87	40.43	10.01	**56.71**	51.09	54.11
wikifacts-sents	24.45	29.53	22.52	16.17	27.44	25.01	29.84	18.15	30.50	28.38	28.60
Avg	45.22	53.32	49.18	47.19	53.69	52.80	49.10	30.37	53.68	58.75	**58.88**

Table 4: Performance comparison across different models and datasets. The best results for each dataset are in bold; the results of the best single models are underlined.

7.2 Recall

Recall quantifies the proportion of relevant documents that are successfully retrieved by the system. It is defined as the ratio of the number of relevant documents retrieved to the total number of relevant documents available. In the context of information retrieval, high recall is crucial to ensure that the system does not miss important information.

The Recall@10 results obtained from the models' inference on the benchmark datasets are shown below.

Model (→)	Lexical	Dense							Re-ranking		
Dataset (↓)	BM25	mE5-large	mE5-base	mE5-small	BGE-M3	USER-BGE-M3	RoSBERTa	LaBSE	BM25+BGE	mE5-large+BGE	BGE+BGE
rus-NFCorpus	16.09	15.68	12.56	12.79	14.93	14.56	13.17	8.57	**16.69**	15.59	14.97
rus-ArguAna	69.70	75.82	64.30	63.87	79.16	75.32	78.52	42.11	76.81	81.01	**81.65**
rus-SciFact	76.63	76.88	76.42	73.46	75.08	70.90	66.61	37.71	79.39	**80.88**	78.58
rus-SCIDOCS	14.48	14.14	12.80	11.14	15.59	14.88	15.34	8.33	15.66	16.34	**17.02**
rus-MMARCO	25.77	52.38	46.68	45.36	46.53	44.42	33.02	15.26	32.32	**55.90**	50.90
rus-MIRACL	31.32	76.70	71.03	68.43	79.59	76.44	63.69	21.16	39.28	81.81	**82.59**
rus-XQuAD	99.58	99.75	99.50	99.50	99.41	99.41	98.91	82.02	99.66	**99.92**	**99.92**
rus-XQuAD-Sentences	91.78	97.44	96.09	95.76	96.30	95.88	95.13	88.40	94.31	**98.07**	97.48
rus-TyDi QA	51.26	79.03	75.34	73.94	78.31	76.55	71.56	42.43	59.07	**83.88**	82.80
SberQuad-retrieval	96.47	93.47	92.14	90.71	91.94	91.42	88.15	58.84	**97.32**	96.29	94.70
ruSciBench-retrieval	38.63	53.98	47.15	45.40	57.87	54.68	46.91	19.23	45.64	62.51	**66.62**
ru-facts	99.82	**100.00**	**100.00**	99.96	**100.00**	**100.00**	99.96	**100.00**	99.82	**100.00**	**100.00**
RuBQ	52.71	86.26	83.10	81.20	84.06	83.18	81.29	41.47	62.00	**88.84**	86.90
Ria-News	77.84	90.30	84.50	83.73	92.34	92.42	89.41	73.41	82.19	92.44	**93.24**
wikifacts-articles	88.39	77.49	71.44	77.90	82.39	86.72	82.99	53.21	**90.64**	85.84	88.98
wikifacts-para	67.30	60.51	57.65	41.44	62.87	65.81	59.37	17.87	**70.17**	62.48	67.77
wikifacts-sents	35.42	40.65	32.28	24.63	39.89	37.39	**43.22**	26.85	39.83	40.10	41.54
Avg	60.78	70.03	66.06	64.07	70.37	69.41	66.31	43.35	64.75	73.05	**73.27**

Table 5: Performance comparison across different models and datasets. The best results for each dataset are in bold; the results of the best single models are underlined.

Generating Encyclopedic Articles Based on a Collection of Scientific Publications

George Lee
Lomonosov Moscow State University
lee.george@mail.ru

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Alexey Khokhlov
Lomonosov Moscow State University
khokhlov@polly.phys.msu.ru

Abstract

Generating texts that demand high factual accuracy and strict formatting, such as encyclopedic articles, presents numerous challenges: how and where to gather relevant information, how to structure it into a coherent and well-formatted text, and how to ensure that the compiled article does not contain factual errors. We propose a solution to these problems for the Russian language by developing a system for generating encyclopedic articles that extracts the most recent and relevant knowledge from scientific publications in the online library eLIBRARY.RU and structures it as a single context for input into a generative model. To evaluate both the impact of the extracted knowledge on the content of the final texts and the overall quality of generation, we considered several prompting strategies, some of which do not use the context found in publications, and compared these approaches using automatic metrics and human expert evaluation. We hope that the created framework will become a reliable reference material for scientists researching new and relevant topics in their field of expertise.

Keywords: Encyclopedic Article Generation, Scientific Publications, eLIBRARY.RU, Retrieval-Augmented Generation, Large Language Models

DOI: 10.28995/2075-7182-2025-23-206-223

Порождение энциклопедических статей по коллекции научных публикаций

Ли Г. Ф.
МГУ им. Ломоносова
Москва, Россия
lee.george@mail.ru

Лукашевич Н. В.
МГУ им. Ломоносова
Москва, Россия
louk_nat@mail.ru

Хохлов А. Р.
МГУ им. Ломоносова
Москва, Россия
khokhlov@polly.phys.msu.ru

Аннотация

Порождение столь требовательных по фактологической точности и строгости оформления текстов, как энциклопедические статьи вызывает множество трудностей: как и откуда собирать релевантную информацию, как структурировать ее в связный и грамотно оформленный текст, и как удостовериться, что составленная статья не содержит фактических ошибок. Мы предлагаем решение этих проблем для русского языка, разработав систему генерации энциклопедических статей, извлекающую самые последние и релевантные знания из научных публикаций онлайн-библиотеки eLIBRARY.RU и структурирующую их как единый контекст на вход генеративной модели. В целях оценки как влияния извлеченных знаний на содержание итоговых текстов, так и качества генерации в целом мы рассмотрели несколько стратегий промптирования, часть которых не использует найденный в публикациях контекст, и сравнили эти подходы на автоматических метриках и оценках людей-экспертов. Мы надеемся, что созданный фреймворк сможет стать надежным справочным материалом для ученых при исследовании новых и актуальных тем в области их экспертизы.

Ключевые слова: порождение энциклопедических статей, научные публикации, e.LIBRARY.RU, поисковая дополненная генерация, большие языковые модели

1 Introduction

Encyclopedic articles are a valuable source of information, storing knowledge in a concise but informative form. Currently, a wide variety of online encyclopedias have been created and maintained, the most famous and comprehensive of which is Wikipedia. However, articles for such Internet sources are written and edited by people, which creates difficulties in timely updating them with current knowledge.

A solution to this problem is the creation of systems for automatically compiling encyclopedic articles. Over the past decades, many works (Sauper and Barzilay, 2009; Liu et al., 2018; Fan and Gardent, 2022) have been published describing such frameworks specifically for Wikipedia due to its useful properties:

- Huge size of the knowledge base — a large volume of data for creating training collections;
- Clearly defined section structure and the presence of an official guide for formatting article texts;
- Active community of editors, which allows for a more objective assessment of the system's output quality.

One of the important sources for forming Wikipedia articles is scientific articles. It can take a considerable amount of time from the start of active use of a scientific term in a particular field until a Wikipedia article about that term appears, which may not be sufficiently detailed. This work investigates the problem of generating an encyclopedic article for terms not present in Wikipedia based on a collection of scientific articles. At the first stage of the collection analysis, target terms are identified. Publications and sentences containing the term are then extracted. At the final stage, the extracted sentences are fed into generative models to generate the article. In the study, we evaluated several language models and prompts for generation. This task formulation can be described as the Retrieval-Augmented Generation (RAG) task, which aims to improve the quality of content generation by generative language models based on found relevant information.

2 Related Work

In 2009, C. Sauper and R. Barzilay developed an algorithm for "smart" data extraction from scientific texts annotated with thematic sections (Sauper and Barzilay, 2009). This approach uses the Yahoo search engine as a knowledge base. The extraction of relevant context occurs through queries containing the target term and popular topics in the subject area of the term. Despite impressive results for its time, the "smart" extraction algorithm has significant limitations regarding the training dataset, requiring that the corpus texts be annotated by topic.

In the work (Banerjee and Mitra, 2016), the WikiWrite system is described to automatically generate full-length Wikipedia articles. This approach retrieves information about a target entity from the Web and then uses classifiers trained on the content of similar articles to assign web-retrieved content to relevant sections of the Wikipedia article. A key innovation is the two-step integer linear programming model, which synthesizes disjointed content into coherent, human-like summaries for each section. Although the authors reported that the majority of created articles were retained in the online encyclopedia, they noted that it was difficult for the Wikiwrite to meet the requirements of encyclopedic tone of Wikipedia.

In 2018, the Google Brain team developed a framework for generating preambles for English-language articles in the style of Wikipedia formatting (Liu et al., 2018). Unlike the previous approach, the system extracts context both through Internet queries and through the reference section available for each Wikipedia article. The main innovation in their research was the use of an optimized generative model based on the Transformer architecture (Vaswani, 2017), capable of processing contexts up to 11 thousand tokens long. It is worth noting that the task of generating preambles is significantly simpler compared to the task of generating complete articles, which we pursue in this study.

A. Fan and C. Gardent conducted research on generating Wikipedia-style biographies (Fan and Gardent, 2022), in which a BART-like model (Lewis, 2019) writes complete articles sequentially, section by section. Distinctive features of their system include a citation module that preserves references to the documents used in the extracted context and generates a literature section based on them, as well as a cached generation method that reduces the amount of repetitive information between adjacent sections. The main drawback of this framework is the BART model's input size limitation of 1024 tokens, which significantly reduces the potential amount of information transmitted through the extracted context.

Currently, approaches to generating encyclopedic articles based on the application of large language models (LLMs) are actively being researched. In work (Gao et al., 2024), LLMs are tested in the task of generating reviews dedicated to well-known terms in natural language processing, such as *word2vec*. The testing is conducted in zero-shot (task description without examples), one-shot (based on a single example article), and descriptive prompt formats, where the main sections of the review are specified and explanations are given for each section. To improve the quality of the generated articles, the publications referenced in the corresponding Wikipedia articles can be used. The study shows that the GPT-4 model generates high-quality reviews comparable to expert reviews. However, it should be noted that in this case, reviews are generated for well-known concepts for which there is a large volume of literature.

In the work (Shao et al., 2024), the STORM system is described to help write high-quality Wikipedia articles. Upon receiving a topic for writing an article, the STORM system queries the LLM to find related topics, retrieves articles on related topics from Wikipedia, and extracts tables of contents from them.

In the work (Balepur et al., 2023), an approach is considered to generate descriptive texts on a given topic using the IRP (Imitate, Retrieve, Paraphrase) method. The first component, Imitator, plans the future content of the text. The second component, Retriever, searches for texts based on the formulated query. The last component, Paraphraser, adjusts the extracted text to match the style of the text being created.

In 2024, J. Zhang et al. conducted the research on generating full-length Wikipedia articles on recent real world events (Zhang et al., 2024). The authors developed three RAG frameworks: Retrieve-then-Read for reranking related documents from the Web and simply reading the top ones for generation, Plan-Retrieve-Read (PRR) which uses LLMs' planning capabilities and a multi-stage reranking strategy to outline and generate articles section by section, and a fine-tuned Retrieve-then-Read strategy for Wikipedia generation. The evaluation of generated articles, conducted using LLM-based metrics, showed that hierarchical-based methods (PRR) can produce more comprehensive content, while fine-tuned methods achieve better verifiability. Unlike their study, our system uses a different two-stage reranking technique, experiments with a single RAG framework but multiple prompting strategies and utilizes human expert evaluation since LLMs are known to have certain biases when assessing machine-generated texts (Gao et al., 2024).

The novelty of our paper lies in the pioneering generation of Wikipedia articles for the Russian language and exploring dictionary entry generation in the constrained context of scientific publications.

3 System for Generating Encyclopedic Articles

Our study is based on the following principles:

- *Use of Russian-language scientific publications as an information source.* Presumably, the first information about new terms appears in scientific articles, where the authors describe these concepts and set the context for their use. Extracting information from original publications is a challenging task due to the lack of preliminary filtering and formatting by subject area experts, but it significantly speeds up the knowledge accumulation process by eliminating the need for such experts;
- *Generation using large language models (LLMs).* Training generative models independently not only requires a large annotated dataset but often falls short in quality compared to using the best available LLMs. Moreover, modern models can follow user instructions to adjust their output and work with contexts of enormous size. Therefore, we decided to train only the knowledge extraction system and, instead of training a generative system from scratch, to configure the optimal prompting strategy for a ready-made model.

The developed generation system can be logically divided into separate stages (see Figure 1).

At the first stage, a collection of scientific articles is gathered, and terms for which an encyclopedic article should be written are extracted.

At the second stage, for the target term, sentences from the collected scientific articles that are most relevant to the encyclopedic article are extracted, i.e. sentences that convey the most general information about the term. Two principles are used to select such sentences: automatic sentence classification based on a specially trained classifier, as well as a special list of markers.

At the third stage, the extracted sentences are fed into a generative language model along with a prompt that formulates the task of generating an encyclopedic article. As a result of the final stage, an encyclopedic article is generated for the given term.

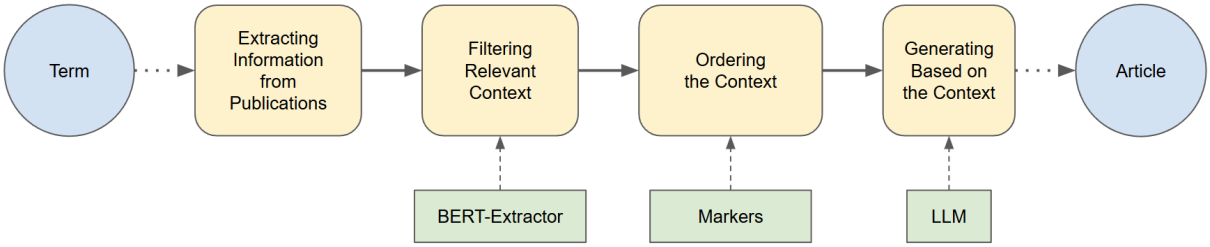


Figure 1: The process of generating an encyclopedic article for a term. Yellow indicates processing stages, green indicates the tools used.

3.1 Sentence Filtering

By a unit of useful information, we mean a separate sentence relevant to the input term. To assess relevance, all candidate sentences from the corpus must first be extracted and processed. The general algorithm for finding sentences can be described as follows:

1. *Selection of articles mentioning the input term*;
2. *Splitting the content of suitable publications into separate sentences*;
3. *Selection of sentences with a given syntactic structure.* For further processing, sentences with a verb predicate and sentences resembling definitions in structure (X is Y) are selected;
4. *Cleaning sentences of insignificant fragments.* Insignificant fragments include section headings, formulas, and hyperlinks, i.e., elements that do not carry significant information but complicate the determination of sentence relevance. Detection of such fragments is done using regular expressions.

The output of the above algorithm is a list of cleaned sentences potentially relevant to the input term. Next, we will consider these stages in more detail.

3.2 Selection of Sentences Relevant for the Encyclopedic Article

Sentences mentioning a term can convey general information about it, useful for inclusion in an encyclopedic article (hereinafter referred to as *encyclopedic sentences*), or specific information related to a particular study. To select encyclopedic sentences, automatic classification based on a transformer neural network model is used. To train the classifier, it was necessary:

1. to create a training dataset based on a given collection of publications;
2. to determine the most suitable model for this dataset and the task.

Training Dataset

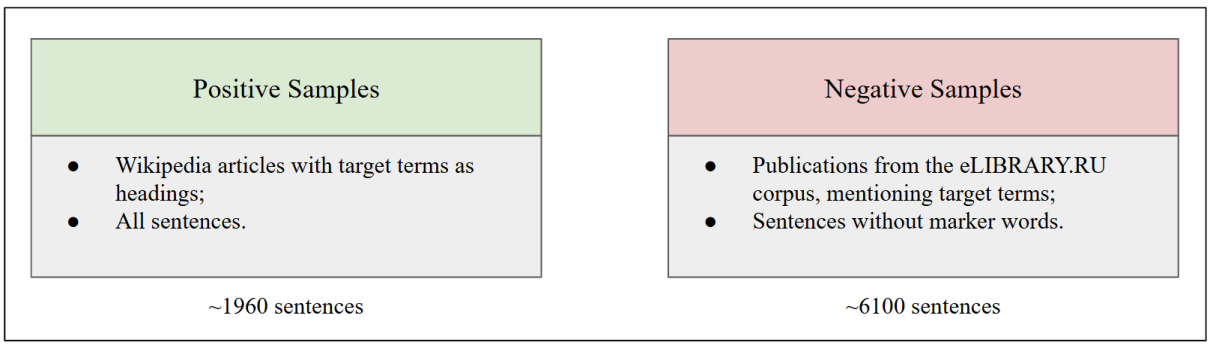


Figure 2: Training data for the task of classifying encyclopedic sentences.

The created dataset consists of automatically annotated sentences for 50 terms that have articles in Wikipedia. Positive sentences are taken from the corresponding Wikipedia articles. The set of negative sentences was formed from scientific articles containing these terms (see Figure 2).

We assumed that the majority of sentences from scientific articles containing the target term relate to a specific study rather than to encyclopedic information. Thus, sentences from articles constitute the set of negative examples for classification. Sentences that potentially could contain encyclopedic information were not included in the negative set. Such sentences usually contain certain markers for conveying basic information about a term and the corresponding concept, i.e. definitin, components and structure, purpose and use, comparisons, as well as markers emphasizing the general significance of the statement contained in the sentence (see Table 1).

Category	**Vocabulary (# words)**	**Representatives**
Definitions	11	вариант, вид, означает, определение, определить, определять, представлять, разновидность, термин, тип, это
Structure	13	включать, включить, входит в состав, компонент, конструкция, модуль, принадлежать, содержат, состав, состоять, структура, часть, элемент
Usage	14	давать возможность, дать возможность, использование, использовать, использоваться, позволить, позволять, предназначить, применение, применить, применять, применяться, служить, удобный инструмент
Distinctive Features	19	больше, больший, меньше, меньший, недостаток, неэффективность, неэффективный, особенность, отличие, отличительная особенность, похож, преимущество, сравнение, сравнивать, сравнительный, сравнить, сходство, эффективность, эффективный
Additional Information	15	В настоящее время, известно, как правило, как принято, общеизвестно, общепринятый, обычно, по большей части, следует считать, считается, таким образом, традиционно, традиционный, часто, чаще всего
Total	**76**	

Table 1: Marker words for encyclopedic sentences, divided by semantic categories.

As a result, a training dataset of 8060 sentences was obtained, in which 1960 sentences are positive examples (filtered sentences from Wikipedia articles), and 6100 sentences are negative examples (filtered sentences from scientific articles). This dataset can be significantly expanded in the future based on the analysis of other terms and corresponding scientific publications and Wikipedia articles.

Comparison of Classifiers for Extracting Encyclopedic Sentences

On the collected training dataset, the quality of several models from the BERT family was evaluated:

- *DeepPavlov/rubert-base-cased* — a model from DeepPavlov based on the multilingual BERT-base, trained on Russian Wikipedia and news data (Kuratov and Arkhipov, 2019);
- *ai-forever/ruSciBERT* — a Russian version of SciBERT (Beltagy et al., 2019), developed by Sber AI in collaboration with the Laboratory of Machine Learning and Semantic Analysis at the AI Institute of Moscow State University (MLSA Lab) for analyzing scientific publications in Russian (Gerasimenko et al., 2022);
- *ai-forever/ruRoberta-large* — a Russian version of the RoBERTa model (Liu, 2019) from the SberDevices team (Zmitrovich et al., 2023);
- *mlsa-iai-msu-lab/sci-rus-tiny* — a small RoBERTa architecture model developed by MLSA Lab for obtaining embeddings of scientific texts in Russian and trained on data from the eLIBRARY.ru

portal;

- *mlsa-iai-msu-lab/sci-rus-tiny3* — an improved and closed version of the previous model.

The characteristics of the classifiers are provided in Appendix A.

The main goal of this stage is to obtain a sentence classifier with minimal error in identifying positive sentences, i.e. sentences carrying encyclopedic information. Within the study, two classification tasks were considered:

1. *Unconditional* — only the sentence itself is fed into the language model, and the model estimates the probability of this sentence to belong to encyclopedic article texts;
2. *Conditional* — as additional information, the target term is fed into the model along with the sentence, separated by the SEP token: ⟨*target term*⟩[**SEP**]⟨*sentence*⟩. In this task, the language model predicts the probability that the sentence belongs to an encyclopedic article dedicated to the target term.

In both cases, contextualized embeddings are generated for the input tokens and the CLS token. For classification, the models *DeepPavlov/rubert-base-cased*, *ai-forever/ruSciBERT*, and *ai-forever/ruRoberta-large* use only the obtained representation of the CLS token, while *mlsa-iai-msu-lab/sci-rus-tiny* and *mlsa-iai-msu-lab/sci-rus-tiny3*, trained to generate semantic embeddings of scientific texts, create an embedding of the entire sentence.

Model	Precision (unconditional)	Precision (conditional)
DeepPavlov/rubert-base-cased	0.9499	**0.9956**
ai-forever/ruSciBERT	0.9618	0.9916
ai-forever/ruRoberta-large	**0.9833**	0.9941
mlsa-iai-msu-lab/sci-rus-tiny	0.9649	0.9642
mlsa-iai-msu-lab/sci-rus-tiny3	0.9338	0.945

Table 2: Precision estimates of various classifiers in the encyclopedic sentences classification task.

It should be noted that the quality of the generative module directly depends on the quality of the extracted sentences, so the precision of classification is a much more significant indicator than its recall. Thus, we can determine the best model using two precision metrics, each corresponding to its classification task (see Table 2).

As can be observed, the *ai-forever/ruRoberta-large* model showed better results than models pretrained on scientific corpora (*ai-forever/ruSciBERT*, *mlsa-iai-msu-lab/sci-rus-tiny*, *mlsa-iai-msu-lab/sci-rus-tiny3*), which can be explained by its size. The final system uses this model because of its high scores on both metrics.

Ordering Encyclopedic Sentences

The extracted encyclopedic sentences can be fed into the generative model without changes. However, we decided to additionally modify the context by ordering the sentences according to the previously mentioned marker words. Presumably, the presence of narrative order in the input context will help the LLM better focus on important fragments for generation.

The sorting of sentences occurs in two stages:

1. *By the semantic segments of the present marker words (primary sorting).* Each semantic segment represents a section of the encyclopedic article. Structurally, it can be described as the following sequence: *definitions, structure, usage, distinctive features, additional information*. As can be seen, the most general categories that can suit terms of different types and different subject areas are chosen as sections. The specified order of sections is used in the primary sorting. If no marker word is encountered in a sentence, it is added to the very end of the sorted sequence;
2. *By the confidence level of the extractor model (secondary sorting).* Almost always after the first stage, segments with many suitable sentences are obtained. To order the context within the sections, a secondary sorting by descending classifier confidence is used.

The relevant sentences ordered in this way will be called the RAG-context of the target term. An example of the joint work of the encyclopedic sentence extraction model and the sorting procedure is provided in Appendix B.

3.3 Context-Based Generation

The responses of the generative RAG-system depend not only on the RAG-context fed into it but also on the used large language model and the type of instructions (prompts) given to the model. The prompt sets the layout of the final text, establishing requirements for content, narrative style, and formatting. The language model uses this layout to generate a response, based on its own and RAG-knowledge, and adhering to the limitations set by its developers. Next, we will consider the LLMs used in the study and the types of prompts.

Used LLMs

We evaluated the quality of generation among three large language models: *Claude-3-sonnet* (Anthropic, 2024), *Meta-Llama-3-70B-Instruct* (Dubey et al., 2024), *Mixtral-8x22b-instruct-v0.1* (Team, 2024). All of them are medium-sized models, providing a balance between speed and quality.

Feature	Claude-3 (Sonnet)	Llama-3 (70B)	Mixtral (8x22b)
Parameter Count	?	70 B	141 B
Context Window (# tokens, 10^3*)*	200	8.192	64
Knowledge Cutoff	August 2023	December 2023	October 2021
Multimodality	+	-	-
Open Source	-	+	+

Table 3: Features of the used LLMs.

The used LLMs were chosen due to their significantly varying properties (see Table 3), which allowed for a more diversified comparative output and thus simplified the identification of certain patterns.

It is worth mentioning that all the presented models are hosted on different cloud platforms, which may indirectly affect the generated articles. For instance, services for Claude-3 and Mixtral forcibly limit the context window of the models to several thousand tokens, which can lead to unwanted shortening of long articles, while the Llama-3 service does not have such a limitation. Moreover, Llama-3 is the only LLM in our work the workflow of which could be automated through the GradioAPI.

Prompt Instructions Types

Each prompt strategy was created to test some aspect of generation (see Table 4):

- *Basic (Simple prompt)* — general knowledge of the LLM about the target term;
- *Detailed (Detailed prompt)* — model's ability to change the output format according to the user's needs and structure knowledge by sections;
- *Basic/Detailed + RAG* — combining the general knowledge of the LLM with knowledge extracted from scientific articles;
- *Basic + Russian language* — controlling the LLM's proneness to switching to its "default language" when responding to unfamiliar topics;
- *Encyclopedic (Encyclopedic prompt)* — semantic consistency and content completeness of sentences in the RAG-context;
- *Encyclopedic + target term* — model's ability to adapt information from the RAG-context to the target term;
- *Freestyle (Freestyle prompt)* — proficiency in composing a coherent text from the RAG-context.

4 Experiments

4.1 Quality Metrics

The quality of the generated texts can be assessed in different ways. In this work, we used two types of automatic metrics: *absolute* and *relative*. Absolute indicators describe the characteristics of the text, while relative indicators compare the model's prediction with some reference:

Prompt Type	Contents
Basic *+RAG* *+forced lang.*	Твоя задача - ответить на запрос пользователя.[Пиши на русском языке.][Релевантная информация: <RAG-context>] Расскажи о следующем термине: <target term>.
Detailed *+RAG*	Ты - большой эксперт в области физики и химии. Твоя задача - написать основную информацию о запрашиваемом термине. Структурируй ответ по разделам: Определение, Структура, Использование, Отличительные особенности, Дополнительные сведения. Придерживайся указанного порядка разделов. Разделы не являются обязательными. Если считаешь, что на какой-то раздел недостаточно информации, не пиши его. Главный критерий качества твоего ответа - фактологическая точность. Пропускай то, в чем не уверен, не лги пользователю. Старайся писать на русском языке.[При ответе помимо собственных знаний используй предоставленную релевантную информацию. Если считаешь, что какая-то часть релевантной информации не подходит, не используй ее. Релевантная информация: <RAG-context>] Расскажи о следующем термине: <target term>.
Encyclopedic *+target term*	Построй текст энциклопедической статьи на основе заданных предложений.[Статья должна быть посвящена следующему термину: <target term>.] Предложения: <RAG-context> Ответ:
Freestyle	Собери связный текст из заданных предложений. Предложения: <RAG-context> Ответ:

Table 4: Types of prompt instructions and their contents.

- Absolute metrics:
  - *Generation length* — the average number of words in the generated articles;
  - *Latin symbols count* — the average number of Latin characters in the texts, shows how often the LLM switches to its "default language", i.e., English;
- Relative metrics:
  - *ROUGE-Precision* (Lin, 2004) — similarity between two texts based on matching parts.

Matching parts can be:
  * unigrams → *ROUGE-1*;
  * bigrams → *ROUGE-2*;
  * the longest common subsequence → *ROUGE-L* at the level of the entire text, *ROUGE-Lsum* at the level of individual sentences;

Relative metrics were measured when comparing with three variants of reference texts:

1. *The full text of the corresponding Wikipedia article* — similarity between the generated and editor-written articles;
2. *RAG-context* — the activity of using the information extracted from the corpus by the model;
3. *Publications from the eLIBRARY.RU corpus mentioning the target term* — the degree of correspondence of the generation to the latest knowledge in the subject area.

4.2 Input Data Description

The corpus of publications was composed of 15000 scientific articles in physics, downloaded from the electronic library eLIBRARY.RU. During the analysis of the collection, about 4000 articles turned out to be written not in Russian (mainly in English and less often in French) and were removed from the corpus.

From the remaining collection, based on statistical indicators, 20 candidate terms of average frequency were extracted (see Table 5).

Wiki Terms	Non-Wiki Terms
1. acoustic emission;	1. pump wave;
2. quantum dot;	2. dynamic recrystallization;
3. internal waves;	3. ionospheric turbulence;
4. Rayleigh wave;	4. quantum beats;
5. gauge bosons;	5. optical rectification;
6. radioactive tracers;	6. parasitic resonance;
7. singlet oxygen;	7. electronic concentration;
8. Stokes number;	8. chaotic synchronization;
9. Rayleigh number;	9. memory function;
10. heavy water.	10. impact layer.

Table 5: Candidate terms for evaluation. Half of them do not have a corresponding Wikipedia article.

4.3 Results

The evaluation was performed for selected terms. The examples of generated articles are provided in Appendix C. The statistics for each variant of the reference text are presented in separate tables, the best indicators for each individual model are underlined, and the best among all three LLMs are highlighted in bold.

Table 6 shows the absolute statistical indicators of the generated articles: the length of the generated text and the number of Latin characters in the texts. From this table, it can be seen that:

- The use of RAG-context and both requirements (to write the text in Russian and for the target term) reduces the size of the final article, regardless of the model and prompt type;
- LLM's occasional switching to the "default language" can be eliminated with the help of RAG-context as well as the explicit requirement to write in Russian;
- On average, Mixtral generates the shortest texts. This is explained by the fact that Claude-3 and Llama-3 usually additionally format their responses (for example, breaking articles into semantic sections and developing each segment into a whole paragraph). Thus, in the case of a detailed prompt, where formatting rules are explicitly set, the difference between the models becomes less noticeable;
- Llama-3 is more prone to switching to English when responding to unfamiliar topics than other models — so, in the case of a basic prompt, it can write the entire article text in English.

Configuration		Generation Length	Latin Count
Claude-3 (Sonnet)	*Basic*	233.85	5.75
	Basic+Lang	213.95	1.15
	Basic+RAG	193.15	8.9
	Detailed	199.9	7.7
	Detailed+RAG	196.6	4.55
	Encyclopedic	239.55	0.5
	Encyclopedic+Term	218.9	0.85
	Freestyle	231.4	0.3
Llama-3 (70B)	*Basic*	246.6	225.35
	Basic+Lang	241.35	20.7
	Basic+RAG	156.6	3.35
	Detailed	**255.85**	16.55
	Detailed+RAG	217.75	1.5
	Encyclopedic	233.8	0.85
	Encyclopedic+Term	226.3	3.9
	Freestyle	249.75	0.7
Mixtral (8x22B)	*Basic*	146.9	0.55
	Basic+Lang	106.15	0.4
	Basic+RAG	138.5	0.45
	Detailed	245.3	1.3
	Detailed+RAG	215.3	1.5
	Encyclopedic	212.4	**0.2**
	Encyclopedic+Term	199.85	0.85
	Freestyle	155.6	0.3

Table 6: Absolute metrics of the generated articles.

Configuration		ROUGE			
		1	2	L	Lsum
Claude-3 (Sonnet)	*Basic*	0.575	0.1786	0.3091	0.5465
	Basic+Lang	0.5773	0.1838	0.3158	0.5441
	Basic+RAG	0.6049	0.2021	0.3307	**0.5771**
	Detailed	0.5808	0.1742	0.3074	0.5533
	Detailed+RAG	0.589	0.1794	0.3109	0.562
	Encyclopedic	0.5351	0.1741	0.2798	0.5003
	Encyclopedic+Term	0.5457	0.1787	0.2933	0.516
	Freestyle	0.534	0.1725	0.289	0.4982
Llama-3 (70B)	*Basic*	0.5537	0.1868	0.3122	0.526
	Basic+Lang	0.523	0.1628	0.2967	0.4981
	Basic+RAG	0.5821	0.2136	0.3496	0.555
	Detailed	0.507	0.1679	0.2867	0.486
	Detailed+RAG	0.509	0.1712	0.2811	0.4828
	Encyclopedic	0.519	0.1639	0.2765	0.4937
	Encyclopedic+Term	0.5315	0.1784	0.2953	0.5052
	Freestyle	0.5144	0.1538	0.2724	0.4819
Mixtral (8x22B)	*Basic*	0.599	0.21	0.3744	0.5673
	Basic+Lang	**0.6055**	0.2008	0.379	0.5741
	Basic+RAG	0.5972	**0.2157**	**0.3876**	0.5652
	Detailed	0.4919	0.1484	0.2765	0.4738
	Detailed+RAG	0.5469	0.1821	0.3136	0.5225
	Encyclopedic	0.5548	0.1947	0.306	0.5226
	Encyclopedic+Term	0.5331	0.1839	0.2961	0.501
	Freestyle	0.5551	0.181	0.3268	0.5219

Table 7: Quality assessment of generation when compared with Wikipedia articles.

Configuration		ROUGE			
		1	2	L	Lsum
Claude-3 (Sonnet)	*Basic*	0.3144	0.0649	0.1665	0.2863
	Basic+Lang	0.3756	0.1156	0.2039	0.3427
	Basic+RAG	0.3949	0.104	0.1981	0.3613
	Detailed	0.378	0.1048	0.1986	0.3469
	Detailed+RAG	0.4139	0.1421	0.2161	0.3832
	Encyclopedic	0.524	0.3393	0.314	0.4994
	Encyclopedic+Term	0.6003	0.4443	0.3552	0.5816
	Freestyle	0.6932	0.6529	0.4334	0.6877
Llama-3 (70B)	*Basic*	0.2943	0.0592	0.1638	0.2681
	Basic+Lang	0.3104	0.0653	0.1782	0.2833
	Basic+RAG	0.6277	0.3868	0.4195	0.5958
	Detailed	0.2974	0.0637	0.1621	0.2697
	Detailed+RAG	0.5119	0.314	0.31	0.4829
	Encyclopedic	0.6755	0.5593	0.5103	0.6593
	Encyclopedic+Term	0.5538	0.3834	0.3821	0.5353
	Freestyle	0.9429	**0.935**	**0.8159**	0.9428
Mixtral (8x22B)	*Basic*	0.3847	0.0787	0.2283	0.351
	Basic+Lang	0.4258	0.086	0.2585	0.3915
	Basic+RAG	0.5493	0.3097	0.3716	0.5137
	Detailed	0.2994	0.0597	0.1708	0.274
	Detailed+RAG	0.4518	0.2345	0.2753	0.4229
	Encyclopedic	0.6802	0.5516	0.485	0.6634
	Encyclopedic+Term	0.6463	0.4748	0.4593	0.623
	Freestyle	**0.9653**	0.9074	0.6459	**0.9595**

Table 8: Quality assessment of generation when compared with RAG-contexts.

Configuration		ROUGE			
		1	2	L	Lsum
Claude-3 (Sonnet)	*Basic*	0.9492	0.6162	0.7783	0.943
	Basic+Lang	0.9594	0.6588	0.7977	0.9527
	Basic+RAG	0.973	0.6795	0.8261	0.9676
	Detailed	0.9682	0.6569	0.8157	0.9638
	Detailed+RAG	0.9744	0.6615	0.8177	0.9712
	Encyclopedic	0.9911	0.7821	0.8243	0.9869
	Encyclopedic+Term	0.9957	0.8265	0.8545	0.9926
	Freestyle	0.9868	0.8811	0.8496	0.9852
Llama-3 (70B)	*Basic*	0.8654	0.517	0.7227	0.8584
	Basic+Lang	0.9485	0.5788	0.7909	0.9422
	Basic+RAG	0.9831	0.7864	0.8632	0.9801
	Detailed	0.9488	0.5466	0.7648	0.9429
	Detailed+RAG	0.9847	0.717	0.8248	0.9821
	Encyclopedic	0.9936	0.8639	0.8438	0.9922
	Encyclopedic+Term	0.9855	0.7777	0.8334	0.9828
	Freestyle	0.9808	0.9534	0.847	0.9784
Mixtral (8x22B)	*Basic*	0.9658	0.6448	0.8518	0.9621
	Basic+Lang	0.9758	0.6552	0.8924	0.9689
	Basic+RAG	0.992	0.7653	0.8814	0.9886
	Detailed	0.9701	0.6274	0.8163	0.9675
	Detailed+RAG	0.979	0.6989	0.835	0.9761
	Encyclopedic	0.9929	0.8378	0.8558	0.9909
	Encyclopedic+Term	0.9911	0.8309	0.8658	0.9885
	Freestyle	**0.9986**	**0.9693**	**0.9075**	**0.9983**

Table 9: Quality assessment of generation when compared with eLIBRARY.RU publications.

Table 7 shows the similarity measures of the generated texts with Wikipedia articles for terms that have such articles. Table 8 contains the similarity measures of the generated articles with the context extracted from scientific articles. Table 9 shows the similarity measures of the generated articles with the full scientific publications in which the target term was found.

The study of the statistical characteristics of the generated articles revealed the following patterns and properties of LLMs:

- One of the reasons for the high ROUGE scores of Mixtral is the brevity of the texts it generates. This is a known problem of metrics based solely on N-gram precision;
- The use of RAG-context increases the values of ROUGE regardless of the model, slightly when evaluated on Wikipedia articles and significantly on the original publications and the context itself. It is worth noting that adding the requirement to generate for the target term in the encyclopedic prompt improves the metrics only for Claude-3 — this suggests that statistical metrics do not necessarily have to correlate with the level of narrative coherence;
- The encyclopedic and freestyle prompts have the highest scores for ROUGE when evaluated on the RAG-context and original documents, the latter prompting strategy is especially good for these metrics. Such a result is expected, since with this strategy the model essentially compiles a story from the sentences of the RAG-context, simply reordering them and almost not adding new tokens;
- The detailed prompt performed worse than the basic one when evaluated on Wikipedia articles — perhaps the standard generation templates in LLMs are sufficiently similar to the Wikipedia format;
- Claude-3 is the most conservative model, it noticeably uses the proposed context less when generating a response. Perhaps this is influenced by its "closedness" (recall that Llama-3 and Mixtral are open-source models);
- Relatively low ROUGE-2 and ROUGE-L scores when evaluated on Wikipedia articles show that the similarity between the generated and reference articles is observed at most at the level of individual words or short common word sequences. This means that LLMs do not explicitly try to copy Wikipedia when responding to queries.

4.4 Human Evaluation

The main drawback of all used automatic metrics is that they evaluate only the statistical characteristics of texts, which say almost nothing about the factual accuracy and logical coherence of what is written. A solution to this problem is the involvement of human experts into the evaluation (Van Der Lee et al., 2019).

As part of our study, we conducted a survey consisting of the following parts:

1. **parallel comparison (60 questions)** — the expert is offered two articles written for the same target term by different models with different prompting strategies (see Figure 3). The task of the respondent is to choose the text that they liked more. There are no specific quality criteria in this task, i.e., the choice is made solely based on the subjective considerations of the expert;
2. **standalone evaluation (40 questions)** — in each question, the respondent is offered one article (see Figure 4)). The task of the expert is to evaluate it according to three criteria:
    - *factual accuracy* — the correspondence of the article's content to real facts;
    - *semantic consistency* — the coherence and logic of the narrative, the completeness of thought;
    - *compliance with the format of an encyclopedic article* — the degree of similarity (in structure and visual design) of the generated article to real encyclopedic articles written by people.

The experts from the Faculty of Physics at Lomonosov Moscow State University were invited to assess the generated articles: 13 experts for the parallel part and 11 for the standalone one. Based on the assessments obtained during the parallel comparison stage, ratings of the best models and prompting strategies were calculated (see Table 10). The scores of the candidates were calculated according to the following procedure:

- *Respondent's choice of candidate i* — the model and prompt corresponding to it receive 1 point. The opponent j does not receive a win;
- *Tie* — both candidates do not receive points;
- *Both are bad* — both opponents are penalized, receiving -1 point.

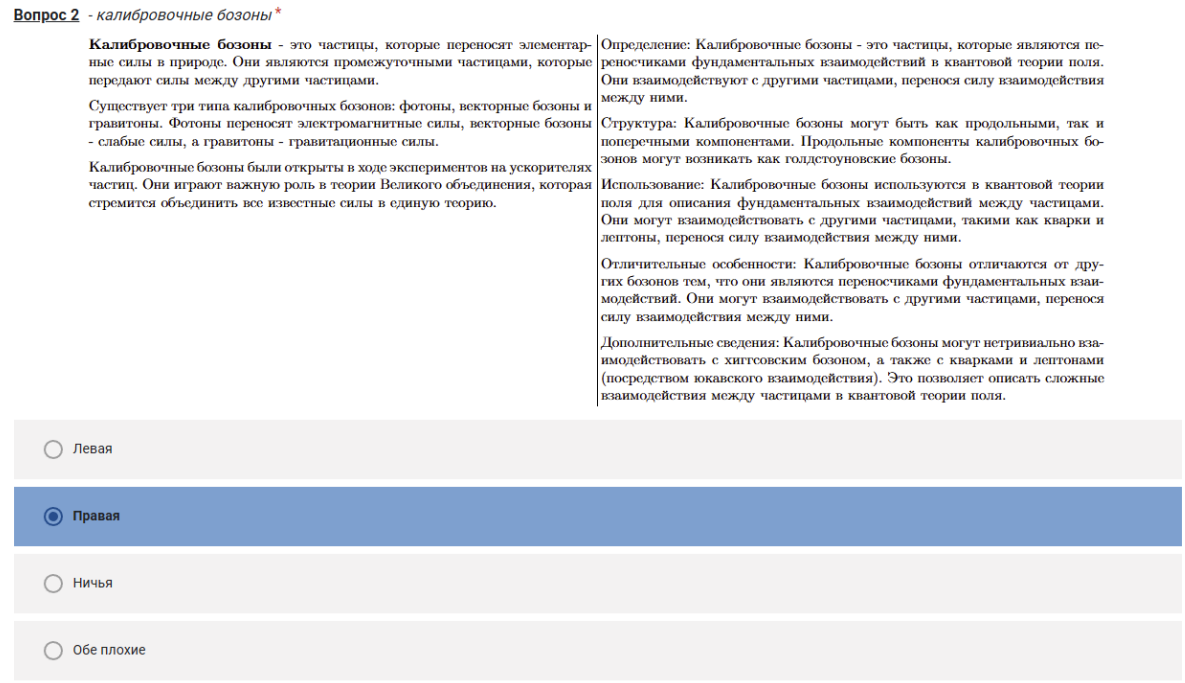


Figure 3: Example of parallel comparison for the term "calibration bosons".

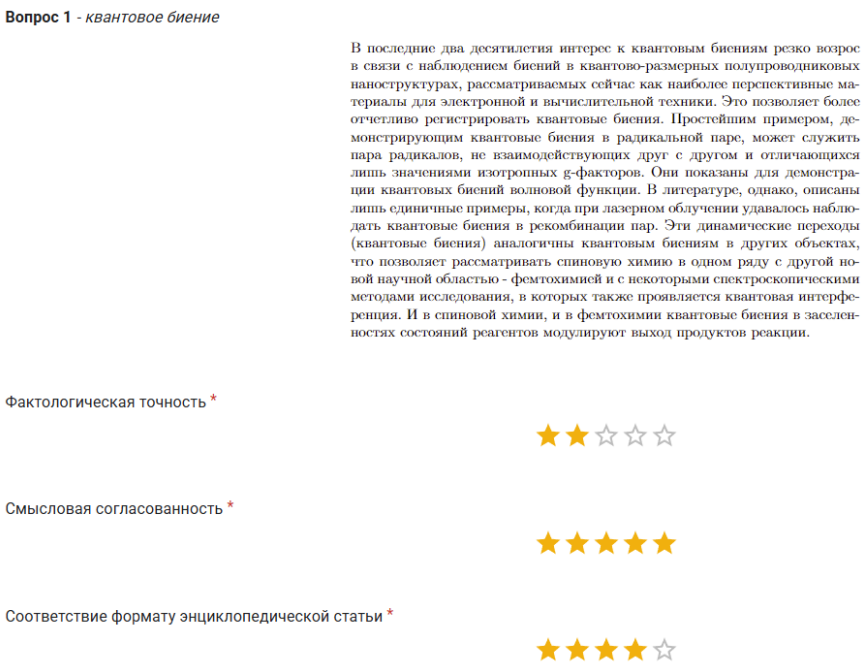


Figure 4: Example of standalone evaluation for the term "quantum beats".

Looking at the Table 10, it becomes clear that the experts significantly preferred the responses of the "closed" Claude-3 (Sonnet) over the responses of open-source models. As for the prompting strategies, the most successful was the detailed prompt. The use of RAG-context generally makes the generated articles more attractive, but does not guarantee this. Surprisingly, encyclopedic prompts received mostly mixed or even negative feedback. It can be assumed that this is due to both insufficient precision in the formulation of these prompts and the lack of explicitly defined text formatting rules in them.

To determine the optimal generation configuration, the characteristics of the top-2 models and top-3 prompts in the ranking were investigated during the standalone evaluation stage (see Table 11).

From the obtained assessments, the following conclusions can be drawn:

- Claude-3 generates more plausible texts, while Llama-3 generates texts that are more similar in style to encyclopedic articles;
- RAG-context improves the accuracy of the model's responses but worsens their logical coherence. This observation can be explained by the semantic heterogeneity of the extracted relevant sentences, which complicates their integration into a single text.

Rating	Model	Score
1	**Claude-3 (Sonnet)**	**91**
2	Llama-3 (70B)	23
3	Mixtral (8x22b)	-44

Rating	Prompt Type	Score
1	**Detailed**	**50**
2	Basic+RAG	46
3	Detailed+RAG	27
4	Basic	16
5	Basic+Lang	14
6	Encyclopedic+Term	0
7	Encyclopedic	-15
8	Freestyle	-68

Table 10: Models and prompts ratings based on expert evaluation.

Configuration		Factual Accuracy	Semantic Consistency	Format Compliance
Claude-3 (Sonnet)	*Basic+RAG*	4.09	3.45	3.14
	Detailed	4.11	**4.26**	3.82
	Detailed+RAG	**4.36**	4	3.63
Llama-3 (70B)	*Basic+RAG*	3.09	2.91	2.82
	Detailed	3.91	3.82	**3.98**
	Detailed+RAG	3.97	3.51	3.69

Table 11: Results of the standalone evaluation for the best rated configurations.

Thus, the choice of the optimal configuration depends on the importance of each criterion in the task at hand: for the best generation accuracy, Claude-3 with a detailed prompt and RAG-context is suitable; for generating articles that are most similar to encyclopedic ones, Llama-3 with a detailed prompt is recommended; and as a compromise, Claude-3 with a detailed prompt can be chosen.

5 Conclusion

In this work, we developed a system for generating Russian-language encyclopedic articles. The framework is based on the extraction and processing of knowledge from scientific publications of the electronic library eLIBRARY.RU to form a current and relevant context for generative models.

The resulting system can be improved in many ways. Thus, in the future, it is planned to improve the BERT-extractor by including meta-information about the target term in the context (for example, the presence of marker words in the sentence) and replacing the extraction method itself with the "small-to-big" retrieval approach. Both proposals will add additional information for the language model necessary for a more thoughtful and high-quality selection of the RAG-context. We also see the possibility of improving the framework by modifying the optimal prompt strategies — at the moment, the LLM is too inconsistent in its writing style, which is undesirable for scientific articles with fairly strict formatting rules. By adding more specific text formatting rules and several examples of correctly written articles to the prompt, the model will be able to better adjust the generation to real publications.

Acknowledgments

The study was carried out with the financial support of the Interdisciplinary Scientific and Educational School of Lomonosov Moscow State University "Mathematical methods of analysis of complex systems", agreement No. 23-ШО5-21. We thank eLibrary for providing the collections of scientific papers.

References

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*, 1.

Nishant Balepur, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Expository text generation: Imitate, retrieve, paraphrase. *arXiv preprint arXiv:2305.03276*.

Siddhartha Banerjee and Prasenjit Mitra. 2016. Wikiwrite: Generating wikipedia articles automatically. // *IJCAI*, P 2740–2746.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Angela Fan and Claire Gardent. 2022. Generating full length wikipedia biographies: The impact of gender bias on the retrieval-based generation of women biographies. *arXiv preprint arXiv:2204.05879*.

Fan Gao, Hang Jiang, Rui Yang, Qingcheng Zeng, Jinghui Lu, Moritz Blum, Tianwei She, Yuang Jiang, and Irene Li. 2024. Evaluating large language models on wikipedia-style survey generation. // *Findings of the Association for Computational Linguistics ACL 2024*, P 5405–5418.

Nikolai Aleksandrovich Gerasimenko, Aleksander Sergeevich Chernyavsky, and MA Nikiforova. 2022. Ruscibert: A transformer language model for obtaining semantic embeddings of scientific texts in russian. // *Doklady Mathematics*, volume 106, P 95–96. Springer.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Mike Lewis. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. // *Text summarization branches out*, P 74–81.

Peter J Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating wikipedia by summarizing long sequences. *arXiv preprint arXiv:1801.10198*.

Yinhan Liu. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Christina Sauper and Regina Barzilay. 2009. Automatically generating wikipedia articles: A structure-aware approach. // *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, P 208–216.

Yijia Shao, Yucheng Jiang, Theodore A Kanell, Peter Xu, Omar Khattab, and Monica S Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. *arXiv preprint arXiv:2402.14207*.

Mistral AI Team. 2024. Cheaper, better, faster, stronger. `https://mistral.ai/news/mixtral-8x22b/`.

Chris Van Der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. 2019. Best practices for the human evaluation of automatically generated text. // *Proceedings of the 12th International Conference on Natural Language Generation*, P 355–368.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Jiebin Zhang, Eugene J Yu, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Weimin Xiong, Xiaoguang Li, Qun Liu, et al. 2024. Wikigenbench: Exploring full-length wikipedia generation under real-world scenario. *arXiv preprint arXiv:2402.18264*.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, et al. 2023. A family of pretrained transformer language models for russian. *arXiv preprint arXiv:2309.10931*.

Appendix A. Properties of Classifiers and Their Training.

Model	Parameter Count	Batch Size	# Epochs	Unconditional Train Time	Conditional Train Time
DeepPavlov/rubert-base-cased	180 M	**16**	3	2.15 min	2.2 min
ai-forever/ruSciBERT	123 M	8	3	2.85 min	2.92 min
ai-forever/ruRoberta-large	**355 M**	4	3	**14.58 min**	**27.26 min**
mlsa-iai-msu-lab/sci-rus-tiny	23 M	**16**	**10**	1.41 min	1.46 min
mlsa-iai-msu-lab/sci-rus-tiny3	23 M	**16**	**10**	1.46 min	1.53 min

Table 12: Characteristics of extractor models and their training process.

Appendix B. Context Filtering by the Extractor Model.

Segment	Context
Definitions	1. Это оборудование позволяет получать большое количество данных в ходе мониторинга и затем проводить полный анализ сигналов акустической эмиссии, чтобы определить изменение показателя b для всего периода исследований. 2. Акустическая эмиссия (АЭ) - явление, сопутствующее многим физическим процессам в твердом теле. 3. Каждый раз при возникновении акустической эмиссии это событие может привести либо к созданию новой элементарной трещины, либо к распространению существующей трещины на некоторую фиксированную величину (квант разрушения).
Structure	4. Поэтому можно считать, что спектральный состав сигнала акустической эмиссии является структурно чувствительной характеристикой.
Usage	5. Применение этого метода ограничивается возможностями выделения сигналов акустической эмиссии из помех, то есть данный метод чувствителен только на заключительных этапах жизни образца, когда размеры дефектов уже достаточно велики и приближаются к критическим значениям.
Additional Information	6. Акустическая эмиссия обычно проявляется при пластической деформации твердых материалов, но в последнее время появились публикации об использовании этого метода и при изучении физико-химических процессов, проходящих в жидкости.
Unidentified	7. Пассивный метод контроля состояния объекта состоит в использовании сигнала акустической эмиссии, излучаемого при деформации объекта контроля, или сигнала естественного акустического фона окружающей среды, отраженного объектом. 8. Полагаем, что такое деление условно и непрерывная и импульсная акустическая эмиссия представляют поток сигналов, различающиеся амплитудой и частотой появления. 9. В технике широко используется спектральный анализ акустической эмиссии шумов, возникающих при эксплуатации агрегатов и устройств, и эмиссии, возникающей при нагружении материала. 10. Метод акустической эмиссии (АЭ) для исследования свойств материалов получил широкое распространение во второй половине прошлого столетия. 11. Таким образом, трещина выступает в качестве рода, в то время как наименьшее событие акустической эмиссии играет роль вида. 12. Такой подход позволяет эффективно сопоставлять эти параметры, особенно при сравнении структуры алюминия, деформации и акустической эмиссии. 13. Разработана методика определения активационных параметров - энергии активации и активационного объема, основанная на анализе акустической эмиссии, зависимой от температуры или механического напряжения.

Table 13: Selected relevant sentences for the term "acoustic emission".

Appendix C. Generation Samples.

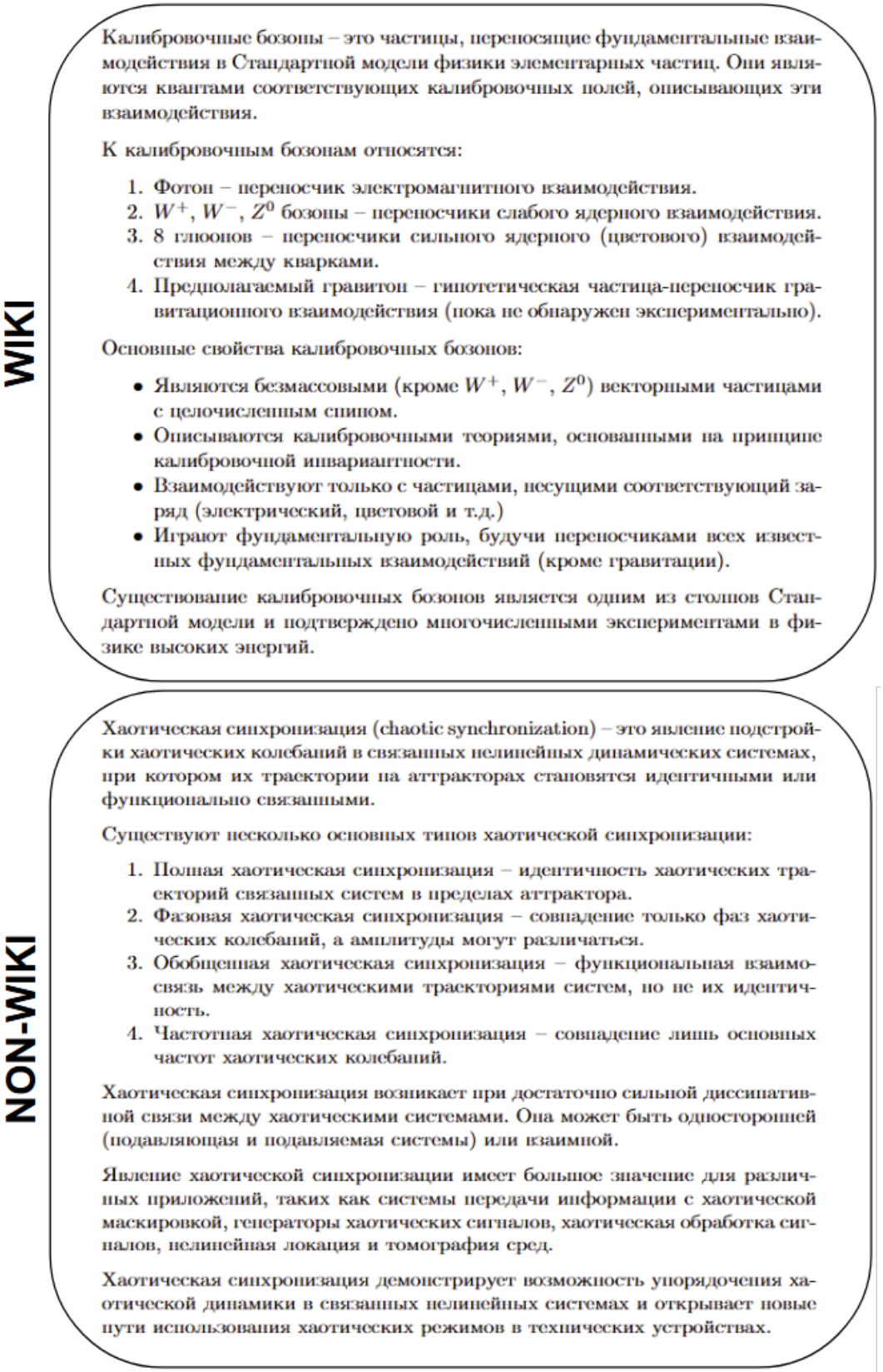


Figure 5: Generated texts for terms with and without corresponding Wikipedia articles. Claude-3 model with Basic+RAG prompt was used as a configuration.

Causal Models and Adversarial Training: Selecting the right properties for robust non-topical text classification

Mikhail Lepekhin
Moscow Institute of Physics and Technology
Russia
lepehin.mn@phystech.edu

Serge Sharoff
University of Leeds
UK
s.sharoff@leeds.ac.uk

Abstract

The vast majority of datasets for non-topical classification of texts contain distribution shifts. In most cases, those are topical shifts. Their presence in the data forces the classifiers to fit topics-related features instead of focusing on those relevant for the target class. It causes a dramatic decrease in the accuracy of the trained models when the test data are taken from a different data source. To address this problem, we experiment with two techniques: causal models and adversarial domain adaptation. In our work, we apply CausalLM, Adversarial Domain Adaptation (ADA), and Energy-based ADA (EADA) for gender classification and compare the results. The results are novel for the non-topical classification task. We show that both causal and adversarial methods manage to make the model more resilient to the distribution shifts although it causes a decrease of accuracy when tested on the domain prevailing in the training dataset. Moreover, we describe the first attempt to reduce the impact of topical shifts in the task of non-topical classification with usage of causal methods. Besides, we provide a link to the GitHub repository with the code of our experiments to ensure their reproducibility: https://github.com/MikeLepekhin/CausalAndAdversarialMethods.

Keywords: adversarial, causal models, gender classification, non-topical classification, bert
DOI: 10.28995/2075-7182-2025-23-224-233

Каузальные модели и состязательные методы: выбор правильных свойств для надежной нетематической классификации текстов

Аннотация

Подавляющее большинство наборов данных для нетематической классификации текстов содержат смещения. В большинстве случаев это тематические сдвиги. Их наличие в данных стимулирует классификаторы выучивать признаки, релевантные для предсказания тем, вместо фокусирования на признаках, относящихся к целевому классу. Это приводит к резкому снижению точности обученных моделей, когда тестовые данные берутся из другого источника данных. Для решения этой проблемы мы экспериментировали с двумя методами: каузальными (причинно-следственными моделями) и состязательными методами доменной адаптации для классификации гендера автора текста. Мы рассматриваем CausaLM, Adversarial Domain Adaptation (ADA) и Energy-Based ADA (EADA) и сравниваем результаты. Мы показываем, что как причинно-следственные, так и состязательные методы позволяют сделать модель более устойчивой к изменениям распределения, хотя это приводит к снижению точности при тестировании на текстах из источника, преобладающего в обучающем датасете. Кроме того, мы предоставляем ссылку на репозиторий GitHub с кодом наших экспериментов, чтобы обеспечить их воспроизводимость.

Ключевые слова: adversarial, каузальные модели, классификация гендера, нетематическая классификация, bert

1 Introduction

Non-topical text classification includes a wide range of tasks aimed at predicting a text property that is not connected directly to a text topic. For instance, predicting a text style, politeness, difficulty level, the age or the first language of its author, etc. Solutions for these tasks are applied in many areas such as information retrieval, language teaching, or linguistic research (Luu and Malamud, 2020).

Unlike topical classification of texts, non-topical classification has a handful of additional difficulties. The most significant of them is that the target variable in these tasks is more complex than a topic. For example, every topic has a set of keywords, and, therefore, one can define whether a text belongs to a topic or not based on the occurrence of the keywords in the text. In contrast, the gender of the author cannot be defined just by a set of keywords. The classification of texts by genres and difficulty level has the same issue.

Complexity of the target variable in tasks of non-topical classification makes it vulnerable to distribution shifts. It implicitly pushes the classifiers to train on the irrelevant data features instead of the relevant ones. One of the most popular issues of non-topical classification of texts is the presence of topical shifts (biases) in the data (Sharoff et al., 2010).

One of the techniques that can potentially mitigate the topical biases of the non-topical text classification is causal models (Feder et al., 2020), (Maiya, 2021) because they have a functionality to make the classifiers more sensitive to the relevant features and to attend less to those that influence both the target variable and the text distribution, causing a spurious association. In causal inference, such features are called *confounders*. For example, (Feder et al., 2020) proposes a causal loss, which contains a negative summand corresponding to the head of the classifier related to the confounder. The efficiency of causal approaches for model de-biasing for the task of sentiment analysis is shown in (Feder et al., 2020). We assume that the approach can also be successful in application to more complex data distribution shifts in non-topical classification tasks.

Another important algorithm is Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017). It uses an adversarial loss to make the classification features less dependent on the domain of the training data. It supposes training a feature extractor, a domain discriminator, and a target classifier. The feature extractor and target classifier are trained to achieve high accuracy for the classification of the target class and at the same time deceive the domain discriminator to make it impossible to differentiate two domains. In contrast, the domain discriminator intends to classify the text domain correctly.

Energy-based ADA or EADA (Zou et al., 2021) is a modification of ADA. The authors show that that energy-based models exhibit free energy biases when training and test data come from different distributions and present a novel loss combined with an active learning strategy to assist knowledge transfer in the target domain, dubbed active domain adaptation. They show that domain gap can be implicitly diminished by aligning the free energy of target data compact around the source domain via a regularization term.

In our work, we compare causal models based on the CausaLM framework, adversarial models based on Adversarial Domain Adaptation (ADA) (Tzeng et al., 2017) and Energy-based ADA (EADA) (Zou et al., 2021) with the baseline of BERT-based models. We show that usage of both causal and adversarial methods helps to increase the accuracy on the dataset under-represented in train and thereby reduce model reliance on the distribution shifts.

In this study we:

1. Show that the BERT classifiers are sensitive to the source of the training data (subsection 6.1);
2. apply causal and adversarial mechanisms in order to decrease the deterioration of classification accuracy on the texts from the source under-represented in the train; (subsection 6.2);
3. make an ablation study (subsection 6.3)

2 Related Studies

The problem of domain adaptation has a long history of research.

Some approaches (Basile, 2020) propose direct manipulations on the textual embeddings. In contrast with out study, (Basile, 2020) does not apply adversarial methods in any form and instead modifies the embeddings of the *weird* words - the words specific to the target domain.

We use the architecture and mechanism from CausaLM (Feder et al., 2020). However, the authors do not have an objective to maximize the classification accuracy with the causal loss. They estimate

significance of multiple textual features instead. In contrast, we apply the causation mechanism in order to incent our classifiers to pay less attention to the confounders. We use the causal loss which consists of a positive summand for the cross-entropy of gender classification and a negative summand for cross-entropy of the confounder.

(Zhou and He, 2023) is a more advanced causal method which adjusts a model for latent covariates and takes into account the non-confounding covariates, which are relevant only to either the treatment or the outcome. Similarly to CausaLM (Feder et al., 2020), the authors only intend to estimate the causal effect of the covariate and do not increase accuracy of the model. However, similarly to CausaLM (Feder et al., 2020), this method could be potentially adjusted for raising the classification accuracy in tasks of non-topical classification.

Another widely used framework for causal models is CausalNLP (Maiya, 2021). Similarly to CausaLM (Feder et al., 2020), its general objective is estimation of causal dependences of the target variable on the confounders.

We use both Adversarial Domain Adaptation (ADA) and Energy-based Adversarial Domain Adaptation (EADA) from (Zou et al., 2021). However, the authors solve a different task in their paper. They focus on transferring knowledge from a label-rich domain (source domain) to a label-scarce domain (target domain) for pervasive cross-domain for text classification, whilst our main objective is to minimise effect of the domain-related features.

(Han et al., 2021) propose a novel meta-learning framework integrated with an adversarial domain adaptation network, aiming to improve the adaptive ability of the text classifiers. The authors add a new component to the BERT-based model - meta-knowledge generator on the base of BiLSTM. The goal of this module is not only to make the final classification results better, but also to confuse the domain discriminator as much as possible. Unlike our study, (Han et al., 2021) mostly focused on the tasks of topical classification. The authors solve the task of sentiment analysis on the basis of Amazon Reviews, and classify the news from the datasets Reuters-21578 and 20 Newsgroups.

3 Causal and adversarial modification of the BERT-based architectures

3.1 CausaLM

Figure 1 shows the architecture implemented in (Feder et al., 2020). It is based on the BERT architecture (Devlin et al., 2018), but includes additional layers on top of it - for classification of the confounder and masked language model.

$$L(\theta_{bert}, \theta_{mlm}, \theta_{nsp}, \theta_{cc}, \theta_{tc}) =$$
$$= \frac{1}{n}\Bigg(\sum_{i=1}^{n} L^i_{mlm}(\theta_{bert}, \theta_{mlm}) + \sum_{i=1}^{n} L^i_{nsp}(\theta_{bert}, \theta_{nsp})$$
$$+ \sum_{i=1}^{n} L^i_{cc}(\theta_{bert}, \theta_{cc}) - \lambda \sum_{i=1}^{n} L^i_{tc}(\theta_{bert}, \theta_{tc}) \Bigg)$$

Where θ_{bert} denotes all the BERT parameters, except those devoted to θ_{mlm}, θ_{nsp}, θ_{tc} and θ_{cc}. L_{cc} - cross-entropy loss for classification of the target variable. In our case, it is the gender of the text author. L_{tc} is the loss of the confounder classification. λ is a hyper-parameter which controls the weight of the adversarial task.

The model in CausaLM is trained in 3 stages:

1. Training the model on MLM;
2. Training the confounder classifier;
3. Training to classify the target class (in our case, it is the gender of the author) with loss including Cross-Entropy components for MLM, confounder, and the target class label.

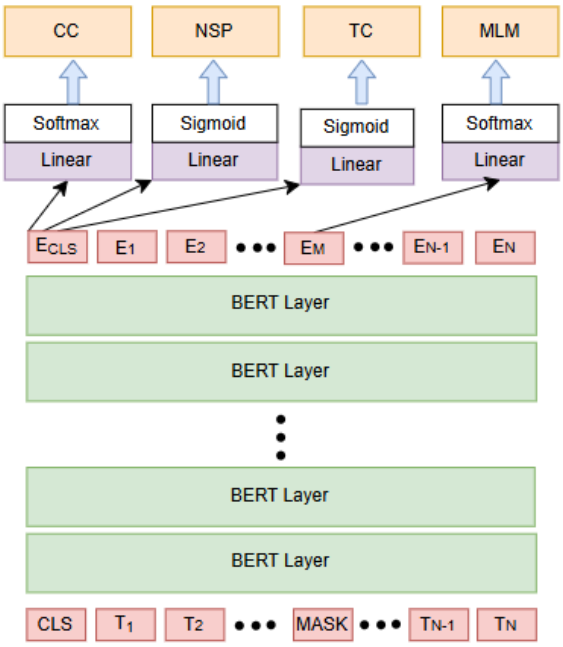


Figure 1: Architecture of a BERT-based causal model

3.2 Adversarial Domain Adaptation

ADA method belongs to Unsupervised Domain Adaptation (Ramponi and Plank, 2020). It shows promising performance in numerous NLP tasks in recent years (Tzeng et al., 2017).

It usually consists of a shared feature extractor $f = G_f(x)$, a label predictor $y = G_y(x)$ and a domain discriminator $d = G_d(x)$. In addition to the standard full supervision learning process in the source domain, a minimax game is designed between the feature extractor f and the domain discriminator d. The domain discriminator d aims to distinguish the domain label between source and target, meanwhile the feature extractor f is trained to deceive the feature discriminator d. This adversarial training process can be formulated as

$$\min_{G_f, G_y} L_y(X_s, Y_s) - \lambda L_f(X_s, X_t),$$

$$\min_{G_d} L_d(X_s, X_t),$$

where L_y is the cross-entropy classification loss for the target label (in our study, it is the gender of the text author). L_f is the loss of the feature extractor. It denotes the cross-entropy of the classification of the text source. Both L_y and L_f are calculated and optimised with freezing of weights of the domain discriminator. L_d is similar to L_f. However, when it is calculated and optimised, the weights of the feature extractor and the label predictor are frozen.

3.3 Energy-based Adversarial Domain Adaptation

The adversarial training objective of three modules forms a minimax game, that is defined by:

$$\min_{G_f, G_y} L_{CE}(X_s, Y_s) + \gamma L_{AE}(X_t),$$

$$\min_{G_a} \big(L_{AE}(X_s) + max(0, m - L_{AE}(X_t))\big),$$

$$\text{where } L_{AE}(x_i) = ||G_a(G_f(x; \theta_f); \theta_a) x_i||$$

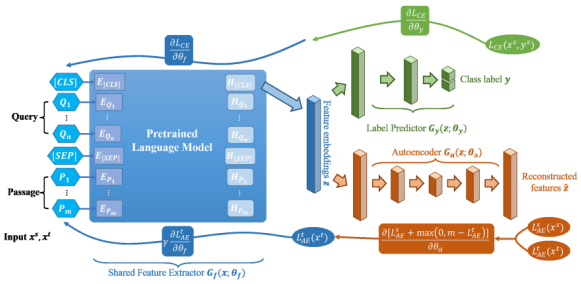


Figure 2: Architecture of a model for Energy-based Adversarial Domain Adaptation. The original image: (Zou et al., 2021)

Dataset	M freq	W freq	mean len	len perc 10	len perc 25	len median	len perc 75	len perc 90
MAIL	3236	6764	217	71	83	115	188	370
AWD	5984	4016	84	13	21	39	76	144

Table 1: Datasets for training and testing the classifiers

where γ is a hyperparameter to control the effectiveness of G_a. L_{CE} is cross-entropy loss for the target label classifier. L_{AE} is the loss of auto-encoder. m is the margin between the representations from the source domain and the target domain. The autoencoder G_a can be considered as an energy function that associates lower energies to the observed samples in a binary classification problem (Zou et al., 2021). The shared feature extractor G_f maps both labeled source data X_s and unlabeled target data X_t to a latent feature space. Both G_f and the label predictor G_y are trained with full supervision using the labeled data in the source domain.

The architecture is shown in Figure 2.

4 Experiments

The main metric we use to compare the models in the experiments is accuracy.

Our goal is to train a reliable classifier for which a domain shift deteriorates the model accuracy as little as possible. We compute the accuracy on the texts from , then compute the accuracy on the texts from the under-represented data source. We degine the difference of accuracies on the texts from the data source prevailing in the training dataset, and on the texts from the data source under-represented in train as *the difference of accuracies*. We denote it as δ (or delta).

We train a set of BERT-based models for CausaLM, ADA, and EADA. We use multilingual BERT with base configuration (12-layer, 768-hidden, 12-heads, 125M parameters, google-bert/bert-base-multilingual cased in HuggingFace) as a baseline for all the experiments. We compare it with the large version (24-layer, 1024-hidden, 16-heads, 355M parameters, google-bert/bert-large-multilingual-cased in HuggingFace). In all our experiments, learning rate=10^{-5} is used, since this value is proposed in (Sun et al., 2019) and (Zou et al., 2021).

5 Data

In our study, we work on the task of gender classification of the text author. It is a kind of non-topical classification, since the target classification variable is not a topic or a topical feature but a more complicated concept which cannot be described by certain keywords..

For the experiments, we use two datasets. The first one contains texts from Mail.Ru Blogs, the second one does from AWD. Each dataset contains nearly 10000 texts. In Table 1, the distribution of the genders and the text lengths is shown. In the Mail.Ru Blogs dataset, around 32% texts are written by men, and

mail share	acc
25	0.901
50	0.931
75	0.920
90	0.905

Table 2: BERT accuracies when trained to predict the source of the texts

the rest are written by women. In AWD, the gender distribution is different, since around 60% texts are written by men. Moreover, the texts from these datasets have different length distributions.

In all the experiments, we take 8000 examples to the train and 2000 texts to the test. Both Mail.Ru Blogs and AWD are Russian social media platforms. However, their content and target audience are different. Mail.Ru Blogs is a general-purpose platform which includes a wide variety of topics including sport, politics, technology, health, science, tourism and so on. At the same time, AWD is a platform about tourism. It causes presence of significant topical shifts in the AWD dataset.

These datasets have the etalon labels for the genders of the text authors, since the gender label is indicated by users of the platforms themselves.

6 Results and discussion

6.1 Prediction of the Text Domain

We use the number of dataset (mail or awd) as the confounder for both adversarial methods and CausaLM. It is a binary label. 0 means awd, 1 means mail. To make sure that the distribution of texts in these two data sources differs significantly, we conduct a range of experiments with classifiers of the source of the text. We indicate how the accuracy of the source classifiers is affected by shares of the data sources in the training dataset. We train multilingual bert classifiers on datasets containing 10%, 25%, 50%, and 75% of the Mail.Ru texts.

Table 2 shows that for each train/test split proportion the accuracy of the source classifier is higher than 90%. It means that the text distributions in Mail.Ru and AWD differ a lot and even BERT is able to notice the difference. We assume that it is caused by topical shifts in AWD given the specialization of this website. Moreover, the high accuracies shows vulnerability of the base BERT model to such a kind of topical shifts.

6.2 Causal and Adversarial Methods

The difference of the accuracy on the overrepresented data source in the train data and the accuracy on the underrepresented one, called *delta* is used as the second key metric to evaluate the vulnerability of the model to the topic shifts in the test data.

We show in Table 3 that Adversarial Domain Adaptation (ADA) helps to decrease the accuracy delta between the test on the text source prevailing on train and testing on the texts from the under-represented source. The increase in γ corresponds to the decrease in delta. At the same time, the accuracy on the test dataset, prevailing on train, decreases whilst the accuracy on the shifted dataset increases. It makes models more resilient to the domain shifts.

As we show in Table 3, dependence of the CausaLM performance on the λ is similar to that for the adversarial methods. The higher the λ, the lower the λ and the higher the accuracy on the texts from the source under-represented on the train.

In Table 3, it is clear that CausaLM achieves a lower delta than the adversarial methods. However, the decrease in accuracy on the texts from the over-represented source is more significant than that for the ADA method.

metod	m	mail share	λ	mail acc	awd_acc	delta
BERT	-	25	0	0.801	0.772	-0.029
ADA	-	25	0.05	*0.810*	0.761	-0.049
ADA	-	25	0.2	**0.818**	0.759	*-0.059*
EADA	4	25	0.05	0.806	0.753	-0.053
EADA	4	25	0.2	0.809	0.746	**-0.063**
BERT	-	75	0	0.838	0.716	0.122
CausaLM	-	75	0.05	0.785	0.725	*0.060*
CausaLM	-	75	0.2	0.781	*0.728*	**0.054**
ADA	-	75	0.05	0.830	0.725	0.105
ADA	-	75	0.2	0.825	**0.731**	0.094
ADA	-	75	0.5	0.815	0.719	0.096
EADA	2	75	0.05	0.819	0.688	0.131
EADA	2	75	0.2	0.819	0.673	0.146
EADA	4	75	0.05	0.819	0.692	0.127
EADA	4	75	0.2	0.815	0.685	0.130
EADA	8	75	0.05	0.832	0.694	0.138
EADA	8	75	0.2	0.823	0.685	0.138
BERT	-	90	0.0	0.840	0.702	0.138
ADA	-	90	0.05	0.836	*0.711*	0.125
ADA	-	90	0.2	0.825	**0.712**	*0.113*
ADA	-	90	0.5	0.694	0.600	**0.094**
EADA	4	90	0.05	0.832	0.699	0.133
EADA	4	90	0.2	0.806	0.600	0.206

Table 3: The results of the Adversarial Domain Adaptation. We compare the models on accuracy. The models with the highest accuracy on the source under-represented in train are highlighted in bold. The models with the second highest accuracy are highlighted in italic. For delta, we select the same way, but the smallest values.

We also find Table 3 that the accuracy on the texts from Mail.Ru is higher than that on the texts from AWD even if AWD prevails on the train. It can be explained by higher length of the texts from Mail.Ru. It is known (Baillargeon and Lamontagne, 2022) that the text length affects the accuracy of transformer-based models. It results in negative values of δ. In this case, the value of δ is less informative than the accuracy on Mail.Ru Blogs - the data source under-represented in the train dataset. The highest and the second highest accuracy on Mail.Ru Blogs is attained with ADA. EADA shows a result worse that ADA but better than the base BERT.

For the base BERT, ADA and EADA we also apply a one-sided t-test to evaluate the statistical significance of the improvements in the accuracy achieved by usage of the adversarial methods. For each setting (25% mail + 75% awd, 75% mail + 25% awd, and 90% mail + 10% awd), we make 5 random splits of mail and awd to train and test in the corresponding proportions and calculate the accuracy on each test sample. We calculate the p-values and find out that it is lower than 0.05 for ADA on the splits 75% mail + 25% awd, and 90% mail + 10% awd. There is not statistical significance for the results in the setting of 25% mail + 75% awd, since the p-value is 0.18. We suppose that it happens because of the length distribution of the AWD dataset prevailing in this setting.

In most cases, the increase of the accuracy on the test dataset is statistically significant, that reiterate applicability of the methods of adversarial domain adaptation for combatting the effects of the topical shifts.

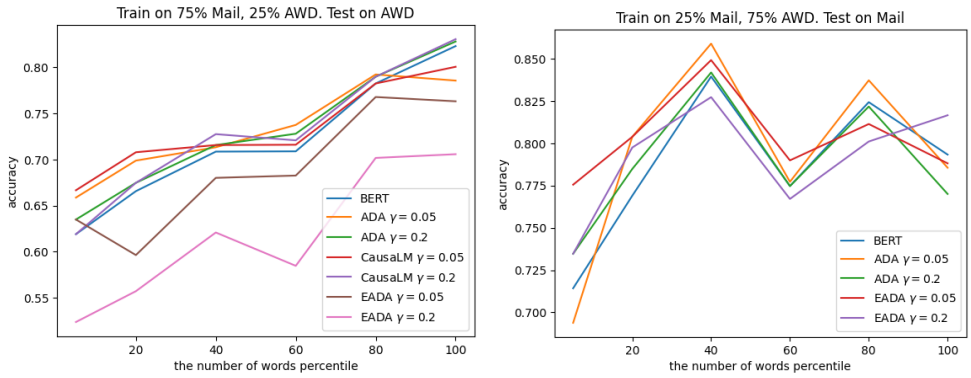


Figure 3: Effect of the text length on the performance of the causal and adversarial methods. Splits: 75% mail + 25% awd and 25% mail + 75% awd

6.3 Ablation Study

We show in Figure 3 and Figure 3 how the model performance is affected by the length of the texts. The texts are split to the groups of the same size by the number of words of them. It can be seen that for the splits with Mail.ru share equal to 75, the ADA model with $\gamma = 0.05$ beats the BERT-based baseline on every text length. However, when we take 25% of the Mail.ru texts and 75% of the AWD texts to the train, it gets less stable on the shortest and the longest texts.

In contrast to ADA, EADA has an additional hyperparameter m, the margin between the representations from the source domain and the target domain. The default value recommended in (Zou et al., 2021) is 4. We also try $m = 2$ and $m = 8$ for train 75% Mail.Ru + 25% AWD. Since, they all show a worse result than $m = 4$.

EADA attains higher accuracy than the baseline only when trained on with 25% mail and 75% awd.

The plot for training on 75% Mail.ru and 25% AWD shows a clear pattern that the model accuracy has a strongly positive correlation with the length of the text in the test. Moreover, it can be seen for all the models we trained on this mail/awd split. It confirms the conclusions from (Baillargeon and Lamontagne, 2022).

But this pattern is not that clear when the models are trained on 25% Mail.ru + 75% AWD and tested on Mail.ru. We assume that it can be caused by the difference in the length distributions of the Mail.Ru and AWD datasets, since the classifiers see too few long texts during the training.

7 Training Time

model	train data	epoch time
BERT	mail	**117**
ADA	mail	*128*
EADA	mail	132
CausaLM	mail	234

Table 4: Training time of all the models on the Mail.Ru Blogs dataset. The model that is trained the fastest is highlighted in bold. The second fastest model is shown in italics.

We calculate the time required for training model of each architecture we mention in our experiments. All the models were trained on single Nvidia TITAN RTX-based GPU. Available GPU capacity: 24 GB.

Table 4 shows that addition of the adversarial loss (both ADA and EADA) to the BERT models does not increase the training time significantly. However, the situation is completely different for CausaLM, where training is done in 3 stages.

The inference time of the classifiers depends solely on the model architecture and is not affected by adding of causal losses in any framework. Hence, there is no statistical significant in difference between evaluation times for the causal models and the non-causal ones.

EADA (Zou et al., 2021) has an additional hyperparameter, m - the margin between the representations from the source domain and the target domain. The default value proposed in (Zou et al., 2021) is 4. We carry out experiments with $m = 2$ and $m = 8$ for $75\%mail + 25\%awd$ to potentially improve the accuracy on the test. However, the default value $m = 4$ turns out to be the best, so we fix it for the splits $25\%mail + 75\%awd$ and $90\%mail + 10\%awd$.

We show Table 3 that the addition of causal loss improves the accuracy of the gender classification task in awd in case when the texts from mail.ru prevail in the training. It partially matches our expectations that causal models are capable of improving accuracy when tested on a dataset from a different domain.

8 Conclusion

By conducting a range of experiments, we come to the following conclusions:

1. The topical shift make a considerable impact on non-topical classification tasks.
2. The addition of either adversarial or causal loss slightly increases the model accuracy when tested on texts from a different domain.
3. Although the delta is decreased more by usage of CausaLM, the ADA method is able to do it without that significant decrease of accuracy on the source dataset.
4. The time spent on training causal models using ADA is significantly less than when using CausaLM. It makes usage of ADA more rational in terms of the required time and computational resources.
5. Although it is a more complex algorithm, EADA shows an improvement only on the split with 75% mail and 25% awd data. Hence, the method turns out to be less efficient than ADA.

Our results indicate that both causal and adversarial methods are useful for improving the quality of non-topical text classification in the presence of topical shifts and domain changes. Overall, the use of ADA can be recommended for cases when there is a significant domain shift on test data. This is an important practical result, given the prevalence and relevance of non-topical classification tasks in the modern world.

References

Jean-Thomas Baillargeon and Luc Lamontagne. 2022. Assessing the impact of sequence length learning on classification tasks for transformer encoder models. *The Florida AI Research Society*.

Valerio Basile. 2020. Domain adaptation for text classification with weird embeddings. *CEUR-WS*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2020. Causalm: Causal model explanation through counterfactual language models. *arXiv preprint arXiv:2005.13407*.

ChengCheng Han, Zeqiu Fan, Dongxiang Zhang, Minghui Qiu, Ming Gao, and Aoying Zhou. 2021. Meta-learning adversarial domain adaptation network for few-shot text classification. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*.

Alex Luu and Sophia A. Malamud. 2020. Non-topical coherence in social talk: A call for dialogue model enrichment. *ACL*.

Arun S. Maiya. 2021. Causalnlp: A practical toolkit for causal inference with text. *arXiv preprint arXiv:2106.08043*.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in nlp—a survey. *Coling*.

Serge Sharoff, Zhili Wu, and Katja Markert. 2010. The Web library of Babel: evaluating genre collections. // *Proc Seventh Language Resources and Evaluation Conference, LREC*, Malta.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? *arXiv preprint arXiv:1905.05583*.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. 2017. Adversarial discriminative domain adaptation. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuxiang Zhou and Yulan He. 2023. Causal inference from text: Unveiling interactions between variables. *Findings of the Association for Computational Linguistics: EMNLP 2023*.

Han Zou, Jianfei Yang, and Xiaojian Wu. 2021. Unsupervised energy-based adversarial domain adaptation for cross-domain text classification. *ACL 2021*.

RuOpinionNE-2024: Extraction of Opinion Tuples from Russian News Texts

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Natalia Tkachenko
Lomonosov Moscow State University
nataliya.m.tkachenko@gmail.com

Anna Lapanitsyna
Lomonosov Moscow State University
anna.lapachka@gmail.com

Mikhail Tikhomirov
Lomonosov Moscow State University
tikhomirov.mm@gmail.com

Nicolay Rusnachenko
Bauman Moscow State Technical University
rusnicolay@gmail.com

Abstract

In this paper, we introduce the Dialogue Evaluation shared task on extraction of structured opinions from Russian news texts. The task of the contest is to extract opinion tuples for a given sentence; the tuples are composed of a sentiment holder, its target, an expression and sentiment from the holder to the target. In total, the task received more than 100 submissions. The participants experimented mainly with large language models in zero-shot, few-shot and fine-tuning formats. The best result on the test set was obtained with fine-tuning of a large language model. We also compared 30 prompts and 11 open source language models with 3-32 billion parameters in the 1-shot and 10-shot settings and found the best models and prompts.

Keywords: Structured Sentiment Analysis, Opinion Tuples, Named Entity, News Texts
DOI: 10.28995/2075-7182-2025-23-234-244

RuOpinionNE-2024: извлечение кортежей мнений из новостных текстов на русском языке

Лукашевич Н.В.
МГУ им. Ломоносова
Москва, Россия
louk_nat@mail.ru

Ткаченко Н.М.
МГУ им. Ломоносова
Москва, Россия
nataliya.m.tkahenko@gmail.com

Лапаницына А.М.
МГУ им. Ломоносова
Москва, Россия
anna.lapachka@gmail.com

Тихомиров М.М.
МГУ им. Ломоносова
Москва, Россия
tikhomirov.mm@gmail.com

Русначенко Н.Л.
МГТУ им. Н.Э. Баумана
Москва, Россия
rusnicolay@gmail.com

Аннотация

В этой статье описано новое тестирование по извлечению мнений из русскоязычных новостных текстов, организованное в рамках серии тестирований Dialogue Evaluation. Задача тестирования состоит в извлечении кортежей мнений для заданного предложения; кортежи состоят из источника мнения, объекта мнения, оценочного выражения и тональности источника по отношению к объекту. Всего на тестирование было подано более 100 прогонов. Участники экспериментировали в основном с большими языковыми моделями в форматах zero-shot, few-shot и fine-tuning. Лучший результат на тестовом наборе был получен на основе дообучения (fine-tuning) большой языковой модели. Также было проведено сравнение 30 промптов и 11 языковых моделей с открытым исходным кодом размером 3-32 миллиардов параметров в 1-shot и 10-shot режимах и выявлены лучшие модели и промпты.

Ключевые слова: таргетированный анализ тональности, именованная сущность, новостные тексты

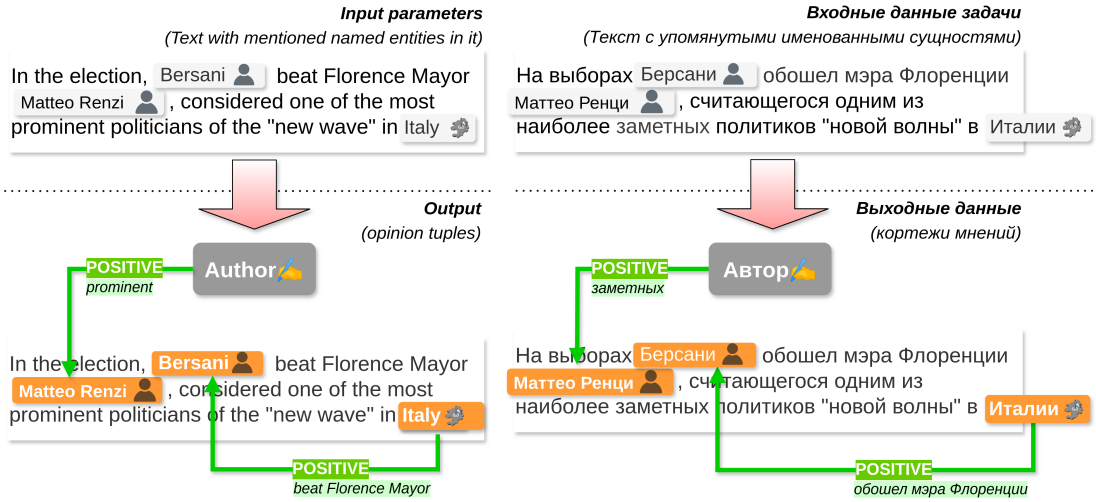


Figure 1: Example of opinion extraction with explanations according to the RuOpinionNE-2024 task. Two opinion tuples are shown: (Italy, Bersany, beat Florence Mayor, positive) and (AUTHOR, Matteo Renzi, prominent, positive)

1 Introduction

Sentiment analysis is one of the most actively developing areas in natural language processing. At the beginning of studies in this area, the task was to extract the overall sentiment of a user review or message in a social network. Currently, there is a wide variety of problem statements, including the identification of sentiment toward a specific entity, an aspect (part or characteristic) of this entity, and a related task of extracting a position (stance) on a certain discussion topic. Modern methods allow us to get closer to a more complete opinion recognition (Liu, 2012), extracting opinion tuples, which include a source of the opinion, an object of the opinion, sentiment, and other opinion components (Barnes et al., 2022).

In this paper, we present the RuOpinionNE shared task on extraction of opinions from Russian news texts organized in the framework of the Dialogue evaluation competitions. The aim of the task is to extract opinion tuples for a given sentence; the tuples are composed of a sentiment holder, its target, an expression and sentiment from the holder to the target. News texts are a very interesting source for extracting opinions because such texts contain numerous entities, which can be the holder or the target of opinion or both. In the same sentence, positive and negative attitudes can be met. At the same time, most entities are mentioned in neutral contexts. The extraction of such tuples from news texts allows a better understanding of attitudes between different subjects discussed in the current flow of news. Figure 1 depicts an example of opinion tuples extracted from a Russian text (right) along with the translated version of the related example (left). In the example, the author is positive to Matteo Renzi, because he is mentioned as a prominent politician. In addition, Italy is positive to Bersani because Italians voted for him.

2 Related Work

The extraction of opinions from texts should aim to identify tuples that comprise several components such as the source of the opinion, the target, maybe the aspect (part, characteristics) of the target and the sentiment (Liu, 2012). Currently, in numerous sentiment-oriented works, the ultimate task is often reduced to simpler subtasks such as general sentiment of a text (sentence), target-oriented sentiment (the identification of the opinion source is not included in the task), and others (Barnes et al., 2021).

Barnes et al. (Barnes et al., 2021) suggest extracting opinions as tuples (h, t, e, p) where h is a holder who expresses a polarity p towards a target t through a sentiment expression e. Extracted tuples can be

represented as a graph comprising a set of labeled nodes and a set of unlabeled edges connecting pairs of nodes. Empty holders and targets are allowed. The authors used a reimplementation of the neural syntactic parser by Dozat and Manning (Dozat and Manning, 2022) to extract such tuples from texts. The base of the network model is a bidirectional LSTM (Hochreiter, 1997; Schuster and Paliwal, 1997) (BiLSTM), which creates contextualized representations $[c_1, \ldots, c_n] = BiLSTM(w_1, ..., w_n)$, where w_i is the concatenation of a word embedding, POS tag embedding, lemma embedding, character embedding created by a character-based LSTM for the i'th token, and also BERT (Devlin et al., 2019) contextualized embeddings. The obtained embeddings are then processed by two feedforward neural networks (FNN), creating specialized representations for potential heads and dependents as $h_i = FNNhead(c_i)$ and $d_i = FNNdep(c_i)$ respectively. The scores for each possible relation between found items are computed by a final bilinear transformation.

In 2022, the multilingual contest on the extraction of opinion tuples was organized at the SemEval 2022 conference (Barnes et al., 2022). The evaluation collection comprises mainly review datasets and a single news collection MPQA. The organizers provided two baselines in the contest: 1) a dependency graph prediction model, and 2) a sequence-labeling pipeline based on three separate BiLSTM models to extract holders, targets, and expressions. The outputs of the models were used to train a relation prediction model that determines sentiment relations between found entities. The best results in the competition were achieved by applying a modified version of the above mentioned sentiment graph analysis model with contextualized embeddings obtained with $\text{RoBERTa}_{\text{Large}}$ (Liu et al., 2019).

In aspect-oriented sentiment analysis, several tuple configurations are used to structure the content. Such configurations are: Aspect-Sentiment Triplet Extraction (ASTE), which should produce {aspect, opinion, sentiment category} triplets, such as "⟨menu, great, positive⟩"; and Aspect-Sentiment Quadruplet Extraction/Prediction (ASQE/ASQP) that outputs {aspect, opinion, aspect category, sentiment category} quadruplets, such as "⟨menu, great, general, positive⟩" (Hua et al., 2024).

The authors of (Zhang et al., 2024) tested zero-shot large language models (LLM) in various sentiment analysis tasks, including aspect tuple extraction: Aspect-Sentiment Triplet Extraction (ASTE) and Aspect-Sentiment Quadruplet Extraction/Prediction (ASQE/ASQP). They found that zero-shot prompting of larger LLMs has very low results in these tasks, fine-tuned T5_{large} model (Raffel et al., 2020) (small LLM) achieved much higher results: about 24 percent points for ASTE on average and 30 percent points for ASQE/ASQP.

For Russian, there were some studies on targeted sentiment analysis, such as aspect-based sentiment analysis (Loukachevitch et al., 2015; Chumakov et al., 2023) or entity-oriented sentiment analysis (Loukachevitch and Rubtsova, 2015; Golubev and Loukachevitch, 2020; Pronoza et al., 2021; Golubev et al., 2023). In 2023, the RuSentNE competition devoted to extraction of sentiment tovards named entities in news texts was organized. The best approaches on the RuSentNE evaluation were BERT-based ensembles (Devlin et al., 2019; Golubev et al., 2023). Currently, the best result in this formulation of the problem has been achieved with fine-tuning of the Flan-T5 model (Rusnachenko et al., 2024). In (Rusnachenko et al., 2019), the authors studied the extraction of triples {holder, target, sentiment} at the document level in the RuSentRel dataset, which comprised 73 analytical texts on international relations. The current work is the first study on extracting sentiment tuples at the sentence level for the Russian language.

3 RuSentNE Corpus

The RuOpinionNE-2024 dataset is based on the annotated RuSentNE corpus, which includes news texts written in Russian. In the first stage of the annotation, the RuSentNE texts were annotated with named entities, including such types as PERSON, ORGANIZATION, PROFESSION, COUNTRY, CITY, NATIONALITY and so on. In the next stage, positive or negative relations between entities (including the author's position towards mentioned entities) were established. Finally, for each annotated attitude, the evidence (that is, an expression that serves as a key element for extracting the opinion) was carefully annotated and checked (Figure 2). The annotations allowed us to create a dataset, in which sets of opinion tuples (holder, target, sentiment, expression) are assigned to each text fragment.

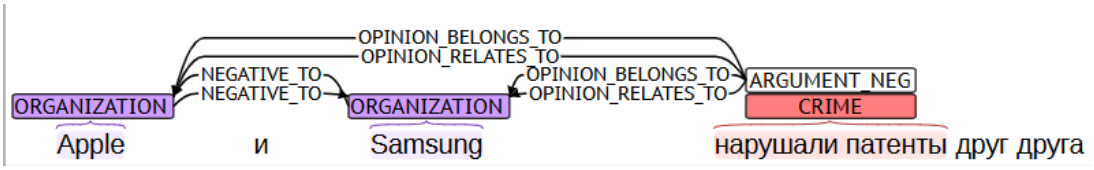


Figure 2: Example of opinion annotation for the RuOpinionNE-2024 task

The source of the opinion (Opinion holder) can be: (i) the author of the text, (ii) the author of the quote, or (iii) another mentioned entity. The opinion can be *implicit*, i.e. expressed through the actions of the entity. For example, in the sentence "X fired Y", it is assumed that there is an implicit negative opinion of X towards Y.

The first version of the created corpus became a basis for the RuSentNE-2023 open evaluation, which required the recognition of sentiment in relation to a given named entity (Golubev et al., 2023). The task was to predict whether the content of a given sentence is positive or negative towards a given entity (Golubev et al., 2023).

4 RuOpinionNE-2024 Dataset and Task Description

The task of the RuOpinionNE-2024 evaluation is as follows: for given fragments of news articles (mainly sentences), it is necessary to extract all tuples: (H, T, P, E), where H is an opinion holder who expresses a polarity P towards a target T through a sentiment expression E. There can be sentences without any tuple. Holders and Targets can be entities of the following types: PERSON, ORGANIZATION, COUNTRY, CITY, REGION, PROFESSION, NATIONALITY, IDEOLOGY. Pronouns are not included in tuples. Holders can also be empty (NULL) for general opinion or AUTHOR for the author's opinion. Targets and expressions are never empty. Fragmented entities are possible. The polarity represents sentiment label which could be positive or negative. Table 1 shows examples of tuples extracted from the example sentence (Figure 2).

Holder	Target	Polarity	Expression
Apple	Samsung	NEG	violated patents
Samsung	Apple	NEG	violated patents

Table 1: Examples of sentiment opinions in RuOpinionNE-2024 dataset (Figure 2)

Table 2 contains the distribution of opinion tuples in train, validation and test sets of the RuOpinionNE-2024 dataset. It can be seen that about 40% of sentences do not contain opinion tuples. Sentences with opinions contain two opinion tuples on average. In about 20% of annotated tuples, the holder is absent, and about 10% of tuples are annotated as the author's opinion.

Type	Stage	#Sent.	#Sent. w/o sentiment	#Tuples	#Tuples w/o holder	#Tuples with author as holder
train	Train	2556	1062	2904	484	173
validation	Development	1316	547	1612	436	196
test	Final	804	216	1172	334	158
Total		4676	1825	5688	1254	527

Table 2: Distribution of sentiment scores in training, validation and test sets.

The data was prepared in the JSON lines format[1], where a line corresponds to a sentence with found opinion tuples or with an empty set of opinion tuples.

[1] https://jsonlines.org/

The main metric for the task is the F1 measure. When evaluating recall, each reference tuple was compared with the entire list of predicted tuples for the corresponding text fragment. The reference tuple was considered found if among the predictions at least one tuple detected, for which:

- 1) holder, target, expression intersect with the reference holder, target, expression by at least one token, respectively (it is necessary that each component is at least partially detected and all of them are combined into a single tuple);
- 2) the polarity of the same tuple coincides with the polarity of the reference.

The cases of holder = AUTHOR or NULL were processed as special tokens that required complete match. For example, for holder overlap:

- holder_overlap = (number of common tokens of the reference and the prediction)/(number of reference tokens).

Target overlap and expression overlap were calculated in a similar way. If there were several options, the best tuple was selected. If the tuple was found, a weighted score was calculated for it, which reflected the coverage of the reference tuple by the predicted one, averaged over the three components. Weighted score and recall were calculated as follows:

- weighted_score = (holder_overlap + target_overlap + expression_overlap)/3, if each overlap >0 and the polarity matches, and 0 otherwise;
- recall = (sum of weighted_score over all reference tuples)/(number of reference tuples).

Precision was calculated in the same way, but for it, the coverage of the set of predicted tuples by the set of reference tuples was estimated. Both metrics were calculated not for sentences separately, but for the entire dataset, and then combined into the F1 metric.

5 Results of Evaluation

As a baseline (baseline_model), we adopt Qwen2.5-32B-Instruct model (Qwen et al., 2025) (32 billion parameters) in the few-shot format (10 examples). The complete textual prompt utilized for inferring results is shown in Appendix A.

In our competition, we received more than 100 submissions. Table 3 shows the results of the models at the development stage, table 4 presents the results of the models at the final stage of the competition. In the following, we describe the approaches of participants.

Participant	F1
Zholtikov_Michail	0.34
msuai	0.28
iarnv	0.21
baseline_model	0.17

Table 3: Results of the Development evaluation stage of RuOpinionNE-2024

Participant	F1
VatolinAlexey	0.41
RefalMachine	0.35
msuai	0.33
iarnv	0.28
Zholtikov_Michail	0.24
baseline_model	0.24
vitalymegabyte	0.20
utmn	0.11

Table 4: Results of the Final evaluation stage of RuOpinionNE-2024

VatolinAlexey. The participant experimented with unsupervised and supervised approaches. In an

unsupervised approach, the LlaMA-3.3-70B[2] (et. al., 2024) model was used. The model could not calculate exact indices of the items for opinion tuples in the substring, therefore, the final prompt did not contain the task to determine the indices. Instead, the algorithm was written for fuzzy search of a substring in a given sentence. The best result of the unsupervised experiments was F1 = 0.22.

As a supervised approach, the QLora (Dettmers et al., 2023) adapter for the LLaMA-3.3-70B model with FP4 precision was trained on the training data. Only text was fed as input without any prompts. The output was similar to what is described in the task, i.e. a JSON with a list of opinions. The model could produce different results after each run due to randomization, so the model was run several times and the results were aggregated. Several aggregation strategies were tested: (1) most frequent answer (`most_common`), (2) N most frequent answers (`most_common_n`), (3) combining all answers (`combine`). The best result on the test data was obtained with $N = 5$, strategy `most_common_n` with the result F1 = 0.41.

RefalMachine. The participant experimented with zero-shot, few-shot (Brown et al., 2020) and fine-tuning approaches. The best results were achieved by fine-tuning the Qwen2.5-32B-Instruct model and a prompt in the instruction format, with an F1 score of 0.35 at the Final evaluation stage.

msuai. The participant experimented with unsupervised and supervised formats of the LLM application: Mistral Large 2[3], GPT4-o(OpenAI et al., 2024). Models were used with prompts containing 12-15 examples. The examples were chosen according to semantic similarity with a target sentence calculated using $\text{BERT}_{\text{large-uncased}}$ model for sentence embeddings in Russian language[4] or Sentence RuBERT[5] models. This approach achieved the highest participant F1 score of 0.33 at the Final evaluation stage.

Zholtikov_Michail. The participant experimented with the instruction tuned versions of Mistral-7B-v0.3[6], MetaLlama-3-8B[7] and Vikhr-7B-0.3[8]. The format of experiments: zero-shot and supervised learning. The best results F1=0.34 were obtained by instruction tuned version of Mistral-7B-v0.3 at the Development stage, and F1 = 0.24 at the Final evaluation stage.

vitalymegabyte. The participant applied traditional NER, splitting polar expressions to separate tags: Polar_expression_POS and Polar_expression_NEG. To extract relationships between extracted items, self-attention was used to build the relationship matrix between items.

Thus, we can see that almost all participants experimented with generative models in zero-shot, few-shot or fine-tuning regimes. The best results were achieved by parameter-efficient fine-tuning of a quite large language model LlaMA-3.3-70B-Instruct with 70 billion parameters.

Compared with the results of the Semeval 2022 competition (Barnes et al., 2021), it could be noted that the Semeval 2022 results for the review datasets are much higher, achieving 0.76 F-measure on an English review dataset, but the best results on the MPQA news dataset is approximately the same: about 0.44 F-measure.

6 Experiments with RuOpinionNE Data in the Framework of Student Course

RuOpinionNE data were used as experimental data in the student course "Practical aspects of LLM training", which was taught by Mikhail Tikhomirov at the Faculty of Computational Mathematics and Cybernetics of Lomonosov Moscow State University.

The students constructed 30 prompts and experimented with them in zero-shot or few-shot regimes on the RuOpinionNE data. The collected prompts were fed into eleven language models in 1-shot or 10-shot formats. The following language models were used in experiments:

- LLaMA-3.1-8B-Instruct [9], LLaMA-3-8B-Instruct [10];

[2] `https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct`
[3] `https://huggingface.co/mistralai/Mistral-Large-Instruct-2407`
[4] `https://huggingface.co/ai-forever/sbert_large_nlu_ru`
[5] `https://huggingface.co/DeepPavlov/rubert-base-cased-sentence`
[6] `https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3`
[7] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`
[8] `https://huggingface.co/Vikhrmodels/Vikhr-7B-instruct_0.3`
[9] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`
[10] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

- Mistral-Nemo-Instruct-2407 (12.2B parameters) [11];
- Qwen2.5-32B-Instruct, Qwen2.5-14B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-3B-Instruct [12];
- T-lite-it-1.0 (7.6B parameters)[13];
- OpenChat-3.5-0106 (7B parameters) [14];
- RuAdapt-LLaMA3 (8B parameters) [15] (Tikhomirov and Chernyshov, 2024);
- Saiga-LLaMA3-8B (8B parameters) [16]

model_name	k=10	k=1
Qwen2.5-32B-Instruct	**0.195**	**0.158**
Mistral-Nemo-Instruct-2407	<u>0.190</u>	0.112
Qwen2.5-7B-Instruct	0.184	<u>0.139</u>
Saiga-LLaMA3-8B	0.179	0.091
T-lite-it-1.0	0.157	0.096
LLaMA-3-8b-Instruct	0.153	0.119
Qwen2.5-14B-Instruct	0.145	0.121
Meta-LlaMA-3.1-8B-Instruct	0.141	0.090
RuAdapt-LLaMA3	0.123	0.073
OpenChat-3.5-0106	0.113	0.087
Qwen2.5-3B-Instruct	0.091	0.088

Table 5: Average model quality in 10-shot and 1-shot settings for the RuOpinionNE test set. The best results are in bold, the second best results are underlined

model_name	k=10	k=1
Qwen2.5-32B-Instruct	**0.229**	**0.204**
Mistral-Nemo-Instruct-2407	<u>0.211</u>	<u>0.157</u>
Qwen2.5-7B-Instruct	0.199	0.168
Saiga-LLaMA3-8B	0.193	0.118
LLaMA-3.1-8B-Instruct	0.173	0.110
T-lite-it-1.0	0.171	0.119
LLaMA-3-8B-Instruct	0.169	0.154
Qwen2.5-14B-Instruct	0.169	0.144
RuAdapt-LLaMA3	0.134	0.104
OpenChat-3.5-0106	0.132	0.108
Qwen2.5-3B-Instruct	0.120	0.119

Table 6: Maximum model quality in 10-shot and 1-shot settings for the RuOpinionNE test set. The best results are in bold, the second best results are underlined

Thus, the set of models in experiments comprises a larger model Qwen2.5-32B-Instruct (32 billion parameters), average-sized models with 12-13 billion parameters (Qwen2.5-14B-Instruct, Mistral-Nemo-Instruct-2407), smaller models with 7-8 billion parameters, and the smallest model with 3 billion parameters (Qwen2.5-3B-Instruct). The models in experiments include three models adapted to the Russian language with 7-8 billion parameters: T-lite-it-1.0, Saiga-LLaMA3-8B and RuAdapt-LLaMA3.

Table 5 shows the performance of each model averaged on all prompts. It can be seen that the averaged best results are obtained on 10-shot prompts by larger models Qwen2.5-32B-Instruct and Mistral-Nemo-Instruct-2407; the averaged best results on 1-shot prompts are achieved by Qwen2.5-32B-Instruct. The

[11] https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407
[12] https://huggingface.co/collections/Qwen/qwen25-66e81a666513e518adb90d9e
[13] https://huggingface.co/t-tech/T-lite-it-1.0
[14] https://huggingface.co/openchat/openchat-3.5-0106
[15] https://huggingface.co/RefalMachine/ruadapt_llama3_8b_instruct_extended_lep_ft
[16] https://huggingface.co/IlyaGusev/saiga_llama3_8b

Qwen2.5-7B-Instruct model obtained the best averaged results among all the smaller models.

Among the Russian-oriented models, the best results for the 10-shot setting were achieved by Saiga-LLaMA3-8B. For 1-shot prompts, similar results were obtained by T-lite-it-1.0 and Saiga-LLaMA3-8B. In both cases, the results of the Russian models are worse than the results of the similar-sized model Qwen2.5-7B-Instruct.

Table 6 shows the best results for each model. The best results are obtained by larger models: Qwen2.5-32B-Instruct and Mistral-Nemo-Instruct-2407. Among the smaller models, the Qwen2.5-7B-Instruct model achieved the highest results. Among Russian models, the Saiga-LLaMA3-8B model obtained better results in the 10-shot setting and similar results with T-lite-it-1.0 in the 1-shot setting.

Table 7 presents the results of all instructions averaged on the models. It can be seen that the obtained results are very close to each other in average quality. This means that there does not exist a single prompt, which is best for all or most models.

instruction_id	k=10	k=1
28	0.162	0.113
19	0.159	0.112
17	0.158	0.120
5	0.157	0.103
0	0.157	0.105
7	0.156	0.099
23	0.155	0.108
13	0.154	0.081
12	0.154	0.112
16	0.154	0.118

Table 7: Results of all instructions averaged on the models in 10-shot and 1-shot settings for the RuOpinionNE test set.

7 Analysis of Errors

In this section, we provide an analysis of the submitted results. To explore the cases that arise from disagreement between gold standard annotation and tuples automatically extracted by participating models, we use two different modes for analysis:

- **Extracted opinions mentioned in manual annotation**: absence of annotated tuples in the submitted results.
- **Extracted opinions not mentioned in the gold standard**: analysis of tuples that were found in the submitted results but were not indicated in the gold annotation.

We consider examples from the final stage of the competition. We limit the scope of methods by overviewing LLM-based submissions, according to the Section 5, The analysis includes all submissions that exceed the baseline, as well as the baseline itself (6 models). We analyze examples of annotated opinion tuples that were not identified by all models under consideration, or tuples, which are absent in the annotated data but recognized by at least three models.

We found the following main causes of the differences between annotated tuples and the tuples extracted by the models. First, the context of the sentence is sometimes not sufficient to reveal the sentiment or the holder of the opinion. For example, the sentence "By the way, Moscow occupies fourth place in the list of leaders for 2010" looks very positive towards Moscow but in fact leaders in road congestion (jams) are discussed, which means negative information about Moscow.

Second, in the task, we required participants to distinguish between the author's opinion and the opinion with the unknown holder. We thought that this is important for extracting the author's position. But this requirement was too difficult for models. In some cases, the short context of the sentence did not provide the necessary information. For example, from the sentence "Russians started the Dakar rally as favorites", it is difficult to understand if "favorites" is a general opinion or an author's opinion.

Next, we considered opinions between mention subjects that were not found by all the models. In the following sentence, all models did not find that the attitude of specialists towards UKRSPIRT is negative possibly because of the quite long distance between the mentions:

- "In addition, the very fact of the possibility of alternative supplies will undermine the [monopoly] of the state concern UKRSPIRT, which most often does not agree to increase prices for its products, specialists note."

In the next sentence, the positive attitude from Syria to HAMAS, Hezbollah and Iran is presupposed in the sentence but was not revealed by all models:

- In Tel Aviv, they believe that the price of the withdrawal of Israeli troops from the Golan Heights should be political concessions from Syria - weakening its ties with its strategic ally Iran and ending support for the Lebanese resistance movement Hezbollah and the Palestinian Hamas.

In the following sentence, the models did not reveal the negative attitude from V. But to the United States because of the long distance between entities and non-typical expression of sentiment:

- V. But suggested that "the court will not examine the factual objective side of the case, since the practice of hearing cases on charges of conspiracy in the United States is such that such charges automatically mean guilt."

8 Conclusion

In this paper, we presented the shared task of extracting structured opinions from Russian news texts. The task of the RuOpinionNE competition was to extract opinion tuples for a given sentence; the tuples are composed of a sentiment holder, its target, an expression, and the holder's sentiment towards the target. We prepared a new dataset of annotated opinion tuples. In total, the task received more than 100 submissions. The participants experimented mainly with large language models in zero-shot, few-shot and fine-tuning formats. The best result on the test set was obtained with fine-tuning of a large language model. We also compared 30 prompts and 11 open-source language models with 3-32 billion parameters in the 1-shot and 10-shot settings and found the best model and prompt.

Acknowledgements

The study was conducted under the state assignment of Lomonosov Moscow State University.

References

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. // Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, P 3387–3402, Online, August. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberländer, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. // *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, P 1280–1295.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Stanislav Chumakov, Anton Kovantsev, and Anatoliy Surikov. 2023. Generative approach to aspect based sentiment analysis with gpt language models. *Procedia Computer Science*, 229:284–293.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: efficient finetuning of quantized llms. // *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, P 4171–4186, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Timothy Dozat and Christopher D Manning. 2022. Deep biaffine attention for neural dependency parsing. // *International Conference on Learning Representations*.

Aaron Grattafiori et. al. 2024. The llama 3 herd of models.

Anton Golubev and Natalia Loukachevitch. 2020. Improving results on russian sentiment datasets. // *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, P 109–121. Springer.

Anton Golubev, Nicolay Rusnachenko, and Natalia Loukachevitch. 2023. Rusentne-2023: evaluating entity-oriented sentiment analysis on russian news texts. // *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*.

S Hochreiter. 1997. Long short-term memory. *Neural Computation MIT-Press*.

Yan Cathy Hua, Paul Denny, Jörg Wicker, and Katerina Taskova. 2024. A systematic review of aspect-based sentiment analysis: domains, methods, and trends. *Artificial Intelligence Review*, 57(11):296.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Bing Liu. 2012. *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

Natalia Loukachevitch and Yuliya Rubtsova. 2015. Entity-oriented sentiment analysis of tweets: results and problems. // *Text, Speech, and Dialogue: 18th International Conference, TSD 2015, Pilsen, Czech Republic, September 14-17, 2015, Proceedings 18*, P 551–559. Springer.

Natalia Loukachevitch, Pavel Blinov, Evgeny Kotelnikov, Yulia Rubtsova, Vladimir Ivanov, and Elena Tutubalina. 2015. Sentirueval: testing object-oriented sentiment analysis systems in russian. // *Proceedings of International Conference Dialog*, volume 2, P 3–13.

OpenAI, :, Aaron Hurst, Adam Lerer, and Adam P. Goucher et. al. 2024. Gpt-4o system card.

Ekaterina Pronoza, Polina Panicheva, Olessia Koltsova, and Paolo Rosso. 2021. Detecting ethnicity-targeted hate speech in russian social media texts. *Information Processing & Management*, 58(6):102674.

Qwen, :, An Yang, Baosong Yang, and Beichen Zhang et. al. 2025. Qwen2.5 technical report.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Nicolay Rusnachenko, Natalia Loukachevitch, and Elena Tutubalina. 2019. Distant supervision for sentiment attitude extraction. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, P 1022–1030.

Nicolay Rusnachenko, Anton Golubev, and Natalia Loukachevitch. 2024. Large language models in targeted sentiment analysis for russian. *Lobachevskii Journal of Mathematics*, 45(7):3148–3158.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE transactions on Signal Processing*, 45(11):2673–2681.

Mikhail Tikhomirov and Daniil Chernyshov. 2024. Facilitating large language model russian adaptation with learned embedding propagation. *Journal of Language and Education*, 10(4):130–145.

Wenxuan Zhang, Yue Deng, Bing Liu, Sinno Pan, and Lidong Bing. 2024. Sentiment analysis in the era of large language models: A reality check. // *Findings of the Association for Computational Linguistics: NAACL 2024*, P 3881–3906.

A Appendix A: Prompt for Baseline Method

Original Text of the prompt (Russian)	Translated version of the prompt (English)
Твоя задача состоит в том, чтобы проанализировать текст и извлечь из него выражения мнений, представленные в виде кортежа мнений, состоящих из 4 основных составляющих: 1. Источник мнения: автор, именованная сущность текста (подстрока исходного текста), либо NULL. Key = Source; 2. Объект мнения: именованная сущность в тексте (подстрока исходного текста). Key = Target; 3. Тональность: положительная/негативная (POS/NEG). Key = Polarity; 4. Языковое выражение: аргумент, на основании которого принята результирующая тональность (одна или несколько подстрок исходного текста). Key = Expression; Если источник мнения отсутствует, то Source = NULL. Если источником мнения является автор, то Source = AUTHOR. В прочих случаях поле Source должно полностью совпадать с подстрокой исходного текста. Поля Target, Expression всегда совпадают с подстроками текста. Ответ необходимо представить в виде json списка, каждый элемент которого является кортежем мнений. Каждый кортеж мнений это словарь, состоящий из четырех значений: Source, Target, Polarity, Expression. Для извлечённых Source, Target, Polarity, Expression должно быть справедливо утверждение: На основании выражения Expression можно сказать, что Source имеет Polarity отношение к Target. Ниже представлены примеры выполнения задачи: ***Текст*** Премьер-министр Молдовы осудил террориста за бесчловечные и жестокие действия. Source: Премьер-министр Молдовы, Target: террориста, Polarity: NEG, Expression: бесчловечные и жестокие действия ***Текст*** Знаменитая актриса продемонстрировала человечность и простоту, достойную уважения публики. ***Ответ*** Source: AUTHOR, Target: актриса, Polarity: POS, Expression: продемонстрировала человечность и простоту, достойную уважения публики Проанализируй таким же образом следующий текст. ***Текст*** {text}	Your task is to analyze the text and extract from it expressions of opinion, presented as a tuple of opinions consisting of 4 main components: 1. Opinion source: the author, a named entity in the text (a substring of the source text), or NULL. Key = Source; 2. Opinion object: a named entity in the text (a substring of the source text). Key = Target; 3. Sentiment: positive/negative (POS/NEG). Key = Polarity; 4. Language expression: the argument on the basis of which the resulting sentiment is accepted (one or more substrings of the source text). Key = Expression; If the opinion source is missing, then Source = NULL. If the opinion source is the author, then Source = AUTHOR. In other cases, the Source field must completely match the substring of the source text. The Target, Expression fields always match the substrings text. The answer must be presented as a json list, each element of which is a tuple of opinions. Each tuple of opinions is a dictionary consisting of four values: Source, Target, Polarity, Expression. For the extracted Source, Target, Polarity, Expression, the statement ""Based on the Expression, it can be said that Source has a Polarity relation to Target"" must be true. Below are examples of completing the task: ***Text*** The Prime Minister of Moldova condemned the terrorist for inhumane and cruel actions. Source: Prime Minister of Moldova, Target: terrorist, Polarity: NEG, Expression: inhumane and cruel actions ***Text*** The famous actress demonstrated humanity and simplicity worthy of public respect. ***Answer*** Source: AUTHOR, Target: actress, Polarity: POS, Expression: demonstrated humanity and simplicity worthy of public respect Analyze the following text in the same way. ***Text*** {text}

Table 8: Textual prompt of the baseline submission; {text} refers to the input parameter of the sentence

RuTermEval-2024: Cross-domain Automatic Term Extraction and Classification in Russian scientific texts

Angelina Mamontova
MSU Institute for Artificial Intelligence
mamontova@mlsa-iai.ru

Roman Ischenko
MSU Institute for Artificial Intelligence
ischenko@mlsa-iai.ru

Konstantin Vorontsov
MSU Institute for Artificial Intelligence
vorontsov@mlsa-iai.ru

Abstract

Automatic Term Extraction (ATE) is a critical NLP task for identifying domain-specific terms, which are essential for tasks like information retrieval, machine translation, and ontology construction. Cross-domain nested term extraction further complicates the task, as traditional methods often fail to handle hierarchical term structures and domain variability. This paper introduces both the CL-RuTerm3 dataset, a novel resource featuring nested term annotations across six domains (the main one is computational linguistics, also mathematics, medicine, economics, literature studies, and agrochemistry), and the RuTermEval-2024 competition, designed to evaluate term extraction systems on this data. The CL-RuTerm3 dataset, comprising 1270 abstracts and 15 full-text articles (over 165k tokens with over 37k annotated entities), is the largest of its kind for Russian scientific texts. Terms are classified into three categories based on lexical and domain specificity: specific terms, common terms, and nomens. The dataset's unique features, such as nested term markup and cross-domain coverage, enable more realistic evaluation of ATE systems. The paper concludes with an analysis of participant approaches in the RuTermEval-2024 competition, emphasizing the effectiveness of contrastive learning. This work aims to advance ATE research by providing a robust dataset and fostering discussions on term extraction methodologies.

Keywords: ATE; automatic term extraction; nested term; dataset
DOI: 10.28995/2075-7182-2025-23-245-256

RuTermEval-2024: Кросс-доменное автоматическое извлечение терминов и их классификация в русскоязычных научных текстах

Ангелина Мамонтова
Институт ИИ МГУ
mamontova@mlsa-iai.ru

Роман Ищенко
Институт ИИ МГУ
ischenko@mlsa-iai.ru

Константин Воронцов
Институт ИИ МГУ
vorontsov@mlsa-iai.ru

Аннотация

Автоматическое извлечение терминов (ATE) – одна из важнейших задач NLP, позволяющая выявлять специфические для домена термины, которые используются в задачах поиска информации, машинного перевода и построения онтологий. Кросс-доменное извлечение вложенных терминов еще больше усложняет задачу, поскольку традиционные методы часто не справляются с иерархическими структурами терминов и новыми доменами. В данной статье представлены новый набор данных CL-RuTerm3, содержащий разметку

фрагментов с терминами и включающий тексты шести областей (главной является компьютерная лингвистика, также присутствуют области математики, медицины, экономики, литературоведения и агрохимии), и конкурс RuTermEval-2024, созданный для оценки систем извлечения терминов с использованием этих данных. CL-RuTerm3 включает 1270 аннотаций и 15 полнотекстовых статей (более 165 тысяч токенов с более чем 37 тысячами размеченных терминов) и является крупнейшим в своем роде для русских научных текстов. Термины разделены на три категории в зависимости от своей лексической и доменной специфики: специфические термины, общие термины и номены. Уникальные особенности набора данных, такие как разметка вложенных терминов и включение текстов различных научных областей, позволяют более реалистично оценивать системы ATE. В заключение статьи приводится анализ подходов участников конкурса RuTermEval-2024, подчеркивается эффективность методов контрастивного обучения. Данная работа направлена на развитие исследований в области ATE путем предоставления надежного набора данных и продолжения развития методологии извлечения терминов.

Ключевые слова: автоматическое извлечение терминов; термин; вложенный термин; набор данных

1 Introduction

Automatic Term Extraction (ATE) is an NLP task used to automatically identify and extract domain-specific terms from a collection of texts. These terms typically represent key concepts within a specialized field, such as medicine, engineering or linguistics. As units of knowledge in a specific field of expertise, extracted terms are not only beneficial for terminographical tasks, but also support and improve several complex downstream tasks, e.g., information retrieval, machine translation, topic detection, and topic modeling, etc.

Despite the significant research interest that automatic term extraction has received, it remains a very challenging task. Terms are generally defined as "textual expressions that denote concepts in a specific field of expertise" (Tran et al., 2023); however, such definitions leave room for many questions about the fundamental nature of terms. Some of the most fundamental differences in terms' basic characteristics are term POS-pattern (only nouns and noun phrases or including other POSes), minimum term frequency and length (in number of tokens). More difficult to quantify are differences such as how specialized or domain-specific a lexical unit needs to be before it is considered a term. The lack of agreement among researchers on even basic characteristics of task is a significant hurdle for all aspects of ATE, from data collection to benchmarking and comparative research in general.

We defined a *term* as follows: it is a word or collocation (multiple syntactically connected words) naming a concept, object, feature or action characteristic of a certain scientific area (domain). The main property of a term is its domain specificity, i.e. belonging to a certain domain. Differences in the so-called lexical specificity, i.e. usage only among a limited group of experts, are expressed by assigning of the labeled terms to different classes.

The task of identifying "flat" terms in the document $D = \{w_0, w_1, \dots, w_{n-1}\}$ is a recognizing entities $t = \{w_i, w_{i+1}, \dots, w_{i+m-1}\},\ 1 \le m < n$, moreover if a word is included in one term, it cannot be included in any other term (formula 2.1). By *nested terms* annotation we mean both classical nested entities setups – a term can be a substring of any other term, or terms can only be intersecting (have a common part). In this setup word can be part of several terms, and also an entire term can be included in another one (formula 2.2). Maximum depth (number of terms in one text fragment) is not limited, nor are the classes of nested terms – they can be either the same or different.

$$t \subseteq D \mid t = \{w_i, w_{i+1}, \dots, w_{i+m-1}\},\ D = \{w_0, w_1, \dots, w_{n-1}\}, 1 \le m < n,\ a \ne b \implies t_a \ne t_b \quad (1)$$

$$\forall\, w, a, b\colon\ w \in t_a \mid a \ne b \implies w \notin t_b \quad (2.1)$$

$$\exists\, w \mid w \in (t_a \cap t_b), \qquad a \ne b \quad (2.2)$$

Cross-domain nested term extraction presents unique difficulties. The syntactic and semantic properties of terms can vary significantly between domains (examples of terms in each class and domain can be seen in Table 2). Existing term extraction approaches often struggle to maintain robust performance when applied to diverse datasets due to their reliance on domain-specific heuristics, frequency-based statistical measures, or supervised models trained on limited annotated corpora. Furthermore, the hierarchical nature of nested terms complicates extraction, as traditional ATE methods primarily focus on flat term structures, failing to account for the compositional relationships between short and long term candidates.

domain	specific term	common term	nomen
computational linguistics	генитивная ИГ pluralia tantum «ё»-омограф разрешение омонимии претренированная ruGPT3 модель конструкция с внешним посессором	слог словарь web-браузер татарский язык ключевое слово	C++ тезаурус RussNet Национальный корпус русского языка Идеографический словарь О.С.Баранова
agrochemistry	α-амилаза 9,10-дифенилантрацен N-удобрение H+/OH−-равновесие диоксид кремния 4-польные севообороты с короткой ротацией	почва растение сорт удобрение сельское хозяйство	Реаком-Хелат Бора Канская лесостепь штамм PGPB Pseudomonas plecoglossicida 2,4-D
literature studies	имажинист 6-стопный ямб дискурс телесности образно-мотивный комплекс	перевод глава публикация советский писатель	«Евгений Онегин» А.С. Пушкин Всероссийский союз поэтов газета «Северная пчела»
medicine	HLA-ген фактор фон Виллебранда травма периферической нервной системы her2-позитивный рак молочной железы	лечение мозг болезнь лекарственное средство	SARS-CoV-2 Thymus Serpyllum L. Федеральный регистр доноров костного мозга
economics	ВВП виртуальная валюта рынок труда модель глобального доминирования публично котируемые нефинансовые корпорации	компания ресурсы организация	БРИКС Великая депрессия Новый банк развития
mathematics	базис s-сплайн (0,∞]-значные функции метод наименьших квадратов	формула точка произведение	Галуа монография Л. Фукса «Бесконечные абелевы группы»

Table 1: Examples of annotated terms in each class and domain

Creating domain-specific corpora with annotated terms is time- and effort consuming. When manual term annotation is involved, inter-annotator agreement is notoriously low and there is no consensus about an annotation protocol. This leads to a lack of available resources. Moreover, it means that the few available datasets are difficult to combine and compare, and often cover only a single language and domain. Also, very few ATE datasets include Russian texts, so the creation of such a dataset is relevant. Additional features of our dataset are nested term annotation and including several domains, which allows to considering the term extraction task in a more real-world setting. Of course, we do not claim

to be fully correct about the chosen methodology and the completeness of the dataset, given that it is based on texts of only one language. In order to compare ATE models and discuss the relevant issues using the created dataset, the RuTermEval-2024 competition was prepared within the framework of the Dialogue conference.

2 Related Work

Problems in creating data for an ATE task usually begin with ambiguity in understanding the search object, which strongly depends on the goals of ATE step, i.e., what the identified units are to be used for next, and actual belonging to terms from a linguistic point of view takes a back seat. It is worth noting that the problem of different understanding of the ATE search object also comes from the ambiguous linguistic nature of the term, its "multifacetedness". There are only few large annotated resources available for the task and they are usually monolingual and cover only a single domain. Since term characteristics, and therefore also ATE performance, can vary greatly between languages and domains, this is a serious drawback.

The most widely used datasets are GENIA (Kim et al., 2003) with 2,000 abstracts from MEDLINE database and ACL-RD-TEC 2.0 (Qasemizadeh & Schumann, 2016) with 300 abstracts from the ACL Anthology Reference Corpus, both in English. Although all annotated datasets exist: CRAFT (Cohen et al., 2017), TTC (Gornostay et al., 2012), KAS-term and KAS-biterm (Ljubešić et al., 2018), etc., the general lack of large, multilingual, and multi-domain resources remains a critical limitation. Given that term characteristics – and by extension, ATE performance – vary significantly across languages and domains, this constraint is a considerable drawback.

In the vast majority of datasets, the ATE task is set just in recognition setup. More interesting is the approach implemented in the ACTER dataset. It is a specialized corpora in three languages and four domains, which markup includes four classes. The main division of classes was made by the level of lexical specificity (its comprehensibility only to a domain expert or any person) and domain specificity (belonging to a certain domain or being used in any research): so there appeared special/specific terms – domain-specific and lexically specific, common terms – domain-specific but not lexically specific (comprehensible even to a non-specialist) and out-of-domain terms – specific only lexically, but used in any domain. The Named Entities class was additionally introduced for unique names of objects of any domain (Terryn et al., 2020). Nevertheless, the TermEval shared task (Terryn et al., 2020) did not use multiclass annotations, it was conducted just in an extraction setup. Despite the novelty of the ACTER partitioning approach, it has not received due attention, although, in our opinion, it is the classification of the units marked up in scientific texts depending on their lexical and domain specificity that can help to prepare a universal dataset for the ATE task for its various applications.

To our knowledge, there are no open datasets for multi-domain nested ATE task in Russian scientific articles. There are several annotated corpora for other task and texts, for example, RuDReC corpus (Tutubalina et al. 2021) was made for NER in consumer reviews about pharmaceutical products (RuDReC also includes not-labeled part of health-related user-generated texts from various Internet sources), not scientific articles or its abstracts. The known NEREL-BIO dataset (Loukachevitch et al., 2023) is the largest annotated dataset of scientific articles in Russian (766 abstracts), but it needs a strong adjustment to be used for ATE purposes, since it was created to solve the NER problem and contains 41 semantic classes of the searched units, some of which are not domain-specific. Also, NEREL-BIO is monodomain, and scaling its markup methodology to other domains is hardly possible due to differences in the composition of semantic classes.

It should be noted that due to the complexity of creating ATE annotated datasets, most of them consist only of abstracts, but not the full texts of scientific articles, and the ATE models using existing data usually work within this limitation. We believe this approach is appropriate, but to understand models' capability to solve the problem on full articles, a small number of such texts were included in the development and test sets of first and second tracks.

The RuTermEval shared task aims to provide a valuable new resource while simultaneously advancing understanding of the state-of-the-art in ATE, identifying existing strengths and weaknesses, and inspiring novel approaches in the field. The CL-RuTerm3 dataset includes six domains and enables participants to train and evaluate their systems using diverse and detailed data. Despite using the

TermEval approach, our shared task is the first competition dedicated to the task of classifying terms by their specificity, since TermEval was conducted as an extraction setup (without classification).

3 Dataset

The CL-RuTerm3 dataset of abstracts and full texts of scientific articles was prepared specially for the RuTermEval–2024 shared task. The basis of our dataset were the proceedings of the conference "Dialogue" for 2000–2023 in Russian (1055 abstracts and 15 full-texts articles of computational linguistics domain). To test the scalability of the models to other domains, about 220 abstracts from five additional scientific areas (mathematics, medicine, agrochemistry, literature studies and economics) were also included in dataset. More detailed quantitative description of our dataset is presented in Table 1.

The uniqueness of the material is that linguistics has hardly been considered before in the ATE task (apart from ACL RD-TEC datasets) and is a new material for cross-domain experiments and for analyzing the differences in NLP processing of term systems of different groups of sciences. An additional feature of CL-RuTerm3 is markup of nested terms.

3.1 Term annotation

The markup was conducted by three annotators – specialists in linguistics, information technology and mathematics – single assessor for each document with collective discussion of challenging cases. Quality control was performed by moderator with experience in the field of terminology, ATE and dataset creation. The moderator and assessors created an assessor's guideline with a detailed description of the annotation task and a breakdown of correct and incorrect annotations. This helped assessors to make consistent decisions and make the whole process more transparent. Nevertheless, term annotation remains an ambiguous and subjective task, and we do not claim that ours is the only possible interpretation.

For the task of classifying terms, we proposed to divide them according to the degree of lexical specificity of the term (its familiarity among ordinary people who are not domain experts) and denotatum uniqueness:

1. specific term – terms that are both domain-specific and lexically-specific;
2. common term – domain-specific terms (known and used by non-specialists);
3. nomen – unique names of objects belonging to a particular domain, including nomenclature names (datasets, programming languages, corpora and dictionaries, scientific schools, writers and scientists, etc.).

Classification of terminological units by specificity (lexical and domain) was first applied in the ACTER dataset, in which classification was done in 4 classes (Specific Terms, Common Terms, Out-of-Domain Terms and Named Entities). In contrast to the ACTER, we dropped the Out-of-domain terms class. Units of this class that denote mathematical concepts (e.g., *p-value*, *confidence interval*, etc.), but belong to out-of-domain terms because of their use in any domain were marked by us as specific terms, and general scientific vocabulary that has neither lexical nor domain specificity (*method*, *research*, *experiment*, etc.) was omitted.

Markup was conducted in the format of a sequence labeling task – fragments with a term were identified and classified into three classes (a markup example is shown in Figure 1). In the output, the markup of each text is represented as a list of triples [start_index, end_index, term_class] for each labeled term.

Figure 1: Markup sample from the CL-RuTerm3 dataset

domain	type of texts	text count	token count	annotation count	unique term count	lemmatized term count
computational linguistics	abstracts	1,053	97,296	22,673	10,684	7,812
	full texts	15	32,254	5,190	2,303	1,539
agrochemistry	abstracts	55	10,321	3,110	1,591	1,267
literature studies	abstracts	60	13,211	2,478	1,578	1,218
medicine	abstracts	40	7,825	1,877	1,138	907
economics	abstracts	30	6,559	1,065	747	609
mathematics	abstracts	32	2,924	746	482	312
Σ	full texts	15	32,254	5,190	2,303	1,539
	abstracts	1,270	138,136	31,949	16,002	11,903
	all	1,285	170,390	37,139	17,652	12,938

Table 2: Quantitative characteristics of CL-RuTerm3 dataset

Consider some aspects of markup using the text in Figure 1 as an example. In phrase "National corpus of the Russian language" the following terms are identified: “National corpus of the Russian language” (as unique name of domain-specific product), “corpus”, “Russian language” and “language”.

Words “national” and “Russian” are not annotated because we do not labeled adjectives without the substantive defined by them as independent term. An exception to the rule is that if such a word is not near each other (forming a discontinuous term), the adjective is annotated as a single-word independent term, as can be seen in the example “morphological and taxonomic annotation of texts”, where “morphological” is a term, but “taxonomic” is not.

The phrases “national corpus” and “corpus of the Russian language” are not annotated because they are not separate scientific units that are often reproduced in the domain or have paradigmatic relations with other terms (in contract to “parallel corpus” or “dialect corpus”). If a term is supplemented by a feature, but the meaning of this phrase does not differ in any way from the sum of the meanings of its parts, phrase’ reproducibility in the domain and/or the existence of systemic relations with other terms become significant.

The full description of markup rules requires a separate scientific coverage due to its volume because of the complexity of the markup task and its linguistic multifacetedness. In this article we will mention only the main features of our markup rules. In general, each labeled term should remain a term even out of context, being part of the lexical system of a particular domain.

To the assessor, an expression or word is most likely to be a term if:

- It is a regular name (reproduced as a result of a single act of speech production and predominantly in an observable speech form).

- It has a definition.
- It can be found in dictionaries or ontologies of the relevant domain.
- It is related to the valid term (e.g., they are hyponyms of the same hyperonym or are in a generic-species relationship).
- It is an abbreviation used for the valid term.
- It has an abbreviation used in the particular domain.
- It is an object of research in the relevant domain.

To improve the quality of the final dataset markup, inaccuracies in the identification of term boundaries were corrected. Additional discussions were also carried out on the markup of terms that had a significant number of labels of different classes, as it is preferable for a term to belong to a particular class.

Final dataset consists of 1,285 text in Russian (over 165k tokens) with ≈37k sequence annotations, provided by domain and terminology human experts, more detailed quantitative description of our dataset is presented in tables 2 and 3.

domain	type of texts	annotation count	class	annotation count	lemmatized term count	mean lemmatized term frequency
computational linguistics	abst.	22,673	specific	15,946	6,748	2.385
			common	5,906	593	9.708
			nomen	821	471	1.741
	full texts	5,190	specific	3,826	1,236	3.224
			common	1,005	81	10.444
			nomen	359	222	1.617
agrochemistry	abst	3,110	specific	2,343	1,121	2.090
			common	655	68	9.632
			nomen	112	78	1.436
literature studies	abst	2,478	specific	927	485	1.903
			common	490	59	8.373
			nomen	1,061	674	1.574
medicine	abst	1,877	specific	1,603	851	1.895
			common	234	27	8.296
			nomen	40	29	1.379
economics	abst	1,065	specific	945	560	1.686
			common	54	16	3.438
			nomen	66	33	2.000
mathematics	abst	746	specific	617	270	2.226
			common	101	26	4.500
			nomen	28	16	1.750

Table 3: Distribution of terms in each class and domain

4 Setup

4.1 Tasks

The RuTermEval-2024 Shared task features three sub-tasks:

1. Nested term extraction (in sequence labeling format);
2. Nested term extraction (in sequence labeling format) and classification (labels are specific, common, nomen);
3. Nested term extraction (in sequence labeling format) and classification (labels are specific, common, nomen) in cross-domain task.

All tracks assume a transfer learning task, so the test set in each of them includes such texts, the likes of which were absent in the train set. Thus, in tracks one and two, the test set includes texts of a different genre – in addition to abstracts as in the train set, it includes full articles. In the third track, the test set consists of texts of only domains that were not present in the training set.

4.2 Evaluation

For all tracks, only full term matches were considered.

The metric for the first task – term extraction without class consideration – is the averaged F1-score across all documents, with abstracts and full texts treated separately:

$$F1_{T1} = \frac{1}{2}\left(\frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i} + \frac{1}{n_{ft}}\sum_{i=1}^{n_{ft}} F1_{ft_i}\right)$$

where n_{abst}, n_{ft} represent the number of abstracts and full-text documents, respectively, and $F1_{ann_i}$, $F1_{ft_i}$ denote the F1-score for each individual text.

To evaluate the quality of solutions for the second task, a weighted F1-score across various classes was employed:

$$F1_{T2} = \frac{1}{6}\left(3 * F1_{spec} + 2 * F1_{comm} + F1_{nom}\right)$$

where $F1_{spec}$, $F1_{comm}$, $F1_{nom}$ represent the F1-scores for specific, common terms, and nomens, respectively. The weights for each class were set according to their importance to the overall task.

Each individual $F1_t$ is calculated similarly to the metric from the first task:

$$F1_t = \frac{1}{2}\left(\frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i} + \frac{1}{n_{ft}}\sum_{i=1}^{n_{ft}} F1_{ft_i}\right)$$

where $t \in \{spec, comm, nom\}$.

For the third task, a weighted F1-score analogous to that used in the second task was applied:

$$F1_{T3} = \frac{1}{6}\left(3 * F1_{spec} + 2 * F1_{comm} + F1_{nom}\right)$$

However, since full texts were not available for the third task, each individual $F1_t$ was calculated as the averaged F1-score across all abstracts:

$$F1_t = \frac{1}{n_{abst}}\sum_{i=1}^{n_{abst}} F1_{abst_i}$$

where $t \in \{spec, comm, nom\}$.

For the second and third tracks, an additional F1-score without class consideration (analogous to the first track) was computed. This allowed for a separate determination of the quality of term extraction without classification.

4.3 Baseline

To compare participants' models with the simplest solution baseline was prepared. All the entities labeled as terms were selected with their annotations count, as well as their total occurrence in the corpus (in any grammatical form). Further, the optimal ratio of the number of markings as a term to its frequency in the corpus was empirically selected – 0.35. With the obtained bag of terms (with retained classes), a test set was marked up to obtain the required kind of markup.

4.4 Dataset Splits

The training data is the same for all subtasks, but in first track the term classes were removed. The training data consists of 850 texts (77k tokens, 18k annotations) in computational linguistics domain.

Development and test set for first and second subtasks is also the same, they belong to the train domain. Development data includes 103 abstracts and 10 full-text articles (25k tokens, 5k annotations), test set consists of 100 abstracts and 5 full-text articles (27k tokens, 5k annotations).

The development and test set for the third track consists of abstracts from domains, which were absent in train data. Development data contains 115 texts (24k tokens, 6k annotations). The test set consists of 102 texts (17k tokens, 4k annotations), but more diverse in domain component.

5 Results and Discussion

Seven teams participated in the RuTermEval–2024. Six teams took part in the final testing phase (distribution across tracks: 5, 5, 3). We provide descriptions of the solutions submitted by the top teams. The names are listed as the participants registered them in CodaLab. The final results are presented in separate tables for each track (in the case of multiple submissions, only the best one is considered).

We summarize participating teams' methods. A detailed description of their work can be found in their own articles.

fulstock [LAIR RCC MSU] participant who won all tracks treated nested term extraction as a NER task using the Binder model (Zhang et al., 2023), which employs contrastive learning to extract nested terms. Text sequences and class descriptions are encoded using RuRoberta-large, with embeddings mapped to a shared vector space. The model aligns entities of the same type closer to a class-specific center (defined by entity descriptions) while distancing unrelated subsequences, optimizing for term extraction accuracy.

Rank	Participant and team	Scores					
		with classification			without classes		
		F1-class_weighted (avg. per docs)	Precision (avg. per docs)	Recall (avg. per docs)	F1 (avg. per docs)	Precision (avg. per docs)	Recall (avg. per docs)
Track 1							
1	**fulstock [LAIR RCC MSU]**	–	–	–	**79.40%**	**79.69%**	**80.12%**
2	**VladSemak [VSemak]**	–	–	–	76.85%	79.40%	75.72%
–	***baseline***	–	–	–	61.69%	78.75%	52.68%
3	**ivan_da_marya**	–	–	–	56.19%	68.03%	50.36%
4	ragunna [KiPL SPBU]	–	–	–	53.49%	58.64%	51.83%
5	angyling	–	–	–	53.33%	42.29%	76.23%
Track 2							
1	**fulstock [LAIR RCC MSU]**	**69.97%**	70.45%	**70.53%**	**77.79%**	78.77%	**78.00%**
2	**VladSemak [VSemak]**	69.96%	**72.18%**	69.20%	77.26%	**79.85%**	76.38%
–	***baseline***	58.26%	67.09%	57.03%	61.69%	78.75%	52.68%
3	**VatolinAlexey [ai]**	57.97%	63.18%	58.79%	63.47%	68.97%	60.75%
4	ragunna [KiPL SPBU]	50.43%	55.98%	51.31%	52.04%	57.07%	50.44%
5	angyling	31.37%	27.83%	42.32%	53.33%	42.29%	76.23%
Track 3							
1	**fulstock [LAIR RCC MSU]**	**48.23%**	53.23%	48.85%	**60.38%**	67.90%	56.77%
2	**VladSemak [VSemak]**	46.54%	**57.85%**	46.22%	50.88%	**71.26%**	42.75%
3	**angyling**	43.70%	45.16%	**51.18%**	52.81%	49.36%	**60.74%**
–	*baseline*	22.76%	40.14%	30.93%	11.81%	34.87%	7.59%

Table 4: All participants' scores

VladSemak used the *span classification* approach (Binder) and contrastive learning with two BERT encoders to map text spans and term descriptions into a shared vector space. It evaluates all spans up to a set length, identifying terms via vector similarity. The method maximizes similarity for term spans with their type descriptions and minimizes it for non-terms.

ivan_da_marya team used a span classification approach, generating lemmatized n-grams (length 1 to 4) and encoding them with SBERT. Span and sentence vectors, along with POS tags, were concatenated and fed into a KNN classifier to identify terms.

angyling tested one-shot prompting on the Qwen2.5 (An el at., 2024). The prompt included an explanation of the task with one example from the training set, and an explanation of the correct form of answer inference. Next, all terms extracted by the model were used to final annotation in nested setup. All entities were labeled as specific class.

The majority of submissions of first track demonstrated higher precision than recall. However, in the winning solution, recall played a decisive role in the final evaluation. Notably, the baseline is competitive with top-performing solutions, and approaches provided by other participants failed to this simple model. A similar pattern was observed in the second track. Solutions outperforming the baseline achieved significantly higher recall. Additionally, both leading solutions performed best on specific terms, while their performance on nomens was noticeably weaker. Interestingly, the baseline handled common terms exceptionally well, but underperformed on specific terms because of low recall. As expected, baseline showed the highest precision in identifying specific and nomen terms, and score on common terms was also quite high. All participants' scores per class are summarized in the Table 5.

Participant	**Specific**			**Common**			**Nomen**		
	Micro-average F1	Precision (class-wide)	Recall (class-wide)	Micro-average F1	Precision (class-wide)	Recall (class-wide)	Micro-average F1	Precision (class-wide)	Recall (class-wide)
Track 2									
fulstock	**74.81%**	77.66%	**72.16%**	69.70%	58.80%	**85.57%**	35.15%	49.33%	**27.31%**
VladSemak	74.36%	78.04%	71.01%	**73.19%**	**65.63%**	82.70%	**35.79%**	62.39%	25.09%
baseline	56.55%	**78.10%**	44.32%	70.48%	61.47%	82.57%	32.39%	**70.37%**	21.03%
VatolinAlexey	39.45%	69.52%	27.54%	43.79%	56.47%	35.76%	25.22%	61.43%	15.87%
ragunna	36.44%	52.92%	27.78%	48.13%	62.14%	39.27%	15.85%	45.61%	9.59%
angyling	43.73%	33.26%	63.80%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Track 3									
fulstock	**61.92%**	67.63%	57.09%	17.20%	18.67%	15.94%	**18.08%**	37.21%	**11.94%**
VladSemak	54.60%	**71.23%**	44.27%	18.57%	**37.50%**	12.34%	13.07%	**52.63%**	7.46%
angyling	48.39%	40.88%	**59.30%**	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
baseline	9.08%	42.04%	5.09%	**26.44%**	32.34%	**22.37%**	0.00%	0.00%	0.00%

Table 5: Scores of participants' models in each term class

Note that only models based on contrastive learning approach outperformed the baseline in the first and second tracks, which shows that within the domain particular work, applying to dictionary sources (a specific list of terms) can allow to quickly achieve acceptable results and even be more effective than other methods. Improving the quality requires the use of already more advanced and sophisticated methods, e.g., contrastive learning, and demands further research and experimentation.

It is worth to mention that the approach using generative LLM with just one sample (by angyling) showed very high recall comparable to the leaders, who used all labeled data of the same domain as the test set. In the more equal conditions of the third track, when no one had data from the test domain (but did have data from another domain), generative LLM approach achieved a comparable result to the leaders primarily because of its the best recall, in contrast to the top two participants, whose precision, substantially exceeding their recall values, contributed most to the final scores. Also, this approach performed best scores on the economics domain, which proved to be the most challenging for the other models (see Table 6 for a look at the quality for each domain), although best results on the math and medicine are achieved by the fulstock with using Binder model (contrastive learning).

Rank	Participant and team	Domain	F1-class_weighted (avg. per docs)	Macro-average F1	Precision (avg. per docs)	Recall (avg. per docs)
1	**fulstock [LAIR RCC MSU]**	Mathematics	**51.88%**	**53.02%**	**52.30%**	**53.75%**
		Medicine	**52.86%**	**53.94%**	57.91%	**50.48%**
		Economics	38.18%	40.24%	46.68%	35.36%
2	**VladSemak [VSemak]**	Mathematics	48.93%	49.75%	50.33%	49.18%
		Medicine	49.51%	52.63%	**66.48%**	43.55%
		Economics	40.01%	43.43%	**56.01%**	35.47%
3	**angyling**	Mathematics	47.35%	48.01%	45.65%	50.64%
		Medicine	38.76%	39.93%	35.88%	45.01%
		Economics	**46.38%**	**47.83%**	42.11%	**55.34%**
4	*baseline*	Mathematics	26.20%	30.06%	43.36%	23.00%
		Medicine	26.19%	29.44%	38.03%	24.02%
		Economics	14.52%	16.04%	18.53%	14.14%

Table 6: Scores of participants' models in each domain (Track 3)

6 Conclusion

We presented the RuTermEval-2024 shared task, the first shared task on cross-domain nested term extraction and classification for the Russian language. The CL-RuTerm3 dataset is based on the proceedings of the "Dialogue" conference from 2000 to 2023 (1055 abstracts and 15 full-text articles in the computational linguistics domain). To evaluate the models' generalizability to other domains, approximately 220 abstracts from five additional scientific areas (mathematics, medicine, agrochemistry, literature studies, and economics) were also included. To our knowledge, this dataset exceeds the volume of all existing open datasets for solving a similar problem for Russian scientific texts.

The competition attracted 13 submissions from seven teams, addressing the problem through three subtasks. The best results were achieved by the solution that utilized the Binder model with contrastive learning. All solutions performed well in extracting terms, but they were slightly less effective in classifying the extracted terms. Within a single domain, only models based on contrastive learning outperformed the baseline approach, which relied solely on labeling terms found in the train and dev sets. This suggests that leveraging dictionary sources (or other resources of domain-specific terms) can enable rapid achievement of acceptable results. Further improvements will require the application of more advanced methods, such as contrastive learning and other emerging techniques.

As expected, performance declined when identifying terms from domains absent in the train and dev sets (Track 3). However, a generative LLM approach (with Qwen2.5) achieved strong results, primarily due to high recall despite minimal annotation samples (just one example from the training set). This result highlights the potential of generative LLMs for term extraction in low-resource scenarios and underscores the need for further research and experiments.

The shared task dataset, codebase and other materials are available online on GitHub[*]. We welcome the communities of NLP developers, linguists, and engineers to contribute to further research in the area.

Acknowledgements

The authors are grateful to Julian Serdyuk from Lomonosov Moscow State University for support in researching the discussed theme and Alsu Vagapova from Kazan Federal University for help in annotating the data.

[*] https://github.com/mlsa-iai-msu-lab.

References

[1] An Yang, Baosong Yang et al. Qwen2.5 Technical Report // Computing Research Repository. – 2024. – Vol. arXiv:2412.15115. Access mode: https://arxiv.org/abs/2412.15115.

[2] Cohen K. Bretonnel, Verspoor Karin et al. The Colorado Richly Annotated Full Text (CRAFT) Corpus: Multi-Model Annotation in the Biomedical Domain // Handbook of Linguistic Annotation. – Dordrecht: Springer, 2017. – P. 1379–1394. – DOI: http://dx.doi.org/10.1007/978-94-024-0881-2_53.

[3] Gornostay Tatiana, Gojun Anita et al. Terminology Extraction, Translation Tools and Comparable Corpora: TTC concept, midterm progress and achieved results // LREC 2012 Workshop on Creating Cross-language Resources for Disconnected Languages and Styles (CREDISLAS), Istanbul, Turkey. – 2012. – P. 1–4.

[4] Kim Jin-Dong, Ohta Tomoko, Tateisi Yuka, Tsujii Junichi. GENIA corpus – a semantically annotated corpus for bio-textmining // Bioinformatics. – 2003. – Vol. 19. – Suppl. 1 – P. i180–i182. – DOI: https://doi.org/10.1093/bioinformatics/btg1023.

[5] Ljubešić Nikola, Erjavec Tomaž, Fišer Darja. KAS-term and KAS-biterm: Datasets and baselines for monolingual and bilingual terminology extraction from academic writing // Proceedings of the Conference on Language Technologies and Digital Humanities. – 2018. – P. 168–174.

[6] Loukachevitch Natalia, Manandhar Suresh et al. NEREL-BIO: A Dataset of Biomedical Abstracts Annotated with Nested Named Entities // Bioinformatics. – 2023. – Vol. 39. – Issue 4. DOI: https://doi.org/10.1093/bioinformatics/btad161.

[7] QasemiZadeh Behrang, Schumann Anne-Kathrin. The ACL RD-TEC 2.0: A Language Resource for Evaluating Term Extraction and Entity Recognition Methods // Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16). – Portorož, Slovenia: European Language Resources Association (ELRA), 2016. – P. 1862–1868. Access mode: https://aclanthology.org/L16-1294.pdf.

[8] Terryn Ayla Rigouts, Hoste Véronique, Drouin Patrick, Lefever Els. TermEval 2020: Shared Task on Automatic Term Extraction Using the Annotated Corpora for Term Extraction Research (ACTER) Dataset // Proceedings of the 6th International Workshop on Computational Terminology. – Marseille, France: European Language Resources Association, 2020. – P. 85–94.

[9] Terryn Ayla Rigouts, Hoste Véronique, Lefever Els. In no uncertain terms: a dataset for monolingual and multilingual automatic term extraction from comparable corpora // Language Resources and Evaluation. Springer, 2020. – Vol. 54(6). – P. 385–418. DOI: https://doi.org/10.1007/s10579-019-09453-9.

[10] Tran Thi Hong Hanh, Martinc Matej, Caporusso Jaya, Doucet Antoine, Pollak Senja. The Recent Advances in Automatic Term Extraction: A survey // Computing Research Repository. – 2023. – Vol. arXiv:2301.06767. – DOI: https://doi.org/10.48550/arXiv.2301.06767.

[11] Tutubalina Elena, Alimova Ilseyar et al. The Russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews // Bioinformatics. – 2021. – Vol. 37. – Issue 2. – P. 243–249. DOI: http://dx.doi.org/10.1093/bioinformatics/btaa675.

[12] Zhang Sheng, Cheng Hao, Gao Jianfeng, Poon Hoifung. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning // Computing Research Repository. – 2023. – Vol. arXiv:2208.14565. Access mode: https://arxiv.org/abs/2208.14565.

BERT-like Models for Automatic Morpheme Segmentation of the Russian Language

Dmitry Morozov
Russian National Corpus &
Novosibirsk State University
Novosibirsk, Russia
morozowdm@gmail.com

Anna Glazkova
Russian National Corpus &
University of Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

Timur Garipov
Russian National Corpus &
Novosibirsk State University
Novosibirsk, Russia
garipov154@yandex.ru

Abstract

Current approaches to automatic morpheme segmentation for the Russian language rely on machine learning, primarily neural network methods. Among the architectures presented, the best results have been achieved using convolutional neural networks and LSTM networks. However, the quality of automatic annotation is far from ideal, especially when dealing with roots that were not present in the training dataset. In this work, we present a new approach to morpheme segmentation based on fine-tuning BERT-like models. Through comparisons using two morpheme dictionaries with different segmentation paradigms, we demonstrated the superiority of our approach over previous ones, including when working with unfamiliar roots. The best result was achieved by fine-tuning the RuRoBERTa-large model: when working with random words, the share of completely correct segmentations increased from 88.5-90.8% to 92.5-93.5%, and when working with unfamiliar roots, it improved from 70.5-72.6% to 74.9-77.2%. Error analysis of the model showed that root nests not encountered in the training dataset can be distributed into two groups during testing: "recognizable", meaning those for which more than 90% of the words are correctly analyzed, and "unknown", meaning those for which the proportion of correct segmentations is less than 10%.

Keywords: automatic morpheme segmentation, Russian language morphology, machine learning, BERT
DOI: 10.28995/2075-7182-2025-23-257-266

BERT-подобные модели для автоматического построения морфемных разборов слов русского языка

Дмитрий Морозов
НКРЯ & НГУ
Новосибирск, Россия
morozowdm@gmail.com

Анна Глазкова
НКРЯ & ТюмГУ
Тюмень, Россия
a.v.glazkova@utmn.ru

Тимур Гарипов
НКРЯ & НГУ
Новосибирск, Россия
garipov154@yandex.ru

Аннотация

Актуальные подходы к автоматическому построению морфемных разборов для русского языка опираются на машинное обучение и, в первую очередь, нейросетевые подходы. Среди представленных архитектур лучших результатов удалось добиться при помощи свёрточных нейронных сетей и LSTM-сетей. Тем не менее, качество автоматической разметки далеко от идеального, в особенности, при работе с корнями, не встретившимися в обучающей выборке. В настоящей работе мы представляем новый подход к морфемной сегментации на базе дообучения BERT-подобных моделей. В ходе сравнения на материале двух морфемных словарей с различной парадигмой членения мы продемонстрировали превосходство нашего подхода над предыдущими, в том числе, при работе с незнакомыми корнями. Лучший результат был достигнут при помощи дообучения модели RuRoBERTa-large: при работе со случайными словами качество разметки выросло с 88,5-90,8% до 92,5-93,5%, при работе с незнакомыми корнями — с 70,5-72,6% до 74,9-77,2%. Анализ ошибок модели показал, что корневые гнёзда, не встретившиеся в обучающей выборке, при тестировании распределяются на две группы: «узнаваемые», то есть такие, для которых более 90% слов разобрано корректно, и «неизвестные», то есть такие, для которых доля корректных разборов составляет менее 10%.

Ключевые слова: автоматическое построение морфемных разборов, морфемика русского языка, машинное обучение, BERT

1 Introduction

Morpheme segmentation refers to the division of a word into several morpheme substrings. Morpheme segmentation is one of the key aspects of studying the morphology of the Russian language, allowing the analysis of how words are formed from minimal meaningful units — roots, prefixes, suffixes, and others. This enables a deeper understanding of the structure of words, revealing patterns in their formation and changes. It is also in demand in the teaching of the Russian language, as many spelling rules rely on the ability of the speaker to identify individual morphemes within a word (Kisselev et al., 2024). Morpheme segmentation also shows promise as a subword tokenization strategy for language models. Recent studies (e.g., (Matthews et al., 2018; Nzeyimana and Niyongabo Rubungo, 2022)) demonstrate that morpheme-based tokenizers can outperform standard Byte-Pair Encoding (BPE) (Gage, 1994) in certain scenarios, yielding measurable improvements in model performance. Finally, morpheme annotation is used in large text corpora (Savchuk et al., 2024), which are employed for linguistic research, to enhance user search capabilities.

Since morpheme segmentation can sometimes be a complex task requiring the attention of professional linguists, morpheme dictionaries are created for practical purposes, such as the Word Formation Dictionary of the Russian Language (Tikhonov, 1990) and the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efremova, 1986). However, it cannot be claimed that such dictionaries are comprehensive. Firstly, morpheme dictionaries typically include segmentations only for lemmata, which is not an ideal solution for the Russian language given its rich morphological possibilities. Secondly, the largest existing morpheme dictionaries contain around 100,000 to 150,000 lemmata, which is significantly smaller than the overall volume of the Russian language; for example, the Main Corpus of the Russian National Corpus (Savchuk et al., 2024) contains 200,000 to 250,000 different lemmata. A further complication arises from the differences in the paradigms of morpheme segmentation used in these dictionaries. Some dictionaries, such as the Word Formation Dictionary of the Russian Language, focus on the transparency of relationships and the presence of word formation chains in modern Russian, while others, like the Dictionary of Morphemes of the Russian Language, pay more attention to the etymology of words and their correspondence with other lexemes of similar structure. As a result, segmentations of the same lemma often differ across dictionaries. Additionally, authors may locally make decisions that contradict the paradigms they declare, complicating the automatic updating of dictionaries, while manual updates require significant regular investments of time and resources due to the emergence of new words in the language.

In this context, approaches based on machine learning may serve as suitable tools. A number of studies have previously been published on the Russian language (Sorokin and Kravtsova, 2018; Bolshakova and Sapin, 2019b; Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2022; Garipov et al., 2024; Morozov et al., 2024), which examine gradient boosting over decision trees, convolutional neural networks (CNN), and LSTM networks, demonstrating high quality in the generated segmentations. At the same time, it was shown (Morozov et al., 2024) that many presented algorithms perform significantly worse with words containing roots that did not appear in the training dataset: about 90% of fully correct segmentations for random words and around 70% for words with unfamiliar roots. Moreover, it has been demonstrated (Sorokin, 2022) that use of embeddings from pre-trained BERT-like models can improve the performance of algorithms, including on data with unfamiliar roots. Although BERT was used in these works only as a source of embeddings for the convolutional network, Peters and Martins showed (Peters and Martins, 2022) that subword transformers can show good results in the task of morpheme segmentation, including on the material of the Russian language. Thus, this raises a question: will the fine-tuning of pre-trained Transformer-based models be effective in the task of morpheme segmentation? In this paper, we build upon this idea and present a family of fine-tuned BERT-like models for the automatic morpheme segmentation of the Russian language lemmata, which surpass all previously presented approaches. Our models are available for download at `https://ruscorpora.ru/license-content/neuromodels`. The increase in the proportion of fully correct segmentations reached up to 3% for random words and up to 4.5% for words with unfamiliar roots.

2 Related Work

Morpheme segmentation algorithms can be classified according to two criteria. The first of these criteria is the type of segmentation. Two variants of segmentation are distinguished: surface and canonical (Cotterell et al., 2016). The distinction of canonical segmentation lies in the fact that it not only involves breaking a word into morphemes but also reconstructing the original form of the morpheme. For example, for the word "funniest" the surface segmentation is *funn-i-est*, while the canonical segmentation is *fun-y-est*. The second criterion is whether the algorithm's tasks include determining the types of morphemes. Most studies on the Russian language focus on surface segmentation with the identification of morpheme types. A notable exception is the SIGMORPHON 2022 competition (Batsuren et al., 2022), in which the task of canonical segmentation was addressed without determining the type.

The most studied architecture for the Russian language is an ensemble of convolutional neural networks, first presented in the work (Sorokin and Kravtsova, 2018). This model operates in a character-level BMES annotation format, where each letter is assigned a label indicating its position within the morpheme (B for beginning, M for middle, E for end, and S for single letter morpheme) and a label corresponding to one of seven morpheme types: ROOT, PREF (prefix), SUFF (suffix), END (ending), POST (postfix), LINK (linking vowel), HYPH (hyphen). The quality of the model is measured using F-score, precision, and recall for morpheme boundaries, character-level accuracy, and the proportion of fully correct segmentations. Based on the dataset relying on the Word Formation Dictionary of the Russian Language (Tikhonov, 1990), the authors achieve approximately 87% fully correct segmentations and about 96% correctly labeled letters. Within a series of papers (Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2019b), this architecture was compared with models based on gradient boosting over decision trees and with LSTM networks using two datasets: the dataset from the original paper and the CrossLexica dictionary (Bolshakov, 2013). The comparison showed a slight advantage of the ensemble of three LSTM networks over a similar ensemble of convolutional neural networks. An alternative approach is the use of generative models. For example, at the SIGMORPHON 2022 competition, the subword transformer DeepSPIN-3 (Peters and Martins, 2022) demonstrated the best performance across multiple languages, including Russian. In (Morozov et al., 2024), some approaches were compared using three datasets: one based on the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efremova, 1986) and the two previously mentioned, with corrections made to the annotation errors of the morpheme types in the dataset based on the Word Formation Dictionary of the Russian Language. The best results were demonstrated by the ensemble of convolutional neural networks. Additionally, comparing the algorithm's results with expert evaluation indicated that when tested on random words, the algorithms are close to achieving expert-level performance.

At the same time, Morozov et al. (Morozov et al., 2024) studied a key drawback of existing approaches: due to the lack of semantic consideration when dealing with roots that did not appear in the training dataset, the quality of annotation sharply declines: around 90% fully correct segmentations for random words, about 70% for words with unfamiliar roots for CNN and LSTM approaches. For DeepSPIN-3 the situation is even worse: 80% correct segmentations on average and only 14% when working with unfamiliar roots. This shortcoming can be partially addressed by enriching the feature description with embeddings from a pre-trained BERT-like model (Sorokin, 2022; Morozov et al., 2024), but the improvement is relatively modest. Pranjic et al. (Pranjić et al., 2024) proposed a binary classifier based on the Glot500 model (Imani et al., 2023) that determines the presence of a morpheme boundary in a word at the character level. This method is difficult to consider practically significant due to the extremely high computational costs (for a word of length N, the model needs to be run N-1 times). At the same time, the potential of using pre-trained Transformer-based language models as an end-to-end solution remains insufficiently explored.

3 Datasets

We used two datasets, which were previously utilized in (Garipov et al., 2024; Morozov et al., 2024): Morphodict-T, based on the Word Formation Dictionary of the Russian Language (Tikhonov, 1990), and Morphodict-K, based on the Dictionary of Morphemes of the Russian Language (Kuznetsova and Efre-

	Morphodict-T	Morphodict-K
Unique words	95,895	75,649
Unique morphemes	15,899	8,079
Unique roots	15,253	7,148
Average morphemes per word	3.86	4.12
Average morpheme occurrence	23.29	38.56
Average root occurrence	7.54	12.24
Average root length in symbols	5.52	4.62

Table 1: Brief characteristics of the datasets

mova, 1986). Besides differences in the vocabulary, the source dictionaries rely on different paradigms of morpheme segmentation, resulting in a significantly higher fragmentation of morphemes in Morphodict-K. More detailed differences between the datasets are described, for example, in (Morozov et al., 2024). The use of the Morphodict-T dataset allows for better comparison of the obtained results with those from the previous works that utilized a similar dataset (Sorokin and Kravtsova, 2018; Bolshakova and Sapin, 2019a; Bolshakova and Sapin, 2019b). Morphodict-K, on the other hand, enables us to verify whether the results obtained are significantly related to the specifics of the particular dataset. This is particularly important in our case, as we do not claim to develop a universal morpheme segmentation algorithm, which is most likely impossible due to disagreements in defining morpheme segmentation itself (Iomdin, 2019). However, within this study, we investigate whether it is possible to automatically extend annotations from a specific morpheme dictionary to out-of-vocabulary words. Brief characteristics of Morphodict-T and Morphodict-K are provided in Table 1.

4 Models and Experimental Setup

We utilized two BERT-like models pre-trained for the Russian language: RuBERT-base-cased (Kuratov and Arkhipov, 2019) (hereinafter referred to as RuBERT) and RuRoBERTa-large (Zmitrovich et al., 2024) (hereinafter referred to as RuRoBERTa). The choice of these models is due to their well-established performance and application in a wide range of practical tasks. We framed the original task as a specific case of Named Entity Recognition: the model was provided with a sequence of letters forming a word, and the task of the model was to determine the class of each letter in the sequence according to the BMES annotation with types (an example is shown in Figure 1). We tested two options for forming input sequences. In the first case, we provided only the letters and their classes, effectively preventing the model's tokenizer from functioning. In the second case, the starting element of the sequence was the original lemma, which was assigned a special class of "0" (Figure 2). Our hypothesis was that having the complete lemma would allow the model to take into account the semantics of the word and improve its performance with roots that were not present in the morpheme training dataset but had been encountered by the model during the pre-training phase. These models are further designated by the index "lex".

We fine-tuned the models using two different methods for splitting the dataset: fully random split of the dataset and split-by-roots, where words with the same root could not appear in different folds (words with multiple roots were excluded from the dataset in advance). The first method of forming folds allows for an average assessment of the model's annotation quality, while the second one evaluates the annotation quality when dealing with unfamiliar roots. We applied 5-fold cross-validation for a more accurate assessment of the automatic labeling quality.

The models were fine-tuned for 30 epochs, using a learning rate of 4e-5 for RuBERT-base-cased model and 5e-6 for RuRoBERTa-large one. To assess quality, we used the metrics proposed in (Sorokin and Kravtsova, 2018) and supplemented in (Garipov et al., 2024): F-score, precision, and recall for morpheme boundaries (these metrics are further designated by the index "full"); F-score, precision, and

Figure 1: Example of input sequence

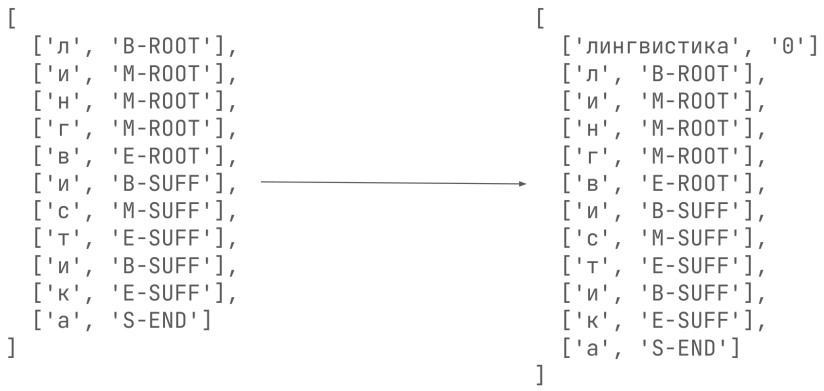


Figure 2: Example of adding lemma to input sequence

recall for root boundaries only (these metrics are further designated by the index “root”); as well as character-level accuracy (Accuracy) and word-level accuracy (WordAccuracy).

5 Results and Discussion

During the training process, we evaluated the quality of the models using word-level accuracy (WordAccuracy). In the case of the random split, the model quality changed significantly during the first few epochs: the difference between the first and tenth epochs is 15-20 percentage points. In contrast, for the split-by-roots mode, such an increase is not observed: the difference between the models after one epoch of training and after 30 epochs is only 2-3 percentage points. Several examples of the dependence of WordAccuracy on the number of epochs are shown in Figure 3.

Tables 2, 3 present the results obtained for the random split experiment using the Morphodict-T and Morphodict-K datasets, respectively, Tables 4 and 5 show results for the split-by-root experiment. In addition to the results for the fine-tuned models, the tables also include the results for the CNN ensemble as the best model from (Morozov et al., 2024). Each cell contains the average metric value measured during cross-validation, with the standard deviation indicated. The best metric value in each row is highlighted in gray.

The results obtained from the two datasets are largely similar. In the case of the random split, BERT-like models demonstrate higher quality than CNN ensemble, with RuRoBERTa-based models performing better than RuBERT-based ones. The superiority of RuRoBERTa models over the ensemble of convolutional neural networks is about 3 percentage points. For the RuBERT model, the one trained without using lemmata surpasses the $\text{RuBERT}_{\text{lex}}$ model in both Accuracy and WordAccuracy for both datasets.

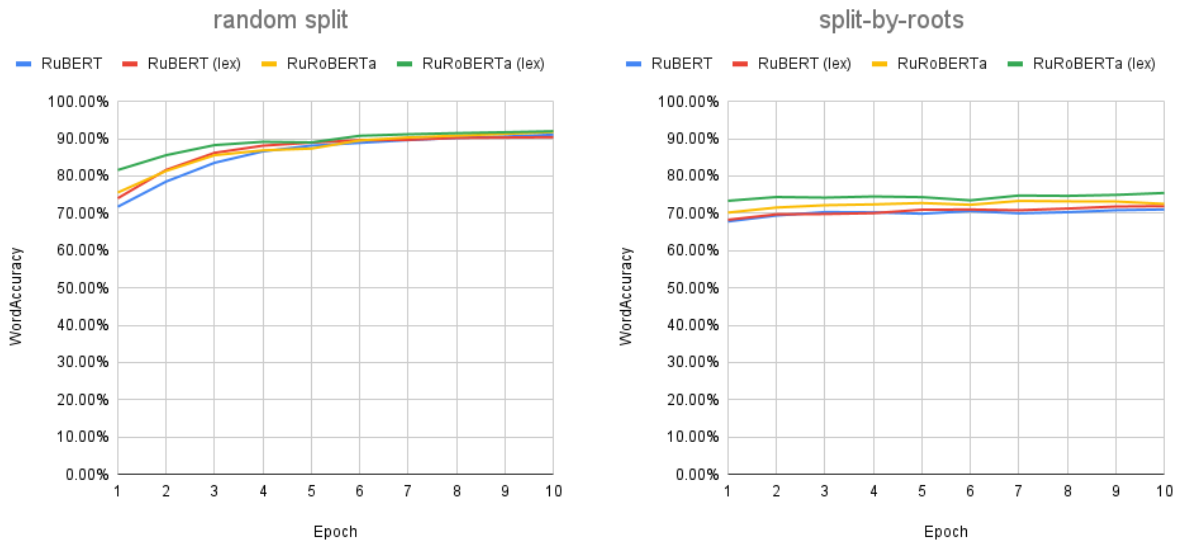


Figure 3: WordAccuracy of the models during the first ten epochs of fine-tuning on the Morphodict-K dataset

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	98.60 ± 0.05	98.18 ± 0.11	98.57 ± 0.17	98.69 ± 0.03	97.79 ± 0.13
$\text{Recall}_{\text{full}}$	98.67 ± 0.05	98.32 ± 0.07	98.56 ± 0.18	98.84 ± 0.04	98.38 ± 0.06
$\text{F-score}_{\text{full}}$	98.63 ± 0.03	98.25 ± 0.07	98.56 ± 0.17	98.76 ± 0.02	98.09 ± 0.05
$\text{Precision}_{\text{root}}$	95.80 ± 0.11	94.61 ± 0.25	95.45 ± 0.44	96.13 ± 0.14	94.19 ± 0.29
$\text{Recall}_{\text{root}}$	95.57 ± 0.12	94.59 ± 0.20	95.18 ± 0.57	96.00 ± 0.08	93.99 ± 0.21
$\text{F-score}_{\text{root}}$	95.68 ± 0.11	94.60 ± 0.21	95.31 ± 0.50	96.07 ± 0.11	94.08 ± 0.25
Accuracy	97.59 ± 0.06	96.84 ± 0.11	97.39 ± 0.31	97.78 ± 0.05	96.61 ± 0.10
WordAccuracy	91.77 ± 0.19	89.59 ± 0.29	91.09 ± 1.01	92.47 ± 0.14	88.49 ± 0.28

Table 2: Results for the Morphodict-T dataset, random split

In the case of RuRoBERTa, this dependency does not hold: for Morphodict-T, $\text{RuRoBERTa}_{\text{lex}}$ performs better, while for Morphodict-K, the model qualities are nearly identical. When using the split by roots, the superiority of RuRoBERTa-based models over the others is confirmed, with $\text{RuRoBERTa}_{\text{lex}}$ outperforming the CNN ensemble by 4.5 percentage points across both datasets. It is also noteworthy that the drop in quality between the random split and the split by roots for models using lemmata is reduced: for CNN, the gap is about 18 percentage points, for $\text{RuBERT}_{\text{lex}}$ about 17.5 percentage points, and for $\text{RuRoBERTa}_{\text{lex}}$ between 16.5 and 17.5 percentage points. Performance in root extraction also improved, as evidenced by the F-score for root morphemes: for $\text{RuRoBERTa}_{\text{lex}}$, this metric is 1.5-2 percentage points higher than that of the CNN ensemble.

The $\text{RuRoBERTa}_{\text{lex}}$ models significantly outperform the CNN ensemble and RuBERT-based models. In comparison with RuRoBERTa models, $\text{RuRoBERTa}_{\text{lex}}$ models excel in the split-by-roots case, while in the random split case, they outperform for the Morphodict-T dataset and show nearly identical results for the Morphodict-K dataset. Therefore, the $\text{RuRoBERTa}_{\text{lex}}$ models should be considered the most promising in terms of practical application. Consequently, we conducted an analysis of the segmentations made by these models and compared them with the previous baseline, the CNN ensemble.

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	99.00 ± 0.03	98.68 ± 0.09	99.05 ± 0.05	99.04 ± 0.06	98.58 ± 0.11
$\text{Recall}_{\text{full}}$	99.11 ± 0.03	98.81 ± 0.04	99.14 ± 0.03	99.17 ± 0.03	98.74 ± 0.09
$\text{F-score}_{\text{full}}$	99.05 ± 0.02	98.74 ± 0.03	99.09 ± 0.04	99.10 ± 0.04	98.66 ± 0.03
$\text{Precision}_{\text{root}}$	97.30 ± 0.11	96.15 ± 0.17	97.44 ± 0.20	97.37 ± 0.12	96.26 ± 0.13
$\text{Recall}_{\text{root}}$	97.24 ± 0.08	96.19 ± 0.18	97.43 ± 0.12	97.35 ± 0.08	96.22 ± 0.12
$\text{F-score}_{\text{root}}$	97.27 ± 0.08	96.17 ± 0.17	97.43 ± 0.16	97.36 ± 0.10	96.24 ± 0.11
Accuracy	98.10 ± 0.05	97.52 ± 0.08	98.19 ± 0.06	98.19 ± 0.06	97.40 ± 0.03
WordAccuracy	93.39 ± 0.14	91.51 ± 0.24	93.61 ± 0.16	93.54 ± 0.16	90.82 ± 0.13

Table 3: Results for the Morphodict-K dataset, random split

	RuBERT	$\text{RuBERT}_{\text{lex}}$	RuRoBERTa	$\text{RuRoBERTa}_{\text{lex}}$	CNN
$\text{Precision}_{\text{full}}$	95.27 ± 0.26	95.62 ± 0.42	94.67 ± 0.78	96.09 ± 0.34	94.46 ± 0.50
$\text{Recall}_{\text{full}}$	93.58 ± 0.50	94.33 ± 0.35	95.31 ± 0.90	95.08 ± 0.28	94.96 ± 0.45
$\text{F-score}_{\text{full}}$	94.42 ± 0.32	94.97 ± 0.32	94.98 ± 0.18	95.59 ± 0.22	94.71 ± 0.34
$\text{Precision}_{\text{root}}$	81.52 ± 0.58	82.47 ± 0.80	81.91 ± 0.48	84.37 ± 0.62	81.96 ± 0.73
$\text{Recall}_{\text{root}}$	81.47 ± 0.54	82.38 ± 0.84	82.26 ± 0.28	84.30 ± 0.61	81.98 ± 0.72
$\text{F-score}_{\text{root}}$	81.50 ± 0.56	82.42 ± 0.82	82.08 ± 0.34	84.34 ± 0.61	81.97 ± 0.72
Accuracy	89.58 ± 0.57	90.61 ± 0.52	90.15 ± 0.38	91.70 ± 0.34	90.16 ± 0.53
WordAccuracy	69.26 ± 1.76	72.12 ± 1.64	71.10 ± 1.36	74.95 ± 1.43	70.53 ± 1.79

Table 4: Results for the Morphodict-T dataset, split by roots

Despite the lower proportion of incorrect segmentations, the remaining annotation errors in the case of the random split exhibit a nature similar to the errors of the CNN ensemble (Morozov et al., 2024): insufficient consideration of semantics by the model and difficulties in distinguishing prefixes and roots in the case of borrowed morphemes. An example of the first type of error is the segmentation for the adverb "передом'' (*in front*): *пере:PREF/дом:ROOT*, whereas the correct segmentation is *пе-ред:ROOT/ом:SUFF*. It is important to note that in the Russian language, there is both the prefix 'пере-' (*over*) and the root '-дом-' (*house*), which is likely the cause of the error. An example of the second type is the segmentation of the word "примадонна" (*primadonna, diva*): although the morpheme boundaries are correctly annotated as *прима:ROOT/донн:ROOT/а:END* (in contrast to the CNN segmentation *при:PREF/мадонн:ROOT/а:END*), the type of the first morpheme in the Morphodict-K dataset is indicated as ***PREF***. When working with unfamiliar roots, compared to the CNN ensemble, the most notable change is the reduction in the number of cases with excessive separation of the prefix from the root, for example:

1. "кооперация" (*cooperation*)
   - Morphodict-K: *ко:PREF/опер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
   - CNN: *ко:PREF/о:PREF/пер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
   - $\text{RuRoBERTa}_{\text{lex}}$: *ко:PREF/опер:ROOT/а:SUFF/ц:SUFF/и:SUFF/я:END*
2. "огниво" (***flint***)
   - Morphodict-K: *огн:ROOT/ив:SUFF/о:END*
   - CNN: *о:PREF/гни:ROOT/в:SUFF/о:SUFF*
   - $\text{RuRoBERTa}_{\text{lex}}$: *огн:ROOT/ив:SUFF/о:END*

For more general conclusions, we randomly selected 50 errors made by $\text{RuRoBERTa}_{\text{lex}}$ for each dataset and each dataset split strategy. The analysis revealed that the most common errors were related to root extraction: either an unnecessary prefix or suffix is included in the root, or conversely, an excessive

	RuBERT	RuBERT$_{lex}$	RuRoBERTa	RuRoBERTa$_{lex}$	CNN
Precision$_{full}$	95.50 ± 0.27	95.69 ± 0.33	95.78 ± 0.43	96.38 ± 0.44	95.35 ± 0.33
Recall$_{full}$	94.72 ± 0.38	95.13 ± 0.35	94.74 ± 0.35	95.81 ± 0.26	95.04 ± 0.29
F-score$_{full}$	95.11 ± 0.25	95.41 ± 0.26	95.26 ± 0.36	96.09 ± 0.27	95.19 ± 0.27
Precision$_{root}$	81.84 ± 2.21	82.52 ± 2.16	82.37 ± 2.38	84.95 ± 2.01	82.46 ± 2.26
Recall$_{root}$	81.75 ± 2.13	82.67 ± 1.97	82.28 ± 2.47	84.78 ± 1.93	82.54 ± 2.25
F-score$_{root}$	81.79 ± 2.17	82.60 ± 2.06	82.32 ± 2.42	84.87 ± 1.97	82.50 ± 2.25
Accuracy	91.13 ± 0.42	91.67 ± 0.50	91.39 ± 0.62	92.82 ± 0.42	91.30 ± 0.47
WordAccuracy	72.46 ± 1.95	73.99 ± 1.71	73.32 ± 2.24	77.17 ± 1.96	72.63 ± 2.22

Table 5: Results for the Morphodict-K dataset, split by roots

morpheme is extracted from the root. In Morphodict-T, for the random split, there were 39 out of 50 errors regarding root boundaries, while for the split-by-roots, there were 47 out of 50. In Morphodict-K, for the random split, there were 27 out of 50 errors regarding root boundaries, and for the split-by-roots, there were 40 out of 50. The higher number of such errors in Morphodict-T can be attributed to the morpheme segmentation paradigm adopted in (Tikhonov, 1990): roots are longer and more diverse, and historical prefixes and suffixes are often included within the root, which the model fails to recognize. Among other errors, the overwhelming majority are related to the segmentation of suffixes (e.g., *-ничь-* vs *-н-ичь-*, *-урова-* vs *-ур-ова-*, etc.). This is largely associated with inconsistent segmentation in the datasets themselves.

We aimed to verify whether the model is indeed capable of "recognizing" unfamiliar roots or if the distribution of errors among the roots is more random in nature. To do this, we calculated the proportion of correctly annotated words for each root from the test set for the models trained in the split-by-roots mode. The results are presented in the form of a histogram in Figure 4. It is easy to see that the distributions for both datasets are similar and allow us to conclude that for the majority of roots, the model either "recognized" the root and identifies it correctly almost always, or it "did not recognize" it and almost always makes errors. Intermediate cases are quite rare. The question of which specific roots the model "recognizes" and which it does not, and how they relate to the training dataset, merits further investigation, which we plan to conduct in the future.

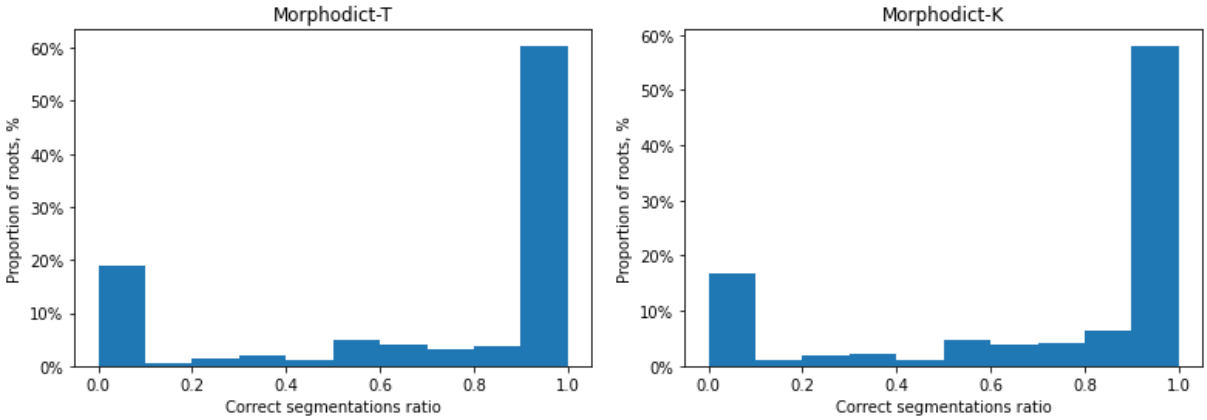


Figure 4: Distribution of the correct segmentations ratio for words containing a specific root

6 Conclusion

Automatic morpheme segmentation of Russian words is a relevant applied task in natural language processing. For a long time, the leading approaches to this task were ensembles of convolutional neural networks and LSTM networks. However, these approaches perform significantly worse with roots that were not present in the training dataset. In this work, we presented a family of fine-tuned BERT-like models for automatic morpheme segmentation that demonstrate better results compared to previous approaches. The model based on RuRoBERTa, trained for character-level annotation of words with the inclusion of the lemma in the input sequence, outperforms the convolutional neural network ensemble by 3% on random words and by 4.5% on words with unfamiliar roots. These results were obtained using two datasets with different paradigms of morpheme segmentation. Error analysis conducted on the best model shows that unfamiliar roots during testing are distributed into two categories: those that the model "recognized", meaning that the words are correctly analyzed more than 90% of the time, and those that the model "did not recognize", meaning that the words are correctly analyzed less than 10% of the time.

Two main limitations of our study should be highlighted. The first pertains to the complexity of training the model: one epoch of fine-tuning of RuRoBERTa on an NVIDIA RTX 4090 GPU took us about 2 minutes, whereas the ensemble of convolutional networks did not require a GPU at all and trained significantly faster. The second limitation relates to the potential application area of the models: the datasets used contain only the lemmata of words, while practical applications often require segmentations of word forms. Additionally, using word forms could enhance the quality of the final model. In the future, we plan to conduct a series of experiments to expand the existing datasets with segmentations of word forms and compare the quality of models trained solely on lemmata with those trained on the expanded data.

References

Khuyagbaatar Batsuren, Gábor Bella, Aryaman Arora, Viktor Martinovic, Kyle Gorman, Zdeněk Žabokrtský, Amarsanaa Ganbold, Šárka Dohnalová, Magda Ševčíková, Kateřina Pelegrinová, Fausto Giunchiglia, Ryan Cotterell, and Ekaterina Vylomova. 2022. The SIGMORPHON 2022 shared task on morpheme segmentation. // Garrett Nicolai and Eleanor Chodroff, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 103–116, Seattle, Washington, July. Association for Computational Linguistics.

I.A. Bolshakov. 2013. Krossleksika: universum sviazi mezhdu russkimi slovami [crosslexica: a universe of links between russian words]. *Biznes-informatika*, №3 (25):12–19.

E. I. Bolshakova and A. S. Sapin. 2019a. Comparing models of morpheme analysis for Russian words based on machine learning. // *Komp'juternaja Lingvistika I Intellektual'nye Tehnologii*, volume 18, P 104 –113.

Elena Bolshakova and Alexander Sapin. 2019b. Bi-LSTM model for morpheme segmentation of Russian words. // Dmitry Ustalov, Andrey Filchenkov, and Lidia Pivovarova, *Artificial Intelligence and Natural Language*, P 151–160, Cham. Springer International Publishing.

Elena I. Bolshakova and Alexander S. Sapin. 2022. Building a combined morphological model for russian word forms. // Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov, Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, *Analysis of Images, Social Networks and Texts*, P 45–55, Cham. Springer International Publishing.

Ryan Cotterell, Tim Vieira, and Hinrich Schütze. 2016. A joint model of orthography and morphological segmentation. // Kevin Knight, Ani Nenkova, and Owen Rambow, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 664–669, San Diego, California, June. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users J.*, 12(2):23–38, February.

Timur Garipov, Dmitry Morozov, and Anna Glazkova. 2024. Generalization ability of CNN-based Morpheme Segmentation. // *2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 58–62.

Ayyoob Imani, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. // Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1082–1117, Toronto, Canada, July. Association for Computational Linguistics.

B. L. Iomdin. 2019. How to define words with the same root? *Russian Speech = Russkaya Rech'*, (1):109–115, February.

Olesya Kisselev, Irina Dubinina, and Galina Paquette. 2024. A corpus-based study on orthographic errors of Russian heritage learners and their implications for linguistic research and language teaching. *Languages*, 9(4).

Y Kuratov and M Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 333–339, Moscow, Russia, June.

A. I. Kuznetsova and T. F. Efremova. 1986. *Dictionary of Morphemes of the Russian Language*. Russkii yazyk, Moscow.

Austin Matthews, Graham Neubig, and Chris Dyer. 2018. Using morphological knowledge in open-vocabulary neural language models. // Marilyn Walker, Heng Ji, and Amanda Stent, *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1435–1445, New Orleans, Louisiana, June. Association for Computational Linguistics.

Dmitry Morozov, Timur Garipov, Olga Lyashevskaya, Svetlana Savchuk, Boris Iomdin, and Anna Glazkova. 2024. Automatic morpheme segmentation for Russian: Can an algorithm replace experts? *Journal of Language and Education*, 10(4):71–84, Dec.

Antoine Nzeyimana and Andre Niyongabo Rubungo. 2022. KinyaBERT: a morphology-aware Kinyarwanda language model. // Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 5347–5363, Dublin, Ireland, May. Association for Computational Linguistics.

Ben Peters and Andre F. T. Martins. 2022. Beyond characters: Subword-level morpheme segmentation. // Garrett Nicolai and Eleanor Chodroff, *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 131–138, Seattle, Washington, July. Association for Computational Linguistics.

Marko Pranjić, Marko Robnik-Šikonja, and Senja Pollak. 2024. LLMSegm: Surface-level morphological segmentation using large language model. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 10665–10674, Torino, Italia, May. ELRA and ICCL.

Svetlana O Savchuk, Timofey Arkhangelskiy, Anastasiya A Bonch-Osmolovskaya, Ol'ga V Donina, Yuliya N Kuznetsova, Ol'ga N Lyashevskaya, Boris V Orekhov, and Mariya V Podryadchikova. 2024. Russian National Corpus 2.0: New opportunities and development prospects. *Voprosy Jazykoznanija*, (2):7–34.

Alexey Sorokin and Anastasia Kravtsova. 2018. Deep convolutional networks for supervised morpheme segmentation of Russian language. // Dmitry Ustalov, Andrey Filchenkov, Lidia Pivovarova, and Jan Žižka, *Artificial Intelligence and Natural Language*, P 3–10, Cham. Springer International Publishing.

Alexey Sorokin. 2022. Improving morpheme segmentation using bert embeddings. // Evgeny Burnaev, Dmitry I. Ignatov, Sergei Ivanov, Michael Khachay, Olessia Koltsova, Andrei Kutuzov, Sergei O. Kuznetsov, Natalia Loukachevitch, Amedeo Napoli, Alexander Panchenko, Panos M. Pardalos, Jari Saramäki, Andrey V. Savchenko, Evgenii Tsymbalov, and Elena Tutubalina, *Analysis of Images, Social Networks and Texts*, P 148–161, Cham. Springer International Publishing.

A. N. Tikhonov. 1990. *Word Formation Dictionary of the Russian language [Slovoobrazovatel'nyi slovar' russkogo yazyka]*. Russkiy yazyk, Moscow.

Dmitry Zmitrovich, Aleksandr Abramov, Andrey Kalmykov, Vitaly Kadulin, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergei S. Markov, Vladislav Mikhailov, and Alena Fenogenova. 2024. A family of pretrained transformer language models for Russian. // Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, P 507–524, Torino, Italia, May. ELRA and ICCL.

“Oh, no! Not on Wednesday! On Thursday!”: interjection in the context of speech disfluency

Podlesskaya V. I.
Institute of Linguistics, RAS, Moscow, Russia
podlesskaya@iling-ran.ru

Abstract

The subject of the study is a special class of uses of the Russian interjection OY, namely, uses in the context of speech disfluencies. The study is based on data from three corpora—the oral and the multimedia (MURCO) subcorpora of the Russian National Corpus and the pilot version of the corpus of oral personal stories “What I saw”. The main types of speech disfluencies in the context of which speakers resort to using the interjection OY, the structure of constructions with OY, as well as the features of their phonetic—segmental and suprasegmental—implementation are investigated. From a quantitative point of view, a comparison of the use of OY in oral public and oral non-public speech is carried out. It is shown that although OY in oral non-public speech occurs, as could be expected, many times more often, the proportion of OY as a marker of a speech disfluency in relation to the total number of uses of OY is almost the same in public and non-public varieties of oral speech. These quantitative data suggest that the general distribution of the use of OY and the distribution of OY as a marker of speech disfluency are not regulated by the same factors. The general distribution is associated with the degree of spontaneity and publicity of speech: interjections as exclamatory discursive markers become more appropriate when the speaker is not constrained by disciplinary guidelines and can allow himself to express emotions and give assessments more freely. And the distribution of OY as a marker of speech disfluency is also associated with the density of speech disfluencies themselves. Of course, this density varies among individual speakers, but at the same time it largely depends on the general principles regulating the generation of unprepared discourse.

Keywords: interjection, speech disfluency, spoken discourse, corpus
DOI: 10.28995/2075-7182-2025-23-267-280

«Ой, нет! Не в среду! В четверг!»: корпусное исследование междометия в контексте речевого сбоя

Подлесская В. И.
Институт языкознания РАН, Москва, Россия
podlesskaya@iling-ran.ru

Аннотация

Предмет изучения — особый класс употреблений русского междометия ОЙ, а именно употреблений в контексте речевых затруднений и сбоев. Материалом послужили данные трех корпусов устной речи — устного и мультимедийного подкорпуса НКРЯ (МУРКО) и пилотная версия корпуса устных личных рассказов «Что я видел». Исследованы основные типы речевых затруднений, в контексте которых говорящие прибегают к использованию междометия ОЙ, структура конструкций с ОЙ, а также особенности их фонетической — сегментной и супрасегментной — реализации. С количественной точки зрения проведено сравнение употребления ОЙ в устной публичной и устной непубличной речи. Показано, что хотя ОЙ в устной непубличной речи встречается, как и можно было ожидать, многократно чаще, доля ОЙ как маркера речевого сбоя по отношению к общему числу употреблений ОЙ оказывается почти одинаковой в публичной и непубличной разновидностях устной речи. Эти количественные данные позволяют предположить, что общее распределение употребления ОЙ и распределение ОЙ как маркера речевого сбоя регулируются не одними и теми же факторами. Общее распределение связано с мерой спонтанности и публичности речи: междометия как эксклamативные дискурсивные маркеры становятся более уместными, когда говорящий не скован дисциплинарными установками и может позволить себе более свободно выражать эмоции и давать оценки. А распределение ОЙ как маркера речевого сбоя связано еще и с плотностью самих речевых сбоев. Разумеется, эта плотность варьируется у индивидуальных говорящих, но при этом она в большой степени зависит от общих принципов, регулирующих порождение неподготовленного дискурса.

Ключевые слова: междометие, речевые сбои, устная речь, корпус

1 Постановка задачи

В данной работе мы исследуем особый класс употреблений русского междометия ОЙ, а именно употребления в контексте речевых затруднений и сбоев. Представление об этом классе случаев дают следующие примеры из устного подкорпуса НКРЯ, в них ОЙ появляется там, где говорящий обнаруживает в собственной речи оговорку, спохватывается и исправляет ее:

(1) Кролик? Недолго был. Потому что мы его отдали в зоопарк/ ***ой***/ ***не*** в зоопарк/ ***а*** в детский садик. У них в звериный уголок/ живой уголок/ точнее. [Разговор студентов о домашних животных (2009)]

(2) Ну короче он месяц назад сломал/ ***ой не*** сломал/ ***а*** вывихнул ногу/ прокатился на скейте. [Телефонный разговор двух подруг (2005)]

Мы рассмотрим основные типы речевых затруднений, в контексте которых говорящие прибегают к использованию междометия ОЙ, структуру конструкций с ОЙ, а также особенности их фонетической — сегментной и супрасегментной — реализации. Материалом исследования послужили данные трех корпусов — устного и мультимедийного подкорпуса НКРЯ (МУРКО) и пилотная версия корпуса устных личных рассказов «Что я видел», далее — «ЧВ» (Коротаев и др. 2024)[1]. Примеры из устного подкорпуса НКРЯ и из МУРКО приводятся в полной графической форме как она дана в НКРЯ. Для примеров из МУРКО при необходимости приводится также просодическая транскрипция части примера с разметкой движения тонов и локализацией фразовых акцентов, а также осциллограмма и интонограмма в формате анализатора PRAAT (Boersma, Weenink 2021). Примеры из корпуса ЧВ приводятся в просодической транскрипции и также, при необходимости, снабжаются визуализацией в PRAAT'e.[2]

В специальной литературе и в словарях употребление междометия ОЙ связывается с реакцией на неожиданно возникшую проблему. Так, в работе Шаронов 2009: 24, 26 указывается, что ОЙ оправдано, когда «сбой, небольшая ошибка в деятельности могут вызвать вскрик досады». Близкую идею формулируют и словари. Так, в Большом толковом словаре русского языка (Кузнецов 2003) указано, что ОЙ употребляется, в частности, как «выражение досады, сожаления, предупреждения о возможных трудностях». В Большом универсальном словаре русского языка (Морковкин и др. 2019) в числе прочих приводится следующее значение ОЙ: «восклицание, к-рое употр. при неожиданно возникших обстоятельствах, препятствующих продолжению разговора или другой совместной деятельности. Син. <*ой-ой-ой* разг.>.□ *Ой, одну минуточку, не кладите трубку, здесь кто-то стучит, я сейчас.*». Междометие ОЙ в том классе употреблений, который рассматривается в данной работе, можно считать частным случаем маркера обнаружения неожиданной проблемы.

Главная особенность этого класса состоит в том, что источником проблемы является сам говорящий. В ходе речепорождения говорящий решает одновременно несколько сложных задач — формирует речевое задание, артикулирует порцию дискурса, осуществляет мониторинг и ревизию артикулированного фрагмента на предмет его соответствия исходному заданию и устраняет, в случае несоответствия, обнаруженную проблему. Таким образом, говорящий выступает, как минимум, в двух лицах — как нарратор и как ревизор сказанного. Для говорящего-ревизора затруднение, которое испытывает говорящий-нарратор оказывается неожиданно возникшей внешней(!) проблемой, вызывающей комплекс речевых реакций: нужно сообщить

[1] Приводимые примеры, если не указано иное, взяты из НКРЯ. Примеры из корпуса «Что я видел» снабжены аббревиатурой «ЧВ».

[2] В просодической транскрипции для указания на направление движения тона иконически используются знаки «/», «\» и «–». Ударный слог слова — носителя рематического акцента подчеркивается. Длина пауз в секундах указывается в круглых скобках; пауза может быть абсолютной или вокализованной: ɥ, ʔ, ə и др. Внутри фигурных скобок приводятся обозначения неречевых вокальных действий: gp — сглатывание, hm — хмыкание, exp — откашливание и др. Сильная редукция отмечается серым цветом шрифта, ускоренное произнесение — курсивом. О других деталях используемой системы просодической транскрипции см. Кибрик, Подлесская 2009. Напомним, что в той версии примера, которая дается по МУРКО, знак «/» имеет другую интерпретацию — там он используется для членения речевого потока. В сегментное наполнение в просодической транскрипции вносятся при необходимости уточнения по сравнению с графическим вариантом МУРКО.

слушающему о неожиданно возникшей проблеме, выразить свое к ней отношение (обычно отрицательное — как и при других проблемах, возникающих неожиданно), и, по возможности немедленно, проблему устранить. Решению этого комплекса задач служат конструкции с междометием ОЙ. Устройство и функционирование конструкций с ОЙ в контексте речевых затруднений будет рассмотрено ниже в следующем порядке: в разделе 2 обсуждаются варианты фонетической реализации междометия ОЙ в контексте речевых затруднений, в разделе 3 рассматриваются типы речевых затруднений, которые обнаруживаются в конструкциях с ОЙ, в разделе 4 приводятся некоторые количественные результаты по данным МУРКО, в разделе 5 подводятся краткие итоги исследования.

2 Варианты фонетической реализации междометия ОЙ в контексте речевых затруднений

В прототипическом случае в качестве сигнала речевого затруднения ОЙ акцентируется и произносится с нисходящим тоном или с восходяще-нисходящим тоном — с искривлением тона, типичным для эмфазы (Янко 2008: 83– 97). В следующем примере ОЙ возникает в той точке, где говорящий (в своей ипостаси ревизора) обнаруживает, что он (в своей ипостаси нарратора) оговорился и вместо правильного *сорок недель* выпалил *сорок дней*:

(3) Неужели работать сорок часов в неделю/ сорок дней/ ээ ***ой***/ сорок недель в год/ сорок лет за пятнадцать тыщ **—** это нормально? [Дебаты участников предварительного голосования по определению кандидатов в депутаты в Государственную думу в Санкт-Петербурге (02.04.2016) (2016)]

Неужели работать сорок часов в /неде̲лю,
(0.29)
сорок /дней ==
(0.24)
(э 0.69) /\о̲й!,
сорок недель /в го̲д,
сорок \лет за-а пятнадцать /ты̲щ,
это \норма̲льно?!

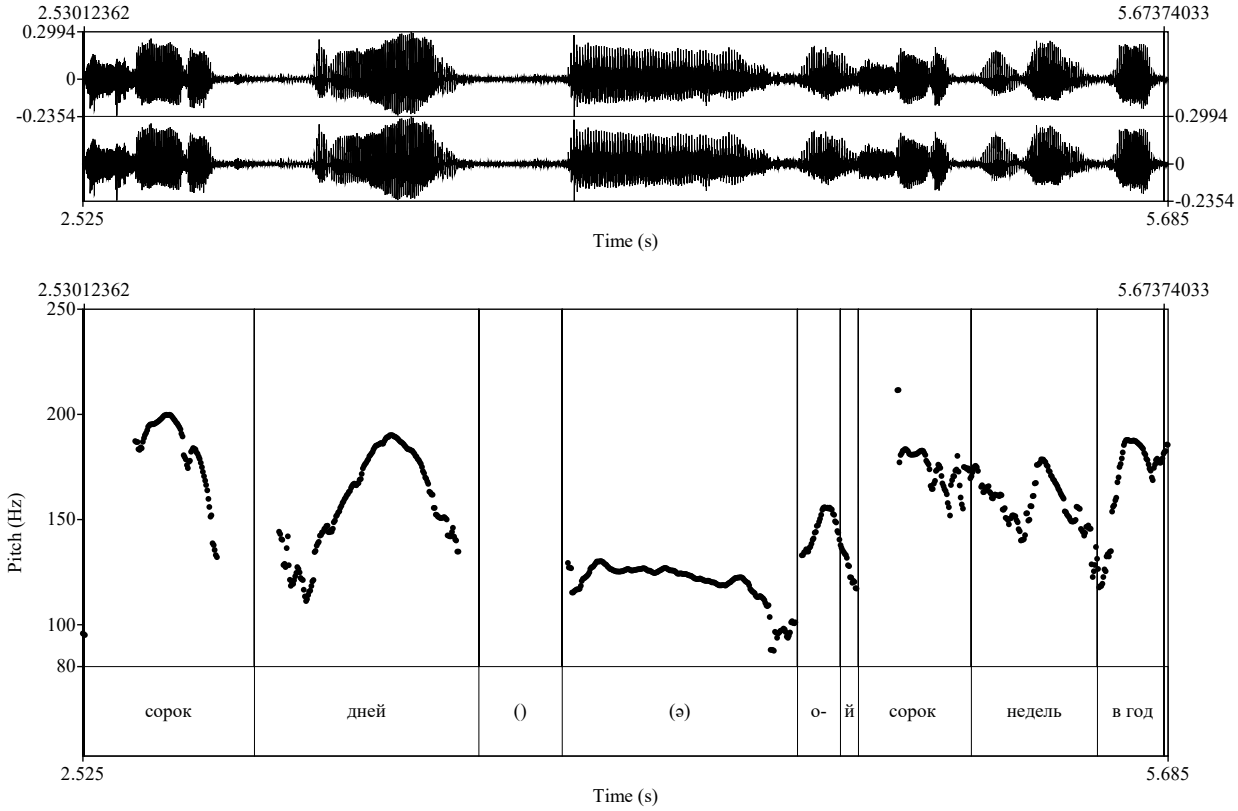


Рисунок 1. Осциллограмма и интонограмма примера (3)

Иногда используются дополнительные средства усиления эмфазы на междометии. Если это усиление связано с подчеркиванием неожиданности и трудности возникшей проблемы, то говорящий может прибегнуть к нефонологическому удлинению гласной и/или добавлению гласного шва-призвука в конце или начале слова (*о-ойᵊ, ᵊой*). Ту же роль играет и редупликация ОЙ, что особенно характерно для женской речи. В следующем примере говорящая обескуражена очевидной бессмысленностью своей собственной оговорки. Помимо редуплицированного ОЙ-ОЙ, здесь обращает на себя внимание разная акцентуация первого, ошибочного, варианта сочиненной именной группы — *животных и растений* — и второго, правильного — *животных и человека*. В первом варианте группа акцентируется интегрально — с единственным акцентом на втором конъюнкте (*растений*), во втором — с раздельными акцентами на каждом из конъюнктов; возникший акцент на слове *животных* призван подчеркнуть, что первый конъюнкт был назван правильно и исправлению подлежит лишь второй — *растений* заменено на *человека*:

(4) Интересно/ ну/ наверное/ это результат/ связанный с клетками животных и растений/ ***ой***/ животных и человека. [Любовь Курганская. «Космический ковчег». Выступление на научном конкурсе Science Slam Samara (2014)]

/Наверное,
это /результат,
тот который связан сᵊ (ə 0.32) клетками животных и \растений.
(0.18)
Ой- \ой!
/Животных (ə 0.36) и-и \человека.

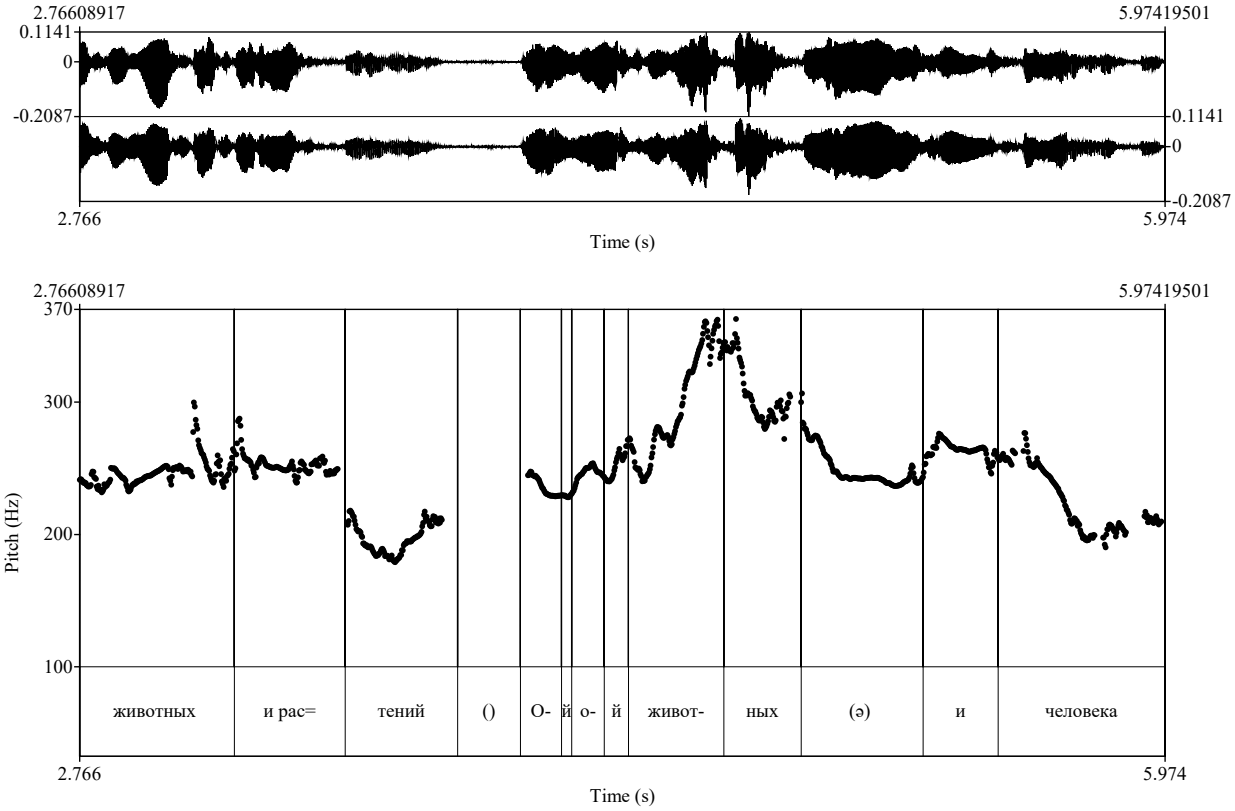


Рисунок 2. Осциллограмма и интонограмма примера (4)

Если усиление эмфазы связано с подчеркиванием досады на возникшую проблему, то используется такое фонетическое средство, как нефонологическое удлинение согласной финали междометия. Удлиняется Й — палатальный аппроксимант ([j] в стандартной записи МФА), Й может произноситься с придыханием (в принятом нами варианте транскрипции обозначается значком «^{h}» в верхнем регистре). Кроме того, [j] может реализоваться как палатальный спирант ([ç] в стандартной записи МФА); этот звук отсутствует в русской фонологической системе, и реализуется подобно немецкому *ch* в словах типа *Ich* или *möchte*. В принятом нами варианте транскрипции специального значка для него нет, приходится использовать «комментарии на полях».

В примере ниже ОЙ произносится как палатальный спирант с удлинением и придыханием. Здесь с помощью ОЙ говорящая выражает досаду в связи с тем, что испытывает трудности с поиском медицинского термина *пластома*. Дополнительно о трудности поиска свидетельствует вздох после междометия (нотируется как {sg}), чуть позже — сверхдлинная пауза 1.91 сек, и, в завершение, повтор слова *пластома* как подтверждение того, что слово в конце концов найдено верно:

(5) Вот моего Сашки одноклассник умер/ писят три года. ***Ой***/ как же диагноз-то было?..
Пластома головного мозга. Пластома. [Беседа с Варварой Петровной за приготовлением лапши (2018)]

/\О̲й-й^{h}!
{sg 0.27}
(0.66)
–Как же \диа̲гноз-то было?
(1.91)
/**Пластома** (0.36) головного \мо̲зга!
(0.95)
\Пласто̲ма.

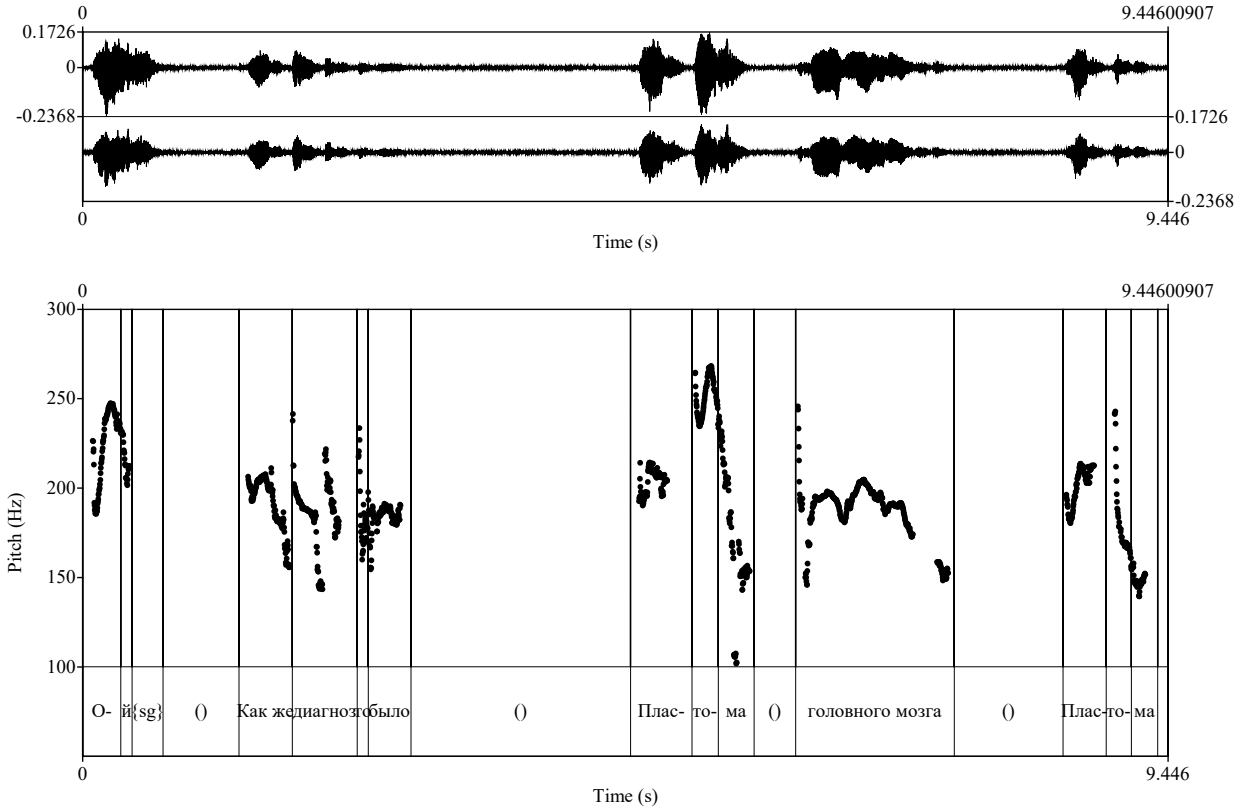


Рисунок 3. Осциллограмма и интонограмма примера (5)

Заметим, что палатальный спирант встречается в реализации ОЙ и вне интересующего нас контекста речевых сбоев, например, в (преимущественно) женском скептическом возражении (типа *Ой-й, я тебя умоляю!*) или при реакции на не очень существенный, но досадный физический раздражитель, как, например, при неприятной обработке незначительной раны (типа *Ой-й, подуй — щиплет!*), о возможных реализациях [j] в конце слова, в том числе, при эмфазе см. Панов 1979: 108-110; Касаткин 2006: 169-170.

Таковы, в кратком изложении, фонетические свойства междометия ОЙ как сигнала речевого затруднения. Перейдем теперь к особенностям его употребления.

3 Функционирование междометия ОЙ в контексте речевого затруднения

Рассмотрим конструкции, которые свидетельствуют о речевом затруднении говорящего, и обсудим, как и когда междометие ОЙ встраивается в такие конструкции. Как мы постараемся показать, употребление ОЙ позволяет проникнуть в «лабораторию говорящего»: локализация ОЙ является достаточно точным индикатором той точки развертывания дискурса, где обнаруживается речевое затруднение. Мы последовательно рассмотрим употребления ОЙ в контексте коррекций, инициированных самим говорящим, случаи дублирования фрагмента, подлежащего коррекции, употребления ОЙ в контексте колебаний при планировании и в контексте коррекций, инициированных собеседником.

3.1 ОЙ в контексте коррекций, инициированных самим говорящим

Самым стандартным случаем является употребление ОЙ при самоисправлении говорящего, инициированного самим говорящим (self-initiated self-repairs, Fox et al. 1996). Обычно ОЙ размещается между репарандумом, т.е. фрагментом, который говорящий (в своей ипостаси ревизора) забраковывает, и репарансом, т.е. фрагментом, который выступает как замена, более соответствующая исходному речевому заданию. Так, в следующем примере репарандум — слово *карьер=* (возникшее, вероятно, в результате оговорки под влиянием предшествующего

упоминания); в точке обнаружения проблемы (оговорки) возникает междометие ОЙ, и далее следует репаранс *бульдозеры*. О серьезности речевого сбоя свидетельствует скрип (ˀ), предшествующий ОЙ, и следующие за ОЙ заполненная пауза (ɐ), абсолютная пауза, чмоканье{sm 0.14}, и снова скрип:

(6) ЧВ

и-и в какой-то /момент выходим на-а горный /карьер,
(0.32)
(ɥ 0.45)
там вся эта /техника работает,
(ɐ 0.34) /\самосва-алы,,,
/\карье-ер= ==
(ˀ 0.20) \ой-й,
(ɐ 0.70)
(0.82)
{sm 0.14}
(ˀ 0.22) /\бульдо-озеры-ы,,.
(ɥ 0.86)
(ɐ 0.57) и вот такое разное горно-транспортное \оборудование.

3.2 Дублирование фрагмента, подлежащего коррекции

Часто при самоисправлении говорящему неудобно работать, ограничиваясь строго забракованным фрагментом, и приходится оперировать более крупной грамматически цельнооформленной единицей и уже внутри нее производить частичную коррекцию. Такой вынужденный повтор правильного фрагмента, синтаксически связанного с забракованным, в англоязычной практике принято называть термином recycling (букв. «вторичная переработка отходов»). Так, в примере (3) выше дублировалась вся количественная и группа, и внутри нее произведена коррекция (*сорок дней* заменено на *сорок недель*); в примере (4) выше дублировалась сочиненная именная группа и внутри нее произведена коррекция только одного конъюнкта (*животных и растений* заменено на *животных и человека*). В примере (7) ниже говорящая повторяет комбинацию «сказуемое плюс подлежащее», заменяя (исправляя) лишь подлежащее (*собирается ли прокуратура* заменяется на *собирается ли администрация Светлогорска*). Как можно увидеть на рисунке 4, повторное вхождение глагола *собирается ли* произносится намного быстрее — продолжительность первого 0.96 сек, второго — как вспомогательного и малоинформативного — всего 0.58 сек т.е. в 1.65 раза короче:

(7) Я вам подавала письмо насчёт аа не-незаконного выделения земель рекреационной зоны в посёлке Донского/ и собирается ли прокуратура/ ***ой***/ собирается ли администрация Светлогорска выполнить предписания прокуратуры о отмене этих постановлений — выделения незаконно этих земель. [Прямая линия с губернатором Калининградской области А.А. Алихановым (2017)]

\собирается ли прокуратура ==
\ой!,
собирается ли администрация-а (ɥ 0.37) /Светлогорска выполнить предписания \прокуратуры?

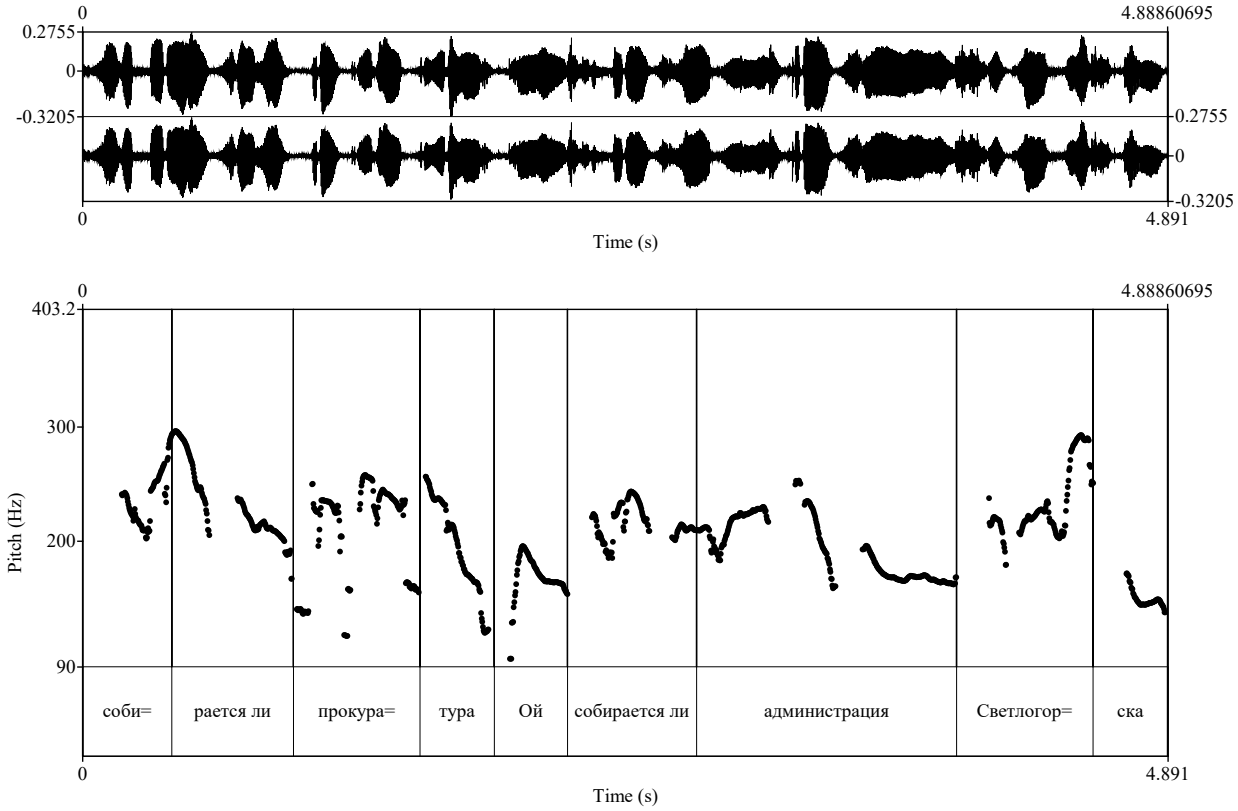


Рисунок 4. Осциллограмма и интонограмма примера (7)

Говорящий как ревизор не всегда срабатывает оперативно, и говорящий как нарратор успевает произнести после репарандума порцию дискурса, которая не подлежит коррекции, и только потом говорящий-ревизор прибегает к ОЙ как к сигналу обнаружения проблемы и дальше исправляет ошибку. Порцию дискурса, которая не подлежала коррекции, приходится после репаранса повторить — чтобы вернуться к правильному развертыванию дискурса. Тогда может возникнуть ситуация дискурсивной непроективности, т.е. последовательности типа «репарандум$_i$ — «правильный» фрагмент$_j$ — ОЙ — репаранс$_i$ — «правильный» фрагмент$_j$», где «парные» элементы снабжены идентичными индексами. В примере (8) ниже эта последовательность выглядит так: *из Первомайского*$_i$ *эти проблемы*$_j$ *ой извините из Новосинеглазово*$_i$ *эти проблемы*$_j$. Здесь репаранс полностью дублирует предложно-падежную форму репарандума (*из* плюс родительный падеж), а правильный фрагмент *эти проблемы* повторяется в неизменном виде:

(8) у нас тоже есть проблемы/ и активисты/ в том числе из Первомайского/ эти проблемы/ ***ой***/ и-из-извините/ из Новосинеглазово/ эти проблемы нам помогают решать. [Ю.Б. Черкасов. Выступление на первомайском митинге в Новосинеглазово (2017)]

У нас тоже есть /проблемы,
(ц 0.22)
и-и /активисты —
в том числе из /Первомайского,
(0.11)
— эти /проблемы ==
(0.15)
–\о-оˀй!,
из= || из= || изв= || \извините!,
из /Новосинеглазово,

{exp 0.21}
эти /проблемы нам помогают \решать.

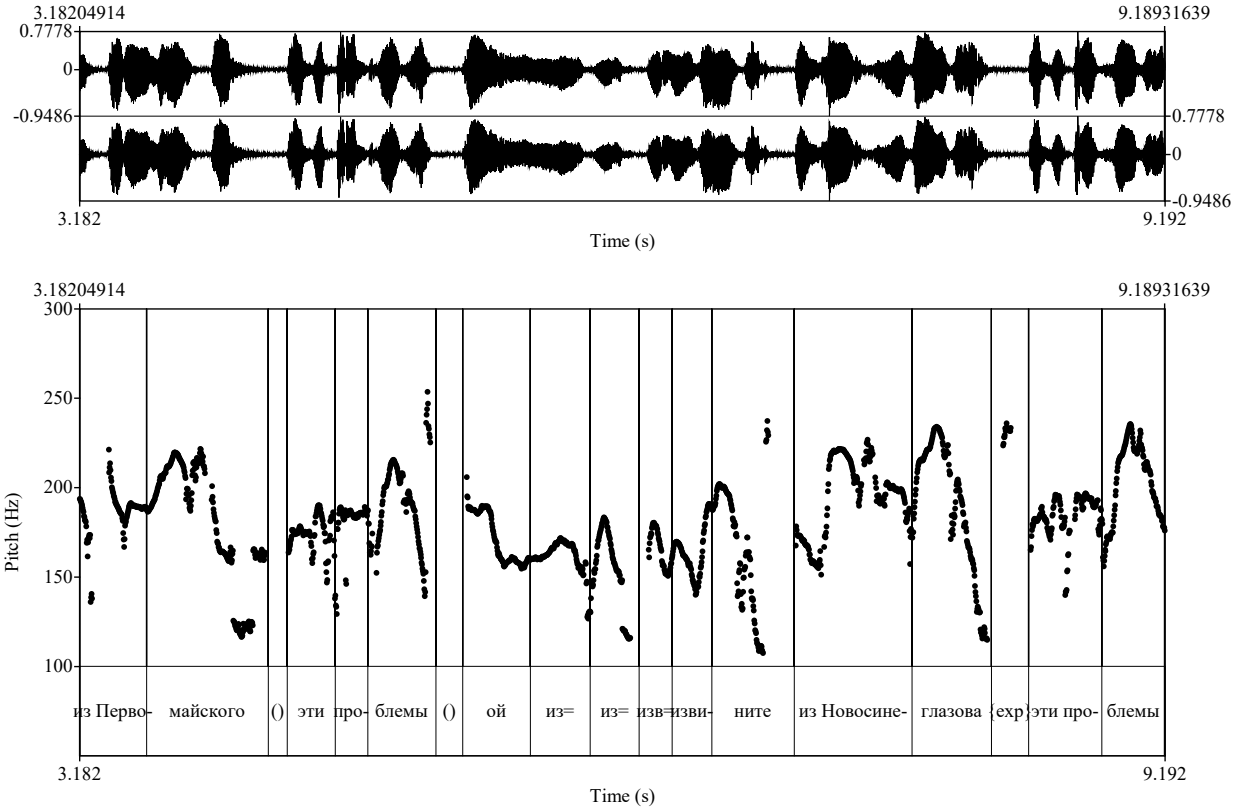


Рисунок 5. Осциллограмма и интонограмма примера (8)

В отличие от варианта recycling, который был проиллюстрирован выше в (7), здесь, в (8), вторично «утилизируется» фрагмент не слева от репарандума, а справа, и этот фрагмент не связан с репарандумом синтаксически. Пример (8) показателен еще и тем, что в нем, помимо ОЙ, имеются и другие сигналы речевого затруднения — это, прежде всего, эксплицитное извинение, размещенное после ОЙ перед репарансом. Кроме того, само извинение реализуется с троекратным повтором начального слога (*из= из= изв=*), а после репарандума перед повтором правильного фрагмента следует откашливание (*из Новосинеглазово {exp 0.21} эти проблемы*)[3]. Как видим на рисунке 5, возвращение к правильному развертыванию дискурса поддерживается просодически тем, что репаранс дублирует просодическое оформление репарандума: *из Первомайского* и *из Новосинеглазово* - основосложные слова с основным ударением на заключительной основе и вспомогательным — на первой; оба реализуются с «двугорбым» контуром: подъем на ударных слогах с последующим падением на заударных, особенно заметным на заключительной основе.

Помимо извинений, ОЙ часто сопровождается маркерами переформулирования *(вернее, точнее, то есть*). Так, в следующем примере (9) используется *то есть*, причем о затрачиваемых усилиях свидетельствуют повтор *то есть* и пауза, заполненная скрипом и «эканием», это позволяет говорящему выиграть дополнительное время на поиск правильного репаранса после того, как ошибка была обнаружена. Заметим попутно, что пример (9) показывает, что извинение может располагаться не только контактно с ОЙ между репарандумом и репарансом, но и после репаранса:

[3] Большое спасибо скрупулезному рецензенту, обратившему внимание на то, что при обращении к видеоклипу этого примера в зоне речевого сбоя у говорящего обнаруживаются еще и интенсивные жесты головы и руки, т.е. имеется и невербальный сигнал сбоя.

(9) У одного из участников либо модель из конструктора Лего собранная/ ну такая/ не совсем тривиальная/ либо разложены фотографии звёздного неба/ ***ой***/ то есть/ то есть/ облачного неба/ простите. Звёздного было бы/ конечно/ ещё круче. [Илья Утехин. Как устроен диалог? Лекция Полит.Ру (2017)]

(в 0.38) /либо /разложены /фотографии звёздного \неба.
(0.57)
\ᵊОй!
То есть || *то есть* (ʔə 0.30) облачного \неба.
\Простите.

Еще одна распространенная тактика самоисправления состоит в том, что после ОЙ следует повтор репарандума с отрицанием, который предшествует репарансу:

(10) ЧВ

и там точно такие же /\белые круглые || (0.53) (ʔ1.06*) белый круглый /пол, на котором белые круглые* ==
\Ой-й!
(0.34)
Они не \круглые.
Просто белые /полки,
(ц 0.27)
и стоит девушка в /белом,

Ту же функцию, что и отрицание, при повторе репарандума может выполнять саркастическое воспроизведение. Так, в следующем примере говорящая ошибочно квалифицировала музей как *муниципапальный*, тут же заметила ошибку, сигнализировала о ней с помощью ОЙ и, прежде чем произвести замену на правильное *корпоративный*, саркастически повторила ошибочное *муниципапальный* в смысле ‘Ну, какой такой муниципальный! Что это я говорю!’. Повтор артикулирован безударно и в высоком темпе (нотировано курсивом в транскрипте), фразовый акцент на ОЙ. Заметим также, что репаранс не удался говорящей с первой попытки — ему предшествует оборванное *зав*= (предположительно — *заводской*):

(11) [Студентка, жен, Студентка] А то есть музей/ сам он э государству принадлежит или он частный?
[Экскурсовод, жен, Работник культуры] Нет/ заводу… заводу… муниципальный/ ой/ муниципальный/ [Нрзбр] ***корпоративный*** музей.
[Студентка, жен, Студентка] Угу… [Разговор в музее (2014)]

\Нет!
\Заводу.
(0.49)
\Заводу.
(0.48)
\Муниципальный.
\Ой *муниципальный*!
Зав= ==
\Корпоративный музей.

3.3 ОЙ в контексте хезитации

До сих пор мы видели, как ОЙ используется в тех случаях, когда говорящий обнаруживает в собственной речи сегмент, не соответствующий исходному замыслу и нуждающийся в редактировании (замене или отмене). Но ОЙ может использоваться как сигнал возникшей проблемы с поиском слова — даже если никаких ошибок, нуждающихся в исправлении, не было сделано. Один такой пример был показан выше в (5), где говорящая пыталась вспомнить трудный для нее медицинский термин. Хезитация может быть связана как с трудностью поиска

подходящей вербализации, так и просто с извлечением из памяти конкретной именованной сущности, как в примере ниже, где говорящий не может вспомнить конкретную фамилию:

(12) Но и-и-интересно/ что я написал тоже соответственно статью/ в своей газете «Сельской нови»/ и када…/ ***ой***/ как же его/ известный-то краевед/ Бравицкий или…/ вот я забыл его// Всё время фамилия вылетает// [О походе по маршруту Арсеньева (2016)]

3.4 ОЙ в контексте коррекций, инициированных собеседником

Во всех предыдущих примерах мы видели, как говорящий (в своей ипостаси ревизора) сам обнаруживает в своей речи проблему и ее устраняет. Но бывают случаи, когда говорящий замечает и устраняет проблему, только когда на нее ему указывает собеседник — это так называемые other-initiated self-repairs (Fox et al. 1996). Обнаружение проблемы «с подсказкой» тоже может сигнализироваться с помощью ОЙ. Подсказка собеседника чаще всего имеет форму переспроса или наводящего вопроса, как в примере ниже:

(13) [Студент2, жен, Студент] Я сёдня хотела купить золотой лак. Но я не нашла его.
[Студент1, жен, Студент] Почему золотой?
[Студент2, жен, Студент] Накрасить мизинец.
[Студент1, жен, Студент] Один мизинец?
[Студент2, жен, Студент] По шен-шую-.
[Студент1, жен, Студент] То есть даже не безымянный/ а мизинец?
[Студент2, жен, Студент] ***Ой***/ безымянный.
[Студент1, жен, Студент] Один? Не два?
[Студент2, жен, Студент] Ну/ можно ещё средний. [Дружеская беседа (2014)]

Говорящий обычно реагирует на подсказку немедленно, часто не дожидаясь ее окончания. Так, в следующем примере говорящий произносит ОЙ и приступает к исправлению уже в середине переспроса. Наложение реплики говорящего и собеседника показано ниже в транскрипте с помощью квадратных скобок, а на рисунке 6 — раздельным аннотированием реплик во временной развертке:

[Зритель №2, муж] У меня вопрос ко всем участникам. Аа скажите/ пожалуйста/ вот мы все знаем/ что есть Фонд обязательного социального страхования/ есть…
[Титов Михаил, муж, телеведущий] Обязательного социального или обязательного медицинского?
[Зритель №2, муж] ***Ой***/ обязательного медицинского/ конечно. [Дебаты участников предварительного голосования по определению кандидатов в депутаты в Государственную думу в Санкт-Петербурге (02.04.2016) (2016)]

Sp1 что есть \Фонд обязательного социального /–страхования,,,
(0.96)
есть ==
Sp2 Обязательного /социального [или обяза][тельного \ме][дицинского?
Sp1 [\Ой-й°!] [(0.29)] [Обязательного \медицинского!

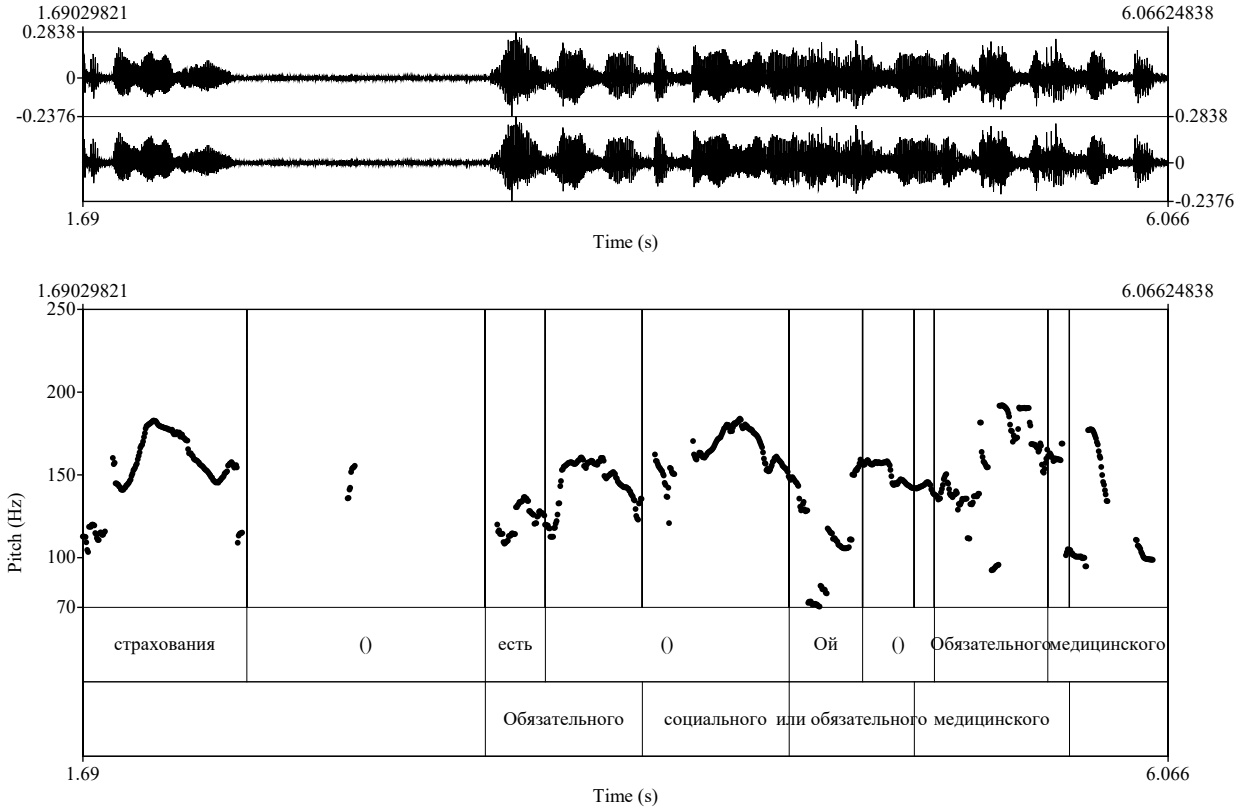


Рисунок 6. Осциллограмма и интонограмма примера (14)

Таковы, в самом общем виде, основные паттерны употребления ОЙ в качестве сигнала о речевом затруднении. В следующем разделе мы покажем, насколько — в количественном отношении — эти паттерны вписываются в общий массив употребления междометия ОЙ.

4 Количественный анализ употреблений междометия ОЙ по данным МУРКО

Для оценки частотности ОЙ как маркера речевого сбоя мы выделили два подкорпуса МУРКО — тексты устной публичной речи (540 текстов 2 078 079 слов) и тексты устной непубличной речи (105 текстов 89 329 слов), исключив из рассмотрения тексты других классов МУРКО (кино и театральная речь, художественное и авторское чтение) — как менее спонтанные.

Частотность междометия ОЙ в целом характеризуется следующими цифрами: в текстах устной публичной речи встретилось 188 вхождений ОЙ, т.е. ipm = 90.5 или 0.0905 на 1000 слов; в текстах устной непубличной речи встретилось 116 вхождений ОЙ, т.е. ipm = 1300 или 1.3 на 1000 слов. Таким образом, ОЙ в устной непубличной речи встречается, как и можно было ожидать, многократно чаще — почти в пятнадцать раз (более точно, в 14.36 раза).

Интересно однако, что доля ОЙ как маркера речевого сбоя по отношению к общему числу употреблений ОЙ оказывается почти одинаковой в публичной и непубличной разновидностях устной речи. Методом сплошной проверки мы установили, что из 188 вхождений ОЙ в текстах подкорпуса устной публичной речи 20 случаев — это ОЙ как маркер речевого сбоя, что составляет 10.6%. Сходным образом, из 116 вхождений ОЙ в текстах подкорпуса устной непубличной речи 16 случаев– это ОЙ как маркер речевого сбоя, что составляет 13.8%. Таким образом «вклад» ОЙ как маркера речевого сбоя в общее число употреблений этого междометия лишь незначительно выше в непубличной устной речи по сравнению с публичной.

Эти количественные данные позволяют предположить, что общее распределение употребления ОЙ и распределение ОЙ как маркера речевого сбоя регулируются не одними и теми же факторами. Общее распределение связано с мерой спонтанности и публичности речи: междометия как эксклamативные дискурсивные маркеры становятся более уместными, когда

говорящий не скован дисциплинарными установками и может позволить себе более свободно выражать эмоции и давать оценки. А распределение ОЙ как маркера речевого сбоя связано еще и с плотностью самих речевых сбоев. Разумеется, эта плотность варьируется у индивидуальных говорящих, но при этом она в большой степени зависит от общих принципов, регулирующих порождение неподготовленного дискурса (см. подробнее Podlesskaya 2015). Возможно, какие-то из этих принципов оказываются не очень чувствительны к мере публичности дискурса и более чувствительны к его спонтанности.

5 Заключительные замечания

Мы проанализировали особый класс употреблений междометия ОЙ — употреблений, в которых говорящий использует это междометие в качестве сигнала о проблеме, возникшей в ходе речепорождения. Эта проблема может быть связана с тем, что говорящий обнаруживает что породил некоторый отрезок дискурса, который его не устраивает или породил отрезок дискурса не в той точке, где это соответствует исходному замыслу — тогда ОЙ возникает совместно с самоисправлением. Проблема может быть связана и с тем, что говорящий испытывает трудности с планируемым фрагментом и ему требуется время, чтобы это планирование осуществить — тогда ОЙ возникает совместно с сигналами хезитации — заполненными паузами, нефонологическими удлинениями звуков и т.д. Междометие ОЙ может возникать и как реакция на реплику собеседника, указывающую на замеченную проблему в речи текущего говорящего — тогда ОЙ возникает как сигнал кооперации — «да, действительно, есть проблема, устраняю».

Способность ОЙ употребляться в качестве срочной реакции на проблему помогает лучше понять общие принципы, регулирующие порождение неподготовленного дискурса. Появление ОЙ помогает оценить, в какой точке речепорождения говорящий обнаруживает проблему, какой участок дискурса говорящий решает забраковать, считает ли он необходимым подвергнуть обработке только ошибочное звено или требуется «захватить» и соседние и т.д. Оценивая окрестность междометия — наличие пауз, обрывов, повторов, изменений в темпе и акцентуации проблемных отрезков, мы можем оценить трудоемкость обработки проблемных участков дискурса. Анализ сигналов, подобных ОЙ, позволяет продвинуться к понимаю когнитивных процедур, обеспечивающих развертывание дискурса.

Благодарности

Я благодарю Ольгу Валентиновну Антонову за помощь с разделом, посвященным сегментной фонетике, и анонимных рецензентов за подробный анализ первоначальной версии этой работы.

Литература

[1] Касаткин Л.Л. (2006). Современный русский язык. Фонетика, Москва: Издательский центр «Академия».
[2] Кибрик А. А., Подлесская В. И. (Ред.) (2009). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. Москва: Языки славянских культур.
[3] Коротаев Н. А., Панышева Д. А., Неверова Е. А., Подлесская В. И. (2024). Корпус «Что я видел» как инструмент анализа панического дискурса // «Слово и жест». Научная конференция, посвященная памяти Е. А. Гришиной («Гришинские чтения»). Москва, 8 февраля 2024 г. Материалы конференции. / Отв. ред. С. О. Савчук. − М.: Институт русского языка им. В. В. Виноградова РАН, 32-37.
[4] Кузнецов С.А. (2003). Большой толковый словарь русского языка. Гл. ред. С. А. Кузнецов. СПб. // Грамота.ру : справ.-информ. портал. — URL: https://gramota.ru/biblioteka/slovari/bolshoj-tolkovyj-slovar [дата обращения: 02.04.2024].
[5] Морковкин В. В., Богачева Г. Ф., Луцкая Н. М. (2019). Большой универсальный словарь русского языка. М.: Словари XXI века.
[6] Панов М.В. (1979). Современный русский язык. Фонетика. Москва: Высшая школа.
[7] Шаронов И.А. (2009). Междометия в языке, в тексте и в коммуникации. Автореферат дисс. … д-ра филол. наук. Москва: РГГУ.
[8] Янко Т. Е. (2008). *Интонационные стратегии русской речи в сопоставительном аспекте*. Москва: Языки славянских культур.

References

[1] Boersma, Paul; Weenink, David. (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.38, 2021. Access mode: http://www.praat.org/

[2] Fox B.A., Hayashi M, Jasperson R. (1996). Resources and repair: a cross-linguistic study of syntax and repair. In: Ochs E, Schegloff EA, Thompson SA, eds. *Interaction and Grammar*. Studies in Interactional Sociolinguistics. Cambridge University Press, 185-237.

[3] Janko T. E. (2008). Intonacionnye strategii russkoj rechi v tipologicheskom aspekte [Intonational strategies in spoken Russian from a comparative perspective]. Moskva: Jazyki Slavjanskix Kul'tur.

[4] Kasatkin L.L. (2006). Sovremennyi russkii yazyk. Fonetika [Modern Russian. Phonetics]. Moskva: Izdatel'skii tsentr «Akademiya»

[5] Kibrik A. A., Podlesskaya V. I. [Eds.] (2009). Rasskazy o snovidenijax: korpusnoe issledovanie usntogo russkogo diskursa [Night Dream Stories: A corpus study of spoken Russian discourse]. Moscow: Jazyki Slavjanskix Kul'tur.

[6] Korotaev N. A., Panysheva D. A., Neverova E. A., Podlesskaya V. I. (2024). Korpus «Chto ya videl» kak instrument analiza panicheskogo diskursa // «Slovo i zhest». Nauchnaya konferentsiya, posvyashchennaya pamyati E. A. Grishinoi («Grishinskie chteniya»). Moskow, February 8, 2024. Materialy konferentsii. / Otv. red. S. O. Savchuk. − M.: Institut russkogo yazyka im. V. V. Vinogradova RAS, 32-37.

[7] Kuznetsov S.A. (2003) Bol'shoi tolkovyi slovar' russkogo yazyka. Gl. red. S. A. Kuznetsov. SPb. // Gramota.ru : URL: https://gramota.ru/biblioteka/slovari/bolshoj-tolkovyj-slovar [access date: 2.04.2024].

[8] Morkovkin V. V., Bogacheva G. F., Lutskaya N. M. (2019). Bol'shoi universal'nyi slovar' russkogo yazyka. M.: Slovari XXI veka.

[9] Panov M.V. (1979). Sovremennyi russkii yazyk. Fonetika [Modern Russian. Phonetics]. Moskva: Vysshaya shkola.

[10] Podlesskaya V. (2015). A corpus-based study of self-repairs in Russian spoken monologues // Russian Linguistics, Volume 39, Issue 1, 2015, 63-79.

[11] Sharonov I.A. (2009). Mezhdometiya v yazyke, v tekste i v kommunikatsii. Avtoreferat diss. … d-ra filol. nauk. Moskva: RGGU.

Intensity and its manifestations in speech and recurrent gestures in spontaneous dialogue

Olga Prokofyeva
Moscow State Linguistic University,
Moscow, Russia
oliviaprok@gmail.com

Maria Kiose
Moscow State Linguistic University,
Institute of Linguistics RAS,
Moscow, Russia
maria_kiose@mail.ru

Anna Leonteva
Moscow State Linguistic University,
Institute of Linguistics RAS,
Moscow, Russia
lentevanja27@gmail.com

Evgeniya Smirnova
Moscow State Linguistic University,
Moscow, Russia
eva.buddaeva@gmail.com

Abstract

The present article explores the category of intensity in multimodal spontaneous dialogues in Russian. We regard intensity as a cognitive category expressed in notional, referential and sign meanings, i.e. in its manifestation degrees (high, medium, and low) and interaction of quantitative and qualitative meanings, referent types and types of lexical meaning. We study multimodal intensity patterns as revealed in the spontaneous dialogue in question, with the main focus on recurrent co-speech gestures, where the latter are attributed the stability of form and function. The research material consists of the recordings of about 3 hours featuring 20 participants, 1082 speech intensifiers, and 392 co-recurrent gestures of 9 groups. The results show, despite the susceptibility to individual preferences, that this is the notional meaning of intensity in speech that is more consistently revealed in recurrent gestures. The presenting recurrent gesture group appears to be most numerically frequent, nevertheless, these are enhancing and locating gestures that could help distinguish between pure quantity and merged quantity-quality cases. In general, exploring the use of gestures as mediated by intensity meanings allowed to specify their discourse functions attributed to intensity.

Keywords: intensity, multimodal behavior, recurrent gesture, speech, spontaneous dialogue
DOI: 10.28995/2075-7182-2025-23-281-290

Интенсивность и особенности ее реализации в речи и рекуррентных жестах в спонтанном диалоге

Ольга Прокофьева
Московский государственный
лингвистический университет,
Москва, Россия
oliviaprok@gmail.com

Мария Киосе
Московский государственный
лингвистический университет,
Институт языкознания РАН,
Москва, Россия
maria_kiose@mail.ru

Анна Леонтьева
Московский государственный
лингвистический университет,
Институт языкознания РАН,
Москва, Россия
lentevanja27@gmail.com

Евгения Смирнова
Московский государственный
лингвистический университет,
Москва, Россия
eva.buddaeva@gmail.com

Аннотация

В статье на материале русского спонтанного диалога исследуются особенности проявления интенсивности в мультимодальном поведении говорящего. Интенсивность рассматривается как когнитивная категория, которая в речи реализуется посредством понятийных, референциальных и знаковых характеристик, обладая разной степенью проявления (высокая, средняя и низкая), специфической особенностью сочетания качественности и количественности для референтов разного типа с учетом типов лексических значений маркирующих ее слов и выражений. В работе в спонтанном диалоге устанавливаются особенности ее мультимодальной реализации с опорой на рекуррентные жесты, демонстрирующие устойчивые паттерны формы и функции. Материалом исследования являются записи 20 участников общей продолжительностью около 3 часов, включающие 1082 интенсификатора и 392 сопровождающих их рекуррентных жеста в 9 группах. В ходе анализа определено, что, несмотря на индивидуальные различия, наиболее устойчивые мультимодальные паттерны устанавливаются в рамках понятийных значений категории. Наиболее частотная группа жестов — презентирующие, однако случаи количественного и количественно-качественного значений интенсификаторов разграничиваются в жестах усиления и местоположения. В целом, результаты уточняют функциональную роль рекуррентных жестов как определяемых характеристиками интенсивности в дискурсе.

Ключевые слова: интенсивность, мультимодальное поведение, рекуррентный жест, речь, спонтанный диалог

1 Introduction

Intensity has long been explored as a speech category, in Russian as well (e.g. [1; 2; 3]); it displays the semantics of quality and quantity and is interrelated with the categories of evaluation, gradation and emotivity. As known, intensity is considered within the frameworks related to emotivity studies, communicative and discursive studies, as well as to the studies in lexical semantics. Additionally, recent studies indicate that it can be manifested in other communicative modes, e.g., in gesture alone or in speech and gesture [4; 5, 6; 7; 8], which means that intensity can be viewed as a speech category, a gesture category, and as a cognitive category structuring both speech and gesture.

In this study, we address intensity as a cognitive category in speech and co-speech recurrent gestures known as "stabilized forms that embody a practical knowledge of dealing with different communicative, interactional, cognitive, and affective tasks" [9, p. 32]. Apart from being conventionalized and culturally shared, they "often work on the level of speech, fulfilling pragmatic functions" [10, p. 1558]. Meanwhile, little is known on how co-speech recurrent gestures shape intensity. The study identifies the regularities which appear in the speech and gesture distribution affected by different characteristics of intensity, such as its degree (high, medium, low), the distribution of its qualitative and quantitative meanings, the object of intensification.

2 Theoretical framework

2.1 Intensity and multimodality

In speech, intensity is commonly examined via its lexical and syntactic markers which can manifest either communicative or discursive meaning. While communicatives serve to express intensity within a speech act or in shaping emotivity, discursive markers serve to foreground the discursive functions of intensity [11]. Therefore, there are studies which explore intensity in expressing pragmaticity and emotivity [1; 12; 2; 3; 13; 14] and which explore its role in structuring discourse [15;16], as well as studies which develop an integral view onto its pragmatic and discursive use [17]. Importantly, intensity can be mediated by evaluation and graduality. In [18], M. Glovinskaya identifies three types of intensifying verbs considering their evaluative component: the ones always modulated by evaluation like *пялиться* (stare), *вырядиться* (be overdressed), non-modulated by evaluation like *наготовить* (еды) (cook all sorts of (food)), *силиться* (take pains), and allowing both depending on the situation like *навалить* (pile up) which can be used to express intensity only or intensity modulated by evaluation. Intensity is frequently explored in adjectives and adverbs as mediated by their graduality expressed in the degrees of comparison, but also in their lexical meaning displaying quality, quantity, evaluation [19]. Syntactically, intensity can be expressed in particles and pronouns where it is modulated by emotionality and foregrounding [20]. Therefore, intensity can be expressed in various linguistic means manifesting lexical, syntactic, pragmatic and discursive meaning and can additionally manifest emotivity and evaluation.

In attribution to gestures, there exist two distinct approaches to considering intensity, with the first approach treating intensity as a singularly gesture category, and the second viewing it as a category structuring both speech and gesture. The first approach explores intensity in gestures via their amplitude and spatial organization [7; 8]. For instance, intensity in gesture was found to modulate discourse efficiency in professional studies [7]. In [8] several singular features of gestures used to manifest emotional intensity were identified, gesture length, pressure and speed. In the second approach, intensity is considered within the framework of co-speech gestural behavior. For example, E. Grishina [4] examined the distribution of gestures with vocal interjection *Oh* (Russian O) and detected that pointing gestures were most frequent. The author also discussed the fist-configuration and its connection to intensity, especially some verbs with suffix -ану- (e.g. *рубануть*) in Russian which can be accompanied by such gestures due to their semantics and intensity that they possess [21]. S. Savchuk and A. Makhova [5] explored gestures mediated by the emotional contexts of different intensity and determined that one emotion can be expressed via different gestures, while one gesture type may relate different emotions. In [6] it was found that the distribution of representational gestures in L1 and L2 is overall mediated by emotional intensity, where fewer representational gestures were used to express intensity in L2.

These findings suggest that intensity can be explored as 1) a speech category, 2) a gesture category, and 3) as a category structuring both speech and gesture. The latter view implies that according to McNeill [22], it can modulate the communicative modes. In this paper, we consider intensity as a cognitive category which shapes the construal of discourse expressing a certain degree of quantity in manifesting the quality of discourse referents. To proceed, we adopt the view developed by O. Iriskhanova, who presumes that a cognitive category can be explored via its epistemic, ontological and semiotic aspects [23], i.e., via its notional, referential and sign meaning. The notional meaning of intensity is demonstrated in its degree of intensifying referents, their features, actions and states in discourse. The referential meaning relates to the type of referent intensified in discourse. The sign meaning describes the role of the intensifying component within the meaning of a word and in the gesture that displays it in discourse. Presumably, speech and gesture may both contribute to shaping intensity in discourse.

2.2 Recurrent gestures

Recurrent gestures bring specific aspects of discourse to attention accompanying negation, disfluency markers, interactional moves, etc., however, they are quite often detached from speech [24]. D. McNeill [25] explores the relationship between mental processes and social interactions through the lens of recurrent gestures. One of the significant findings is the concept of gesture-speech co-expressivity, which suggests that gestures and speech work together to convey meaning. A. Kendon was one of the first authors to describe the physical. i.e. formal aspect of hand gestures, identifying such forms as Open Hand Neutral Open Hand Prone, Open Hand Oblique, etc. with different palm orientation (i.e., away, up, down, vertical) [26]. C. Müller [27] identifies three key recurrent gestures: the Palm-Up-Open-Hand, Holding Away and Cyclic gesture, which highlight their position on a continuum from spontaneous to fully conventionalized expressions. In [28] ring, cyclic, palm up and down open hand, index finger and several others, specific form and movement parameters with a set of related functions are exemplified. H. Ladewig and J. Bressem [29] analyzed gestures using a systematic approach based on four parameters: hand shape, orientation, movement, and position in gesture space across the German speakers. The study revealed that the parameters of orientation, movement, and position showed similar patterns across individuals, which suggests that gestures are not merely individual improvisations but rather standardized forms that are consistently used within a cultural group.

Following this approach, we analyzed the recurrent gestures according to their form and movements, based on the image schemas that shape gestural physical properties. We distinguish seven major types of such schemas in gesture that accompanied the intensity markers in speech. These include presenting, outlining, locating, enacting, opposing, enhancing, away gestures. There are three categories, cyclic gestures, emblems and sway gestures, that were not put into any bigger categories and each presents its own group.

3 Experiment design

3.1 Participants and experiment procedure

The present study is carried out on the data collected within an experiment which recorded spontaneous dialogues between the native Russian speakers as they discussed in pairs a preset series of questions about AI. The corpus is 171 minutes long featuring 20 participants whose speech was recognised, transcribed and segmented by the automated transcription tool Whisper AI integrated into ELAN, and then further divided into elementary discourse units (EDUs) (the term was specifically developed by [30] for spoken discourse studies). Both speech and gesture of each participant were subjected to the analysis of stance-taking in speech including intensity coded Ev2 and form and movement patterns in gesture, where speech and gesture were annotated and categorised separately by two teams of researchers on different tiers in ELAN (Figure 1).

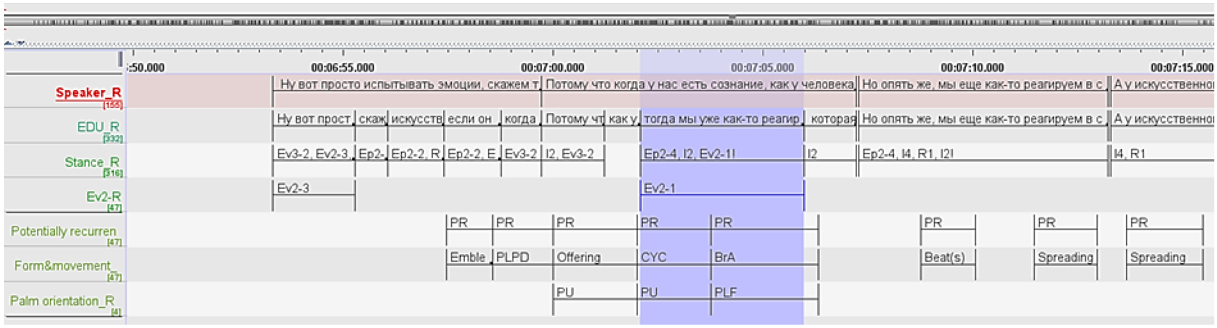


Figure 1: Fragment of analysis in ELAN with EDU *тогда мы уже как-то реагируем на там не знаю любое любую вещь* and two temporally co-occurring gestures highlighted

The data included 1083 cases of intensity (ranging from 29 to 115 cases per participant) accompanied by the total of 392 recurrent gestures (ranging from 5 to 42 gestures per participant) that were temporally aligned with the markers of intensity. Both markers and gestures could appear within one EDU as alone standing units or in clusters of 2-4 for speech or up to 2 in co-speech gestures. Though the speech analysis included putting speech into elementary discourse unit, the speech-gesture analysis was more precise due to the purposes of the current work. The analysis included only those gestures which co-occurred with intensity markers (as it was done in some works by E. Grishina, e.g. in [31]).

The data processing algorithm included 3 steps: frequency analysis of notional, referential and sign meanings in manifesting intensity in speech; frequency analysis of recurrent gestures used as co-speech gestures with the discursive markers of intensity; contingency analysis which allows to identify the regularities in the use of recurrent gestures with notional, referential and sign meanings of intensity.

4 Results and Discussion

4.1 Speech distribution in manifesting intensity

While addressing the notional meaning of intensity, following [1; 2; 3] we identified its two major notional features, namely, 1) degree of intensity and 2) degree of quantity and quality.

First, we assessed the degree of intensity, high, medium, and low intensity like in *в любом случае* (in any case), *достаточно* (enough), and *не очень хорошие* (not very good) respectively. The high degree cases proved to dominate in the corpus (694 cases – 64%). The examples often included highly expressive and emotionally charged lexical units covering reduplication cases (*большой-большой вопрос* – a big-big question) and exaggerations (*увидели там миллиарды люди* – it was seen by billions of people). The medium and low degree cases were scarcer, 152 (14%) and 237 (22%) cases respectively. The medium degree cases also included comparatives (*еще меньше процент даже* – the percentage is even lower), whereas the low degree cases included absolute negation and absence (*ни одного человека с планеты Земля не осталось* – there was no human being from Earth left). Second, we identified the pure quantity cases and those carrying the quantity meaning modified by qualitative features (merged quantity-quality cases). For instance, in *ну их много типа* (well there are many of them) we observe a

case of pure quantity, as *много* (many) is stylistically neutral and tells the interlocutor only about the number of objects. In *хотя бы будет симулировать какие-то эмоции искусственный интеллект* (AI will at least simulate some emotions) the intensifying marker *хотя бы* (at least) adds the attitude of annoyance to the meaning of small quantity. Being spontaneous by nature, the data proved to be rich in the 'emotional' samples of the second group (433, 40% of all intensity cases).

In exploring referential meaning, we presumed following [32] that the intensified referent can be related to either an object, a process (processes proper, actions, states and attitudes) or an attribute (stable and unstable). For instance, in *и ничто не может быть подобно человеку* (and nothing can be similar to a man) in *ничто* (nothing) the referent is of an object type. The use of a process can be exemplified in *просто находясь в атмосфере этой планеты* (simply while being in the atmosphere of this planet), where the marker *просто* (simply) modulates the use of a referent *находясь* (being) which is a state. In *хотя наверное это гораздо легче* (although this is much easier) the marker is *гораздо* (much), it shapes the referent expressed in an attribute *легче* (easier). In multiple cases we observe the presence of several referent types, e.g., in *но это не всегда правда* (this is not always the truth) the referent *truth* (правда) intensified by *не всегда* (not always) manifests both an object and a quality. In *и нет однозначного ответа* (there is no definite answer) the referent expressed in *ответ* (the answer) intensified by *однозначного* (definite) due to its morphological semantics expresses both an object and a process; additionally, the EDU intensifies the referent of an attribute type *однозначного* (definite). In major cases the intensified referents are of an attribute type (520 cases), object referents are less frequent (420 cases), the least frequent are processes (271 cases).

The sign meaning is manifested in direct or indirect meaning of intensity within a word. In the first case this is the main significative meaning of a word, e.g., in *единственный, любой* (only, every), in the second deals the intensifying meaning is connotational, e.g., in *прекрасно* (*прекрасно понимаю;* understand perfectly). Of special interest are the cases of using words or expressions with emotional connotations appearing in adjectives, verbs and nouns, e.g., *облажаться* (fail) for 'a very poor showing', *бредни* (nonsense) for 'unlikely information', *развлекаловка* (amusement) for 'something very frivolous', etc. The intensifying meaning is often implied in single words and in expressions in the context, e.g., in *кожаные мешки* (skin sacks) – which means 'merely humans', where *merely* is an intensifier. The number of cases of direct intensifying meaning is significantly higher (62% of all the examples considered).

4.2 Speech and gesture distribution in manifesting intensity

Next, we address the distribution of recurrent gestures. The first group, presenting gestures, include the hand movements that in their form have a resemblance of giving something to the interlocutor. The second group, outlining, includes gestures which in their movement contain the idea of shaping some idea or object. The third group, locating gestures, are used when the gesture movement is used to put some idea or object in space around the speaker. The fourth group, enacting gestures, include hand movements with the idea of doing an action with some object, e.g., as if cutting, hitting, clapping, grasping or tossing. The fifth group, opposing gestures, include gestures of binary relation, when a speaker repeats the same movement in two opposite planes, as if mirroring, weighing it. The sixth group, enhancing gestures, includes gestures that can be used to highlight some idea (precision grip, straight index held, fist, snapping one's fingers). The seventh group is away gestures, which have the idea of moving something away from the speaker, e.g., sweeping, brushing, throwing, keeping away and back and forth movement. Some gestures (sway, cyclic and emblematic) were left separate, as they do not form any groups nor fall under any of the above categories.

In Table 1 we present the distribution of gestures in the corpus.

	presenting	outlining	locating	enacting	opposing	enhancing	away	sway	cyclic	emblem	All
Total	945	278	286	237	115	225	118	29	173	66	2600
With intensity, abs (ratio)	129 (13.7)	58 (20.9)	34 (11.9)	40 (16.9)	17 (14.8)	41 (18.2)	32 (27.1)	4 (13.8)	23 (13.3)	14 (21.2)	392 (15.1)

Table 1: Distribution of recurrent gestures

With the total number of gestures equal to 2600, 329 gestures (15% out of the whole data set) co-occur with intensity cases in speech. The most frequent gestures are presenting, outlining, enhancing and enacting. Contingency tests showed the differences in the distribution of several gesture groups with and without intensity; these are away-gestures ($\chi2=13.99$, $p<0.001$) and outlining gestures ($\chi2=8.14$, $p=0.005$), which means that these gestures more often than not function as intensifying gestures. The results do not fully corroborate the results described in [4], where the prevalence of pointing gestures (classified as locating in this paper) with emotive words was found. Although locating gestures are rather frequent in our corpus, presenting gestures are 3.4 times more frequent. Meanwhile, the results comply with the findings in [6], which reported that representational gestures were frequent in expressing emotive intensity. Since in the present study representational gestures appear in the recurrent gestures of outlining, locating, enacting, away gestures, we can conclude that they constitute the major part of all the gestures used by the speakers.

One of the regularities found in the multimodal discourse under analysis is an almost identical distribution ratio of high, low and medium degree of intensity markers in speech and co-speech gestures of all types accompanying these markers at 64% vs 21% vs 14%. The same could be said about the distribution of 'merged quantity-quality' vs. 'pure quantity' cases (433 vs. 649 cases in speech and 151 vs. 241 cases in gestures), with their ratio of 40% vs. 60% for both, as well as for the distribution of indirect and direct lexical meanings in the sign meaning of intensity (about 40% vs. 60%). As for referential meaning, we observe 1211 cases, with objects constituting 35%, attributes constituting 43%, and processes constituting 22%. However, with 422 gestures used with different referents, 34% are used with objects, 24% with attributes and 41% with processes; therefore, we observe the redistribution of gestures used with attributes and processes.

4.3 Contingency and variance in multimodal manifestation of intensity

The gesture distribution with notional, referential and sign meaning of intensity is shown in Table 2.

	presenting	outlining	locating	enacting	opposing	enhancing	away	sway	cyclic	emblem	All
Intensity Degree											
high	75	37	20	28	9	24	25	2	18	12	250
medium	19	10	7	4	4	9	2	2	0	0	57
low	35	11	7	8	4	8	5	0	5	2	85
Quantity											
merged	53	26	7	16	5	22	12	1	6	3	151
pure	76	32	27	24	12	19	20	3	17	11	241
Referent											
object	42	24	14	14	7	11	16	3	10	4	145
attribute	33	12	6	13	9	11	9	0	9	0	102
process	55	26	16	18	7	23	13	2	6	9	175
Lexical meaning											
indirect	46	25	17	12	6	15	15	2	7	9	154
direct	83	33	17	28	12	26	17	2	16	5	238

Table 2: Distribution of recurrent gestures accompanying speech intensity markers

In terms of the differences between high and low intensity degree with gestures, we observed the tendency of presenting gestures to differentiate between them, with $\chi2=3.59$, $p=0.059$, which means low-intensity markers more commonly than high-intensity markers co-occur with presenting gestures. Chi-square test carried out on the cases of merged quantity and quality vs. cases of pure quantity demonstrated a significant contingency result for locating ($\chi2=5.055$) and enhancing ($\chi2=4.431$) gesture groups ($p < 0.05$). Therefore, the use of these gestures could serve to identify purely quantitative (Figure 2) or merged quantitative-qualitative meanings of intensity.

Figure 2: Placing down gesture. Locating gesture with a pure quantity case in EDU *что как раз таки больше всего как бы контроля и нет* (that there is exactly kinda no control more than anything)

The speaker, his hand protruded, places it down on the wording *больше всего* (more than anything), as if determining his choice and locating it in discourse. Nevertheless, the cases of locating stay rather scarce (34), with only 11% of all locating gestures accompanying intensifiers.

In exploring both referential and sign meaning contingency with gesture groups, we did not identify any differences in the use of referents with any gesture groups. Overall, the results show that only the notional meaning of intensity manifests several steady speech-and-gesture alignment patterns, whereas referential and sign meanings do not. The results are somewhat contingent with the ideas expressed in [1; 2; 3] which mostly addressed the degree of intensity and the degree of quantity and quality as its two major features. The results shown in Table 2 allowed to observe several regulations in the use of speech markers of intensity and gesture groups. As seen, presenting gestures are rather frequent, especially when dealing with high-degree, pure quantity, object and attributive referents, and direct meaning intensifying cases (Figure 3).

Figure 3: Presenting gesture, view from two angles: *и там будет все так вот по структуре* (it will all be structured like that)

As the speaker raises her hand up, as if measuring the whole scale bottom-up, the presenting gesture co-occurs with the word *всё* (all), which features the high degree of intensity, pure quantity, whereas the referent is an object. Figure 4 illustrates an away gesture, as this group demonstrated the highest co-occurrence with intensifiers (27%).

Figure 4: Away gesture: *что он не сл- не всл- не* ***слепо*** *должен идти и верить всему* – that he should not bl- blindly go and believe everything)

The participant moves the hand back and forth emphasising her negation by an aligned gesture and uses a high-degree intensifier in *слепо* (blindly). Away gestures most frequently accompany a high-degree-intensity, pure-quantity intensifier with direct meaning and its referent being an object.

Overall, the study supports the idea proposed in [25; 27], who attributed the use of recurrent gestures to embodying their social interactions through gestures. Specifically, the study confirms the effectiveness of classification suggested in [29] based on hand shape, orientation, movement, and position in gesture space. Applying it to exploring intensity in discourse helped reveal that this is the notional meaning of intensity which is displayed more systematically in gesture, which evidences of its functional priority in multimodal discourse. The results also contribute to the efficiency of adopting the tripartite view of a cognitive category with its epistemic, ontological and semiotic aspects [23] in exploring its multimodal expression.

5 Final remarks

The present study developed the idea of exploring intensity as a cognitive category expressed multimodally, in speech and co-speech recurrent gesture. The results, as mentioned above, were obtained based on the word-gesture analysis rather than EDU-gesture analysis to establish clear relations between the hand movement and the intensity marker. Adopting the view that a cognitive category is revealed in notional, referential and sign meanings, we hypothesized that the differences in these meanings could also appear in recurrent gestures based on hand shape, orientation, movement, and position in gesture space. The obtained results indicate high frequency of presenting gestures as such gestures can display various functions in comparison to other gesture groups and can be polysemantic due to their form and movement types. The away gestures were shown to be quite rare however they were used more often than expected and thus we can conclude that one of their prime functions could be to intensify the quality of the referent they are used with, especially since in many cases they perform the function of negation, thus highlighting the referent or adding the semantics of negation to it. The main outcome of the paper is that this is the notional meaning of intensity in speech which is more consistently revealed in co-speech recurrent gesture, while referential and sign meanings do not display rigid alignment with recurrent gestures.

Acknowledgements

This research is supported by the Russian Science Foundation, project No. 24-18-00587 “The role of recurrent gestures across languages: The socio-cognitive core of multimodal stance-taking”, https://rscf.ru/project/24-18-00587/.

References

[1] Roman Mrázek. Designators of action intensification in the Russian language. *Sborník prací Filozofické fakulty Brněnské university*, 13 (12), 1964. P. 169–180.

[2] Svetlana E. Rodionova. Semantics of intensity and its expression in the modern Russian language. *Problems in functional grammar. Field structures*, ed. by A.V. Bondarenko. St Petersburg: MAKS Press, 2005. P. 150–166.

[3] Galina I. Kustova. Words with high degree meaning: semantic models and semantic mechanisms (Magn-adjectives). *Word and Language. Collection of articles for the 80th anniversary of Acad. Y. D. Apresyan* / ed. by I. M. Boguslavsky, L. L. Iomdin, L. P. Krysin. Moscow: Yask, 2011. P. 256–268.

[4] Elena A. Grishina. On the correlation between word and gesture (vocal gesture O in oral speech). *Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue",* 8 (15), 2009. P. 80–90.

[5] Svetlana Savchuk and Alexandra Makhova. Multimodal Russian Corpus and its use in emotional studies. *Russian Journal of Communication*, 13 (1), 2021. P. 42–61.

[6] ibrahim Akkan, Cagla Aydin, Sami Gulgoz and Tilbe Göksun. The Role of Hand Gestures on Emotional Intensity and Phenomenological Ratings of Autobiographical Narratives. *Proceedings of the Annual Meeting of the Cognitive Science Society*, 46. 2024. Retrieved from https://escholarship.org/uc/item/7381f12m

[7] James L. Byo. Recognition of Intensity Contrasts in the Gestures of Beginning Conductors. *Journal of Research in Music Education*, 38 (3), 1990. P. 157–163.

[8] Nabil Bin Hannan, Khalid Tearo, Joseph Malloch, Derek Reilly. Once More, With Feeling: Expressing Emotional Intensity in Touchscreen Gestures. *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 2017. P. 427–437.

[9] Silva H. Ladewig. Recurrent Gestures: Cultural, Individual, and Linguistic Dimensions of Meaning-Making. Cienki A, ed. *The Cambridge Handbook of Gesture Studies. Cambridge Handbooks in Language and Linguistics*. Cambridge: Cambridge University Press: 2024. P. 32–55.

[10] Silva H. Ladewig. Recurrent gestures. Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill & Jana Bressem (eds.), *Body-Language-Communication: An International Hand-book on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38.2.)* Berlin, Boston: De Gruyter: Mouton: 2014. P. 1558–1575.

[11] Igor A. Sharonov. Discursive words and communicatives. *Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue"*, 15, 2016. P. 538–547.

[12] Nina D. Arutunova. On the object of general evaluation. *Issues of Linguistics*, 3, 1985. P. 13–24.

[13] Yulia. V. Nikolaeva. *On the history of the word absolutely in the Russian language*. Studi Slavistici, XVII (2), 2020. P. 199–211.

[14] Valentina Y. Apresyan. Implicit aggression in a language. Irina M. Kobozeva, Natalya I. Laufer, Vladimir P. Selegei (eds.), *Computational linguistics and intellectual technologies, Proceedings of the International Conference "Dialogue"*, 2003. P. 32–35.

[15] Irina M. Kobozeva and Leonid M. Zakharov. Why we need a sounding dictionary of discursive words of the Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue"*, 2004. P. 292–297.

[16] Vera I. Podlesskaya and Andrey A. Kibrik. Discursive markers in the structure of oral storytelling: a corpus study. *Computational linguistics and intellectual technologies. Proceedings of the International Conference "Dialogue"*, 8(15), 2009. P. 390–395.

[17] Natalia V. Bogdanova-Beglarian, Olga I. Blinova, Tatiana Yu. Sherstinova, Daria A. Gorbunova, Kristina D. Zaides, Tatiana I. Popova. Pragmatic Markers in Dialogue and Monologue: Difficulties of Identification and Typical Formation Models. *SPECOM 2020. Lecture Notes in Artificial Intelligence, LNAI*, 12335. Springer, Switzerland, 2020. P. 68–78.

[18] Marina Y. Glovinskaya. Hidden hyperbole as expression and justification of speech aggression. Apresyan, Y. D. (ed.) *Sacred Meanings: Word. Text. Culture: Collected Papers in Honor of N.D. Arutyunova*. Moscow: Languages of Slavic Culture, 2004. P. 69–76.

[19] Irina A. Kraeva. Graduality: conceptual foundations and specifics of changes. *RSUH/RGGU Bulletin: Literary Theory. Linguistics. Cultural Studies Series*, 4 (3), 2022. P. 323–337.

[20] Svetlana M. Kolesnikova. *The category of graduality in the modern Russian language.* PhD Thesis, Moscow, 1999.

[21] Elena A. Grishina. *Russian gestures from a linguistic perspective.* Moscow: Languages of Slavic Culture, 2017.

[22] David McNeill. Gesture-speech unity. What it is, where it came from. In S. D. Kelly, R. B. Church, M. W. Alibali (eds.) *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins Publishing Company, 2017. P. 77–101.

[23] Olga K. Iriskhanova. On the concept of creativity and its role in the meta-language of linguistic descriptions. *Cognitive Studies of Language*, 5, 2009. P. 157–171.
[24] Jana Bressem, and Cornelia Müller. The family of AWAY gestures. Cornelia Müller, Alan Cienki, Ellen Fricke, Silva H. Ladewig, David McNeill and Jana Bressem (eds.), *Body – Language – Communication. An International Handbook on Multimodality in Human Interaction*. (Handbooks of Linguistics and Communication Science 38.2.) Berlin/Boston: De Gruyter Mouton, 2014. P. 1592–1604.
[25] David McNeill. Recurrent gestures: How the mental reflects the social. *GEST* 17, 2018. P. 229–244.
[26] Adam Kendon. *Gesture: Visible action as utterance*. Cambridge, UK: Cambridge University Press, 2004.
[27] Cornelia Müller. How recurrent gestures mean: Conventionalized contexts-of-use and embodied motivation. *GEST* 16, 2017. P. 277–304.
[28] Alan Cienki. Gesture and pragmatics: From paralinguistic to variably linguistic. A. Barron, G. Steen, & Y. Gu (eds.), *The Routledge Handbook of Pragmatics.* London: Routledge: 2017. P. 61–68.
[29] Silva H. Ladewig, Jana Bressem. New insights into the medium hand: Discovering recurrent structures in gestures. *Semiotica, 197,* 2013. P. 203–231.
[30] Nikolay A. Korotaev, Andrej A. Kibrik, Vera I. Podlesskaya. Complications in canonical structures: Between mono- and poly-predicative constructions. Andrej A. Kibrik, Vera I. Podlesskaya (eds.), *Night Dream Stories. A corpus study of spoken Russian discourse.* Moscow: Languages of Slavic Cultures: 2009. P. 219–332.
[31] Elena A. Grishina. Vertical axis in gesticulation from the linguistic point of view. *Russian Language and Linguistic Theory*, 1 (27), 2014. P. 42–89.
[32] Elena S. Kubryakova. On the Way of Obtaining Knowledge about Language: Parts of Speech from the Cognitive Point of View. *Language and Knowledge*. Moscow: Languages of Slavic Culture, 2004. P. 29–304.

Structured sentiment analysis using few-shot prompting of an ensemble of LLMs

Petr Rossyaykin
Lomonosov Moscow State University
Moscow, Russia
petrrossyaykin@gmail.com

Abstract

This paper describes our participation in RuOpinionNE-2024 shared task (Loukachevitch et al., 2025). The objective of this task is to extract opinion tuples of the form <holder, target, polarity expression, polarity> from news texts in Russian. We approached this task with few-shot prompting of super large language models (LLMs). The quality of LLMs' predictions was improved in two ways. In the first stage we used prompts with examples which text embeddings were similar to that of the target text. In the second stage we augmented prompts with answers of LLMs from the previous stage, achieving the second-best F1 score in the competition in the post-evaluation stage. Our results show that the addition of answer suggestions to the prompt is particularly useful if they provide novel and variable information.

Keywords: structured sentiment analysis, opinion extraction, few-shot prompting, large language models
DOI: 10.28995/2075-7182-2025-23-291-298

Few-shot prompting ансамбля больших языковых моделей для структурированного анализа тональности

Аннотация

В этой статье мы описываем наше участие в соревновании RuOpinionNE-2024. Задача соревнования состоит в извлечении кортежей мнений вида <источник мнения; объект мнения; выражение, содержащие мнение; полярность> из русскоязычных новостных текстов. Для решения этой задачи мы использовали few-shot prompting больших языковых моделей. Мы использовали два способа улучшения качества их предсказаний. На первом этапе в качестве примеров в промпте мы использовали тексты, векторные представления которых близки к векторному представлению целевого текста. На втором этапе мы добавили к промпту предсказания моделей, полученные на предыдущем этапе. Таким образом, на стадии post-evaluation мы достигли второго результата соревнования по F-мере. Наши результаты показывают, что добавление к промпту предлагаемых вариантов ответа улучшает качество, если они содержат новую и разнообразную информацию.

Ключевые слова: структурированный анализ тональности, извлечение мнений, few-shot prompting, большие языковые модели

1 Introduction

Sentiment analysis has been one of the major tasks in NLP for decades (Pang et al., 2002; Turney, 2002; Wiebe et al., 2005; Liu, 2012; Nakov et al., 2013; Socher et al., 2013; Pontiki et al., 2014; Golubev et al., 2023). Any opinion expressed in a text is instantiated in a polar expression (*e*) and has a holder (*h*), a target (*t*) and a polarity (*p*), so one has to recognize these four entities in order to extract an opinion from a text (Liu, 2012). Structured Sentiment Analysis (SSA) is thus defined as a task of extracting all opinion tuples of the form <*h*, *t*, *e*, *p*> from a given text (Barnes et al., 2021, 3387).

SSA is especially relevant for news texts, because, unlike e.g. reviews, they often do not have a single text-level sentiment but contain different opinions with different holders and/or targets or no opinions at all (Wiebe et al., 2005, 166). Despite this, SSA in news texts remains a particularly understudied subtask

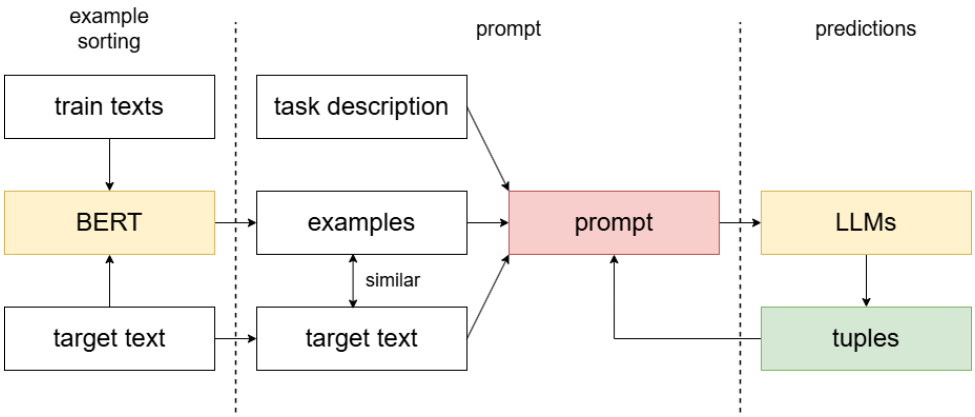


Figure 1: Overview of our approach.

of natural language understanding. For example, only one of seven datasets presented for shared task on SSA at SemEval 2022 contained news texts (Barnes et al., 2022, 1282). Incidentally, all but one of the participants of that task achieved the worst score on that particular dataset (Barnes et al., 2022, 1284), which suggests that news texts constitute a more difficult domain for SSA than reviews.

In this paper we discuss our participation in RuOpinionNE-2024, a shared task on extracting opinion tuples from Russian news texts (Loukachevitch et al., 2025). Our approach consisted in direct generation of opinion tuples using few-shot prompting of super large language models (LLMs). We used two tricks to improve the quality of LLMs' predictions. The first trick was to choose examples according to their semantic similarity to the target example. The second trick was to augment prompts with the answers from the first stage. This can be done cyclically and for any LLM both its own answers and the answers of other LLMs can be included. This allowed us to further improve the quality, although we found out that cyclic application of this procedure is not beneficial because LLMs benefit from novel and variable, rather than similar and recycled, information. Figure 1 schematically presents our method. We describe the method, the models used and the results in more detail in Sections 3–5.

2 Task description and the data

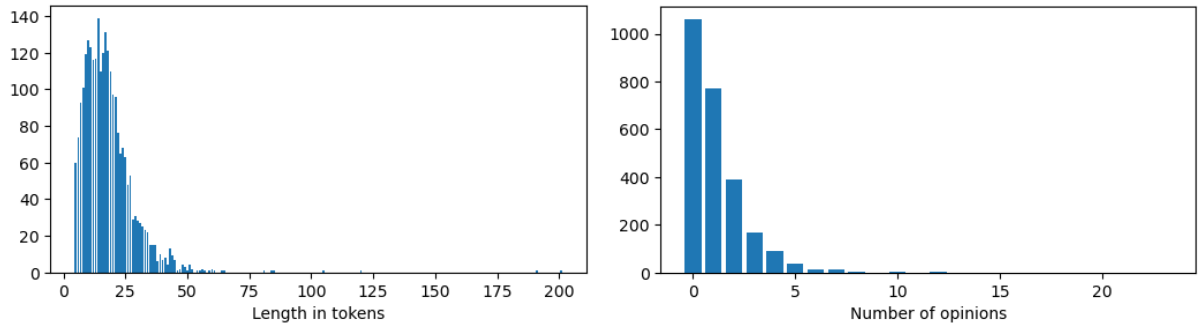
RuOpinionNE-2024 presents the task of opinion tuple extraction from Russian news texts. Opinion tuples have the following form: *<h(older)*, *t(arget)*, *e(xpression)*, *p(olarity)>* (the polarity is either POS(itive) or NEG(ative)). The train set of the competition consists of 2556 texts annotated for structured sentiment and the test set consists of 803 texts. The texts are mostly single sentences taken from news texts in Russian, although there are several texts which contain multi-sentential direct speech. The average length of texts in the train set is 18.1 tokens, i.e. orthographic words, the shortest text is 5 tokens long and the longest one consists of 201 tokens. The distribution of the lengths of texts in the train set is shown in Figure 2a.

Train set texts contain from 0 to 23 opinions, although only 5 texts contain more than 10 opinions and only 38 texts contain more than 5 opinions. 1062 texts contain no opinions. The number of opinions per text is shown in Figure 2b.

It is also of note that there are 2904 opinions in the train set in total and in 1281 cases the opinion holder is covert ('NULL' in 897 cases and 'AUTHOR' in 384 cases).

3 Few-shot prompting approach to SSA

The gist of our approach is to generate opinion tuples directly by prompting super large language models.



(a) Distribution of texts' lengths in the train set

(b) Number of opinions per text in the train set

Figure 2: Basic properties of the train set

3.1 The first stage: basic prompt

The basic prompt we used is given below in the left column with two examples ("shots") in the prompt. The English translation of the prompt is given in the right column.

Ты эксперт в оценке тональности.

Тебе нужно найти все негативные и позитивные отношения между сущностями в тексте и вывести их в следующем формате:

[источник отношения, объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Если источником отношения является автор, то пиши:

['AUTHOR', объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Если выраженного источника нет, то пиши:

['NULL', объект отношения, выражение в тексте содержащее оценку, оценка (POS/NEG)]

Допустимо вернуть пустой ответ:

[]

Не нужно давать пояснений к ответу.

Примеры

Текст: По итогам первого полугодия 2016 года банк занимает 41-е место по размеру активов.

Ответ: []

Текст: Русская Википедия в четвёртый раз получила «Премию Рунета»

Ответ: [['NULL', 'Русская Википедия', 'Премию Рунета', 'POS']]

Текст: (К слову четвертое место в списке лидеров по итогам 2010 года занимает Москва).

Ответ:

You are an expert in sentiment analysis.

You need to identify all positive and negative relations between entities in the text and present them in the following format:

[source of sentiment, target, polar expression, polarity (POS/NEG)]

If the author is the source of sentiment, write:

['AUTHOR', target, polar expression, polarity (POS/NEG)]

If there is no explicit source, write:

['NULL', target, polar expression, polarity (POS/NEG)]

Returning an empty answer is allowed:

[]

Do not provide any explanations in the response.

Examples

Text: As of the first half of 2016, the bank ranks 41st in terms of assets.

Answer: []

Text: The Russian Wikipedia has received the "Runet Prize" for the fourth time.

Answer: [['NULL', 'Russian Wikipedia', 'Runet Prize', 'POS']]

Text: (By the way, Moscow ranks fourth in the list of leaders based on the results of 2010.)

Answer:

This prompt consists of three parts: the instruction, “shots” and the target text, which is *(К слову четвертое место в списке лидеров по итогам 2010 года занимает Москва)* in the given example. We used only the “user” part of the prompt without splitting it into the “system” and the “user” parts, because this resulted in a slightly worse quality during our preliminary experiments.

The “shots” were chosen according to cosine similarity of their text embeddings to that of the target text. We used mean pooling of token embeddings from BERT output to get text embeddings. For each target sentence, we added *n* examples with most similar texts after the instruction, where *n* is the number of shots.

The opinion tuples generated by LLMs given prompts of this kind constituted the predictions in the first stage.

3.2 The second stage: augmented prompt

We augmented the prompt with predictions obtained in the first stage. They were added to the prompt after the list of examples. An example of augmented prompt is given below (left column) together with its translation (right column):

<instruction>
<examples>
Текст, который нужно проанализировать:
(К слову четвертое место в списке лидеров по итогам 2010 года занимает Москва).
Ответы экспертов к этому тексту:
[[’NULL’, ’Москва’, ’четвертое место в списке лидеров по итогам 2010 года занимает’, ’POS’]]
[]
[]
Ты можешь выбрать из этих ответов или ответить по-своему.
Твой ответ:

<instruction>
<examples>
Text to analyze:
(By the way, Moscow ranks fourth in the list of leaders based on the results of 2010.)
Expert answers to this text:
[[’NULL’, ’Moscow’, ’ranks fourth in the list of leaders based on the results of 2010’, ’POS’]]
[]
[]
You can choose from these answers or provide your own answer.
Your answer:

This stage can be repeated cyclically, i.e. the predictions of LLMs from this stage can be added to the prompt instead of predictions from the previous stage or in addition to them, and inference can be rerun with the new prompt.

To sum it up, the prompts have the following form:

Basic prompt (the first stage)	**Augemented prompt (the second stage)**
<instruction>	*<instruction>*
<examples>	*<examples>*
<target text>	*<target text>*
	<answers from the previous stage/iteration>

4 Experiments

4.1 LLMs

Originally we experimented with three super large language models: GPT-4o[1], Grok-Beta[2], and Mistral Large 2 [3]. As a baseline, we predicted opinion tuples with the basic prompt and randomly chosen examples from the train set (Stage 0 in Table 1).

[1]Model name via OpenAI API: `gpt-4o-2024-11-20`.
[2]Model name via xAI API: `grok-beta`.
[3]Model name via Mistral API: `mistral-large-2411`.

In the first stage we used the same prompt with examples chosen according to their semantic similarity to the target text. In the second stage the predictions of all three models were added to the basic prompt to form the augmented prompt. In the third stage we used the answers from the second stage to augment the prompt. Grok and GPT were tested only at stages 0–2.

We also experimented with DeepSeek-V3[4] but its answers were not included in the prompts (at stages 2–3), because we started experimenting with this model at a later moment. As will be shown below, this resulted in this model yielding the best score.

With each of the models, we set the temperature to 0.1, top-p to 0.9 and max tokens to 512. We used 12 examples per prompt. These parameters were chosen heuristically.

4.2 Text embeddings

We tried out two models for text embeddings generation: Sentence RuBERT[5] and SBERT[6]. We achieved slightly better results using Mistral with SBERT text embeddings during our preliminary experiments, so we used these text embeddings for example choice throughout the experiments we report here. However, we also experimented with switching to Sentence RuBERT embeddings after the second stage with two LLMs: Mistral Large 2 and DeepSeek-V3 (Stage 3b in Table 1). Intuitively, this makes sense because examples ranked as the most similar to the target by Sentence RuBERT and SBERT barely overlap. In particular, if 12 most similar examples are taken per target text, the average overlap across all 803 target texts is just 1.84 examples per text.

The results of our experiments are summarized in Table 1. The best result for each model is in bold. The overall best result is 0.363 achieved at Stage 2 by DeepSeek-V3. This was the second-best result in the competition by the time of submission of this paper (April 2024) with the best score being 0.41.

model	stage				
	Stage 0	Stage 1	Stage 2	Stage 3a	Stage 3b (w/ RuBERT)
gpt-4o-2024-11-20	0.239	**0.349**	0.344	–	–
grok-beta	0.245	0.33	**0.349**	–	–
mistral-large-2411	0.23	0.327	0.34	**0.345**	0.343
deepseek-chat	–	0.33	**0.363**	0.356	0.358

Table 1: F1-score on the test set

4.3 English prompt

Following a suggestion of an anonymous reviewer, we also partially replicated the experiments using the English version of the prompt, which can be found in Section 3 above. All else being equal, this resulted in a slight increase of F1-score in 3 out of 4 setups we had tried. This is shown in Table 2.

model	stage	
	Stage 1	Stage 2
mistral-large-2411	0.335	0.344
deepseek-chat	0.34	0.351

Table 2: F1-score on the test set using the English prompt

5 Discussion

Our overall findings are as follows:

1. The choice of semantically similar, rather than random, examples resulted in a significant increase of quality for each of the models.

[4]Model name via DeepSeek API: `deepseek-chat`.

[5]`http://hf.co/DeepPavlov/rubert-base-cased-sentence`

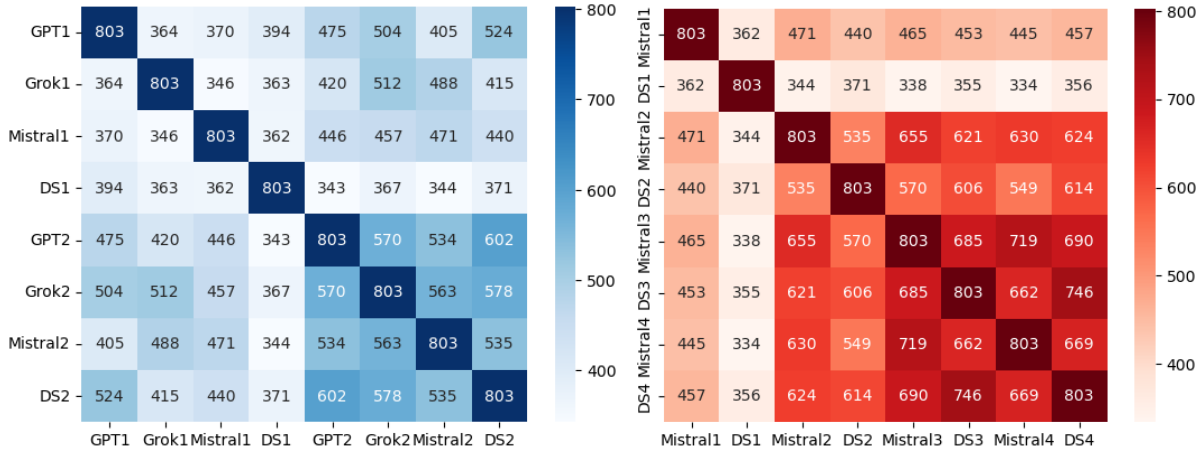
[6]`http://hf.co/ai-forever/sbert_large_nlu_ru`

2. The inclusion of answers from the previous stage improved the quality with 3 out of 4 models. The most significant increase is observed with DeepSeek-V3 which provided our best result in the competition. It should be noted once again, that we did not include the answers of this model in the augmented prompt, so this is the only model which was provided with the answers of three *other* models rather than its own answers and answers of two other models.
3. As for the third stage, it can be seen that it can improve the quality by only a very little margin, e.g. from 0.34 to 0.345 or 0.343 depending on text embeddings using Mistral. So, after the second stage further iterations of predictions using predictions from the previous stage does not seem to make much sense.
4. The best result was achieved with DeepSeek at the second rather than the third stage, although in both cases the model received completely novel information in the form of answer suggestions from Grok, GPT and Mistral and the suggestions at Stage 3 were of slightly better quality.
5. Finally, we observe that after the first stage the scores of three models which were prompted with each others answers converge at almost the same level (0.34–0.35 F1).

Figure 3 shows the matches between models' predictions. These heatmaps explain some of our observations. Firstly, as can be seen in Figure 3a, the models' predictions became more similar to each other in the second stage which explains similar scores achieved by different models at this stage (Observation 5). Secondly, it can be seen that the predictions of the first stage exhibit roughly the same level of disparity (upper left part of Figure 3a). However, DeepSeek's predictions in the first stage are very different from any model's prediction in the second stage. In other words, this LLM benefited from the augmented prompt more than the others and changed its predictions more radically which resulted in a higher improvement of the score. This is arguably due to the absence of DeepSeek's own predictions in the augmented prompt. This is a plausible explanation of Observation 2.

Figure 3a also explains Observation 4. Although predictions of Stage 2 are of higher quality, they are less variable than the predictions of Stage 1 in terms of content. Providing DeepSeek with more variable suggestions of slightly lower quality resulted in better predictions achieved at Stage 2.

As for Observation 3, Figure 3b shows that with further iterations of augmented prompting the models' predictions become increasingly similar which makes cyclic application of this procedure basically useless. In general, our results suggest that LLMs benefit from seeing answers generated by other LLMs but their own answers should not be mixed in.

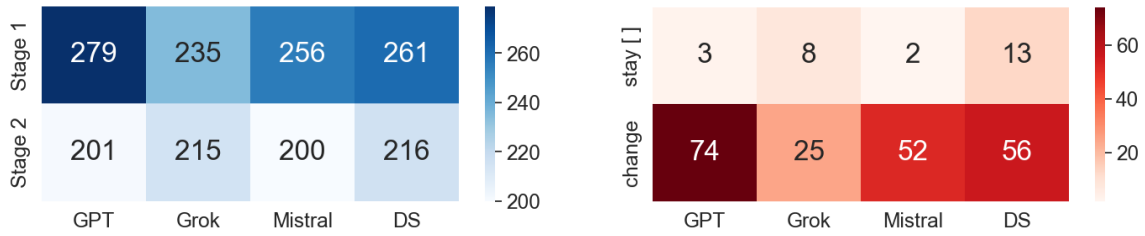


(a) All models, Stages 1–2

(b) Mistral & DeepSeek, Stages 1–3b

Figure 3: Pairwise counts of matching predictions of LLMs

We also observe that in the Stage 2 the number of predicted empty tuples dropped for all the models (Figure 4a). The models tend to switch to a non-empty tuple if at least one answer in the augmented prompt suggests a non-empty tuple. Figure 4b shows how the decisions were taken in such cases. This confirms that the answers suggested in the prompt have a significant influence on the models' output.



(a) Number of empty tuples predicted, Stages 1–2

(b) Resolution of conflicting predictions (Stage 2)

Figure 4: Changes in empty tuple predictions between Stage 1 and Stage 2

6 Conclusion

In this paper we described our participation in RuOpinionNE-2024 (Loukachevitch et al., 2025). Our method consisted in few-shot prompting of LLMs with semantically similar examples and answer suggestions. The contribution of this paper is two-fold. Firstly, we have shown that few-shot prompting of LLMs is a viable approach to SSA which allowed us to achieve the second-best score during the post-evaluation stage. Secondly, we explored interaction of super large language models in this task. In particular, we observed that (i) suggested answers have a significant influence on LLMs' predictions; (ii) LLMs benefit from novel and variable information from other LLMs; (iii) inclusion of LLMs' own answers and recycling of predictions in general do not have a positive impact.

Acknowledgments

The work was financially supported by Non-commercial Foundation for the Advancement of Science and Education INTELLECT.

We would like to thank the instructors at MSU.AI and in particular Alexander Ivchenko, Viktor Nemchenko and Artem Vasiliev. We would also like to thank Nikita Belyakov, Regina Nasyrova, Ksenia Studenikina, and anonymous reviewers of Dialogue 2025 for their valuable comments.

References

Jeremy Barnes, Robin Kurtz, Stephan Oepen, Lilja Øvrelid, and Erik Velldal. 2021. Structured sentiment analysis as dependency graph parsing. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processin*, P 3387–3402. Association for Computational Linguistics.

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. // *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, P 1280–1295. Association for Computational Linguistics.

Anton Golubev, Nicolay Rusnachenko, and Natalia Loukachevitch. 2023. Rusentne-2023: Evaluating entity-oriented sentiment analysis on russian news texts. *Computational Linguistics and Intellectual Technologies*, 130–141.

Bing Liu. 2012. *Sentiment analysis and opinion mining*. Springer Nature.

Natalia Loukachevitch, Natalia Tkachenko, Anna Lapanitsyna, Mikhail Tikhomirov, and Nicolay Rusnachenko. 2025. Ruopinionne-2024: Extraction of opinion tuples from russian news texts. // *Proceedings of International Conference on Computational Linguistics and Intelligent Technologies Dialogue 2025*.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 2013. Semeval-2013 task 2: Sentiment analysis in twitter. // *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, P 312––320. Association for Computational Linguistics, Atlanta, Georgia.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up? sentiment classification using machine learning techniques. // *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)*, P 79—86. Association for Computational Linguistics.

Maria Pontiki, Dimitros Galanis, John Pavlopoulos, Haris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. // *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014*, P 27–35, Dublin.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. // *Proceedings of the 2013 conference on empirical methods in natural language processing*, P 1631–1642, Seattle, Washington. Association for Computational Linguistics.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. // *Proceedings of the 40th Annual Meeting of the Association for Computational Lin- guistics*, P 417–424, Philadelphia, Pennsylvania.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39:165–210.

Methods for Recognizing Nested Terms

Igor Rozhkov
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
fulstocky@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University
Leninskie Gory, 1/4, Moscow, Russia
louk_nat@mail.ru

Abstract

In this paper, we describe our participation in the RuTermEval competition devoted to extracting nested terms. We apply the Binder model, which was previously successfully applied to the recognition of nested named entities, to extract nested terms. We obtained the best results of term recognition in all three tracks of the RuTermEval competition. In addition, we study the new task of recognition of nested terms from flat training data annotated with terms without nestedness. We can conclude that several approaches we proposed in this work are viable enough to retrieve nested terms effectively without nested labeling of them.

Keywords: nested terms, automatic term extraction, contrastive learning
DOI: 10.28995/2075-7182-2025-23-299-311

Методы распознавания вложенных терминов

Рожков И.С.
МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва, Россия
fulstocky@gmail.com

Лукашевич Н.В.
МГУ имени М.В. Ломоносова
Ленинские горы, 1/4, Москва, Россия
louk_nat@mail.ru

Аннотация

В этой статье мы описываем наше участие в конкурсе RuTermEval, посвященном извлечению вложенных терминов. Мы применяли модель Binder, которая ранее успешно применялась для распознавания вложенных именованных сущностей, для извлечения вложенных терминов. Мы получили наилучшие результаты распознавания терминов во всех трех треках конкурса RuTermEval. Кроме того, мы исследуем новую задачу распознавания вложенных терминов из плоских обучающих данных, аннотированных терминами без вложенности. Несколько методов, предложенных нами в данной работе, показывают, что вложенные термины возможно извлекать эффективно без наличия их вложенной разметки.

Ключевые слова: вложенные термины, автоматическое извлечение терминов, контрастивное обучение

1 Introduction

Automatic term extraction is a well-known research task that has been studied for decades. Terms are defined as words or phrases that denote concepts of a specific domain, and knowing them is important for domain analysis, machine translation, or domain-specific information retrieval. Various approaches have been proposed for automatic term extraction. However, automatic methods do not yet achieve the quality of manual term analysis.

During recent years, machine learning methods have been intensively studied (Loukachevitch, 2012; Charalampakis et al., 2016; Nadif and Role, 2021). The application of machine learning improves the quality of term extraction, but requires creating training datasets. In addition, the transfer of a trained model from one domain to another usually leads to degradation of the performance of term extraction. Currently, language models (Xie et al., 2022; Liu et al., 2020) are texted in automatic term extraction.

Available datasets usually contain so-called flat term annotation, that is, an annotated term cannot include other terms. Another formulation of the task assumes that terms can be contained within other terms, such terms are called nested terms. For example, the term "‘infection of upper tract" includes the following term: "infection", "tract", "upper tract". The extraction of nested terms allows the researcher to recognize more terms, but requires the application of special methods.

In 2024 new Russian datasets have been prepared for automatic term analysis in the framework of the RuTerm-2024 evaluation. The datasets are annotated with nested terms. The dataset includes texts from different domains, which allows for the study of model transfer between domains.

In this paper, we consider an approach for the extraction of nested terms and test it in the RuTerm-2024 dataset. In addition, we study the task of the transfer from flat terms to nested ones, that is, we suggest that the training data are annotated only with flat terms, but the trained model should can annotated nested terms. For both tasks, we experiment with the Binder model, which creates representations for target entity types using contrastive learning (Zhang et al., 2023).

2 Related Work

The automatic term extraction task has two main variants: corpus-oriented and document-oriented (Šajatović et al., 2019). In the former case, a system should generate a list of terms for a document corpus; in the latter case, the task is to identify mentions of terms in domain documents. This variant can be classified as a sequence labeling task.

The first corpus-oriented methods utilized statistical measures such as tf-idf, context measures (c-vaule), association measures based on frequencies of component words and co-occurrence frequencies of a candidate phrase (Astrakhantsev et al., 2015). The combination of measures leads to a significant improvement in term extraction performance (Bolshakova et al., 2013; Loukachevitch, 2012), but the transfer of a model trained for one domain to another domain leads to degradation of results (Hazem et al., 2022; Loukachevitch and Nokel, 2013; Bolshakova and Semak,).

In document-oriented term extraction, the authors of (Bolshakova and Efremova, 2015) exploited lexico-syntactic patterns and rules. Currently, the most effective approaches to document-oriented term extraction utilized BERT-based techniques (Lang et al., 2021; Tran et al., 2024).

The extraction of nested terms was studied in several works (Marciniak and Mykowiecka, 2015; Tran et al., 2024; Vo et al., 2022). The authors of (Marciniak and Mykowiecka, 2015) divide longer terms into syntactically correct component phrases and estimate the "unithood" of candidate terms using the NPMI statistical measure (Normalized Pointwise Mutual Information) (Bouma, 2009). Tran et al. (Tran et al., 2024) propose using a novel term annotation scheme for training models. The proposed scheme comprises an additional encoding for nested single terms. Applying the scheme and XLMR classifier, the authors obtain the best results on the ACTER multilingual dataset (Rigouts Terryn et al., 2022).

In a similar task of recognition of nested named entities, various approaches have been proposed (Loukachevitch et al., 2024; Zhang et al., 2023; Zhu et al., 2022a). One of the best methods for nested named entity recognition is Binder (Zhang et al., 2023), based on contrastive learning.

3 RuTermEval Competition and Dataset

As part of the Dialogue Evaluation in 2025, the RuTermEval competition was organized. In this competition, one should design a model capable of extracting nested terms given the labeled data.

There are three tracks in this competition:

- Track 1: Identification of terms
- Track 2: Identification and classification of terms into three classes (specific_term, common_term, nomen);
- Track 3: Identification and classification of terms into three classes (specific_term, common_term, nomen) with the formulation of the transfer learning task to other domains.

The organizers used Codalab to hold the competition. In the first track they used the usual F1 score for measuring the teams' submissions accuracy. For the second and third tracks, organizers used weighted F1 and class-agnostic F1 scores. The weighted F1 considers some classes more important than the others,

	train	dev (track 1 & 2)	dev (track 3)
specific	12664 (69.95%)	3387 (67.77%)	3270 (58.52%)
common	4866 (26.88%)	1275 (25.51%)	1173 (20.49%)
nomen	573 (3.17%)	336 (6.72%)	1145 (20.99%)
total	**18103**	**4998**	**5588**

Table 1: Dataset term class amount and relative count.

nestedness	class	character length			word length		
		min	max	mean	min	max	mean
outermost	specific	2	91	17.58	1	13	1.89
	common	2	41	10.98	1	10	1.36
	nomen	3	95	24.06	1	14	3.19
	all	2	95	17.58	1	14	1.84
inner	specific	1	83	13.95	1	13	1.46
	common	2	31	7.54	1	5	1.12
	nomen	3	11	7.11	1	3	1.22
	all	1	83	13.95	1	13	1.32
overall	specific	1	91	17.83	1	13	1.79
	common	2	41	9.47	1	10	1.25
	nomen	3	95	23.79	1	14	3.16
	all	1	95	15.78	1	14	1.69

Table 2: Train dataset terms length.

thus increasing their value ratio in the total score. The second one treats the task of term extracting as term identification task, i.e. measuring the accuracy of term prediction regardless of their class.

The total table of competitors' results was sorted based on the F1 score in the first track and the weighted F1 score in the second and third tracks.

Three types of terms were considered:

- *specific*: terms that are specific both in-domain and lexically;
- *common*: terms that are specific only in-domain (they can be known and used by non-specialists);
- *nomen*: names of unique objects that belong to a specific domain.

Examples for each of three term types are as follows:

(1) *эпистемической* *модальности*
epistemic modality
'Example of specific term'

(2) *пользователю*
user
'Example of common term'

(3) *Национального* *корпуса* *русского* *языка*
National Corpus of Russian Language
'Example of nomen term'

In Table 1, the frequencies of each term type in each track are shown. It can be seen that specific terms are the most frequent in the dataset. Table 2 demonstrate the length distribution for each term type in the training set. We can see that nomen terms are longest (their length is about 3 words on average), and common terms are shortest (1-2 words on average).

level	specific	common	nomen	total
1 (outermost)	9589 (75.72%)	2734 (39.07%)	564 (96.91%)	12887 (55.26%)
2	2690 (21.24%)	1734 (24.78%)	9 (1.55%)	4433 (19.01%)
3	360 (2.84%)	367 (5.24%)	0 (0%)	727 (3.12%)
4	25 (0.20%)	30 (0.43%)	0 (0%)	55 (0.24%)
5	0 (0%)	1 (0.01%)	0 (0%)	1 (0.01%)
total (inner)	3075 (19.54%)	2132 (30.47%)	9 (1.55%)	5216 (22.37%)
total (overall)	15739	6998	582	23319

Table 3: Nestedness count of train dataset

level	specific	common	nomen	total
1 (outermost)	2606 (62.52%)	749 (41.59%)	269 (66.75%)	3624 (56.87%)
2	668 (16.03%)	423 (23.49%)	67 (16.63%)	1158 (18.17%)
3	92 (2.21%)	87 (4.83%)	0 (0%)	179 (2.81%)
4	18 (0.43%)	12 (0.67%)	0 (0%)	30 (0.47%)
5	3 (0.07%)	4 (0.22%)	0 (0%)	7 (0.11%)
total (inner)	781 (18.74%)	526 (29.21%)	67 (16.63%)	1374 (21.56%)
total (overall)	4168	1801	582	6372

Table 4: Nestedness count of dev dataset (track 1 & 2)

In Tables 3, 4, 5, 6, the levels of term nestedness for different sets and tracks are shown. The maximal level of nestedness in term annotations is 5. The fraction of longest (outmost) entities is about 55-56%. That is, nested terms constitute a significant share of annotated terms.

4 RuTermEval Competition Solution

4.1 Nested Named Entity Extraction

We depicted the term extraction task of the RuTermEval competition as a named entity extraction task, since both tasks are sequence labeling tasks. Moreover, the term extraction task could be framed as a named entity extraction task on a specific domain, that is, terms in the RuTermEval dataset. Thus, terms are onwards considered entities throughout this paper.

More specifically, in entity recognition tasks, for a given text sequence $X = \{x_1, x_2, \ldots, x_n\}$, where n is its length, we need to assign each subsequence $E = \{(x_i, x_{i+1}, \ldots, x_j) \mid 1 \leqslant i \leqslant j \leqslant n; i, j, n \in \mathbb{N}\}$ to its corresponding label $y_{i,j} \in Y$ (if any), where Y is a predetermined set of entity classes.

In the usual setting of the "flat" entity recognition task $\forall e_1, e_2 \in E : e_1 \nsubseteq e_2$, i.e., entities in a sequence cannot contain other entities within them. Moreover, $\forall e_1, e_2 \in E : e_1 \cap e_2 = \varnothing, e_1 \neq e_2$ i.e., such entities do not overlap too. We will denote such labeling as F:

$$F = \{e \in E \mid \forall e_1, e_2 : e_1 \cap e_2 = \varnothing\}.$$

On the other hand, in practice, entities can overlap. In this case, we are having nested labeling, i.e. $\exists e_1, e_2 \in E : e_1 \subset e_2$. We will denote such data as N:

$$N = \{e \in E \mid \exists\, e_1, e_2 : e_1 \subset e_2; \forall\, e_3, e_4 : \text{either } e_3 \subset e_4 \text{ or } e_4 \subset e_3 \text{ or } e_3 \cap e_4 = \varnothing\}.$$

Note: $F \subset N$.

The RuTermEval competition is devoted to the recognition of nested terms. Our solution is based on the Binder model, learned on training data with the ruRoberta-large pretrained language model (Zmitrovich et al., 2023).

level	specific	common	nomen	total
1 (outermost)	2404 (58.12%)	705 (44.48%)	940 (66.86%)	4049 (56.81%)
2	722 (17.46%)	373 (23.53%)	215 (15.29%)	1310 (18.38%)
3	131 (3.17%)	57 (3.60%)	15 (1.07%)	203 (2.85%)
4	13 (0.31%)	10 (0.63%)	3 (0.21%)	26 (0.36%)
5	0 (0%)	0 (0%)	0 (0%)	0 (0%)
total (inner)	866 (20.94%)	440 (27.76%)	233 (16.57%)	1539 (21.59%)
total (overall)	4136	1585	1406	7127

Table 5: Nestedness count of dev dataset (track 3)

inner \ outer	specific	common	nomen
specific	3124	3	358
common	1508	773	281
nomen	4	0	5
total	4636	776	644

Table 6: Inner classes of terms inside outermost terms of train dataset

4.2 Binder Model

The main model used in the competition and in this study was the `Binder` model (Zhang et al., 2023). This model allows for extracting entities (named entities, terms) from sentences based on the so-called descriptions of entities and a contrastive learning method.

The input is the sequence of words $x_1, \ldots, x_n$ and the description of the named entity E_k of class k. Both are fed to the so-called encoders, which represent the original text data as vector representations (the so-called embeddings), enriched with contextual information about each word. The most popular encoder is a model such as `BERT` (Kenton and Toutanova, 2019). The `BERT` model also allows for obtaining vector representation of the whole sentence using a special token `[CLS]`.

Each obtained vector is mapped onto a single vector space using a linear layer. In this space, it is assumed that all entities of the same type should be closer to each other than any other entity, as well as any other subsequences of words of the original sequence.

The main idea of the model is so-called contrastive learning. It consists in the fact that the model learns not from the vector sequences of words and their subsequences but learns to bring entities of one type closer to a certain single center of the most characteristic entity of this type and, conversely, to move away all other uncharacteristic subsequences.

At the prediction stage, the model is required to find out where the boundary between positive and negative mentions is drawn. To do this, based on the `[CLS]` vector of the second encoder used to encode the main sequence, its position in the same vector space is trained, and the distance between it and the anchor is the desired radius within which the mentions are positive. This is the so-called dynamic threshold approach, since it is trained together with the rest of the model, and is not fixed in advance.

4.3 Model parameters

This solution was the same for all three tracks. In the first track, we considered all terms mentions as entities of the dummy class *any*, treating the task of term identification as term classification of a single entity type. In tracks 2 and 3, the Binder model was trained to extract three term types.

We trained this model on 128 epochs on GPU with batch size of 8, maximum sequence length of 192, doc_stride of 16, learning rate of $3 * 10^{-5}$ using AdamW optimizer (Loshchilov and Hutter, 2019) with default settings. These parameters were the same for the solutions for the three tracks.

User	Team Name	Scoreboard F1 score
fulstock (ours)	**LAIR RCC MSU**	**0.7940**
VladSemak	VSemak	0.7685
ivan_da_marya	Ivan da Marya	0.5619
ragunna	KiPL SPBU	0.5349
angyling	-	0.5333
VatolinAlexey	ai	0.0000

Table 7: All teams results on the Track 1 of RuTermEval competition.

User	Team Name	Total score	Weighted F1 score	Class-agnostic F1 score
fulstock (ours)	**LAIR RCC MSU**	**0.6997**	**0.6997**	**0.78**
VladSemak	VSemak	0.6996	0.6996	0.77
VatolinAlexey	ai	0.5797	0.5797	0.63
ragunna	KiPL SPBU	0.5043	0.5043	0.52
angyling	-	0.3137	0.3137	0.53

Table 8: All teams results on the Track 2 of RuTermEval competition.

In Tables 7, 8, 9 we show the results of our submissions, compared to results of other teams.

We can see that the approach of nested term recognition based on the Binder model came out first on all three tracks, overtaking other teams submissions. Though, in the second track the next participant got the score almost identical to ours — with negligibly small difference. Thus, we can consider that our method surpasses the others in first and third tracks, but stays on the same line of prediction performance as the second-best method in the second track.

4.4 Error analysis

We perform error analysis on the predictions of the development dataset.

First, we can see that the model confuses common and specific terms, for example:

(4) *оценке* *сочетаемости* *слов*
assessment of compatibility of words

(5) *синтаксического* *словаря*
syntactic dictionary

The model recognized these terms as common, though they were actually labeled as specific. However, the following examples were labeled common, but the model recognized them as specific.

(6) *повтору*
repeat

(7) *модальная*
modal

The confusion of these classes occurs rather frequently. We believe this is due to their definition: model fails to differ when the found term is used by non-specialists and when it is not. We can also see that in the first track there are no such errors.

Second, from our results, we see that most of the model prediction errors were false negative, i.e., the model retrieves less amount of erroneous terms than not retrieves the labeled ones.

Third, we see some labeling ambiguity in the original data. For example,

User	Team Name	Total score	Weighted F1 score	Class-agnostic F1 score
fulstock (ours)	**LAIR RCC MSU**	**0.4823**	**0.4823**	**0.60**
VladSemak	VSemak	0.4654	0.4654	0.51
angyling	-	0.4370	0.44	0.53

Table 9: All teams results on the Track 3 of RuTermEval competition.

(8) *ядро* *системы* *персонализированного* *синтеза* *речи* *по* *тексту*
core of system of personalized synthesis of speech from text

(9) *системы* *персонализированного* *синтеза* *речи*
system of personalized synthesis of speech

(10) *персонализированного* *синтеза* *речи*
personalized synthesis of speech

were all labeled as specific, but

(11) *системы* *персонализированного* *синтеза* *речи* *по* *тексту*
system of personalized synthesis of speech from text

was not. The model recognized all four of these examples as specific terms, but because the last was not labeled, it resulted in a false negative error.

5 Nested Term Recognition from Flat Supervision Task

Currently, most existing datasets labeled with terms have flat annotations. However, we can see that many terms can be nested. Therefore, we study methods for extracting nested terms from flat-term annotations based on the RuTermEval data.

5.1 Task

We consider the problem of extracting nested data given only flat labeling. That is, for training data only flat data F is given, but for validation and test subsets we have full nested data N. This task was proposed for named entities in (Zhu et al., 2022b).

There are different options available for flat data F. For example, if F is generated from nested data N by deleting all overlaps $e_1 \subset e_2, e_1, e_2 \in N$, we can delete only e_1 or e_2 from N to remove this overlap. Thus, we come to many scenarios of flat data.

We consider the ultimate and most common option — "outermost" flat data. By "outermost" we mean entities that can contain other entities, but are not part of any other entity in data, i.e., "outermost" flat data $F = F_u$ is as:

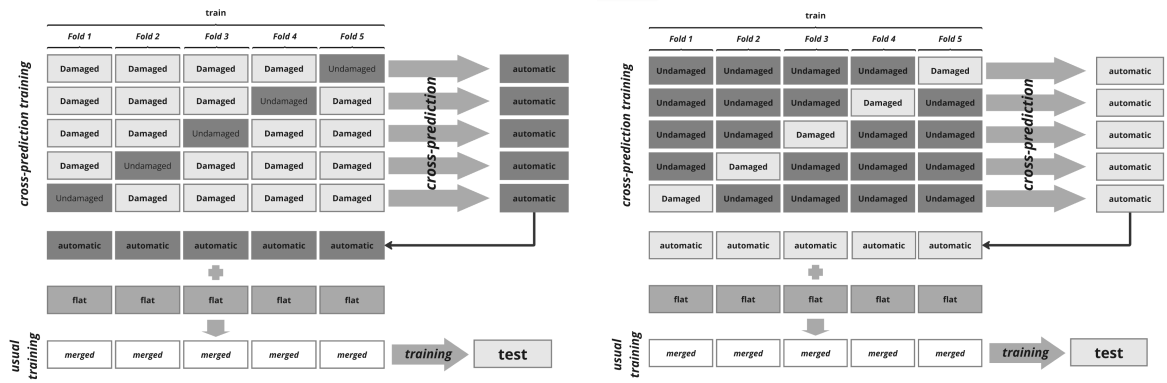
$$F_u = \{e \in N \mid \nexists e' : e \subset e'\}$$

In this work, we propose several methods for the prediction of nested data N given only "outermost" flat data F_u. Furthermore, we denote all remaining nested entities that are not part of outermost flat data as inner entities I: $I = N \setminus F_u$.

5.2 Baselines

Two baselines are used: full nested learning and "pure" flat learning.

Full nested learning is the option to learn a model on all available nested data N to achieve the best results in extracting nested entities as described in Section 4. This baseline would give us the upper limit that should be achieved through flat supervision methods.



(a) ”Early” damage

(b) ”Late” damage

Figure 1: Scheme of damaged cross-prediction method, example of 5 folds.

”Pure” flat learning is the option of learning only on available flat data F without any changes to the learning scheme, model architecture, or data augmentation. This is the simplest approach in flat supervision and is regarded as a bottom baseline.

Both approaches would give us the upper and lower expected limits.

5.3 Automatic pseudo-labeling

In this work, we propose several novel methods grouped by the term ”automatic pseudo-labeling”. All these approaches represent data augmentation techniques that try to leverage the omitted nestedness of the data.

There are four approaches in total, as follows:

- ”Inclusions” (simple and lemmatized);
- ”Damaged cross-validation” (”early” and ”late”).

The explanations are given below.

5.3.1 Inclusions

Consider the flat dataset F_u defined as above. Entities in such dataset can contain inner entities within them; they were either unlabeled or deleted from initial labeling. Since the given text itself was not altered, the entities are still hidden in the text, an we just cannot know where exactly.

In practice, many entities do not appear just once in the text. Each occurrence is called ”mention”, which is widely known in other fields such as entity linking. Furthermore, each entity can have many mentions in the text.

This phenomenon inspired the approach of so-called ”inclusions”. We consider the hypothesis that each mention of the known entities should be tackled as entities too. Inclusions are mentions of known entities inside other ones.

We add such inclusions to the initial flat dataset as a pseudo-labeling, much like data augmentation, and train model on a new dataset. Inclusions are computed only at the training phase.

Another variant of this method additionally exploits the lemmatization of such inclusions. Before we looked for mentions of known entities ”as whole,” i.e., each mention should be written exactly as the original entity itself. But if we split such entities as a list of words, lemmatize and combine them in one unordered set of lemmas, we can extend the definition of inclusions.

From our calculations, there were 1296 simple inclusions (and 14183 entities as a result) and 3681 lemmatized inclusions (and 16568 entities as a result) in the train dataset.

Approaches	dev						test
	F1 micro, %			F1 macro, %			Scoreboard F1 %
	overall	inner	outer	overall	inner	outer	
pure flat	64.73	0.43	71.53	66.23	0.39	73.04	65.10
inclusions	68.28	17.78	**71.65**	67.79	11.74	71.92	67.42
lemm. inclusions	70.51	32.69	71.28	70.47	23.50	72.24	70.26
early damage	69.87	23.76	71.13	72.43	25.65	72.64	71.09
late damage	66.66	10.57	70.06	68.63	10.43	72.31	68.90
lemm. inc. + early dmg	**71.99**	**38.42**	70.52	**74.30**	**36.04**	73.18	**72.81**
lemm. inc. + late dmg	71.16	32.88	71.56	71.91	28.25	**73.26**	71.89
full	79.87	67.73	72.90	81.84	65.20	74.68	79.40

Table 10: Results on track 1.

Approaches	dev						test	
	F1 micro, %			F1 macro, %			w. F1, %	c/a F1, %
	overall	inner	outer	overall	inner	outer		
pure flat	61.01	0.68	67.04	63.61	1.42	69.62	55.04	63.60
inclusions	64.98	17.79	**67.96**	65.50	12.97	69.28	58.64	66.60
lemm. inclusions	66.84	30.80	67.55	67.02	24.06	68.60	59.49	70.21
early damage	65.34	19.34	65.32	66.96	20.86	66.97	60.39	70.03
late damage	63.18	8.63	66.68	64.51	8.68	68.17	58.84	67.29
lemm. inc. + early dmg	**68.64**	**36.40**	66.40	**69.52**	**32.15**	68.20	**63.10**	**73.37**
lemm. inc. + late dmg	67.13	32.97	66.66	68.31	28.07	**69.74**	60.97	71.41
full	76.06	64.76	68.44	77.40	61.29	70.41	69.97	77.79

Table 11: Results on track 2.

5.3.2 Damaged Cross-prediction

Another approach to pseudo-labeling guides the model during training to "look for" more inner entities.

Firstly, flat training data is divided into K folds (K is predefined). These folds are used for cross-validation: $K - 1$ of them are used as a subtraining dataset, while the remaining fold is used as a subdevelopment dataset.

Secondly, we change the given flat data or "damage" it. Only training subset is changed. Here we propose two different methods:

- "Early" damage. We delete labeling of all long flat entities (of length 3 and more) from all subtraining folds, while changing one of the words x_t of such entity to some other character sequence.
- "Late" damage. Same as "early", but we damage the subdevelopment fold, while subtraining folds remain untackled.

From our calculations, out of 12887 of train subset entities, 2053 were damaged, while 10834 others were kept the same.

Thirdly, we train model K times, just like in cross-validation: $K - 1$ subtraining folds are used for training, while the remaining fold is used for testing. For this purpose, in both methods the model tries to predict entities in parts of the initial training dataset. After all K training procedures, we get predictions of the model on all K folds. All these predictions are now considered as additional pseudo-labeling.

Finally, obtained labeling is added to the initial flat training subset. After that, the method stays familiar with usual training and evaluation. Figure 1 depicts the early damage and late damage pseudo-labeling schemes.

Approaches	dev						test	
	F1 micro, %			F1 macro, %			w. F1, %	c/a F1, %
	overall	inner	outer	overall	inner	outer		
pure flat	40.20	1.79	42.36	36.76	1.66	38.67	40.17	51.58
inclusions	43.99	9.99	44.36	39.78	7.78	40.61	42.05	55.54
lemm. inclusions	47.17	18.32	45.10	41.65	14.42	40.80	43.69	57.23
early damage	45.84	13.02	44.44	40.85	9.98	40.68	43.39	55.99
late damage	37.21	2.08	39.26	33.88	1.96	35.95	41.02	54.10
lemm. inc. + early dmg	**48.62**	**20.49**	**46.19**	**43.52**	**15.86**	**42.50**	**45.47**	**58.75**
lemm. inc. + late dmg	46.77	17.50	45.07	41.16	13.79	40.50	43.95	56.94
full	50.35	29.57	46.18	45.16	23.91	42.18	48.23	60.38

Table 12: Results on track 3.

All approaches were conducted in all three tracks, both on development and test subsets. Results are given in 10, 11, 12. In the RuTermEval competition in the second and third tracks two metrics were utilized namely weighted F1 (denoted as w. F1) and class-agnostic F1 (denoted as c/a F1).

First, we can see in all three tracks that pure flat approach extracts almost none of inner entities, while still capable of extracting outermost entities, though underperforming.

Second, as first two tracks consider same data but with different labeling, we see similar results on all approaches. But when task switches from term identification to term classification, all methods perform a bit poorer in general. Moreover, the third track considers the prediction on the different data domain — hence worse results in comparison.

Third, we can see that approaches on development and test sets share similar relative results.

Fourth, we see that all approaches enhance the performance of the inner terms prediction, hence the better result in the competition too.

Fifth, inclusions achieve good boost performance with its pseudo-labeling, from 0.68% to 17.79% inner terms prediction on second track. Moreover, lemmatized inclusions boost the results even more — up to 30.80%.

Sixth, damage cross-prediction methods achieve results around the same level of inclusions. We see that the "late" damage appeared to be poorer in general: model could not really retrieve damaged entities. We believe that this is due change of the context and hence embeddings confusion.

Seventh, merging together results of damaged cross-prediction and lemmatized inclusions, we achieved best inner terms predictions performance, up to 36.40%. On third track, this approach almost achieves the full nested approach, 58.75% to 60.38%. Thus, we see that such simple pseudo-labeling can be both very helpful and viable for recognizing nested terms from flat data.

5.4 Error analysis

As in our RuTermEval solution, we perform the same error analysis of our methods on predictions on the development dataset.

Firstly, we see that the model receives the ambiguity of the predictions because of nestedness. For example,

(12) *лексической* *системы* *языка*
lexical system of language
is labeled as specific and model did not recognize it at all, but

(13) *лексической* *системы*
lexical system
was recognized by model as specific term, though it was not labeled in data.

Secondly, we see another common pattern. Due to much less nested data in all our approaches with pseudo-labeling, model still fails to generalize on inner terms. Nevertheless, with our approach model now does recognize some flat terms that the usual nested approach of our best solution failed to extract instead:

(14) *озимой* *пшеницы*
winter wheat
was extracted correctly in flat approaches, while nested solution did not recognize it (though it was labeled);

(15) *пшеницы*
wheat

(16) *зерна*
grain
were mistakenly extracted as terms by flat approaches, though they were actually not (i.e., not labeled so).

We believe that the model got better trained on longer entities, which resulted in such prediction behaviour.

Still, this hypothesis holds true to erroneous predictions: for longer subsequences where there was no term, the model retrieved them in flat approaches, while the best nested solution did not:

(17) *биологического* *биоцидного* *препарата* *для* *борьбы* *с* *вредными* *членистоногими*
biological biocidal preparation for the control of harmful arthropods

was mistakenly extracted as a term in flat approaches. We see that such errors were the most frequent.

Thirdly, our pseudo-labeling inclusions approach introduced to the trained model that many mentions of some term are always terms too. Thus, there are many mistakes such as:

(18) *почва*
soil
at positions 239, 243; 901, 906; 1265, 1270; 1304, 1309, etc. in text track3-test1-81

(19) *урожайность*
productivity of land
at positions 132, 143; 298, 309; 437, 448; 567, 578; 641, 652, etc. in text track3-test1-81

while nested solution did not produce them.

6 Conclusion

In this work, we present our solution to the RuTermEval competition to all three tracks. Our solution via regarding Nested Term Extraction task as a Nested Named Entity Extraction task appeared to be the most effectively or at least alongside other competitors' solutions — we have got the first place in all three tracks, devoted to different scenarios of term extraction. We used the Binder model, which was previously successfully applied to the recognition of nested named entities, to extract nested terms and obtained the best results of term recognition in all three tracks of the RuTermEval competition.

We describe and motivate a new task — Nested Term Recognition from Flat Supervision. In this task, a model should predict nested terms based only on flat labeling at hand. We propose some approaches to this task and evaluate them in the RuTermEval competition. Of them, the combined lemmatized inclusions with early damage approaches resulted in the best inner term prediction score, coming close to the score on the full nested data.

References

Nikita A Astrakhantsev, Denis G Fedorenko, and D Yu Turdakov. 2015. Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software*, 41:336–349.

Elena I Bolshakova and Natalia E Efremova. 2015. A heuristic strategy for extracting terms from scientific texts. // *Analysis of Images, Social Networks and Texts: 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9–11, 2015, Revised Selected Papers 4*, P 297–307. Springer.

Elena I Bolshakova and Vladislav V Semak. An experimental study on cross-domain transformer-based term recognition for russian.

Elena Bolshakova, Natalia Loukachevitch, and Michael Nokel. 2013. Topic models can improve domain term extraction. // *Advances in Information Retrieval: 35th European Conference on IR Research, ECIR 2013, Moscow, Russia, March 24-27, 2013. Proceedings 35*, P 684–687. Springer.

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, 30:31–40.

Basilis Charalampakis, Dimitris Spathis, Elias Kouslis, and Katia Kermanidis. 2016. A comparison between semi-supervised and supervised text mining techniques on detecting irony in greek political tweets. *Engineering Applications of Artificial Intelligence*, 51:50–57, May.

Amir Hazem, Mérieme Bouhandi, Florian Boudin, and Béatrice Daille. 2022. Cross-lingual and cross-domain transfer learning for automatic term extraction from low resource data. // *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, P 648–662.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. // *Proceedings of naacL-HLT*, volume 1. Minneapolis, Minnesota.

Christian Lang, Lennart Wachowiak, Barbara Heinisch, and Dagmar Gromann. 2021. Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 3607–3620.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. FinBERT: A Pre-trained Financial Language Representation Model for Financial Text Mining. // *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, P 4513–4519, Yokohama, Japan, July. International Joint Conferences on Artificial Intelligence Organization.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization.

Natalia Loukachevitch and Michael Nokel. 2013. An experimental study of term extraction for real information-retrieval thesauri. // *Proceedings of TIA*, P 69–76.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, Igor Rozhkov, Artem Shelmanov, Elena Tutubalina, et al. 2024. Nerel: a russian information extraction dataset with rich annotation for nested entities, relations, and wikidata entity links. *Language Resources and Evaluation*, 58(2):547–583.

Natalia V Loukachevitch. 2012. Automatic term recognition needs multiple evidence. // *LREC*, P 2401–2407.

Malgorzata Marciniak and Agnieszka Mykowiecka. 2015. Nested term recognition driven by word connection strength. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, 21(2):180–204.

Mohamed Nadif and François Role. 2021. Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, 22(2):1592–1603, March.

Ayla Rigouts Terryn, Veronique Hoste, and Els Lefever. 2022. Acter 1.5: Annotated corpora for term extraction research. // *CLARIN Annual Conference Proceedings, 2022*, P 1–4.

Antonio Šajatović, Maja Buljan, Jan Šnajder, and Bojana Dalbelo Bašić. 2019. Evaluating automatic term extraction methods on individual documents. // *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, P 149–154.

Hanh Thi Hong Tran, Matej Martinc, Andraz Repar, Nikola Ljubešić, Antoine Doucet, and Senja Pollak. 2024. Can cross-domain term extraction benefit from cross-lingual transfer and nested term labeling? *Machine Learning*, 113(7):4285–4314.

Chau Vo, Tru Cao, Ngoc Truong, Trung Ngo, and Dai Bui. 2022. Automatic medical term extraction from vietnamese clinical texts. *Terminology*, 28(2):299–327.

Qianqian Xie, Jennifer Amy Bishop, Prayag Tiwari, and Sophia Ananiadou. 2022. Pre-trained language models with domain knowledge for biomedical extractive summarization. *Knowledge-Based Systems*, 252:109460, September.

Sheng Zhang, Hao Cheng, Jianfeng Gao, and Hoifung Poon. 2023. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning, February.

Enwei Zhu, Yiyang Liu, Ming Jin, and Jinpeng Li. 2022a. Recognizing Nested Entities from Flat Supervision: A New NER Subtask, Feasibility and Challenges, November.

Enwei Zhu, Yiyang Liu, Ming Jin, and Jinpeng Li. 2022b. Recognizing nested entities from flat supervision: A new ner subtask, feasibility and challenges.

Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Tatiana Shavrina, Sergey Markov, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian.

Faroese Corpus Development: Strategies for Low-Resource Languages Corpora

Konstantin Satdarov
National Research University Higher School of Economics / Moscow, Russia
kesatdarov@edu.hse.ru

Darya Kharlamova
National Research University Higher School of Economics / Moscow, Russia
dasha.kh18@gmail.com

Andrey Yakuboy
National Research University Higher School of Economics / Moscow, Russia
aiyakuboy@edu.hse.ru

Alisa Lezina
National Research University Higher School of Economics / Moscow, Russia
aalezina@edu.hse.ru

Abstract

This paper presents the development of a comprehensive morphologically annotated corpus for Faroese, a low-resource Germanic language. We describe the creation of a large corpus of contemporary news texts automatically annotated using a custom-trained SpaCy model. The study demonstrates the effectiveness of creating linguistic resources for low-resource languages using minimal initial data. We trained a Transformer-based morphological parsing model on the small but high-quality OFT treebank using 5-fold cross-validation, achieving significant accuracy in morphological tagging and lemmatization. Manual evaluation confirms satisfactory performance of the automatic annotation, though certain challenges remain in distinguishing homonymous word forms across different parts of speech. This research provides a methodological framework for developing comprehensive linguistic resources for other low-resource languages with minimal initial data requirements.

Keywords: Faroese language, morphological annotation, corpus linguistics, low-resource languages, SpaCy, Tsakorpus, corpus development

DOI: 10.28995/2075-7182-2025-23-312-322

Создание фарерского корпуса: подходы для малоресурсных языков

Константин Сатдаров
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
kesatdarov@edu.hse.ru

Дарья Харламова
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
dasha.kh18@gmail.com

Андрей Якубой
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
aiyakuboy@edu.hse.ru

Алиса Лезина
Национальный исследовательский университет «Высшая школа экономики» / Москва, Россия
aalezina@edu.hse.ru

Аннотация

В статье описывается разработка морфологически размеченного корпуса для фарерского языка – малоресурсного германского языка. Описывается создание большого корпуса современных новостных текстов с автоматической разметкой, выполненной с помощью специально обученной модели SpaCy. Мы демонстри-

руем эффективность создания лингвистических ресурсов для малоресурсных языков при минимальном количестве начальных данных. На основе небольшого, но качественно размеченного набора данных OFT была обучена морфологическая модель на базе архитектуры Transformer с использованием кросс-валидации, что позволило достичь высокой точности в морфологической разметке и лемматизации. Ручная оценка подтверждает удовлетворительное качество автоматической разметки, хотя остаются определенные трудности в различении омонимичных словоформ, относящихся к разным частям речи. Данное исследование предлагает методологическую основу для разработки лингвистических ресурсов для других малоресурсных языков при минимальных требованиях к количеству начальных данных.

Ключевые слова: фарерский язык, морфологическая разметка, корпусная лингвистика, малоресурсные языки, SpaCy, Tsakorpus, разработка корпусов

1 Introduction

It is our deep belief that everyone in the world of linguistic studies would benefit if it were easy to develop and deploy automated corpora for minor and endangered languages. This would give theory-grounded researchers (e.g. typologists) more easily accessible data and statistics-based methodology to back up their claims, and the practise-oriented developers would face less difficulties developing linguistic resources (like parsers, automated translation systems, etc.) for the wide support of minor languages. This would, in turn, lead to more chances of preserving endangered languages. However, even given the modern automatisation tools, it is still challenging to create a voluminous and reliable corpus without pouring countless money and efforts into manual annotation. This slows down the industry significantly.

Despite being the primary language of the Faroe Islands, Faroese remains among the low-resource languages in terms of available linguistic data and tools, save for some. This is the main reason why we chose this language as a test subject to develop a new corpus encompassing both historical and modern texts and an automated UD-compliant markup. The goal of this research is to create such a corpus working from the ground up, using as little resources as possible, as if Faroese was a minor language mainly studied through field research and sporadic Internet presence. If this proves to be successful, we will know the exact number of resources one needs to build a reliable base corpus with a universally acknowledged markup system for almost any language.

2 Related work

Currently, there are three major resources for the Faroese language available:

1. OFT [16] is based on texts from the Faroese Wikipedia. The corpus contains approximately 10,000 tokens across 1,208 sentences, with rich morphological annotation. Trosterud's finite-state morphological analyzer and constraint grammar [15] was used to markup the majority of the data, followed by manual verification of the morphological features and addition of dependency annotations according to Universal Dependencies 2.0 guidelines. The morphological tags were then automatically converted to UD-compatible features using a lookup table and set overlap procedure. While the treebank provides high-quality morphological information, its limited size and reliance on Wikipedia as the sole source results in certain limitations, such as an abundance of copular constructions and minimal representation of first and second person forms.
2. FarPaHC [2], a corpus of 40.000 tokens in size, was created through the automatic conversion of the Faroese Parsed Historical Corpus [8] from the Penn Treebank format. While the treebank provides broader coverage than OFT, the automatic conversion process has resulted in less detailed morphological annotations. Notably, the absence of lemma information in FarPaHC limits its utility for training lemmatization models, a crucial component for creating searchable corpora.
3. The news text corpus integrated into the GiellaLT linguistic infrastructure [17]. This corpus consists of approximately 25 million tokens with automatic morphological annotation performed using the system's finite-state grammar. The corpus is built on top of the Korp corpus engine [4] and includes a collection of news texts that have been automatically processed and annotated. However, several limitations affect its utility: the interface may present challenges

for researchers, the grammar-based disambiguation system is still in early development stages, and copyright restrictions prevent full corpus downloads, allowing only partial access to the dataset.

4. Regarding available morphological parsing tools, both Stanza [12] and UDPipe2 [14] offer neural processing pipelines for Faroese, trained on the FarPaHC treebank. However, these tools are limited by their training data: the absence of lemmatization capabilities and rather poor morphological tagset.
5. The GiellaLT system also provides a rule-based finite-state morphological and syntactic parser [15], which consists of a morphological analyzer based on a comprehensive lexicon and morphophonological rules, a constraint grammar disambiguator, and a dependency parser. The morphological component excels in handling regular inflectional patterns and compounds, while the constraint grammar manages morphological disambiguation, especially within noun phrases and prepositional constructions. However, this approach struggles with novel vocabulary, such as loanwords, and relies on a proprietary tagset that needs conversion for integration with multilingual pipelines. The system also shows limitations in handling irregular verbs and certain adjectival forms, including comparatives and superlatives. Moreover, its disambiguation component is still in early development, affecting analysis reliability.

It can be seen that there were no systematic attempts at training a Transformer-based morphological tagger, so our research will cover this aspect, employing the novel neural network along with the usage of SpaCy [7], a powerful tool in NLP that is capable of processing minor languages (to show that, there are official guides on how to train a custom pipeline applicable to low-resource languages). In the current research, we mainly used the OFT treebank and FarPaHC treebank — the former as the source of training data, and the latter as the reference for historical subcorpus markup.

3 Data preprocessing

We decided to use Faroese news websites as the source of our data since they were the most readily available source of contemporary Faroese texts online. The distribution of texts from different sources was as follows in the Table 1:

Website	n, texts
dagur.fo	24796
nordlysid.fo	1246
foroya-landsstyri.fo	4195
in.fo	315

Table 1: Sources and the number of texts gathered from them.

After preprocessing, which included the removal of non-Faroese texts (e.g. news articles with English or Danish translations), the combined total of texts from these sources amounted to 30732. For every entry we also have gathered the following metadata: date of publication, heading, link, author and tags.

Concerning the tags, we faced a problem: although a few sources were used for data retrieval, each of them contained specific tags that were close in meaning, but not interchangeable. Since we wanted to preserve search by tag, we manually unified the closest tags into broader categories.

For instance, Annar ítróttur (‘Other sports’), Fótbóltur (‘Footbal’l), Hondbóltur (‘Handball’), Ítróttur (‘Sports’) were all merged into Ítróttur. On a governmental site foroyalandsstyri.fo tags were representing different ministries (e.g. Umhvørvismálaráðið - Ministry of the Environment, Heilsumálaráðið - Ministry of Health, etc.), so we merged them into Landsstýristíðindi (Government News). However, even though we have merged tags into broader categories, we have preserved the source tags as

searchable secondary tags. As a result, the end corpus supports 8 primary tags and 23 secondary tags. The markup process is illustrated in Figure 1: the secondary tags highlighted with the same color were unified under one primary tag.

Almanna- og mentamálaráðið (Ministry of Social Affairs and Culture)	Landsstýristíðindi (Government News)	Annar ítróttur (Other sports)	Ítróttur (Sports)
Barna- og útbúgvingarmálaráðið (Ministry of Children and Education)		Fótbóltur (Football)	
Fíggjarmálaráðið (Ministry of Finance)		Hondbóltur (Handball)	
Fiskivinnu- og samferðslumálaráðið (Ministry of Fisheries and Transport)		Ítróttur (Sports)	
Heilsumálaráðið (Ministry of Health)		Andlát (Death)	Annað (Other)
Løgmálaráðið (Ministry of Justice)		Annað (Other)	
Umhvørvismálaráðið (Ministry of Environment)		Einsamallur um Atlantshavið í førc (Alone across the Atlantic in a Far	
Uttanríkis- og vinnumálaráðið (Ministry of Foreign Affairs and Labor)		Ferðavinna (Tourism)	
Bilar (Cars)	Vinna (Industry)	Fiskivinna (Fisheries)	
Byggivinnan (Construction Industry)		Lesarabrøv (Letters from readers)	
Innanlands (Domestic)		Lýsingagrein (Advertisements)	
Orkuvinna (Energy Industry)		Merkisdagar (Anniversaries)	
Vinna (Industry)		Minningarorð (Memories)	
Mentan (Culture)	Mentan (Culture)	Uttanlands (Abroad)	Uttanlands (Abroad)

Figure 1: The manual process of tags unification.

4 Models

To expedite the speed of the manual annotation, we trained a Transformer-based morphological parsing model that is capable of performing UD markup of the Faroese texts using a guide provided by SpaCy and recommended for use when working with low-resource languages.

As we did not have enough resources to collect a sufficiently large and consistent corpus that could be used to train a model, we resorted to using OFT Faroese UD treebank. While the dataset was developed as a data collection aimed at testing rather than training, we used it due to the fact that it was a small, but curated and manually checked data collection. This made it the perfect choice for proving the possibility of training an automated SpaCy model on a small amount of data. In the dataset that we chose, there were only 1208 sentences, making it a very small collection.

Due to the size of the data, we trained the model using 5-fold cross-validation [10] to prove the validity of the metrics. Before we started training, we prepared the data as follows: we united every five sentences together and used them as separate data points so that the model could learn how to handle cross-sentence barriers. After that, we separated out the test dataset (73 texts, each containing 5 sentences from the original dataset).

Overall, we performed three experiments. For the first one, we used a default SpaCy-configuration file that was obtained through the official SpaCy config generation utility. We set up the training pipeline to contain the following in-built SpaCy components: "tok2vec", "tagger", "morphologizer", "trainable_lemmatizer", "parser". Here we did not use Transformer, because we wanted to understand the impact this computationally costly component has on the morphological tagging - a task that was known to be solved effectively without the use of neural networks by previous works (see Related Works above). We arrived at the following results for the dev dataset and the test dataset (see Table 2).

Fold number	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	dep_las	sents_f
1	0.77982	0.9263	0.90896	0.82514	0.83162	0.74274	0.99099
2	0.81331	0.9442	0.91449	0.83158	0.81443	0.73625	0.98235
3	0.81328	0.92973	0.90015	0.84214	0.83289	0.75165	0.98246
4	0.81188	0.94621	0.91675	0.82591	0.82301	0.74226	0.96512
5	0.82651	0.94319	0.91961	0.83601	0.84533	0.77172	0.94737
Mean	0.80896	0.937926	0,911992	0.832156	0.829456	0.748924	0.973658
Test	0.8190	0.9396	0.9105	0.8296	0.8207	0.7294	0.9699

Table 2: The results for the heuristics-based SpaCy pipeline

Having obtained the baseline, we performed the second experiment once again using pre-generated SpaCy config and the following pipeline: "transformer", "trainable_lemmatizer", "tagger", "morphologizer", "parser". Bert-base-multilingual-uncased [6] was used as the baseline Transformer architecture. We once again performed the same 5-fold cross-validation and arrived at the following results (see Table 3).

Fold number	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	dep_las	sents_f
1	0.83065	0.96468	0.94737	0.85914	0.8596	0.796	0.96491
2	0.82105	0.95102	0.93878	0.84558	0.85297	0.78229	0.94461
3	0.84253	0.96939	0.95044	0.8812	0.88889	0.82317	0.98817
4	0.82692	0.97584	0.95492	0.8864	0.89864	0.8356	1
5	0.84552	0.9639	0.95523	0.88592	0.88654	0.80779	0.97281
Mean	0.833334	0.964966	0.949348	0.871648	0.877328	0.80897	0.9741
Test	0.8398	0.9657	0.9508	0.8808	0.8918	0.8135	0.9591

Table 3: The results for the Transformer-based SpaCy pipeline

For the final experiment, we performed a bayes-based hyperparameter optimization sweep using the WandB library [3]. To maximize the morph_micro_f metric that reflects the quality of intra-tag F-score, we tried 6 different combinations of hyperparameters. See the different hyperparameter combinations that were explored, results and the reported target metric in Tables 4 and 5 along with Figures 2 and 3 depicting the target metric alignment with different parameter combinations and the hyperparameter importances and correlations with the target metric.

Figure 2 showcases the importance of every metric and its' correlation with the result. The metric's importance was received through WandB logging system. It shows the degree of usefulness of each hyperparameter in predicting the end result (namely, micro_f_score). Correlation with the result indicates the presence of a link between the value of a given parameter and the prediction. The correlation is linear and ranges from -1 to 1, with -1 reflecting negative correlation, 1 signifying positive correlation, and 0 indicating the absence of correlation.

Number of combination/ hyperparameters	components.-morphologizer.label_smoothing	components.tagger.label_smoothing	components.trainable_lemmatizer.-top_k	components.trainable_lemmatizer.-top_k.min_tree_freq	training.-dropout	training.optimizer.learn_rate
1	0.2	0.1	2	3	0.028	0.010
2	0	0.2	5	5	0.195	0.005
3	0	0.1	5	3	0.181	0.002
4	0.2	0.1	3	5	0.263	0.008
5	0.2	0	3	2	0.120	0.004
6	0	0.2	3	5	0.454	0.004

Table 4: The hyperparameter combinations that were explored

Combination	lemma_acc	tag_acc	pos_acc	morph_acc	dep_uas	sents_f
1	0.81004	0.90005	0.8775	0.81394	0.76503	0.96266
2	0.79365	0.91235	0.87955	0.81394	0.74348	0.95
3	0.80594	0.91645	0.8816	0.81292	0.78351	0.92683
4	0.78751	0.91235	0.85392	0.80574	0.75715	0.9465
5	0.82232	0.91645	0.8775	0.82727	0.78957	0.93061
6	0.79263	0.90313	0.88365	0.83137	0.80146	0.975

Table 5: The metrics for different hyperparameter combination

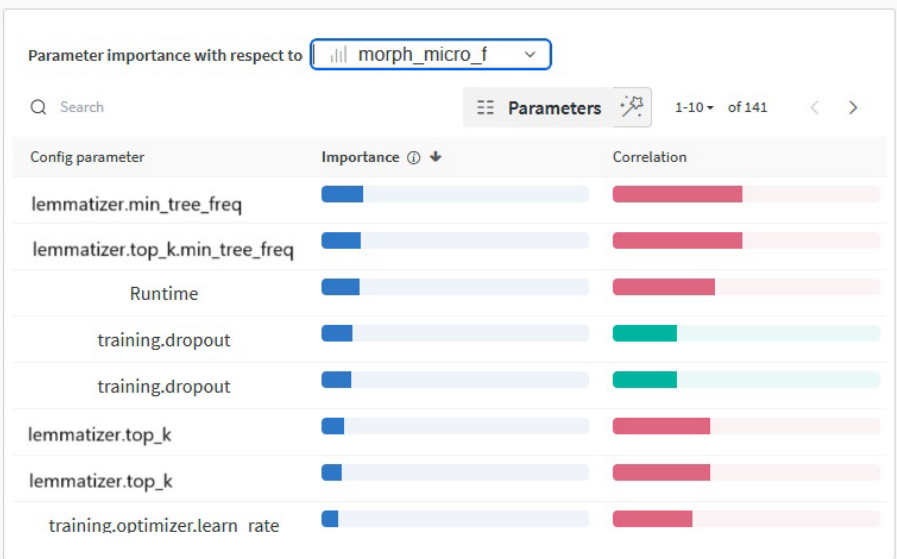


Figure 2: Hyperparameter importance for morph_micro_f.

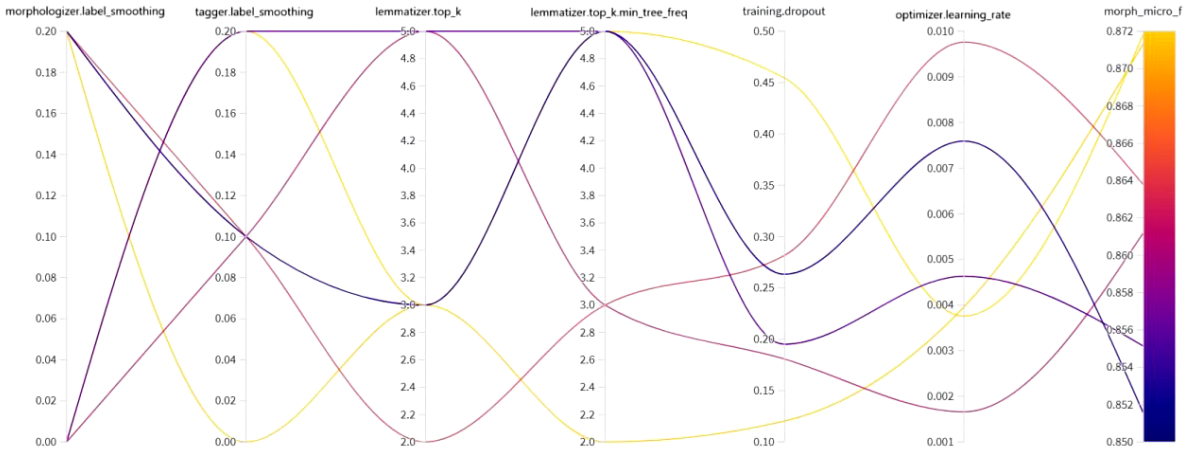


Figure 3: Hyperparameters alignment with different target metric scores.

5 Corpora

Several open-source corpus engines are available for creating searchable corpora from morphologically annotated texts, including NoSketchEngine [13; 9], ANNIS [11], Korp [4], and Tsakorpus [1]. However, none of these platforms natively support corpus creation from CONLL-U files. For our Faroese corpora, we selected Tsakorpus due to the platform's modular design and ElasticSearch configurations reducing resource needs while sustaining comparable performance to alternatives. Its data model, which stores annotated texts in JSON files, accommodates any morphological tagset and metadata schema, making it suitable for conversion from various annotation formats.

Our resource consists of two independent corpora: a large corpus of news texts (approximately 9 million tokens) automatically annotated using our SpaCy model, and a smaller corpus of biblical texts (about 3,000 tokens) with manual morphological annotation. Both corpora were initially annotated in CONLL-U format and converted to Tsakorpus-compatible JSON format using a custom-built universal converter. This converter can process both single CONLL-U files, automatically detecting and separating document boundaries, and directories containing multiple CONLL-U files, with one file per document. Document-level metadata can be extracted either directly from the dataset or from external files. The converter processes morphological annotation by extracting word forms, lemmas, Universal POS tags, and grammatical features, while syntactic markup is omitted. The converter's language-independent design makes it suitable for potential integration into the Tsakorpus platform to facilitate the creation of searchable corpora from other Universal Dependencies treebanks.

We optimized the corpus interface for researchers' usability. One of the most convenient features is the grammar selection table. It allows users unfamiliar with UD tags to select them from categorized lists organized by linguistic categories such as parts of speech, case, number, tense and more. The search functionality includes queries by word form, lemma, grammatical features, multi-word and negative queries, and full-text search. Users can create subcorpora based on metadata: the news corpus supports filtering by source, publication year range, and primary or secondary tags, while the biblical corpus enables chapter-based selection. The platform also provides statistical analysis tools for examining subcorpora.

Both corpora are freely available at www.farcorpus.fikl.ru. The interface is currently available only in English, but the platform's localization mechanisms allow for addition of languages natively supported by Tsakorpus, such as Russian, French, German, and Hebrew. Adding other interface languages would require more effort but remains feasible through the platform's built-in localization framework.

6 Manual quality evaluation

As stated above, the corpus currently includes two types of annotation: manual and automatic. Apparently, the former is seen as an example on which we can rely in order to evaluate the quality of the automatic one. The reason for this is that detailed manual analysis allows to create almost flawless annotation meeting the demands.

Currently, the most concerning weakness is that it is still prone to confuse parts of speech such as nouns, verbs, adjectives and pronouns. However, given high metrics pertaining to the morphological tagging (pos_acc, morph_acc in Table 3) this is not strongly observed on the training data. Thus, we will be able to understand the scale of this issue no sooner than we obtain a bigger testing dataset, the works for developing which are already underway.

This can be explained by the fact that many word-forms of different parts of speech in Faroese have the same endings, which renders the analysis more elaborate. For instance, the -ur ending is both a nominal ending (Nom. Sg., for instance) and a verbal one (3Sg). Thus, though we say that lemmatization and morphological description are fairly good, there are still data that need clarifying and checking. Indeclinable lexemes and frequently attested word-forms have proven to be annotated accurately.

Despite the fact that automatic annotation still needs improvement, we can still claim that the overall performance of the model seems satisfactory. For some sentences, the model can provide accurate morphological descriptions and lemmas of word-forms. The following figure exemplifies an accurately lemmatized sentence, and we agree with the morphological description it provides.

1	Og	og	CCONJ	CC	_	3	cc	_	_	
2	í	í	ADP	Pr	_	3	case	_	_	
3	stundini	stund	NOUN	N	Case=Dat|Definite=Def|Gender=Fem|Number=Sing	4	obl	_	_	
4	fór	fara	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Past	0	ROOT	_	_	
5	spitalskusjúkan	spitalskusjúka	NOUN	N	Case=Nom|Definite=Def|Gender=Fem|Number=Sing	4	nsubj	_	_	
6	av	av	ADP	Pr	_	7	case	_	_	
7	honum	honum	PRON	Pron	Case=Dat|Gender=Masc|Number=Sing|Person=3|PronType=Prs	4	obl	_	SpaceAfter=No	
8	,	,	PUNCT	CLB	_	11	punct	_	_	
9	og	og	CCONJ	CC	_	11	cc	_	_	
10	hann	hann	PRON	Pron	Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs	11	nsubj	_	_	
11	varð	verða	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Past	4	conj	_	_	
12	reinur	reinur	ADJ	A	Case=Nom|Definite=Ind|Gender=Masc|Number=Sing	11	xcomp	_	SpaceAfter=No	
13	.	.	PUNCT	CLB	_	11	punct	_	SpaceAfter=No	

Figure 4: An accurately lemmatized sentence.

The next figure illustrates a sentence which should be further corrected in terms of lemmatization and morphological description.

1	Og	og	CCONJ	CC	_	3	cc	_	_	
2	tá	tá	ADV	Adv	_	3	advmod	_	_	
3	ið	ið	SCONJ	CS	_	0	ROOT	_	_	
4	Jesus	Jesus	PROPN	N	Case=Nom|Gender=Masc|Number=Sing	6	nsubj	_	_	
5	sá	sá	ADV	V	_	6	advmod	_	_	(VERB)
6	trúgv	trúgvur	ADJ	A	Case=Nom|Definite=Ind|Gender=Fem|Number=Sing	3	dep	_	_	(NOUN)
7	teirra	teirra	PRON	Pron	Case=Acc|Number=Sing|Person=3|PronType=Prs	3	obj	_	SpaceAfter=No	
8	,	,	PUNCT	CLB	_	22	punct	_	GEN	
9	sigur	siga	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres	22	nsubj	_	_	
10	hann	hann	PRON	Pron	Case=Nom|Gender=Masc|Number=Sing|Person=3|PronType=Prs	9	nsubj	_	_	
11	við	við	ADP	Pr	_	13	case	MASC		
12	hin	hin	DET	Det	Case=Acc|Gender=Fem|Number=Sing	13	det	_	_	
13	giktsjúka	giktsjúki	NOUN	N	Case=Acc|Definite=Ind|Gender=Fem|Number=Sing	9	obl	_	SpaceAfter=No	(ADJ)
14	:	:	PUNCT	CLB	_	19	punct	_	_	
15	«	«	PUNCT	PUNCT	_	16	punct	_	SpaceAfter=No	
16	Sonur	Sonur	NOUN	N	Case=Nom|Definite=Ind|Gender=Masc|Number=Sing	19	dep	_	_	
17	mín	mín	PRON	Pron	Case=Gen|Number=Sing|Person=1|PronType=Prs	16	punct	_	SpaceAfter=No	(DET)
18	,	,	PUNCT	CLB	_	19	punct	_	_	
19	syndir	synda	VERB	V	Mood=Ind|Number=Sing|Person=3|Tense=Pres	20	amod	_	_	(NOUN)
20	tínar	tínar	PRON	N	Case=Nom|Number=Plur|Person=3|PronType=Prs	9	nsubj	_	_	(DET — tín)
21	eru	vera	AUX	V	Mood=Ind|Number=Plur|Tense=Pres	22	cop	_	_	
22	fyrigivnar	fyrigivnur	ADJ	A	Case=Nom|Definite=Ind|Gender=Masc|Number=Plur	3	dep	_	SpaceAfter=No	
23	!	!	PUNCT	CLB	_	22	punct	_	SpaceAfter=No	(VERB; VerbForm=Part)
24	»	»	PUNCT	PUNCT	_	0	ROOT	_	SpaceAfter=No	

Figure 5: A sentence with suggested corrections.

7 Discussion of the results

Over the course of this research, we managed to achieve remarkable results in every sphere of corpus creation. Let us discuss our findings and the outcomes of the research.

First of all, we have shown that even though Faroese tends to be viewed as a low-resource language, obtaining a significant number of texts from the open access webpages is more than feasible. This method of corpus creation can be viewed as viable, providing an insight into contemporary speech used in different social situations. This method could be used to create more robust corpora for minor languages, because the texts there are not solicited by the researchers.

Second of all, we have created a Transformer-based model capable of reliably processing faroese texts and providing full UD-compliant annotations while using a small corpus of high-quality data. We have performed three experiments. The first experiment showed that the effective high-quality morphological markup was achievable, even given the limited, but high-quality dataset. It also helped point out the areas where the model was the weakest: namely, dep_las metrics. The second experiment demonstrated the importance of using Transformer-based architecture in training language processing pipelines. As shown by the data in Tables 1-4, the introduction of Transformer allowed to significantly improve the scores for the majority of the metrics. As we employed cross-validation in both experiments, we have demonstrated the reproducibility of our results. As for the last experiment, we have shown that the explored hyperparameters do not impact the results significantly. However, the tuning still led to some improvements in the test metrics.

Third of all, we have managed to create a usable corpus that was adequately processed, using only a small and easily achievable number of resources. We have also chosen a corpus engine easy to deploy and maintain that has a rich search and analysis capabilities and rich functionality. Overall, this research shows that it is indeed possible to create a corpus and a model with a very small number of resources. To us, this opens a lot of possibilities for minor language researchers, because, as shown by the authors of [16] and confirmed by the present research, an adequate language dataset can be collected by using an automated model and correcting the errors there. As for the base number of sentences required to train such a model, we have demonstrated that the default SpaCy pipeline can be trained on approximately 1200 sentences - a moderate dataset that can be collected over the course of multiple field researches. Thus, it stands to reason that using such a pipeline a lot of minor languages can be processed and the data on them made publicly available for the advance of the linguistic studies.

However, there are some limitations that need to be addressed in the subsequent research. First, due to the differences in markup systems, we could not directly compare the quality of our model against the existing solutions, like [15]. This issue can be addressed by converting the tags used in each system and using the conversions for meaningful comparative analysis. Second, as we used the treebank initially aimed at testing, we would like to change this as soon as possible. At the moment, we are working to create a new small Faroese training corpus of a similar size that will be used to retrain the model from scratch. This will allow to use the OFT treeback for its intended purpose of benchmarking.

Lastly, while SpaCy provided an effective framework for our research due to its ease of setup, moderate computational requirements, and streamlined deployment capabilities, several alternative approaches merit consideration for future work. The selection of SpaCy allowed us to focus on data collection and annotation quality rather than complex infrastructure development, which was crucial given our limited initial resources. However, more sophisticated frameworks such as Stanza [12] and UDPipe2 [14] have demonstrated superior performance in morphological and syntactic analysis across numerous languages and could potentially yield improved results for Faroese with appropriate training data. Additionally, recent advances in transformer architectures such as XLM-RoBERTa [5] offer promising alternatives that could enhance tagging accuracy through their cross-lingual transfer capabilities.

We did not pursue these approaches at this stage due to the substantial computational resources required for fine-tuning such models and because current multilingual models lack explicit Faroese language data.

An additional limitation that we would like to address is the possibility of using data from the closely related languages. The Faroese language belongs to the North Germanic language branch. However, unlike the Mainland Scandinavian languages, i.e. Danish, Swedish, and Norwegian, Faroese and Icelandic are characterised by arguably complex morphology; therefore, the range of the donor languages is limited to Icelandic. Nonetheless, despite being partially intelligible in writing, they also exhibit morphological differences. For instance, the Faroese verbs do not conjugate for person in plural, whereas the Icelandic verbs do. Another limitation concerns morphonological alternations, which appear to be different in both languages.

We should also briefly mention that we have endeavoured to work with the language data as if we did not have other opportunities and sources. Our work showcases what can be done supposing any limitations arise during the creation of a corpus for a low-resource language.

8 Conclusion

Over the course of this work, we managed to create a Transformer-based model for automated UD-compliant Faroese texts markup and developed a corpus of modern and historical Faroese texts using as little pre-existing resources as possible. This research stands to confirm that given modern linguistic technologies and language learning models and pipelines capabilities, it is possible to create a full-fledged corpus and a model without manually annotating copious amounts of language data.

It is possible to bring the research even further and explore the possibilities of data augmentation strategies through the integration of rule-based morphological parsers or machine-readable dictionaries. Such approaches could potentially provide a cost-effective method for expanding the training data volume and, consequently, enhancing model performance. Additionally, we could explore the options of implementing syntactic annotation into the existing corpus. While the development of such corpora presents additional complexities, such resources could substantially advance Faroese language research.

References

[1] Arkhangelskiy Timofey A. Electronic corpora of the albanian, kalmyk, lezgian, and ossetic languages // Automatic Documentation and Mathematical Linguistics. — 2012. — Vol. 46. — P. 118–123.

[2] Arnardóttir Þórunn, Hafsteinsson Hinrik, Sigurðsson Einar Freyr, Bjarnadóttir Kristín, Ingason Anton Karl, Jónsdóttir Hildur, Steingrímsson Steinþór. A Universal Dependencies Conversion Pipeline for a Penn-format Constituency Treebank // Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020). — Barcelona, Spain, 2020. — P. 16–25.

[3] Biewald Lukas. Experiment Tracking with Weights and Biases. — 2020. — Access mode: https://www.wandb.com/

[4] Borin Lars, Forsberg Markus, Roxendal Johan. Korp – the corpus infrastructure of Språkbanken // Proceedings of LREC 2012. — Istanbul: ELRA, 2012. — P. 474–478.

[5] Conneau A. et al. Unsupervised cross-lingual representation learning at scale //arXiv preprint arXiv:1911.02116. – 2019.

[6] Devlin Jacob, Chang Ming-Wei, Lee Kenton, Toutanova Kristina. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // CoRR. — 2018. — Vol. abs/1810.04805. — Access mode: http://arxiv.org/abs/1810.04805

[7] Honnibal Matthew, Montani Ines, Van Landeghem Sofie, Boyd Adriane. spaCy: Industrial-strength natural language processing in python. — Zenodo, Honolulu, HI, USA, 2020.

[8] Ingason Anton Karl, Rögnvaldsson Eiríkur, Sigurðsson Einar Freyr, Wallenberg Joel C. Faroese Parsed Historical Corpus (FarPaHC) 0.1. — CLARIN-IS, 2012.

[9] Kilgarriff Adam, Baisa Vít, Bušta Jan, Jakubíček Miloš, Kovář Vojtěch, Michelfeit Jan, Rychlý Pavel, Suchomel Vít. The Sketch Engine: ten years on // Lexicography. — 2014. — Vol. 1, No. 1. — P. 7–36.

[10] Kohavi Ron. A study of cross-validation and bootstrap for accuracy estimation and model selection // Proceedings of the 14th International Joint Conference on Artificial Intelligence. — Montreal, Canada, 1995. — Vol. 14, No. 2. — P. 1137–1145.

[11] Krause Thomas, Zeldes Amir. ANNIS3: A new architecture for generic corpus query and visualization // Digital Scholarship in the Humanities. — 2016. — Vol. 31, No. 1. — P. 118–139.

[12] Qi Peng, Zhang Yuhao, Zhang Yuhui, Bolton Jason, Manning Christopher 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. — Online, 2020. — P. 101–108.

[13] Rychlý Pavel. Manatee/Bonito-A Modular Corpus Manager // RASLAN. — 2007. — Vol. 2007. — P. 65–70.

[14] Straka Milan. UDPipe 2.0 Prototype at CoNLL 2018 UD Shared Task // Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies. — Brussels, Belgium, 2018. — P. 197–207.

[15] Trosterud Trond. A constraint grammar for Faroese // Constraint Grammar and robust parsing: Proceedings of the NODALIDA 2009 workshop. — 2009. — Vol. 8. — P. 1–7.

[16] Tyers Francis M., Sheyanova Mariya, Martynova Alexandra, Stepachev Pavel, Vinogradovsky Konstantin. Multi-source synthetic treebank creation for improved cross-lingual dependency parsing // Proceedings of the Second Workshop on Universal Dependencies (UDW 2018). — 2018. — P. 144–150.

[17] Wiechetek Linda, Hiovain-Asikainen Katri, Mikkelsen Inga Lill Sigga, Moshagen Sjur, Pirinen Flammie, Trosterud Trond, Gaup Børre. Unmasking the Myth of Effortless Big Data - Making an Open Source Multilingual Infrastructure and Building Language Resources from Scratch // Proceedings of the Thirteenth Language Resources and Evaluation Conference. — Marseille, France, 2022. — P. 1167–1177.

Comparing Transformer-Based Approaches for Term Recognition in Russian texts

Semak V. V.
Lomonosov Moscow State University
Moscow, Russia
vlad.semakk@gmail.com

Bolshakova E. I.
Lomonosov Moscow State University
HSE, Moscow, Russia
eibolshakova@gmail.com

Abstract

The paper describes an experimental comparative study of three transformer-based approaches applied for automatic term recognition (ATR) in Russian texts. The approaches include sequential labeling of word tokens in a given text, phrase classification with context sentence enclosing the phrase, and text span prediction by using vector representations obtained by contrastive learning. The BERT-based models were trained and evaluated on the data of RuTermEval-2024 competition for nested term identification and classification, which encompassed three tasks: binary term identification, term recognition with classification (into one of predefined types), and cross-domain term recognition. The experiments have shown that the span prediction models based on contrastive learning outperform the other models across all three RuTermEval tasks, but at the same time demonstrate the most significant decrease in quality in the cross-domain task.

Keywords: automatic term recognition; ATR; nested terms; transformer-based term recognition models; RuTermEval competition

DOI: 10.28995/2075-7182-2025-23-323-330

Сравнение подходов на основе трансформеров для распознавания терминов в русскоязычных текстах

Семак В. В.
МГУ имени М. В. Ломоносова
Москва, Россия
vlad.semakk@gmail.com

Большакова Е. И.
МГУ имени М. В. Ломоносова
НИУ ВШЭ, Москва, Россия
eibolshakova@gmail.com

Аннотация

В статье описывается экспериментальное сравнительное исследование трех подходов на основе трансформеров, примененных для автоматического распознавания терминов в русскоязычных текстах. Подходы включают последовательную разметку слов-токенов в заданном тексте, классификацию фраз с учетом включающего фразу контекстного предложения, и предсказание спанов (отрезков) текста с использованием векторных представлений, полученных методом контрастивного обучения. Модели на основе архитектуры BERT были обучены и протестированы на данных соревнования RuTermEval-2024 по идентификации и классификации вложенных терминов, которое охватывало три задачи: бинарную идентификацию терминов, распознавание терминов с классификацией (на один из предопределенных типов) и кросс-доменное распознавание терминов. Эксперименты показали, что модели классификации спанов, основанные на контрастивном обучении, превосходят другие модели на всех трех задачах RuTermEval, но в то же время демонстрируют наиболее значительное снижение качества в кросс-доменной задаче.

Ключевые слова: автоматическое распознавание терминов; вложенные термины; модели извлечения терминов на основе трансформеров; соревнование RuTermEval

1 Introduction

Automatic term recognition (ATR), or automatic term extraction [5, 10, 14] is a natural language processing task aiming to recognize in texts and extract terms — words and phrases denoting concepts of specialized problem domains (e.g., *refactoring*, *legacy code, transfer learning, surface tension force*). With rapid development of scientific domains and emergence of new terms, ATR became essential for such applications as automated construction of terminology dictionaries, thesauri, glossaries and subject indexes, as well as machine translation of scientific texts. Despite being studied for some decades, ATR remains a challenging task, since proposed methods still can not achieve human-like performance (which is explained by complex semantic nature of terminology and high variability of term features across different domains).

The traditional and well-studied statistical approach to ATR [5, 14] relies on extraction from text phrases matching some grammatical patterns (such as *ADJECTIVE+NOUN* for terms like *linear function* and *magnetic field*) and then ranging the obtained list of extracted term-candidates by frequency-based statistical measures (e.g., MI or C-value) in order to obtain true terms on the top of the ranked list. Such techniques is sense-agnostic and give only 30–60% average precision for extracted terms, requiring human efforts to select terms from ranked term-candidates list or additional domain-specific heuristics for ranking and selecting terms.

Machine learning (ML) for automatic term extraction began to be applied since the 2000s, leveraging traditional supervised methods (like logistic regression, random forests, CRF and so on) to automatically integrate multiple term features into a binary classifier to categorize term-candidates as terms or non-terms. Such classifiers take into account many term features: grammatical, statistical, orthographical, and contextual, thus achieving higher precision of term extraction compared to models based on the traditional approach. In particular, in the work [7] machine-learning model trained with about 130 various features showed F1-score of 55% against 28% for traditional ATR methods.

Although machine learning approach improves the quality of term recognition, the trained machine classifiers remain sensitive to domain differences, showing worse results for texts in another (different of target) problem domain.

Recent papers [4, 6, 9, 11] have proposed new ML approaches to term recognition, leveraging modern deep learning transformer-based techniques (mainly BERT [2]) in order to further increase ATR quality and to tackle the problem of cross-domain transfer for trained classifiers.

The works [9, 11] apply sequential labeling of word tokens in the text being processed, this approach was initially-proposed and widely-used for named entity recognition (NER) in texts, well-known NLP problem [3, 12] close to ATR. Sequential labeling implies token-level classification, when each word token is assigned a label indicating whether it is part of term (or part of named entity in NER). Another approach presented in [4, 6] relies on binary classification of phrases (term-candidates) preliminary extracted from text, which is somewhat similar to traditional ML classifiers for ATR, but the phrases are classified together with their enclosing sentence (i.e. with their context). Experimental results showed that ATR models built within these approaches achieve up to 69-75% of F1-score for term extraction, the results depend on datasets and specialized domains taken for training, It should be noted that most transformer-based models were built and evaluated with the data from ACTER corpus [8] that contains detailed manual annotations of terms and encompasses terms from texts in three natural languages (English, French, and Dutch) and for four specialized domains. For Russian, for a long time there was no freely available corpus (or even dataset) suitable for training various ATR models and for comparing approaches. Recent paper [1] reports a comparison of ATR approaches, in relation to Russian texts, but only for a dataset with terms in domains of mathematics and programming.

In this work, a comparative study of the transformer-based ATR approaches was conducted with annotated data created for RuTermEval-2024 competition devoted to nested term identification and classification. The competition provides datasets with manually annotated terms including named entities, nested terms (such as *differential equation* and *linear differential equation*) were annotated as well. RuTermEval-2024 include three tasks (each task implies identification of nested terms): (1) binary term identification, (2) term recognition with classification (into one of predefined types), (3) cross-domain term recognition and classification. Besides two above-described ATR approaches, namely sequential labeling of tokens and phrase classification with context sentence, we also considered the approach that gives state-of-the-art results for nested NER, that is text span prediction by using vector representations obtained by contrastive learning [13].

The experiments with BERT-based models trained and evaluated in the considered approaches have shown that the span prediction models based on contrastive learning outperform the other models across all RuTermEval tasks, but at the same time demonstrate the most significant decrease in quality in the cross-domain task. In the competition, our best models achieve F1-score of 77% for binary term identification (Task 1), 70% for term recognition with classification (Task 2), and only 47% for cross-domain recognition with classification (Task 3), placing second for each task.
The next section presents a short overview of the transformer-based approaches studied in our experimental study. Then, we describe implementation of our BERT-based models for term recognition and discuss results of experiments with them. The final section presents conclusions.

2 Approaches for ATR

Transformer-based phrase classification with context sentence was initially proposed for term recognition in the paper [4]. In this approach, various phrases (e.g., n-grams) are first extracted from the text being processed, together with their corresponding context sentences, and then each phrase-sentence pair is categorized by BERT-classifier, for determining whether the phrase is term or not. For training such binary classifier, besides positive phrase-sentence pairs taken from a dataset with annotated terms, negative pairs of non-terms and their enclosing sentences are to be formed.

In the work [4] n-grams were taken as term-candidates for classification, and necessary sets of positive and negative sample pairs were built with multilingual ACTER corpus [8] (negative samples were formed by random selection of n-grams). The trained binary transformer-based classifiers (specifically, RoBERTa for English and CamemBERT for French were used) were evaluated and compared with XGBoost model, which was built by traditional ML techniques with the same data, through feature engineering with certain linguistic and statistical term features. Experiments demonstrated significant advantage of transformer-based phrase-sentence classification approach: XGBoost model showed high precision but low recall, thus giving about 27% F-score, while the BERT-based classification models showed in average 48%. Contextualized embeddings of BERT-based classifiers obviously enhance detection of terms, thus increasing precision of their prediction, and this transformer-based approach also eliminates huge feature engineering. Another important circumstance is equal handling of all extracted term-candidates, which enables recognition of nested terms, but at the same time, extracting from the text all possible phrases (as term-candidates) is quite expensive computational task.

Sequence labeling approach to ATR does not require preliminary extraction of terms candidates, instead, terms occurrences are directly detected in texts by classifier performing tagging of text tokens (similar to the task of named entity recognition [12]). As a rule, BIO tag scheme is used, and each word token in the text is labeled by the classifier with one of three tags: B (beginning of term), I (inside term word) or O (word not belonging to term). The paper [9] describes such sequence labeling classifiers built with recurrent neural network (RNN) and with certain types of input word embeddings, and their comparison is reported, with an analogous classifier built with CRF (conditional random field) method and a vast set of linguistic and statistical features. All classifiers were trained on data from multilingual ACTER corpus, the best performance with 75% F-score was achieved by RNN model with multilingual BERT embeddings and additional using training samples from several languages.

In the work [11], sequential labeling approach for ATR was applied to Slovenian texts, BERT-based models were trained on RSDO5 corpus containing manually annotated data from four specialized domains (biomechanics, chemistry, veterinary science, and linguistics). Experiments in cross-domain settings were conducted, when models were trained in one specialized domain and tested in another. Evaluation of implemented cross-domain models (12 classifiers) showed F-scores ranging from 64% to 71%, which means possibility of domain transfer for term recognition models.

These two above-described approaches were studied and compared in cross-language experiments in the work [6] exploiting pretrained multilingual model XML-RoBERTa and ACTER corpus. Trained binary classifiers for phrase-sentence pairs demonstrated F1-score up to 58%, while the token labeling approach showed the highest F1-score of 69.8%. At the same time, analogous comparison for Russian texts performed in recent work [1] showed some advantage of binary phrase-sentence classification (70,5% against 52,5% of F1-score), so the question about superiority of one approach over another remains open. It should be noted that for recognizing nested terms, the sequential labeling approach has limitations compared with phrase-sentence classification, not all possible cases of nested terms may be recognized, or more complicated (than BIO-scheme) token labeling is required.

The problem of recognizing nested entities in texts is quite important for named entities (NE), and attempts to tackle the problem has led to paradigm shift in NER, from predominant sequence labeling to span prediction (see, for example, [3]). The new paradigm implies identifying NE as text spans (fixed by their boundaries) and also classifying them into predefined set of entity types (person, location, and so on). The state-of-the-art results for NER were obtained within this research approach in the work [13] that describes Binder, new architecture and framework for training NER models.

Binder (BI-encoder for NameD Entity Recognition) performs NER as classification of text spans by using their vector representations obtained by contrastive learning. Two pretrained BERT encoders are employed to separately map into the same vector space text spans to be classified and also textual descriptions of named entity types (e.g., "person" entity type may be described as "individuals or groups of individuals identified through proper nouns"). The common vector space makes it possible to calculate similarity between representations of the spans and the considered entity types. Contrastive learning follows two objectives: (1) maximizing similarity scores between representations of entity types and named entity spans, and (2) minimizing similarity scores between representations of entity types and spans for non-entities (and also between representations of entities and inappropriate entity types), thus separating non-entity spans from entity mentions. The contrastive learning with these objectives is supplemented with dynamic thresholding loss to work out specific dynamic threshold to distinguish entity spans from non-entity ones.

The described Binder's approach makes it easy to handle both nested and flat NE, and trained Binder models outperform previous NER models, across several benchmarks, achieving up to 90% F1-score for nested NER (e.g., on ACE2005 and GENIA corpora) and 95% F1-score for flat NER.

Thus, the above-described transformer-based approaches and models for ATR have been tested on various datasets for several languages (but few experiments for Russian), sometimes showing contradictory results in quality, which most likely depends on language, specialized domains of processed texts, and evidently data for training. So we have chosen two approaches — phrase classification with context sentence and sequential labeling of tokens — for our experimental study within RuTermEval-2024 competition for Russian texts. As the competition was devoted to nested terms, we also considered in our comparison Binder's span prediction approach (which has not previously been used for ATR), with its adaptation to Russian and to tasks of the competition. In particular, instead of textual descriptions of named entity types, text definition of concept "term" was used, as well as textual descriptions of term types (classes).

3 Experiments and Results

3.1 RuTermEval Tasks and Dataset

RuTermEval-2024 competition on nested term identification and classification in Russian scientific texts includes three tasks (challenges). Binary term identification (Task 1) aims to detect position boundaries of all terms in a given text. Term recognition with classification (Task 2) requires not only to detect terms' positions but also to classify each term into one of three predefined classes:

- specific terms − terms that are lexically specific and belonging to the considered scientific domain (e.g. *идиома − Eng. idiom, эпистемическая модальность − Eng. epistemic modality*);
- common terms − terms that are domain specific and at the same time may be known and used by non-specialists, as they consist of words of general lexicon (e.g. *словарь − Eng. dictionary, носитель языка − Eng. native speaker*);
- nomens − names of domain-specific objects, such as names of databases and datasets, programming languages, corpora, dictionaries, programs and software packages, etc. (e.g. *Интернет − Eng. Internet, тезаурус Роже − Eng. Roget's thesaurus*).

Cross-domain term recognition and classification (Task 3) involves detecting all term positions in the text and their classification into one of the above-described classes, which is similar to Task 2, but for texts from scientific domains different from the domains used for training ATR models.

For evaluating participant's term recognition models, RuTermEval-2024 provided CL-RuTerm3 dataset with annotated terms, including nested terms, such as *interrogative pronoun — pronoun*). The dataset contains: 1) training subset, which covers 850 abstracts and certain full-text Russian scientific

papers in domains of linguistics and computational linguistics; 2) the subset for development (testing and evaluation of models), with 103 abstracts and 10 full-text papers in the same domain; 3) the subset for testing and evaluation in Task 3, which includes 100 abstract and texts from two different problem domains (literary studies and agricultural technology).

We compared some characteristics of CL-RuTerm3 with ACTER corpus that is the most often used in research papers on ATR. The training subset of CL-RuTerm3 is comparable in size with ACTER subcorpora (about 60 K tokens in each specialized domain and language), but the subset significantly differs in diversity of included terms (e.g., English subcorpus of ACTER on heart failure domain includes 2519 unique terms with nomens, while CL-RuTerm3 subset contains 7832). This is most likely explained by broader domain of CL-RuTerm3 (two wide specialized domains: linguistics and computational linguistics) and also by fragmentary nature of it (mainly abstracts of papers not full paper texts). Our analysis revealed that most terms of the training subset are encountered in texts rarely or extremely rarely − see Fig. 1, demonstrating distribution for frequencies of unique terms' occurrences, across term classes. Nomens are the most poorly represented in the texts (296 nomen were found only once). This feature may evidently influence quality of the trained models.

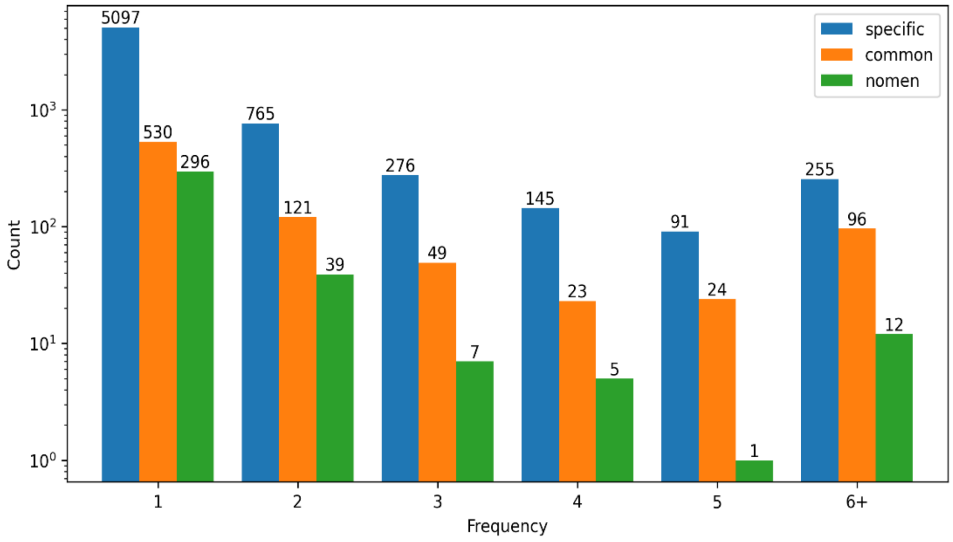


Figure 1: Statistics of term occurrences for term classes in training CL-RuTerm3 subset

3.2 Training of Models

For our experiments, we used pretrained ruBert-base [14] model from SberDevices — BERT [1] initially trained for Russian on texts from Russian Wikipedia, news, books, web, film subtitles and so on. For each of RuTermEval-2024 tasks, ruBert-base models were fine-tuned according to the considered approaches:

- classification of phrase-sentence pairs, when the fine-tuned model classifies a phrase from a given input pair as term or non-term (binary classification) or additionally produces particular term class (specific, common, nomen) for the recognized term (multiple classification);
- sequential labeling of words, when each word token in a given text is classified according to BIO tagging scheme (as part of term or not), or additionally produces particular class (specific, common, nomen) for such recognized token;
- span prediction based on contrastive learning, when for a given text the trained model processes all possible text spans with length up to predefined value and recognizes class for each span: binary (is it term or non-term) or particular term class (specific, common, nomen).

Eventually, nine ATR models were trained for experiments (three models for each of the three competition's tasks), all of them were fine-tuned for 10 epochs with AdamW optimizer, batch size of 8, and learning rate 5e-5. Among data provided for training in the competition, 10% portion was randomly selected for validation during training itself, and the model with the best F1-score on the validation data among all epochs was taken as the final model.

For classification of phrase-sentence pairs, positive examples were constructed with annotated terms from the training subset of CL-RuTerm3 dataset, while negative examples were generated by randomly selecting n-grams taken from the subset and consisting of adjectives and nouns, the number of negative examples was equal to the number of positive examples. The models for phrase-sentence classification and for sequential labeling of tokens were developed with the transformers library[1].

Span prediction models based on contrastive learning were implemented with the Binder framework[2] developed and described in [12], which requires to set for training, besides other parameters (e.g. maximum span length was 30 tokens of the BERT-model), natural language descriptions of units being detected and classified: that is text description for concept "term" (Task 1) and text descriptions of terms' classes (Task 2 and Task 3). We have adapted for this purpose descriptions of RuTermEval tasks given by organizers of the competition. For example, text description for "specific term" was taken as *"Термин, обозначающий концепцию или явление, специфичное для определенной области и её лексики" (Eng. "Term denoting a concept or phenomenon specific to a particular domain and its lexicon"*).

3.3 Results in Competition

In the competition, quality of the trained models was evaluated by the F1-score. For Task 2 and Task 3, two modes for evaluation were exploited: classified, i.e. with term classes (F1-score is calculated based on predicted terms' positions and classes); and unclassified, i.e. without term classes (F1-score is calculated only based on predicted terms' positions, predicted terms' classes are ignored). The results of our models are presented in Table 1 (the unclassified mode for Task 2 is identical to Task 1).

Approach of Model	Task 1	Task 2		Task 3	
		Classified	Unclassified	Classified	Unclassified
Phrase classification	0.44	0.41	0.44	0.33	0.37
Sequential labeling	0.57	0.55	0.57	0.40	0.44
Span prediction	**0.77**	**0.70**	**0.77**	**0.47**	**0.51**

Table 1: Quality of term recognition in Leader board

The experiments demonstrate that the models of span prediction approach consistently and noticeably outperform the models of alternative approaches (sequential labeling and phrase-sentence classification), across all RuTermEval tasks. Specifically, the span prediction models achieved 77% F1-score for binary term identification (Task 1), 70% for term recognition with classification (Task 2), and 47% for cross-domain term recognition and classification (Task 3). Models of all the considered approaches exploit BERT contextual embeddings, so the advantage of the span prediction approach is most likely explained by contrastive learning for spans' representations in the common vector space.

Though the span prediction approach demonstrates better quality across all tasks, it also shows the highest sensitivity to task complexity: additional term class prediction required in Task 2 led to 4-7% decrease of F1-scores (for other approaches decrease is 2-3%), while the cross-domain Task 3 resulted in a dramatic 23-26% drop in F1-scores (8-15% for other models).

One can notice evident correlation between task complexity and quality of the models: simple term identification (Task 1) yields the best results, additional term class prediction (specific terms, common terms, or nomen in Task 2) slightly degrades the quality (2-7% across the models), but change of specialized domain (cross-domain term recognition and classification, Task 3) proved particularly challenge, with F1-scores dropping by 8-26% compared to Task 2.

The results obtained in our work correspond to the second place in RuTermEval competition for all its tasks, and the gap from the winner is quite insignificant in Task 2 (F1-scores 0.6996 versus 0.6997) and relatively small in Task 1 (0.769 versus 0.794).

[1] https://huggingface.co/docs/transformers/index

[2] https://github.com/microsoft/binder

3.4 Comparing Precision and Recall for Models and Term Classes

Besides F1-score that measures averaged quality of the trained models, precision and recall estimate additional important characteristics, so we also have evaluated them for our models on the development subset of CL-RuTerm3, the only available for this purpose. The results are given in Table 2 for all term classes in total, and in Table 3 for the classes separately. For sequential labeling model, recall is consistently worse than precision, this may be explained by difficulties of the approach to recognize nested terms. For phrase classification, situation varies for different Tasks and term classes, while for span classification, recall and precision often are approximately the same, except Task3, when recall drops sharply (probably due to good adaptation to the training subset).

Approach of Model	Task 1			Task 2			Task 3		
	P	R	F1	P	R	F1	P	R	F1
Phrase classification	0.67	0.68	0.66	0.36	0.52	0.40	0.27	0.28	0.26
Sequential labeling	0.77	0.57	0.65	0.68	0.52	0.58	0.51	0.24	0.32
Span classification	**0.83**	**0.78**	**0.79**	**0.79**	**0.74**	**0.76**	**0.63**	**0.25**	**0.33**

Table 2: Quality of term recognition on development subset

The data in Table 3 for three considered term classes shows that in both Tasks specific terms are recognized by sequential labeling model somewhat better than common terms, but for phrase classification the situation is opposite. As for span classification model, in Task 2 it gives about equal results, but it has worse quality for common terms. Nomens are recognized very poorly by all models, the worse result (0.03 of F1-measure, in cross-domain Task 3) belongs to span classification model, in this term class it unexpectedly loses to other approaches. Such worse quality for nomens is likely caused by their diversity (personal names, names of text corpora, databases, etc.) and by simultaneously rare frequency in the dataset used for training.

Approach of Model	Specific			Common			Nomen		
	P	R	F1	P	R	F1	P	R	F1
Task 2									
Phrase classification	0.35	0.50	0.37	0.41	0.53	0.42	0.11	0.17	0.12
Sequential labeling	0.68	0.56	0.60	0.60	0.39	0.44	**0.30**	**0.21**	**0.23**
Span classification	**0.77**	**0.75**	**0.75**	**0.76**	**0.77**	**0.75**	0.26	0.15	0.18
Task 3									
Phrase classification	0.23	0.23	0.21	0.29	**0.33**	**0.26**	0.17	**0.17**	**0.15**
Sequential labeling	0.55	0.26	0.34	0.25	0.14	0.15	**0.22**	0.10	0.13
Span classification	**0.64**	**0.31**	**0.40**	**0.36**	0.12	0.16	0.19	0.02	0.03

Table 3: Quality of recognition for term classes

Our analysis of errors in term classes recognized by the models revealed that all they more often tend to get confused with classification between specific and common terms, e.g., specific term *часть речи* (*parts of* speech) was incorrectly categorized as common, while *деловая коммуникация* (*business communication*) was classified as specific instead of common. However, distinguishing between these classes is difficult even for a terminologist, due to the lack of clear boundaries between term domains.

4 Conclusion

The described comparative study of modern transformer-based approaches to automatic term recognition and classification, which has been applied to Russian texts, has revealed the advantage of span prediction by contrastive learning, but at the same time this approach may be particularly vulnerable to domain shift and task complexity. Nevertheless, further experiments for the considered

approaches are needed, relying on representative annotated datasets of various sizes and with texts of various specialized domains. Term annotations is a complex task even for humans (especially identification of term class), and datasets for training often vary significantly across domains and languages, and in some cases different approaches may be more effective, not only span prediction. The more challenging research task is cross-domain term recognition, which is more necessary in applications because of the lack of terminologically annotated data for new specialized domains.

References

[1] Bolshakova E.I., Semak V.V. Methods and means of term extraction from texts for terminological tasks [Metody i sredstva izvlecheniya terminov iz tekstov dlya terminologicheskikh zadach] // Programming products and systems [Programmnye produkty i sistemy]. — 2025. — V. 38 — №. 1. — pp. 5–16.

[2] Devlin, J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 NAACL Conference: Human Language Technologies, Minneapolis — 2019 — pp. 4171–4186.

[3] Fu, J., Huang X., Liu P. SpanNER: NamedEntity Re-/Recognition as Span Prediction // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int. Joint Conference on Natural Language Processing — 2021 — pp. 7183–7195.

[4] Hazem, A., Bouhandi, M., Boudin, F., Daille, B. TermEval 2020: TALN-LS2N System for Automatic Term Extraction // Proceedings of the 6th International Workshop on Computational Terminology — 2020 — pp. 95–100.

[5] Korkontzelos, I., Ananiadou, S. Term Extraction // Oxford Handbook of Computational Linguistics — 2014 — 2nd Ed. — Oxford University Press, Oxford — pp. 991–1012.

[6] Lang, C. et al.: Transforming term extraction: Transformer-based approaches to multilingual term extraction across domains // Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 — 2021 — pp. 3607–3620.

[7] Terryn, A.R. et al. Analysing the impact of supervised machine learning on automatic term extraction: HAMLET vs TermoStat // Proceedings of the Int. Conference on Recent Advances in Natural Language Processing (RANLP 2019) — 2019 — pp. 1012-1021.

[8] Terryn, A.R. et al. TermEval 2020: Shared task on automatic term extraction using the annotated corpora for term extraction research (ACTER) dataset // 6th International Workshop on Computational Terminology (COMPUTERM 2020) — 2020 — pp. 85-94.

[9] Terryn, A.R., Hoste V., Lefever E. Tagging terms in text: A supervised sequential labelling approach to automatic term extraction // Terminology. Int. Journal of Theoretical and Applied Issues in Specialized Communication — 2022 — Issue 28, Vol. 1 — pp. 157–189.

[10] Tran, H.T.H. et al. The recent advances in automatic term extraction: A survey // arXiv preprint — 2023 — Access mode: https://arxiv.org/abs/2301.06767

[11] Tran, H.T.H., Martinc M., Doucet A., Pollak, S. A transformer-based sequence-labeling approach to the slovenian cross-domain automatic term extraction // Slovenian Conference on Language Technologies and Digital Humanities — 2022 — Ljubljana — pp. 196-204.

[12] Yadav, V., Bethard S. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models // Proceedings of the 27th International Conference on Computational Linguistics — 2018 — ACL, Santa Fe, USA — pp. 2145–2158.

[13] Zhang, S. et al. Optimizing Bi-Encoder for Named Entity Recognition via Contrastive Learning // The Eleventh International Conference on Learning Representations — 2022 — Access mode: https://arxiv.org/abs/2208.14565

[14] Zhang, Z., Iria J., Brewster, C., Ciravegna, F. A Comparative Evaluation of Term Recognition Algorithms // Proc. of the Sixth Int. Conf. on Language Resources and Evaluation (LREC'08) — 2008 — pp. 2108–2111.

[15] Zmitrovich, D., et al. A Family of Pretrained Transformer Language Models for Russian // arXiv preprint — 2023 — Access mode: https://arxiv.org/abs/2309.10931

“Okie dokie, here’s the no-cap truth!”: Everyday Russian Youth Speech in Corpus Representation (Structure and Application of the ESC Sound Corpus)

Sherstinova T. Yu., Melnik A. G., Petrova I. A., Azarevich K. I., Melkozerova V. I., Chepovetskaya S. V.
Laboratory for Language Convergence, National Research University Higher School of Economics — St. Petersburg
123A Griboedov Canal Embankment, St. Petersburg, 190068, Russia
tsherstinova@hse.ru, melnik-a@esc-corpus.ru, {iapetrova_2, kiazarevich, vimelkozerova, avchepovetskaya}@edu.hse.ru

Abstract

The article focuses on the creation and potential applications of the Everyday Student Conversations Russian speech corpus (ESC corpus), referred to as KURS corpus in Russian. The corpus is being developed based on a modified methodology of the “one day of speech” recording approach, adopted from the ORD Corpus, with an emphasis on capturing contemporary youth speech. The article highlights the key aspects of the corpus creation process, with special attention to its structure, data collection and processing methodology, as well as the functionality of the corpus online demo version. The primary aim of the project is to explore linguistic changes in the youth environment, create a resource for scientific and applied research, and develop a big data collection of everyday speech for machine learning and advanced artificial intelligence systems. For linguists, the corpus offers unique opportunities to study sociolinguistic, pragmatic, and phonetic features of speech, making it a valuable tool for analyzing contemporary discourse.

Keywords: Russian language; everyday speech; youth speech; corpus linguistics; sociolinguistics; pragmatics; speech technologies; oral discourse

DOI: 10.28995/2075-7182-2025-23-331-344

«Оки доки, вот вам без рофла!»: Повседневная речь молодежи в корпусном представлении (структура и назначение звукового корпуса КУРС/ESC)

Шерстинова Т. Ю., Мельник А. Г., Петрова И. А., Азаревич К. И., Мелкозерова В. И., Чеповецкая С. В.
Лаборатория языковой конвергенции, Научно-исследовательский университет «Высшая школа экономики» — Санкт-Петербург
Санкт-Петербург, 190068, наб. канала Грибоедова, д.123, лит. А
tsherstinova@hse.ru, melnik-a@esc-corpus.ru, {iapetrova_2, kiazarevich, vimelkozerova, avchepovetskaya}@edu.hse.ru

Аннотация

Статья посвящена принципам создания и возможностям звукового Корпуса устной речи студентов и молодежи КУРС/ESC (Everyday Student Conversations в английском переводе). Корпус разрабатывается на основе модифицированной методики записи «одного речевого дня», заимствованной из Корпуса ОРД, с акцентом на фиксацию современной речи молодежи. В статье описываются ключевые аспекты процесса создания корпуса, при этом основное внимание уделяется структуре корпуса, методологии сбора и обработки данных, возможностям демоверсии сайта корпуса. Основная цель проекта — изучение изменений в языке молодежной среды, а также создание ресурса для научных и прикладных исследований, формирование «больших данных»

повседневной речи для задач машинного обучения и создания систем сильного искусственного интеллекта. Для лингвистов корпус предлагает уникальные возможности для изучения социолингвистических, прагматических и фонетических характеристик речи, что делает его важным инструментом для анализа современного дискурса.

Ключевые слова: русский язык; повседневная речь; речь молодежи; корпусная лингвистика; социолингвистика; прагматика; речевые технологии; устный дискурс

1 Введение

В статье дается описание нового мультимедийного лингвистического ресурса — Корпуса устной речи студентов и молодежи (КУРС/ESC), содержащего реальные звукозаписи речевого общения, записываемые в течение определенного, достаточно длительного промежутка времени (чаще всего в течение суток или «речевого дня»).

История лонгитюдных звукозаписей повседневной речи берет свои истоки в подходе, впервые реализованном в исследованиях языкового существования (gengo seikatsu), инициированных в конце 1940-х гг. Японским национальным институтом исследования японского языка (NLRI) [4]. Разнообразные по методике проекты были направлены на фиксацию речевого поведения информантов «в течение 24 часов» [5] или сбор речевого материала в естественных условиях общения (напр., для японского Корпуса ESP, входящего в крупный проект JST/CREST, посвященный обработке эмоциональной речи) [6]. По похожей методике записывались данные и для Британского национального корпуса [7]. В отечественном языкознании метод записи речи в течение «одного речевого дня» применялся при создании корпуса ОРД, когда волонтеры-информанты получали на сутки профессиональный диктофон, фиксирующий все их общение в течение этого периода времени [8], [9].

Создание этого корпуса методологически основано на подходе, использованном разработчиками ОРД с небольшими модификациями. Необходимость создания нового ресурса повседневной речи, сфокусированного именно на речи молодежи, объясняется тем, что с момента последних записей ОРД прошло уже 7 лет, с тем пор язык претерпел определенные изменения, наблюдаемые главным образом в речи молодежи и студенчества. Создание корпуса направлено на фиксацию текущего языкового состояния с целью выявления новой разговорной лексики, а также новых значений и коннотаций известных языковых единиц.

2 Методология сбора данных

Проект по получению молодежных звукозаписей по методике «Одного речевого дня» и их корпусному представлению был инициирован в 2023 г. Лабораторией языковой конвергенции НИУ ВШЭ — Санкт-Петербург при поддержке Программы фундаментальных исследований НИУ ВШЭ. Информанты записывают свою речь в течение обычного дня на профессиональный диктофон. Основной единицей записи, получаемой от информантов, является их «речевой день» — совокупность аудиофайлов, отражающих их вербальное взаимодействие с момента пробуждения утром до отхода ко сну вечером.

Как и при сборе данных для корпуса ОРД [10], основным условием остается информированное согласие всех основных участников разговора на проведение записи. Информантам предлагается заранее уведомить своих потенциальных собеседников о предстоящей записи за день до ее начала, чтобы тематика разговоров не сводилась преимущественно к обсуждению деталей звукозаписи. В настоящее время информантом для корпуса может стать любой волонтер, чье речевое поведение осуществляется в русскоязычной (или преимущественно в русскоязычной) речевой среде, даже если русский язык не является для него родным.

Для участия в записи добровольцы сначала заполняют первичную анкету, отражающую их намерение[1], после чего с ними связывается представитель разработчиков, чтобы согласовать наиболее удобный день для получения оборудования. Одновременно с передачей диктофона осуществляется инструктаж, который включает в себя как технические моменты использования аппаратуры, так и организационные вопросы, включая заполнение информационных согласий, анкет и ознакомление с правилами осуществления записи и ее использования в корпусе. В день или

[1] Возможность заявить о желании принять участие в звукозаписи скоро появится на сайте корпуса.

сразу после звукозаписи «речевого дня» информанты заполняют дневник, в котором отмечаются основные события, темы, условия и участники разговоров, а также отвечают на вопросы социологической анкеты про себя и своих собеседников, проходят психологическое тестирование и тест на пассивный словарный запас. Записи передаются в корпус в анонимизированном виде — информация о говорящих кодируются, для открытого представления расшифровок на сайте все личные имена заменяются.

Звукозаписи, полученные от информантов, передаются на экспертное прослушивание и сегментирование на макроэпизоды [12], одновременно происходит их аннотирование с точки зрения типа коммуникации, места коммуникации и социальной роли информанта, а также нормализация этих параметров для корпусного представления.

Отсегментированные на макроэпизоды звуковые файлы становятся основной единицей представления и описания данных в корпусе. Они обрабатываются одной из систем автоматического распознавания речи (см. п. 3.3). После этого автоматически распознанные записи снова передаются экспертам, которые проверяют и исправляют ошибки распознавания и диаризации (атрибуции речи по говорящим). Это наиболее трудоемкий этап создания корпуса, так как системы автоматического распознавания речи до сих пор недостаточно хорошо работают на зашумленных повседневных звукозаписях [15], а лексическое своеобразие молодежной коммуникации еще более усложняет эту задачу. Выверенные расшифровки после этого направляются на токенизацию, лемматизацию и грамматическую разметку.

Несмотря на то, что общая методология сбора данных была заимствована из Корпуса ОРД, методику создания Корпуса КУРС/ESC отличает несколько ключевых моментов:

1) В корпус принимаются как записи целых «речевых дней», так и их фрагменты или отдельные разговоры.
2) У респондентов появилась возможность записи «речевого дня» на собственные смартфоны (при условии высокого качества встроенного диктофона).
3) Произошел полный отказ от бумажной документации: все анкеты, инструкции и согласия заполняются в электронном виде онлайн.
4) Транскрибирование аудиозаписей проводится в полуавтоматическом режиме: сначала создаются черновые транскрипты с использованием систем распознавания речи, которые затем проверяются и корректируются экспертами.
5) При первичной расшифровке транскрипты звукозаписей используют стандартную пунктуацию и членение на «предложения».
6) Планируется привлечение краудсорсинга на запись речевого материала.

3 Структура корпуса КУРС/ESC

Структура звукового корпуса КУРС состоит из следующих основных блоков:

1) Звукозаписи, полученные от информантов.
2) Метаинформация к звукозаписям, описывающая как участников разговора, так и условия коммуникации.
3) Расшифровки звукозаписей, выполненные автоматическим методом.
4) Расшифровки звукозаписей, прошедшие экспертные проверку.
5) Лингвистическая аннотация.
6) Информационная система и утилиты обработки данных.

3.1 Звукозаписи

Для работы используются два вида профессиональных звукозаписывающих устройств: Roland R09-HR и Zoom H1n. Формат выходного файла: WAV, стерео, 16 бит, 44 100Гц. Поскольку в корпус принимаются также записи, выполненные на собственную звукозаписывающую технику информантов, включая мобильные телефоны, их оригинальные записи могут отличаться по формату, количеству каналов и частоте дискретизации. В последнем случае осуществляется их конвертация в указанный выше формат.

3.2 Метаинформация

Метаинформация к звукозаписям, описывающая как участников разговора, так и условия коммуникации, содержит: 1) социологическую анкету информанта и коммуникантов, 2) дневник речевого дня, 3) результаты двух тестов, выполненных информантом (психологического теста и на пассивный словарный запас).

Дневник речевого дня описывает основные события в виде следующей таблицы: 1) код информанта, 2) код коммуниканта, 3) кем приходится этот коммуникант информанту, 4) имя коммуниканта (этот параметр подлежит анонимизации), 5) когда (и сколько) происходило общение, 6) где происходило общение, 7) основные темы разговора. На основе информации из «дневника» осуществляется нарезка речевого материала на макроэпизоды.

Социологическая анкета информанта предполагает ответы на следующие вопросы: 1) пол, 2) возраст, 3) место рождения, 4) текущее место жительства, 5) проживание в других местах (где и как долго), 6) родной язык, 7) языки, которыми владеет информант (с указанием уровня), 8) языки, используемые в семейном общении и с друзьями, 9) профессия родителей, 10) учебное учреждение, 11) образовательная программа или специальность, 12) образовательная ступень, 13) курс, 14) место работы (если есть), 15) вопросы и комментарии, 16) согласие на публикацию записей. Подобная информация по возможности заполняется на всех основных коммуникантов.

Психологическое тестирование осуществляется с помощью пятифакторного опросника личности 5PFQ, позволяющего оценить степень выраженности каждого из пяти факторов «большой пятёрки»: экстраверсии, доброжелательности, добросовестности, эмоциональной стабильности и открытости новому опыту [16]. Кроме того, информантам предложен тест на определение пассивного словарного запаса [17].

Данные, полученные из тестов, заносятся в информационную базу данных и могут быть использованы для изучения социолингвистической и психолингвистической вариативности.

3.3 Автоматические расшифровки звукозаписей

В настоящее время для автоматического распознавания звукового материала используются две ASR системы: 1) Whisper от OpenAI [13], обученная на письменных текстах и порождающая «олитературизированные» расшифровки, напоминающие скорректированные тексты интервью, и 2) акустическая модель, разработанная отечественной компанией ООО НТР [14], в которой нет языковой модели, что позволяет получать звукозапись, похожую на упрощенную фонетическую транскрипцию. Для чернового распознавания речи используется Whisper, а акустическая модель находит свое применение при анализе редуцированных форм. Качество итоговой расшифровки в значительной степени зависит от качества исходной звукозаписи, ее фоновой зашумленности и количества одновременно говорящих. В достаточно «чистых» непродолжительных звукозаписях с хорошим уровнем громкости полезного сигнала показатель WER (доля неправильно распознанных слов) может быть небольшим (напр., 10 %), с ухудшением качества сигнала и увеличением количества говорящих этот показатель может возрастать до 90 %. На момент подготовки статьи, ручная коррекция требуется в 35-40 % словоупотреблений. Для повышения качества автоматического распознавания осуществляется поиск более точных моделей, а также проводится их дообучение (fine-tuning) на собранной звуковой выборке.

Транскрипты звукозаписей автоматически членятся на реплики говорящих, которые могут содержать неограниченное количество фраз. Если фразы одного говорящего идут без длительных пауз, то они заносятся в одну реплику. Транскрипты выгружаются в табличном виде, при этом указываются имя эпизода (звукового файла), начало реплики от начала эпизода, результаты диаризации[2] (выгружаемые в виде *speaker_N*, где *N* — порядковый номер говорящего от 0 до общего количества говорящих в данном эпизоде), и собственно транскрипт реплики.

[2] Диаризация речи — это автоматический процесс сегментации аудиозаписи по говорящим, в ходе которого звуковой поток разбивается на фрагменты, соответствующие индивидуальным репликам, и каждому фрагменту присваивается идентификатор условного участника

3.4 Расшифровки звукозаписей, проверенные экспертами

Расшифровки звукозаписей, выполненные автоматическим методом, проходят экспертную корректуру, в процессе которой звукозапись прослушивается, проверяется корректность идентификации говорящих, заполняется поле SCode, содержащее уникальный код информанта в корпусе, и выверяется корректность расшифровки. Итоговый транскрипт выгружается в базу данных в табличном виде (см. рис. 1).

SFName	Time	Speaker	SCode	Text
escAF001_06m	00:00:01	speaker_0	AF001	Привет.
escAF001_06m	00:00:03	speaker_1	XF001	Как дела?
escAF001_06m	00:00:05	speaker_0	AF001	Средненько.
escAF001_06m	00:00:07	speaker_1	XF001	Что такое?
escAF001_06m	00:00:10	speaker_1	XF001	О, Король и Шут.
escAF001_06m	00:00:11	speaker_1	XF001	У меня подруга снималась там в массовке.
escAF001_06m	00:00:13	speaker_0	AF001	Прикольно.

Рисунок 1: Пример выгрузки транскрипта системы распознавания речи (фрагмент)

В случае наложения речи используется символ @, обозначающий смену говорящего, код второго говорящего указывается в дополнительном столбце.

Пример автоматической расшифровки и результат ее коррекции представлен в табл. 1.

ASR расшифровка	Экспертная коррекция
Рассказывала короче Сон, смотри как ты знаешь я читаю так титанов и поэтому мне теперь часто снятся титаны. Ну потому что они с**и страшные сегодня не исключение и я там непосредственно участвовала в событиях, выдавала себя заикаю. Но то же потом. Сначала, конечно, все отодрались, там что то там превращался, еще какие то люди бегали, еще какие то титаны там, короче, просто мясо, заварушка, крошево какое то, огонь это все на каком то острове происходит. Ехали туда какие то пожарные, посмотрели такие ничего не можем сделать, уехали, короче, вот. То есть просто полное мясо из каких то людей. Вот и мне в какой то момент понадобилось, ну, куда то уйти, я не знаю почему. То есть я там не то чтобы я там что то помогала как то, я просто бегала и пыталась не попасть под руку Тим. Вот. а потом, когда я вернулась, уже все закончилось. все почему то сидят за столом, вроде как наши победи я такая ну здорово.	Рассказываю, короче, сон... Свой. Ээ, как ты знаешь, я читаю "Атаку Титанов", и поэтому мне теперь часто снятся Титаны, ну, потому что они, с**и, страшные. Вот, сегодня не исключение, и я там непосредственно участвовала в событиях, выдавала себя за Микасу. Но это уже потом. Сначала, конечно, все подрались. Там что-то Эрен в титана превращался, еще какие-то люди бегали, еще какие-то титаны там, короче, просто мясо, заварушка, крошево какое-то, огонь, это все на каком-то острове происходит. Приехали туда какие-то пожарные, посмотрели такие: "Ничего не можем сделать", уехали, короче, вот. То есть просто полное мясо из каких-то людей. Вот и мне в какой-то момент понадобилось, ну, куда-то уйти, я не знаю почему. То есть я там не то чтобы я там что-то помогала как-то, я просто бегала и пыталась не попасться под руку Титану. Вот, а потом, когда я вернулась, уже все закончилось. Все почему-то сидят за столом, вроде как наши победили, я такая: "Ну, здорово".

Таблица 1: Сопоставление автоматической и экспертной транскрипции (фрагмент монолога о сне, эпизод *escAF001_03*)

3.5 Лингвистическая аннотация

Токенизация выполняется по пробелам, лемматизация и грамматическая разметка выполняются в MyStem [19]. Выбор этого морфологического парсера обусловлен, во-первых, тем фактом, что именно с его помощью выполнена разметка Национального корпуса русского языка, а во-вторых — высокими показателями эффективности разметки по сравнению с другими парсерами для русского языка [26]. Неологизмы и особенности лексики молодежного сленга отмечаются на отдельном уровне аннотации. В поле комментариев указываются внешние неречевые события, которые могут быть полезны для понимания происходящего.

3.6 Информационная система и утилиты обработки данных

Информация из всех модулей корпуса сводится в единую базу данных, обеспечивающую поиск и фильтрацию записей по запросу пользователя. Одновременно разрабатываются программные средства, осуществляющие ряд вспомогательных функций, таких как анонимизация личных имен и персональной информации, цензурирование (замена звездочками непечатной лексики), конвертация форматов аннотации (например, для перевода табличной транскрипционной записи в формат ELAN [18], используемый в корпусе ОРД), а также конвертация нестандартных форматов звуковых файлов, полученных от информантов.

4 Текущая статистика корпуса КУРС/ESC

К началу 2025 г. получены записи от 90 информантов в возрасте от 16 до 27 лет, средний возраст информантов составляет 22 года. В настоящее время информанты представляют студентов 10 вузов Санкт-Петербурга, юношей и девушек преимущественно гуманитарных специальностей. Примерно половина информантов родилась в Санкт-Петербурге. Среди других упомянутых мест рождения встречаются Красноярск (4,26 % информантов), Екатеринбург, Тюмень, Челябинск (по 3,19 %) и т.д. Кроме того, среди волонтёров, записавших свою речь для корпуса, есть граждане государств СНГ (Казахстана, Белоруссии, Кыргызстана и Узбекистана), при этом в большинстве случаев их родной язык — русский.

На настоящий момент суммарная длительность всех исходных записей составляет 1178 часов. Из них примерно 800 часов записано в 2024 году. В среднем от каждого информанта было получено приблизительно по 12 часов записи речи.

Автоматическое распознавание речи выполняется по мере сегментации. Примерно половина полученных записей прошла через процесс ручного сегментирования, благодаря чему было получено 952 макроэпизода. Экспертная проверка и коррекция транскриптов выполнена для 258 макроэпизодов, что соответствует 84 часам звучания. Сбор новых звукозаписей продолжается. Корпус КУРС/ESC планируется к публикации в открытом доступе для тех материалов, которые имеют разрешение на публикацию.

5 Демоверсия сайта

В настоящее время сайт корпуса КУРС/ESC работает в тестовом режиме по адресу https://esc-corpus.ru. Для отработки возможностей сайта отобраны 38 коммуникативных макроэпизодов, содержащих 12,85 часов речевого материала, полученного от 59 говорящих. Расшифровки для этих эпизодов содержат 59 162 словоформы, относящихся к 4601 реплике.

При отборе данных для сайта приоритетно отбирались звукозаписи, для которых есть разрешение для открытой публикации звукового контента, что дает возможность совместить расшифровку звукозаписей с их звуковым оригиналом[3].

Полученная подборка звукозаписей отражает многогранность повседневной жизни петербургских студентов, включая личное, профессиональное и социальное общение. Сюда входят голосовые сообщения друзьям, обсуждение фильмов, шопинг, разговоры с близкими, включая братьев, сестер, партнеров, друзей и подруг, а также взаимодействие с домашними питомцами и даже с самим собой. Представлены эпизоды учебной и профессиональной деятельности, (например, проведения уроков), а также обсуждения бытовых тем, таких как текущие расходы, питание, здоровье и косметика. Включены звукозаписи, сделанные в общественных местах, таких как кафе, магазины, транспорт, а также представлены моменты уединения, например, размышления вслух и игры в одиночестве.

На рис. 2 представлен принтскрин результатов поиска по словоформе/подстроке с возможностью просмотра кратких метаданных об участниках разговора и особенностях коммуникации.

[3] Публикация звуковых материалов запланирована на 3-й квартал 2025 года.

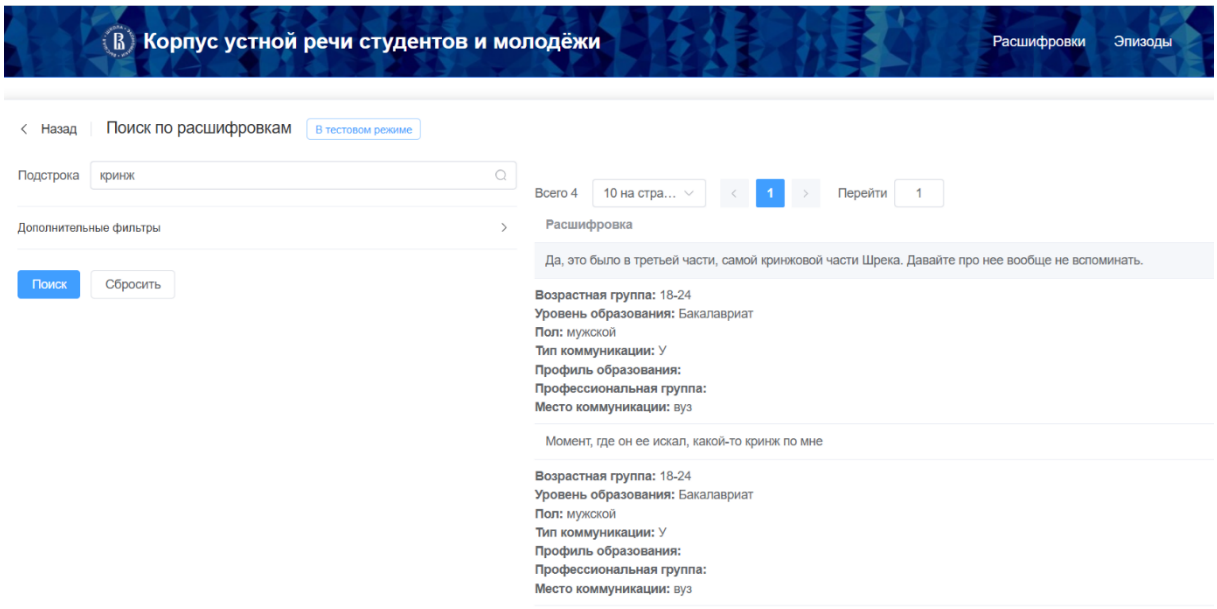


Рисунок 2: Поиск по подстроке/словоформам в расшифровках корпуса КУРС/ESC

На вкладке «Эпизоды» можно использовать фильтры по коммуникативным ситуациям (рис. 3), а также перейти к полным текстам расшифровок для каждого эпизода (рис. 4).

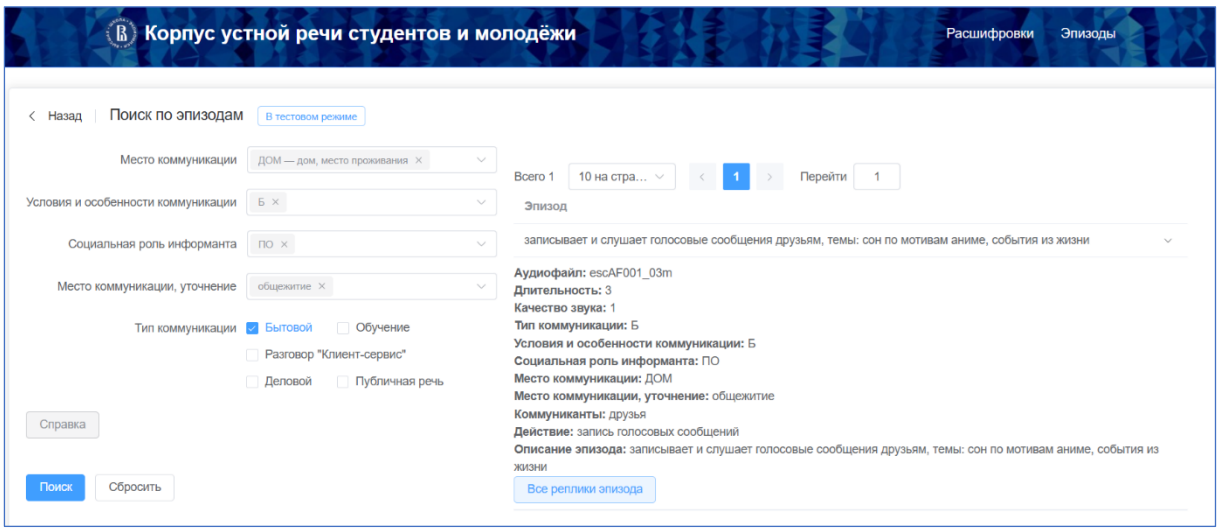


Рисунок 3: Поиск эпизодов по условиям коммуникации в корпусе КУРС/ESC с переходом на полный текст расшифровки (кнопка «Все реплики эпизода»)

Прости меня. Я думал, что я записывал аудио. Смотри, вы до начала кинопоказа лучше пообедайте, потому что столовая работает до семи. Вы подходите, говорите, что на вас заложено питание в рамках день ментального здоровья. У них там листочек в клеточку, список из десяти человек, среди них есть Эмма%, есть Ваня%. И просто расписываетесь, набираете на триста пятьдесят рублей все, что хотите.	›	XM001	00:02:07
Девочки, прикиньте, у меня сегодня питание в столовке на триста пятьдесят рублей, вот это вообще!	›	AF001	00:02:49
И как мне в этом качестве смотреть? Вы с дуба рухнули?	›	AF001	00:02:59

Рисунок 4: Выгрузка полного текста эпизода с указанием кода говорящего и временной метки (фрагмент)

В 2025 году поиск по транскриптам корпуса будет расширен за счет поиска по леммам и грамматическим категориям, а планируется выборочное размещение звукового контента. Количество эпизодов, представленных на сайте, также будет увеличено.

Корпус КУРС/ESC планируется расширять с использованием краудсорсинговых возможностей. Для этого на сайте будет организована возможность «донировать» собственные материалы (звукозаписи и их транскрипты), подобно тому, как это сделано в Американском национальном корпусе [20] или отечественных ресурсах: корпусе дневников и воспоминаний «Прожито» [21] и корпусе открыток «Пишу тебе» [22].

6 Возможности и научная новизна корпуса КУРС/ESC

Материалы корпуса КУРС предназначены для решения следующих актуальных задач:

1) Фиксировать ускользающую звучащую материю, отражающую как текущее состояние языка повседневного общения, так и принятые на данный момент в обществе особенности коммуникативного поведения. Звукозаписи повседневной речи могут стать ценным архивом для сохранения языкового и культурного наследия, документируя изменения в языке и коммуникативном поведении различных групп общества.

2) Служить основой для проведения фундаментальных исследований языка повседневного общения и устного дискурса в целом — изучать особенности фонетики, просодии, лексики, синтаксиса, семантики, прагматики, структуру и законы коммуникации, а также социолингвистическую вариативность этих явлений, гендерные и возрастные особенности, диалекты и жаргоны, а также изменения в языке под влиянием технологий и глобализации.

3) Применяться для решения целого спектра прикладных задач — от уже упомянутого обучения больших языковых моделей, систем распознавания и синтеза речи, обеспечивая модели аутентичными данными, отражающими разнообразие акцентов, интонаций и речевых паттернов, до создания реалистичных учебных материалов, которые помогают учащимся адаптироваться к живой, естественной речи, включая изучение интонации, ритма и типичных ошибок носителей. Данные повседневного общения могут использоваться для создания приближенных к реальности диалогов для чат-ботов и виртуальных сред самого разного назначения, где требуется высокая степень достоверности воссоздания речевых взаимодействий.

4) Использоваться для решения научных и практических задач в ряде дисциплин, изучающих поведение человека в обществе и взаимодействие социальных групп, таких как социология, антропология, культурология, а также психология, лингвистика и межкультурная коммуникация, предоставляя уникальные данные для анализа социальных и культурных процессов.

Научная новизна корпуса КУРС/ESC по сравнению с существующими корпусами русской повседневной речи заключается в следующем:

1) До настоящего времени не существовало специализированного корпуса, посвящённого современной студенческой и молодёжной устной речи, записанной в естественных условиях после 2020 года. С учётом стремительных языковых изменений, вызванных цифровизацией, глобализацией и влиянием соцсетей, КУРС/ESC фиксирует новую языковую реальность, которую невозможно смоделировать по более ранним корпусам (например, ОРД).

2) Совмещение полевых методов и автоматических технологий распознавания речи. Впервые для корпуса русской повседневной речи внедрена полуавтоматическая схема обработки данных, включающая комбинацию двух разных ASR-систем (одна — с языковой моделью, другая — без нее) с последующей экспертной верификацией, что позволяет ускорить процесс транскрипции.

3) Помимо лингвистической разметки и социолингвистической информации о респондентах, в корпус впервые включены результаты теста на пассивный словарь и пятифакторного опросника личности, что открывает новые перспективы для междисциплинарных исследований — в частности, для анализа взаимосвязей между языковым поведением и личностными характеристиками говорящих.

4) Благодаря методологической преемственности с корпусом ОРД, новый корпус позволяет проводить сопоставление данных с временным интервалом более 15 лет, что создает уникальную возможность отслеживания динамики развития повседневного дискурса в русской устной речи.

5) Проект ориентирован на открытое распространение данных, а также на формирование модели участия пользователей в наполнении корпуса. Такая модель ранее не применялась в русскоязычных аудиокорпусах и позволяет значительно расширить репрезентативность собранного материала.

7 Некоторые наблюдения о лексических особенностях речи современных студентов

Несмотря на тот факт, что корпус КУРС/ESC находится в состоянии разработки, полученные данные можно использовать для выявления наиболее характерных особенностей современной молодежной речи. Покажем это на примере лексики.

Материалом для представленного в данном разделе исследования послужили текстовые расшифровки 203 коммуникативных макроэпизодов [12] молодежной студенческой коммуникации, которые содержат в общей сложности 283643 словоупотреблений. Эти коммуникативные эпизоды были сначала автоматически расшифрованы с помощью акустической модели распознавания, разработанной российской компанией ООО НТР [14], после чего была выполнена их экспертная проверка. В выборке представлена речь 212 студентов (156 девушек и 56 юношей). Относительная несбалансированность корпусного материала по половому признаку носит временный характер и объясняется тем, что запись корпуса была начата на гуманитарных факультетах, где большинство студентов традиционно составляют девушки.

Рассмотрим наиболее частотную лексику молодежной коммуникации. В табл. 2 представлен список из 100 самых частотных слов (без лемматизации), которые покрывают 53% всей лексики в проанализированной выборке. Для каждого слова приводятся его абсолютная и относительная частоты, а также накопленная частота (кумулятивный процент). В качестве «реферативного корпуса» для сравнения используется выборка из корпуса ОРД, описанная в статье [24], состоящая из 152 макроэпизодов и содержащая речь 209 человек, записанную в Санкт-Петербурге в 2007 и 2010 гг. (общим объемом подкорпуса в 232370 словоупотреблениях), и соответствующий ей частотный список наиболее употребительных слов повседневной устной речи [там же].

Список самых частотных слов молодежной речи демонстрирует преобладание личных местоимений (*я*, *ты*, *он*, *она*, *мы* и их словоизменительных форм), служебных слов (союзов, предлогов, частиц), дискурсивных и прагматических маркеров, и во многом напоминает список наиболее частотной лексики, полученный на материале корпуса ОРД. Как и для языка повседневного общения в целом, значительную часть этого списка составляют дискурсивные и прагматические маркеры, играющие важную роль в структурировании и управлении дискурсом (такие как *вот*, *ну*, *так*, *короче*, *ладно*). Это говорит о том, что молодежный лексикон не только насыщен субъективными и эмоциональными элементами, но и активно использует средства организации речи, направленные на поддержание контакта и управления коммуникативным процессом.

Также заметно широкое использование частиц и вводных слов, придающих речи оттенки модальности и оценки: *просто*, *типа*, *кажется*, *наверное*, *конечно*. Они выполняют функции уточнения, смягчения высказывания или маркирования неуверенности, что типично для разговорной речи молодых людей, особенно в неформальной обстановке.

Ранг	Слово	Абс. ч.	%	Кумул.%	Ранг	Слово	Абс. ч.	%	Кумул.%
1	я	8350	2,94	2,94	51	тебе	755	0,27	44,60
2	не	7342	2,59	5,53	52	может	743	0,26	44,86
3	что	6618	2,33	7,87	53	за	727	0,26	45,12
4	ну	6144	2,17	10,03	54	можно	724	0,26	45,37
5	да	5725	2,02	12,05	55	тебя	716	0,25	45,63
6	вот	5514	1,94	13,99	56	тут	644	0,23	45,85
7	и	5367	1,89	15,89	57	потом	642	0,23	46,08
8	в	5216	1,84	17,73	58	такой	642	0,23	46,31
9	это	4867	1,72	19,44	59	к	585	0,21	46,51
10	а	4809	1,70	21,14	60	ой	585	0,21	46,72
11	у	3712	1,31	22,45	61	такая	585	0,21	46,93
12	там	3513	1,24	23,68	62	только	583	0,21	47,13
13	на	3281	1,16	24,84	63	его	578	0,20	47,33
14	то	3113	1,10	25,94	64	хорошо	554	0,20	47,53
15	как	3013	1,06	27,00	65	такое	539	0,19	47,72
16	ты	2711	0,96	27,96	66	короче	535	0,19	47,91
17	так	2706	0,95	28,91	67	из	531	0,19	48,10
18	все	2395	0,84	29,75	68	чтобы	527	0,19	48,28
19	с	2373	0,84	30,59	69	раз	518	0,18	48,46
20	мне	2049	0,72	31,31	70	где	515	0,18	48,65
21	он	1899	0,67	31,98	71	для	507	0,18	48,82
22	просто	1837	0,65	32,63	72	до	498	0,18	49,00
23	есть	1823	0,64	33,27	73	даже	487	0,17	49,17
24	нет	1715	0,60	33,88	74	была	473	0,17	49,34
25	меня	1700	0,60	34,48	75	конечно	472	0,17	49,51
26	мы	1625	0,57	35,05	76	что-то	466	0,16	49,67
27	она	1624	0,57	35,62	77	наверное	453	0,16	49,83
28	но	1508	0,53	36,15	78	быть	446	0,16	49,99
29	еще	1478	0,52	36,68	79	почему	441	0,16	50,14
30	типа	1377	0,49	37,16	80	этот	435	0,15	50,30
31	бы	1270	0,45	37,61	81	про	434	0,15	50,45
32	они	1268	0,45	38,06	82	о	424	0,15	50,60
33	угу	1252	0,44	38,50	83	кажется	419	0,15	50,75
34	по	1195	0,42	38,92	84	ничего	418	0,15	50,89
35	же	1174	0,41	39,33	85	блин	414	0,15	51,04
36	очень	1163	0,41	39,74	86	давай	407	0,14	51,18
37	если	1059	0,37	40,12	87	ее	407	0,14	51,33
38	уже	1051	0,37	40,49	88	знаешь	394	0,14	51,46
39	вообще	1030	0,36	40,85	89	был	391	0,14	51,60
40	знаю	1025	0,36	41,21	90	их	383	0,14	51,74
41	потому	1005	0,35	41,56	91	от	382	0,13	51,87
42	было	973	0,34	41,91	92	тогда	379	0,13	52,01
43	тоже	965	0,34	42,25	93	здесь	376	0,13	52,14
44	надо	957	0,34	42,59	94	такие	376	0,13	52,27
45	когда	951	0,34	42,92	95	два	375	0,13	52,40
46	вы	859	0,30	43,22	96	какой	373	0,13	52,53
47	нас	828	0,29	43,52	97	эти	372	0,13	52,67
48	сейчас	800	0,28	43,80	98	ладно	368	0,13	52,80
49	или	766	0,27	44,07	99	вас	361	0,13	52,92
50	будет	758	0,27	44,33	100	нужно	353	0,12	53,05

Таблица 2: Наиболее употребительные слова устной студенческой речи

В таблице 3 представлены списки наиболее частотных слов, по употреблению которых современная студенческая речь больше всего отличается от общего частотного списка [24]. Первые четыре столбца показывают списки слов, ранг или доля которых в молодежной речи существенно больше, чем для общего списка. Любопытно, что в этих группах слов лидируют два — *типа*[4] и *короче*[5], которые в 2016 г. возглавляли список слов, характерных для мужской речи [там же].

Частотная лексика, чаще встречающаяся в студенческой речи				**Частотная лексика, реже встречающаяся в студенческой речи**			
По разности рангов		**По разности долей**		**По разности рангов**		**По разности долей**	
типа	>50	что	5520	значит	-120	вот	-3846
короче	34	типа	>4855	говорит	-86	так	-2690
из	30	я	3832	чего	-49	а	-2511
потому	28	то	3310	здесь	-46	там	-2342
для	22	просто	2710	говорю	-44	да	-2031
такая	19	не	2646	вам	-32	сейчас	-1987
быть	18	очень	1720	ага	-28	угу	-1830
такое	17	была	1668	ли	-23	значит	-1811
чтобы	17	мне	1660	давай	-22	ну	-1608
просто	16	потому	1611	вас	-20	говорит	-1468
очень	15	наверное	1597	сейчас	-20	нет	-1386
даже	15	почему	1555	знаешь	-18	здесь	-1157
можно	14	как	1543	один	-16	говорю	-852
когда	13	про	1530	кто	-11	надо	-727
раз	12	о	1495	эти	-10	чего	-705
до	11	ты	1480	угу	-10	ага	-613
или	10	кажется	1477	ничего	-10	давай	-579
за	10	ее	1435	надо	-10	знаешь	-543

Таблица 3:Динамика встречаемости наиболее частотной лексики (топ-100) в речи современных студентов по сравнению с повседневной речью корпуса ОРД

Также можно отметить увеличение частоты *такая* и *такой* в функции ксенопоказателей (*я такая…* и т. п.), которое закономерно снижает долю *говорю* и *говорит*. В речи студентов уменьшается роль метакоммуникатива *знаешь*, прагматического маркера *значит*, маркеров обратной связи *угу* и *ага*, частиц *давай* и *да*, а также делимитативного маркера *вот*. *Здесь* и *сейчас* также, неожиданно, теряют популярность в молодежной среде.

Возможно, часть отмеченных особенностей касаются не только молодежной речи, но и языковых тенденций в целом. Однако для уточнения этого вопроса необходимо расширение корпуса с привлечением информантов всех возрастных групп.

Характерной для молодежной студенческой речи является употребление английских заимствований. Ниже приводится список наиболее частотных заимствований с указанием абсолютной частоты и ipm (в скобках), их можно условно поделить на 4 группы:

1) лексика, связанная с передачей эмоций и реакций: *окей* — 42 (148), *вау* — 19 (67), *трэш* — 6 (21), *кринж* — 30 (106), *кринжовый* — 5 (18),

2) лексика, относящаяся к обучению, коучингу и бизнесу: *коучи* — 27 (95), *коуч* — 24 (85), *коучинг* — 14 (49), *коучинга* — 7 (25), *коворкинг* — 6 (21), *прокторинг* — 6 (21), *оффер* — 9 (32),

3) лексика, связанная с интернет-активностями и мультимедиа: *видос* — 22 (78), *мем* — 17 (60), *загуглить* — 11 (39), *ватсап* — 10 (35), *рилс* — 7 (25), *сторис* — 7 (25),

4) лексика, связанная с модой и молодежной культурой: *свитшот* — 14 (49), *бейби* — 5 (18), *вайб* — 5 (18).

[4] https://www.ord-multimedia-dict.com/tipa

[5] https://www.ord-multimedia-dict.com/koroche

Следует отметить, что некоторые из этих слов, такие как *окей*, *вау*, *загуглить*, *оффер* и др. в настоящее время используются не только в молодежной студенческой речи, но и в более широком контексте современного общения, выходя за рамки возрастных и социальных групп.

Представленную в данном разделе информацию следует рассматривать как предварительную. По мере увеличения объема расшифрованной части корпуса эти данные будут скорректированы.

8 Заключение

Создание звукового корпуса повседневной речи, собранной «полевым способом», представляет собой достаточно трудоемкий и неизбежно долгий процесс. На настоящем этапе работы в фокусе интереса разработчиков находится в первую очередь лексико-грамматические, социолингвистические и прагматические характеристики молодежного дискурса. Фонетические и просодические особенности устной речи пока остаются за кадром нашей работы. Мультимедийный характер исходного материала позволяет расширять корпус и в этом направлении. Представляется, что для аннотирования речевых данных на фонетическом уровне также уже можно использовать современные речевые технологии — от акустических моделей, записывающих речь «как она слышится», что позволяет получить упрощенную фонетическую транскрипцию [23], до использования полифункциональных систем извлечения речевых характеристик речи, таких как openSMILE [25].

Совмещение технологий новейших систем распознавания и обработки речи с корпусными методами представляется более чем перспективным. Результаты такого технологического объединения позволят не только вывести традиционные корпусные лингвистические исследования на уровень «больших данных» в том числе и для устной речи, но и смогут служить в качестве обучающих данных для больших языковых моделей и создания сильного искусственного интеллекта [1], [2], [3]. В данном случае «большие данные» относятся не к текущим масштабам создаваемого ресурса, а к особенностям его структуры и потенциала, предназначенного для восполнения нишевого дефицита русскоязычных разговорных ресурсов, необходимых для развития прикладных речевых технологий.

Корпус КУРС/ESC представляет собой многоуровневый массив аутентичной повседневной русской речи с богатыми аннотациями (социо-, психолингвистическими, лексико-грамматическими), который может использоваться: 1) для дообучения и тестирования моделей автоматического распознавания речи (ASR) на реальных, зашумленных и спонтанных данных; 2) как источник редуцированной, разговорной и сленговой лексики для адаптации языковых моделей и синтеза речи; 3) в качестве «золотого стандарта» для оценки качества NLP-систем, работающих с устной речью. По количеству записанного речевого материала этот ресурс уже превосходит известный CSJ корпус спонтанной японской речи, который содержит 650 часов звукозаписей [27].

Запись повседневной русской речи методом суточной фиксации проводится в Санкт-Петербурге спустя 15 лет после первых записей корпуса ОРД и основана на схожей методологии. Это позволяет выстраивать серии записей этих двух корпусов в единую лонгитюдную цепочку, предоставляя уникальную возможность отслеживать изменения и развитие повседневного русского дискурса. Представляется, что подобные серии звукозаписей могут быть продолжены в будущем, что обеспечит возможность долгосрочного мониторинга и анализа динамики языковых и коммуникативных явлений.

Благодарности

Статья подготовлена в рамках проекта «Текст как Big Data: методы и модели работы с большими текстовыми данными» (ФИ-2025-32) для выполнения фундаментальных проектов тематического плана научных исследований, предусмотренных Государственным заданием Национального исследовательского университета «Высшая школа экономики» на 2025 год.

References[6]

[1] Muehlhauser Luke. What Is AGI? // Machine Intelligence Research Institute. — August 11, 2013. — Access mode: https://intelligence.org/2013/08/11/what-is-agi/.

[6] References, Scopus version

[2] Antebi L. General Artificial Intelligence // Artificial Intelligence and National Security in Israel. — Institute for National Security Studies, 2021. — Pp. 59–60.
[3] Chim J., Ive J., Liakata M. Evaluating Synthetic Data Generation from User Generated Text // Computational Linguistics. — 2024. — Pp. 1–44.
[4] Shibamoto Janet S. Japanese Sociolinguistics // Annual Review of Anthropology. — 1987. — Vol. 16. — Pp. 261–278.
[5] Shibata Takeshi. Study of Language Behavior over 24 Hours [Issledovanie yazykovogo sushchestvovaniya v techenie 24 chasov]. Translated by V.M. Alpatov // Alpatov V.M. (ed.). Linguistics in Japan [Yazykoznanie v Yaponii]. — Moscow: Raduga, 1983. — Pp. 134–141.
[6] Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). — Eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. — Lisbon, 2004. — Pp. 183–186.
[7] Reference Guide for the British National Corpus. — Access mode: http://www.natcorp.ox.ac.uk/docs/URG.xml. — Last accessed: 10/01/2025.
[8] Everyday Speech Corpus (ORD) Homepage. — Access mode: https://ord.spbu.ru/.
[9] Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // TSD 2009. — Eds. Matoušek, V., Mautner, P. — Lecture Notes in Artificial Intelligence (LNAI), vol. 5729. — Springer, Berlin-Heidelberg, 2009. — Pp. 250–257.
[10] Sherstinova T., Kolobov R., Mikhaylovskiy N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level // SPECOM 2023. — LNCS, vol. 14338/14339. — Springer, 2023. — Pp. 43–56.
[11] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9811. — Springer, Switzerland, 2016. — Pp. 659–666.
[12] Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus // SPECOM 2015. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9319. — Springer, Switzerland, 2015. — Pp. 268–276.
[13] Whisper Model Homepage. — Access mode: https://openai.com/index/whisper/.
[14] NTR Homepage. — Access mode: https://ntr.ai. — Last accessed: 2024/07/25.
[15] Kolpashchikova E.O. Writer Robin Dranattagor: Testing the Whisper Model on Russian-Language Spoken Speech [Pisatel' Robin Dranattagor: aprobatsiya modeli Whisper na russkoyazychnoi zvuchashchei rechi] // Socio- and Psycholinguistic Studies [Socio- i psikholingvisticheskie issledovaniya]. — 2023. — Issue 11. — Pp. 23–27.
[16] Khromov A.B. Five-Factor Personality Questionnaire: A Training Manual [Pyatifaktornyi oprosnik lichnosti. Uchebno-metodicheskoe posobie]. — Kurgan: KSU, 2000.
[17] Golovin G.V. Measuring Passive Vocabulary in Russian [Izmerenie passivnogo slovarnogo zapasa russkogo yazyka] // Socio- and Psycholinguistic Studies [Socio-i psikholingvisticheskie issledovaniya]. — No. 3. — 2015. — Pp. 148–159.
[18] ELAN Homepage. — Access mode: http://tla.mpi.nl/tools/tla-tools/elan. — Last accessed: 2024/07/25.
[19] MyStem Homepage. — Access mode: https://yandex.ru/dev/mystem/.
[20] American National Corpus (ANC) Homepage. — Access mode: https://anc.org/contribute/texts/.
[21] Prozhito Project Homepage [Proekt "Prozhito"]. — Access mode: https://prozhito.org/.
[22] Pishu Tebe Project Homepage [Proekt "Pishu tebe"]. — Access mode: https://pishutebe.ru/.
[23] Sherstinova T., Mikhaylovskiy N., Kolpashchikova E., Kruglikova V. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations // 35th Conference of Open Innovations Association (FRUCT). — Tampere, Finland, 2024, April 24–26. — Pp. 253–258.
[24] Шерстинова Т. Ю. Наиболее употребительные слова повседневной русской речи (в гендерном аспекте и в зависимости от условий коммуникации) [The Most Frequent Words in Everyday Spoken Russian (in the gender dimension and depending on communication settings)] // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Москва, 1-4 июня 2016 г.). Вып. 15 (22). М.: Изд-во РГГУ, 2016. С. 616-631.
[25] Eyben F., Weninger F., Gross F., Schuller B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor // Proceedings of the 21st ACM International Conference on Multimedia (MM '13). — 2013. — Pp. 835–838.
[26] Kotelnikov E., Razova E., Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? // Proceedings of the Conference on Artificial Intelligence and Natural Language — 2018. — Pp. 131–142.
[27] The corpus of Spontaneous Japanese: https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html — Last accessed: 28/03/2025.

References[7]

[1] Muehlhauser Luke. What Is AGI? // Machine Intelligence Research Institute. — August 11, 2013. — Access mode: https://intelligence.org/2013/08/11/what-is-agi/.
[2] Antebi L. General Artificial Intelligence // Artificial Intelligence and National Security in Israel. — Institute for National Security Studies, 2021. — Pp. 59–60.
[3] Chim J., Ive J., Liakata M. Evaluating Synthetic Data Generation from User Generated Text // Computational Linguistics. — 2024. — Pp. 1–44.
[4] Shibamoto Janet S. Japanese Sociolinguistics // Annual Review of Anthropology. — 1987. — Vol. 16. — Pp. 261–278.
[5] Сибата Такэси. Исследование языкового существования в течение 24 часов. Перевод В. М. Алпатова // Алпатов В. М. (сост.) Языкознание в Японии. Общая редакция И. Ф. Вардуля. — М.: Радуга, 1983. — С. 134–141.
[6] Campbell N. Speech & Expression; the Value of a Longitudinal Corpus // Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04). — Eds. M.T. Lino, M.F. Xavier, F. Ferreira, R. Costa, R. Silva. — Lisbon, 2004. — Pp. 183–186.
[7] Reference Guide for the British National Corpus. — Access mode: http://www.natcorp.ox.ac.uk/docs/URG.xml. — Last accessed: 10/01/2025.
[8] Everyday Speech Corpus (ORD) Homepage. — Access mode: https://ord.spbu.ru/.
[9] Asinovsky A., Bogdanova N., Rusakova M., Ryko A., Stepanova S., Sherstinova T. The ORD Speech Corpus of Russian Everyday Communication "One Speaker's Day": Creation Principles and Annotation // TSD 2009. — Eds. Matoušek, V., Mautner, P. — Lecture Notes in Artificial Intelligence (LNAI), vol. 5729. — Springer, Berlin-Heidelberg, 2009. — Pp. 250–257.
[10] Sherstinova T., Kolobov R., Mikhaylovskiy N. Everyday Conversations: a Comparative Study of Expert Transcriptions and ASR Outputs at a Lexical Level // SPECOM 2023. — LNCS, vol. 14338/14339. — Springer, 2023. — Pp. 43–56.
[11] Bogdanova-Beglarian N., Sherstinova T., Blinova O., Ermolova O., Baeva E., Martynenko G., Ryko A. Sociolinguistic Extension of the ORD Corpus of Russian Everyday Speech // SPECOM 2016. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9811. — Springer, Switzerland, 2016. — Pp. 659–666.
[12] Sherstinova T. Macro Episodes of Russian Everyday Oral Communication: towards Pragmatic Annotation of the ORD Speech Corpus // SPECOM 2015. — Lecture Notes in Artificial Intelligence (LNAI), vol. 9319. — Springer, Switzerland, 2015. — Pp. 268–276.
[13] Whisper Model Homepage. — Access mode: https://openai.com/index/whisper/.
[14] NTR Homepage. — Access mode: https://ntr.ai. — Last accessed: 2024/07/25.
[15] Kolpashchikova E.O. Pisatel' Robin Dranattagor: aprobatsiya modeli Whisper na russkoyazychnoi zvuchashchei rechi // Socio- i psikholingvisticheskie issledovaniya (Socio- and Psycholinguistic Studies). — 2023. — Issue 11. — Pp. 23–27.
[16] Хромов А. Б. Пятифакторный опросник личности. Учебно-методическое пособие. — Курган: КГУ, 2000.
[17] Головин Г. В. Измерение пассивного словарного запаса русского языка // Социо-и психолингвистические исследования. — № 3. — 2015. — С. 148–159.
[18] ELAN Homepage. — Access mode: http://tla.mpi.nl/tools/tla-tools/elan. — Last accessed: 2024/07/25.
[19] MyStem Homepage. — Access mode: https://yandex.ru/dev/mystem/.
[20] American National Corpus (ANC) Homepage. — Access mode: https://anc.org/contribute/texts/.
[21] Проект «Прожито» Homepage. — Access mode: https://prozhito.org/.
[22] Проект «Пишу тебе» Homepage. — Access mode: https://pishutebe.ru/.
[23] Sherstinova T., Mikhaylovskiy N., Kolpashchikova E., Kruglikova V. Bridging Gaps in Russian Language Processing: AI and Everyday Conversations // 35th Conference of Open Innovations Association (FRUCT). — Tampere, Finland, 2024, April 24–26. — Pp. 253–258.
[24] Sherstinova T. The Most Frequent Words in Everyday Spoken Russian (in the gender dimension and depending on communication settings) // Komp'juternaja Lingvistika i Intellektual'nye Tehnologii, pp. 616 - 631.
[25] Eyben F., Weninger F., Gross F., Schuller B. Recent developments in openSMILE, the Munich open-source multimedia feature extractor // Proceedings of the 21st ACM International Conference on Multimedia (MM '13). — 2013. — Pp. 835–838.
[26] Kotelnikov E., Razova E., Fishcheva I. A Close Look at Russian Morphological Parsers: Which One Is the Best? // Proceedings of the Conference on Artificial Intelligence and Natural Language — 2018. — Pp. 131–142.
[27] The corpus of Spontaneous Japanese: https://clrd.ninjal.ac.jp/csj/misc/preliminary/index_e.html — Last accessed: 28/03/2025.

[7] References, РИНЦ version

Evaluating the Pragmatic Competence of Large Language Models in Detecting Mitigated and Unmitigated Types of Disagreement

Valery Shulginov
Laboratory of Linguistic Conflict Resolution Studies and Contemporary Communicative Practices,
HSE University / Moscow, Russia
MIPT / Moscow, Russia
shulginov.va@mipt.ru

Hasan Berkcan Şimşek
Centre for Sociocultural and Ethnolinguistic Studies,
HSE University / Moscow, Russia
hasanberkcansimsek@gmail.com

Sergei Kudriashov
HSE University / Moscow, Russia
xenomirant@gmail.com

Renata Randautsova
Laboratory of Linguistic Conflict Resolution Studies and Contemporary Communicative Practices,
HSE University / Moscow, Russia
randovtsovarn@gmail.com

Sofya A. Shevela
HSE University / Moscow, Russia
MIPT / Moscow, Russia
sofyashevela@gmail.com

Abstract

This study presents a framework for evaluating the effectiveness of language models (LLMs) in detecting disagreement across a wide range of pragmatic strategies, from mitigated forms to overt verbal aggression. Special attention is given to complex cases of implicit manifestations of irony and sarcasm, which pose significant challenges for both automated analysis and interpersonal communication. Experimental testing of LLMs was conducted in two types of tasks: binary classification for identifying disagreement and classification of specific strategies for its expression. The results showed that large multilingual models outperformed other models, especially in binary classification. However, models that focus primarily on the Russian language, such as GigaChat and YaGPT, tend to interpret irony and sarcasm more accurately and have a higher result density. Comparative analysis with human judgments revealed that, despite progress, the accuracy of sarcasm detection by LLMs still lags significantly behind human judgments. The results suggest a need for further optimization of LLMs to improve their pragmatic competence in real communicative situations.

Keywords: LLMs evaluation; benchmark; pragmatics; disagreement; irony; sarcasm; QUD
DOI: 10.28995/2075-7182-2025-23-345-360

Оценка прагматической компетенции больших языковых моделей в выявлении смягченных и несмягченных типов несогласия

Валерий Шульгинов
Лаборатория лингвистической конфликтологии и современных коммуникативных практик,
НИУ ВШЭ / Москва, Россия
МФТИ / Москва, Россия
shulginov.va@mipt.ru

Хасан Беркджан Шимшек
Центр социокультурных и этноязыковых исследований факультета гуманитарных наук,
НИУ ВШЭ / Москва, Россия
hasanberkcansimsek@gmail.com

Сергей Кудряшов
НИУ ВШЭ/ Москва, Россия
xenomirant@gmail.com

Рената Рандовцова
Лаборатория лингвистической конфликтологии и современных коммуникативных практик,
НИУ ВШЭ / Москва, Россия
randovtsovarn@gmail.com

Софья Шевела
НИУ ВШЭ / Москва, Россия
МФТИ / Москва, Россия
sofyashevela@gmail.com

Аннотация

В настоящем исследовании представлена методика оценки эффективности больших языковых моделей (БЯМ) в области выявления несогласия, включая широкий диапазон стратегий его выражения — от смягченных форм до явной вербальной агрессии. Особое внимание уделяется сложным случаям имплицитных проявлений иронии и сарказма, представляющим значительные трудности как для автоматического анализа, так и для межличностного общения. Экспериментальное тестирование БЯМ проводилось в двух типах задач: бинарная классификация для идентификации несогласия и классификация конкретных стратегий его выражения. Результаты показали, что большие мультиязычные модели демонстрируют преимущество над другими моделями, особенно в рамках бинарной классификации. Тем не менее БЯМ, ориентированные преимущественно на русский язык, например, GigaChat и YaGPT, склонны более точно интерпретировать иронию и сарказм и характеризуются более высокой плотностью результатов. Сравнительный анализ с оценками ассесоров показал, что несмотря на достигнутый прогресс, точность определения сарказма у БЯМ по-прежнему существенно уступает человеческим оценкам. Результаты исследования указывают на необходимость дальнейшей оптимизации БЯМ для повышения их прагматической компетенции в реальных коммуникативных ситуациях.

Ключевые слова: оценка БЯМ; бенчмарк; прагматика; несогласие; ирония; сарказм; вопрос над дискуссией

1 Introduction

In recent years, there has been an accelerated advancement in artificial intelligence systems and large language models (LLMs). They are increasingly employed in tasks involving direct human interaction through chatbots, search engines, and web browsers (Min et al., 2023; Sravanthi et al., 2024). The rapid integration of LLMs into everyday life raises critical questions regarding the assessment of these models' communicative skills, particularly their ability to interpret not only the literal and direct meaning of an utterance, but also those expressed more implicitly through various pragmatic strategies.

In this context, a pivotal skill warranting thorough examination is pragmatic competence, which refers to the ability to understand and use language effectively in specific social situations (see Taguchi, 2009). Previous studies have assessed individual pragmatic skills such as the understanding of implicatures and presuppositions (Qi et al., 2023; Hu et al., 2023; Ruis et al., 2023) and the interpretation of figurative language (Tong et al., 2021; Liu et al., 2022; Gu et al., 2022). In a large-scale evaluation, the Pragmatic Understanding Benchmark (PUB) (Sravanthi et al., 2024) is introduced, which is composed of a total of

fourteen tasks focusing on four major pragmatic phenomena: implicature, presupposition, reference, and deixis. As a result, aggregators of pragmatic tasks are formed, allowing for a comprehensive study of LLM competencies.

In this vein, we ground our approach in politeness theory, which posits that adult members of society navigate social interactions to maintain 'face'—their desired public self-image. As Brown and Levinson (1978/1987, p. 61) note, "people cooperate (and assume each other's cooperation) in maintaining face in interaction, such cooperation being based on the mutual vulnerability of face". Face consists of a positive aspect (the need to be accepted, respected, or valued by others) and a negative aspect (the desire for autonomy and freedom from imposition). In communication, various politeness strategies are employed to mitigate threats to these face needs, such as avoiding direct demands (protecting negative face) or offering compliments (reinforcing positive face).

Within this framework, we center our analysis on disagreement, as it exemplifies a speech act that can often trigger confrontation. In some contexts, disagreement is "consensual" (Bolander & Locher, 2017, p. 608). It is a natural and expected response that can enhance social interaction (Schiffrin, 1984), especially when it is preferred or permitted (Kakava, 2002). However, disagreement often "does not leave us cold," as it evokes negative emotions such as annoyance, anger, or contempt, and can lead to conflict (Langlotz & Locher, 2012, p. 1591). In its most basic form, disagreement involves the expression of an opposing opinion without intending negativity, using straightforward and unmitigated phrases that convey a contrary stance, such as 'I don't agree', 'I don't think so', or simply 'no', which can be termed direct disagreement. However, disagreement can also be expressed in many other ways. Understanding and recognizing different strategies for expressing disagreement can enable models to interact more effectively with users and to anticipate the direction of communication.

Assessing the degree of consensus and conflict in disagreement requires considering both the speaker's verbal behavior and an understanding of normative elements, relationships between parties, and community-level expectations (Angouri & Locher, 2012; Bolander & Locher, 2017). In this paper, we limit our analysis to verbal strategies used in disagreement, without addressing extra-linguistic cues related to consensus and conflict. Our goal is to evaluate how effectively different LLMs detect disagreement at the linguistic level, particularly when encountering its various forms of expression. To achieve this, we develop a nuanced benchmark that captures the distinct linguistic strategies used in disagreement.

2 Various Types of Disagreement

As noted by Benz and Jasinskaja (2017), each sentence in discourse typically responds to a (often implicit) Question Under Discussion (QUD), either by providing an answer or by introducing another question that helps in addressing the original one. The way a sentence is structured and interpreted is often influenced by the QUD it engages with. Similarly, each turn in a conflictual sequence addresses a specific QUD – whether directly answering the existing one or shifting it. For example, when Speaker A makes an arguable statement in Turn 1, they implicitly raise a QUD that Speaker B may either agree or disagree with in Turn 2. The structure of disagreement is shaped by the evolving QUDs, with each turn potentially redefining the scope of the issue at hand. While Turn 3 and subsequent turns are also important for the development of potential conflict episodes, Turns 1 and 2 constitute the minimum necessary sequence for the realization of a disagreement. Therefore, the two-turn structure forms the foundational pattern essential for LLMs to identify disagreement and its potential progression into conflict. For this reason, this study limits its analysis to the two-turn structure of disagreement, focusing on how the QUD initiated in Turn 1 is responded to in Turn 2.

Linguistic strategies in disagreement can range from unmitigated to mitigated (Kakava, 1993a, as cited in Kakava, 2002). Unmitigated disagreement is a direct and explicit expression of opposing viewpoints, with little or no attempt to soften the statement, making it potentially face-threatening. Choosing unmitigated disagreement can reflect a desire to be rude, disruptive, or hurtful (Locher, 2004), but it can also occur in contexts of consensual disagreement, where the primary concern is to engage in debate and defend one's opinions (e.g., Kotthoff, 1993). On the other hand, mitigated types of disagreements involve using politeness strategies to soften the potential face-threatening effect of the disagreement.

To develop a typology of disagreement strategies suitable for benchmarking, we first conducted a systematic literature review focusing on studies that propose taxonomies of disagreement and prioritize

comprehensive classification within their respective contexts (e.g., Culpeper, 1996; Muntigl & Turnbull, 1998; Locher, 2004; Shum & Lee, 2013). The aim was to synthesize different patterns of disagreement and establish a consolidated framework reflecting well-documented disagreement types. This approach prioritizes cross-contextual robustness to support benchmarking applications, though we recognize that domain-specific adaptations and culturally situated disagreement practices may extend beyond the scope of this typology. Disagreement types used in the benchmark are outlined below (see Appendix 1 for two-turn examples for each type):

Polite disagreement. A person mitigates their speech to avoid or reduce the face-threatening effect. Mitigation strategies include using hedges (e.g., 'well', 'I think'), modal auxiliaries (e.g., 'may', 'could'), emphasizing subjectivity through personal reasoning (e.g., 'it's just hard for me') (Locher, 2004, pp. 114-133), or concessive constructions such as 'yes, but…' (Uzelgun et al., 2015). Additionally, one can offer an alternative claim without directly opposing (cf. *counterclaim*, Muntigl & Turnbull, 1998), express regret, be less direct (Leech, 1983), or frame disagreement as questions (Locher, 2004).

Interrogative Disagreement. A person uses interrogatives to challenge a claim, usually by demanding evidence and implying that the other party is unable to do so (cf. *challenge*, Muntigl & Turnbull, 1998; *objections in the form of a question*, Locher, 2004; *raising rhetorical questions*, Shum & Lee, 2013) through statements like "What do you know about it?" (Langlotz & Locher, 2012, p. 1594).

Repetitive Disagreement. A person shows disagreement by repeating or reconstructing previous comments. Repetition can emphasize negative attitudes and increase imposition on the listener, boosting impoliteness (Bousfield, 2008, see also Locher, 2004, p. 123). It helps sustain and "anchor" disagreement (Kakava, 2002, p. 1560).

Referential disagreement. A person bolsters their disagreement by drawing on personal experiences and/or presenting external evidence, such as expert opinions or statistics (cf. *giving personal experience* and *giving facts*, Shum & Lee, 2013).

Irrelevance-Based Disagreement. A person communicates to another through a meta-dispute-act that their previous statement is not relevant to the current discussion, using expressions such as "It doesn't matter" or "You're straying off topic" (Muntigl & Turnbull, 1998, p. 229). This type of disagreement challenges the social skill of making relevant claims (Langlotz & Locher, 2012).

Reprimand and Profanity. Strong disapproval is expressed through reprimands or profanity, fostering conflict (cf. *using short vulgar phrases* and *reprimanding, giving negative comments*, Shum & Lee, 2013). This involves asserting one's stance as the correct one, while accusing the other party in a personalized manner, typically using pronouns like 'I' and 'you/your' to cast them as wrongdoers.

Ironic Disagreement. An ironic remark passes a negative judgment, but it directs the disagreement not at the interlocutor but at the situation as a whole, does not involve a face-threatening act, is not aggressive, and can be open to interpretation (Witek 2022). For instance, the statement "What nice weather!" can be made when it is windy and pouring rain.

Sarcastic Disagreement. A person makes a sarcastic remark in response to another person's statement to express disagreement (cf. *sarcasm or mock politeness*, Culpeper, 1996). It is a face-threatening act "with the use of politeness strategies that are obviously insincere" (Culpeper, 1996, p. 356). Sarcasm is perceived as more deliberate than irony, as it typically has a target, is more aggressive and offensive, and is conveyed in a cutting tone that is rarely equivocal (Lee, 1998; Reyes et al., 2013; Aniruddha et al., 2015; Sykora et al., 2020).

3 Methodology

Our dataset involved eight categories of tasks that encompass various strategies for expressing disagreement, with each category represented by a set of 50 assignments. We evaluated the capabilities of several LLMs using our benchmark in a zero-shot setting. Prompts consisted of two parts: a) the task description with output options provided as numbers or distinct single words, b) a minimal communication context for model analysis and evaluation. The configuration of the test dataset was as follows:

Polite Disagreement (PolD) involved responses where disagreement is mitigated through politeness strategies, with the objective for LLMs to identify phrases containing disagreement.

Interrogative Disagreement (InterD) featured responses in the form of questions, where only some express disagreement, requiring identification of such intent.

Repetitive Disagreement (RepD) included consistently rephrased responses, with only a portion indicating disagreement, aiming to identify these instances.

Referential Disagreement (RefD) encompassed responses containing references to documents, past experiences, or general knowledge, with some expressing disagreement, necessitating LLMs to identify disagreement.

Irrelevance-Based Disagreement (IrrBD) involved responses expressing disagreement with the QUD or indicating its irrelevance, requiring classification by LLMs.

Mitigated, Direct, Reprimand, and Profanity (MDRP) always contained disagreement, which was either mitigated, direct, or reprimanding, with the objective to categorize disagreement.

Ironic Disagreement (IronyD) always contained disagreement, which can be ironic or direct, necessitating identification of irony.

Sarcastic Disagreement (SarcD) always contained disagreement, which was either sarcastic or direct, requiring LLMs to identify sarcasm.

The dataset was divided into two parts: (1) tasks that required a binary choice between the presence and absence of disagreement (Pold, InterD, RepD, RefD), and (2) tasks where disagreement was present but required correct classification (IrrBD, MDRP, IronyD, SarcD).

The models under investigation included *Claude 3.5 Sonnet*, *Gemini Pro*, *GPT-4o*, *Llama-3 405B* (Grattafiori et al., 2024), *Mistral Nemo*, *GigaChat-Pro*, *YandexGPT-Pro*, and *Saiga Llama-3 70B*. Models were prompted via the respective APIs with default generation parameters and *temperature set to zero* in order to ensure greedy generation of specified answers. We are aware of the potential compromises of this approach, as models' performance may be subject to change dependent on various context and prompt parameters. However, as we are more interested in evaluating discriminative abilities of the model at identifying certain pragmatic patterns, and various sampling strategies, despite their ability to provide more thorough investigation of real-world performance, tend to get lower scores on most discriminative tasks (Song et al., 2024), we have chosen to lean towards deterministic generation with a prompt fixed for all the models. As some APIs don't provide access to output logits, the answers were evaluated in a generative approach by parsing them for the correct output structure and target answer structure, specified in the prompt. Also, as the tasks in the benchmark required either binary classification or ternary classification (for one task) with balanced class distributions, we measured accuracy as the generalization score and normalized Levenshtein similarity as the measure of ability to follow the provided instructions and produce structured output. Testing was conducted in September 2024 with the latest model versions provided via API.

To evaluate the consistency between the LLMs and human judgments, sarcasm identification, a particularly interesting and debatable task, was selected. A survey was conducted using tasks from the benchmark, where 11 evaluators assessed whether the provided context contained a sarcastic disagreement or not across 50 samples. The evaluators, mostly university students fluent in Russian, were presented with the same task description used for zero-shot LLM prompting.

4 Results and Analysis

4.1 Model evaluation

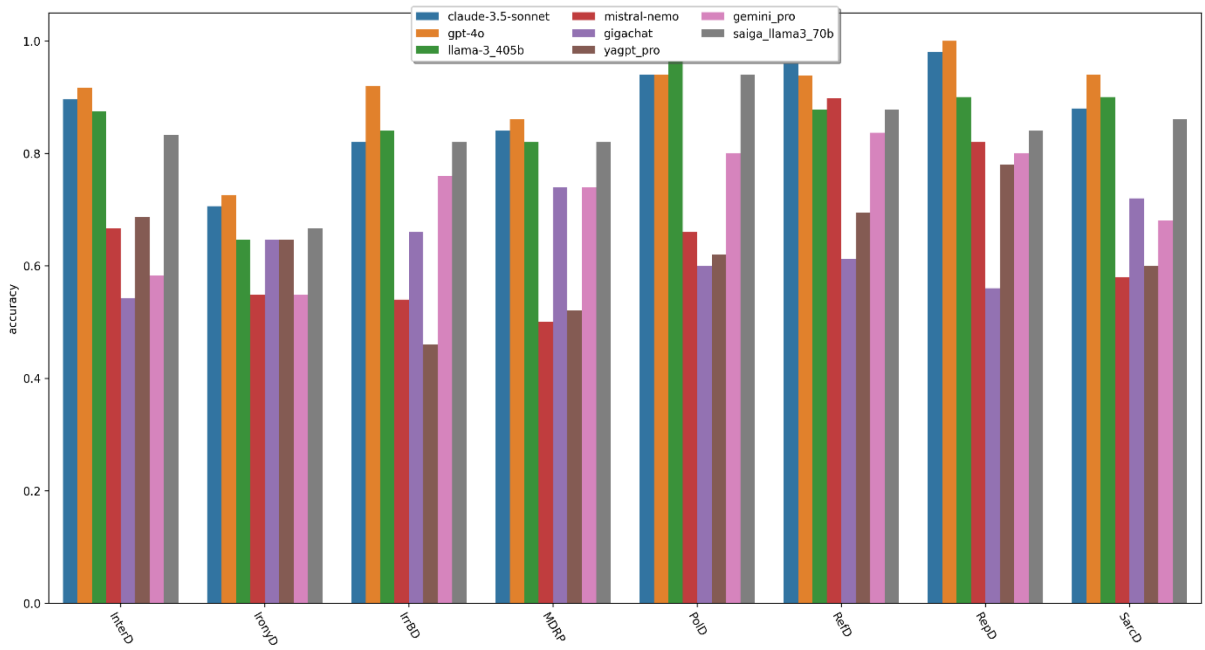


Figure 1: Overall accuracy per task

	claude-3.5-sonnet	gpt-4o	llama-3-405b	mistral-nemo	gigachat	yagpt-pro	gemini-pro	saiga-llama3-70b
Interrogative Disagreement	0.896	0.917	0.875	0.667	0.542	0.688	0.583	0.833
Ironic Disagreement	0.706	0.725	0.647	0.549	0.647	0.647	0.549	0.667
Irrelevancy claim	0.820	0.920	0.840	0.540	0.660	0.460	0.760	0.820
Mitigated/Direct/Reprimand and Profanity	0.840	0.860	0.820	0.500	0.740	0.520	0.740	0.820
Polite Disagreement	0.940	0.940	0.980	0.660	0.600	0.620	0.800	0.940
Referential disagreement	0.959	0.939	0.878	0.898	0.612	0.694	0.837	0.878
Repetition and Rewording	0.980	1.000	0.900	0.820	0.560	0.780	0.800	0.840
Sarcastic Disagreement	0.880	0.940	0.900	0.580	0.720	0.600	0.680	0.860
Detection tasks	0.944	0.949	0.908	0.761	0.578	0.696	0.755	0.873
Classification tasks	0.811	0.861	0.802	0.542	0.692	0.557	0.682	0.792
Overall score	0.878	0.905	0.855	0.652	0.635	0.626	0.719	0.832

Table 1: Accuracy of evaluated models

As previously mentioned, our dataset evaluated two types of skills: 1) binary choice between the presence and absence of disagreement, and 2) classification of disagreement strategies. As shown in Table 1, multilingual LLMs (Claude 3.5, GPT-4o, Llama-3, Gemini-Pro) consistently outperform in solving each class of tasks. Fine-tuning a pre-trained model on Russian data (Saiga-Llama-3) yields similar results. In contrast, a model primarily oriented towards the Russian language (GigaChat) shows significantly better results in the task of disagreement classification and achieves the highest scores in the MDDP task (Figure 1), which is one of the most challenging for other models. Another distinguishing feature of models focused on the Russian language (GigaChat, YaGPT) is their enhanced ability to accurately identify ironic disagreement: they demonstrate comparable results with other tasks. For other models, the performance decline in detecting irony can reach 20-25% below the average across all tasks.

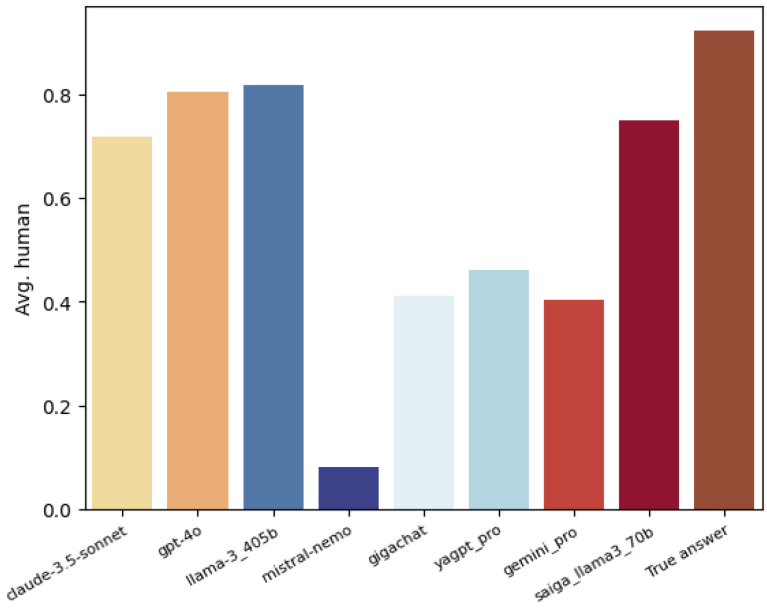


Figure 2: Correlation between model prediction and avg. human judgement

We selected the SarcD task set for comparative evaluation using annotators. It should be noted that sarcasm detection proved to be an easier task for the models than irony detection – all LLMs, except YaGPT, showed higher accuracy in solving this type of task. We aggregated the scores of all experts and used the resulting class to measure the correlation between the average human annotator, our target class, and the model scores (see Figure 2). The results indicate that the accuracy in sarcasm detection, even among leading large language models (Claude 3.5, GPT-4o, Llama-3, Saiga-Llama-3), is significantly lower than the responses of annotators. This is remarkable as it shows that the benchmark is capable of identifying common pragmatic beliefs, which is crucial to provide positive user experience and ensure alignment of these models with human intentions.

4.2 Levenshtein distance evaluation

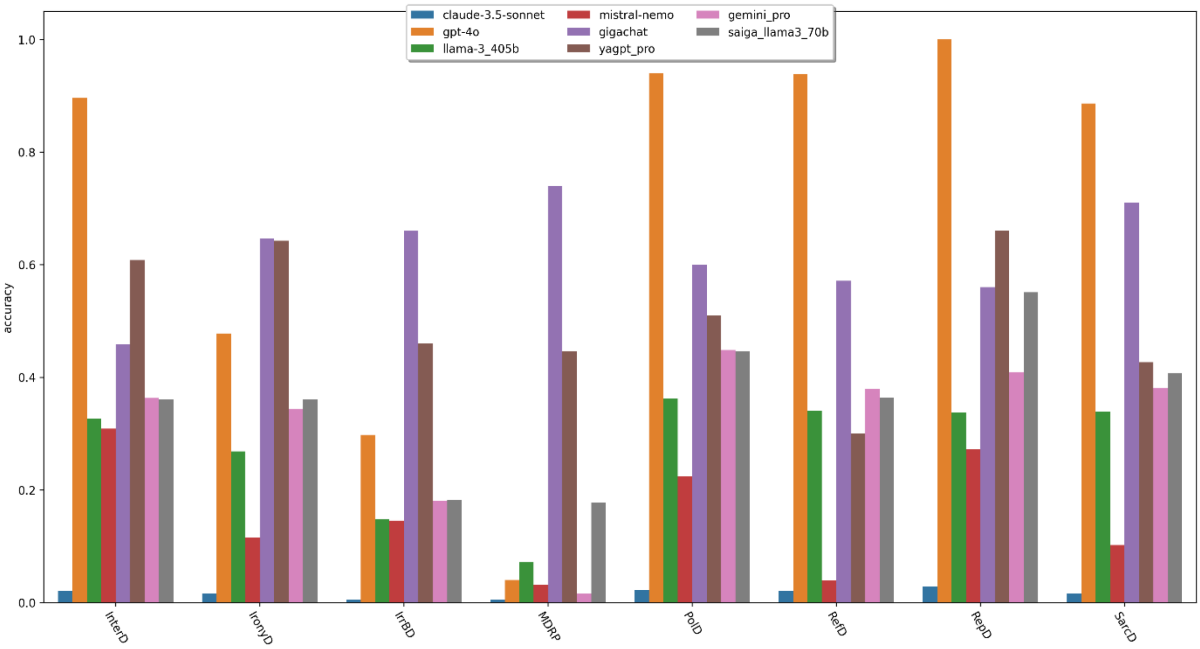


Figure 3: Overall normalized Levenshtein similarity per task

	claude-3.5-sonnet	gpt-4o	llama-3-405b	mistral-nemo	gigachat	yagpt-pro	gemini-pro	saiga-llama3-70b
Interrogative Disagreement	0.020	0.896	0.326	0.309	0.458	0.608	0.363	0.361
Ironic Disagreement	0.016	0.478	0.268	0.115	0.647	0.642	0.343	0.360
Irrelevancy claim	0.005	0.298	0.148	0.145	0.660	0.460	0.181	0.182
Mitigated/Direct/Reprimand and Profanity	0.005	0.040	0.072	0.031	0.740	0.446	0.016	0.177
Polite Disagreement	0.022	0.940	0.362	0.224	0.600	0.510	0.449	0.446
Referential disagreement	0.020	0.939	0.341	0.040	0.571	0.301	0.379	0.364
Repetition and Rewording	0.028	1.	0.337	0.273	0.560	0.661	0.409	0.551
Sarcastic Disagreement	0.016	0.886	0.339	0.102	0.710	0.427	0.380	0.408
Detection tasks	0.023	0.944	0.342	0.212	0.547	0.520	0.400	0.430
Classification tasks	0.010	0.426	0.207	0.098	0.689	0.494	0.230	0.282
Overall score	0.016	0.685	0.274	0.155	0.618	0.507	0.315	0.356

Table 2: Normalized levenshtein similarity of evaluated models

The evaluation based on Levenshtein distance (Fig. 3) exhibited a similar pattern with minor variations. Our aim was to assess whether the model could adhere to instructions directly without requiring additional explanations or arguments. If the results closely align with accuracy, it indicates that the model strictly follows the prompt, as the discrepancy between the expected answer and the model's output is minimal or nonexistent. Conversely, if accuracy is high, but the Levenshtein score is low (particularly evident in models like Claude-3.5, which tend to elaborate on the reasoning process), it suggests that while the model provides the correct answer, it also includes explanations or additional formatting that were removed during our post-processing. If both accuracy and Levenshtein scores are low, it typically indicates a fundamental misunderstanding of the task by the model, as observed with models like Mistral-Nemo.

4.3 Bias Detection

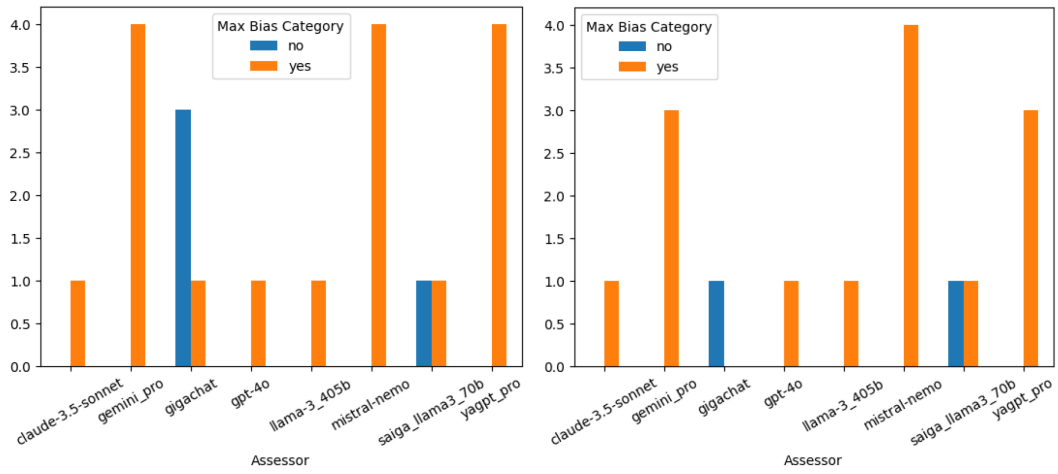


Figure 4: a) Number of tasks where stat. sign. biases were observed
b) Number of tasks where stat. sign. biases were observed when the sample contains negative particles. Max. possible – 6.

We conducted a statistical analysis to test whether the models we evaluated have biases towards detecting or neglecting disagreement. Thus, we constrained the analysis to the subset of tasks involving disagreement detection with a binary answer and conducted a chi-squared test with the initial hypothesis being the uniform distribution as each task contained an equal number of affirmative and negative answers. The graph in Fig. 4a represents the number of tasks where statistically significant biases were detected. Interestingly, most of the models tend to respond to the prompts affirmatively. Moreover, only GigaChat and Saiga models appear to display sensitivity to the task in this manner and display different biases. Overall, all the models display some level of bias, at least on one of the benchmarks.

We also conducted the same test to measure the extent to which negative particles affect the biases of models. Thus, we extracted a subset of samples that included negative particles in the context and applied the test to those. The results are provided in Fig. 4b. Contrary to possible expectations, the outputs for these subsets of samples don't differ much from the whole benchmark tasks and still contain statistically significant deviations towards negative answers only for the same models as in the general setting.

4.4 Error Analysis

The conducted study identified communicative contexts in which the majority of LLMs in our dataset exhibited erroneous results.

Firstly, these responses often express the speaker's uncertainty. They may include clarifying questions aimed at elucidating the situation, yet they support the speaker's intention and do not contradict it. For example, errors were observed in situations requiring clarification of a bus stop name, with 6 out of 8 language models making mistakes. Similarly, 6 out of 8 LLMs made errors in discussions about the benefits of dairy products. Furthermore, all LLMs failed in clarifying the size of clothing demonstrated by the interlocutor.

Secondly, the responses are often expressive reactions that involve a negative evaluation of the interlocutor's proposition, yet without an explicit intention to disagree. For example:

- Lately, restaurants have started serving beetles too [V posledneye vremya nachali v restoranakh i zhukov podavat].
- Has exoticism really gotten to this point? [Neuzheli ekzotika i do takogo doshla?] (6 out of 8 LLMs made errors)

In this instance, it can be hypothesized that the perception of disagreement was prompted by the presence of the lexeme 'neuzheli' ('really'), along with the association of emotional communication with negative speech behavior.

Finally, the tasks for classifying direct and ironic disagreement caused difficulties. In most cases, the models showed a tendency to identify irony even in the absence of irony in the context. For example, the response 'Ya by tak ne skazal' ('I wouldn't say that') was classified as ironic in 7 out of 8 LLMs. In the study sample, there were three cases in which all models made a mistake, five cases in which seven models made an error, and another five cases in which six models gave an incorrect estimate.

5 Conclusions

The study underscores the significance of assessing pragmatic competence in LLMs, focusing on their ability to detect and classify disagreements in communication. Evaluating LLMs within the framework of politeness theory highlights the way models interpret direct and indirect opposing viewpoints, alongside their potential implications for human interaction. Disagreement in communication manifests itself across a spectrum, ranging from polite and mitigated forms to unmitigated and potentially offensive ones, such as those involving reprimands, profanity, or sarcasm. This study's benchmarking approach therefore aimed to capture these linguistic complexities while examining the models' capacity to navigate them effectively.

While the models tested – including multilingual models and those fine-tuned for Russian – demonstrated varying levels of success across tasks, several consistent patterns emerged. Multilingual models like Claude 3.5 and GPT-4o, as well as fine-tuned ones like Saiga-Llama-3, performed well overall, but language-specific models such as GigaChat displayed particular strengths in nuanced tasks like detecting ironic disagreement. However, irony detection proved challenging for most models, showcasing their limitations in understanding subtle pragmatic cues.

The observed discrepancies between model performance and human judgments further demonstrate the challenges inherent in teaching artificial systems to interpret elements of pragmatics. For instance, sarcasm detection, although it can be seen as a more straightforward task than irony detection, revealed significant gaps in the models' capacities when compared to human annotators. Additionally, error analysis highlighted situations where most LLMs faltered, particularly in distinguishing disagreement from mere uncertainty or emotional evaluations. This reflects the complexity of conversational pragmatics, where context, tone, and cultural nuances play key roles – areas in which LLMs often lack nuanced understanding.

Importantly, biases were also identified in models' responses, such as a tendency to favor affirmative answers or misinterpret context containing negative particles. These biases, while varying across models, highlight the ongoing need for refinement in LLM training, particularly in ensuring balanced output sensitivity and improved detection of linguistic subtleties.

References

[1] Aaron Grattafiori et al. The Llama 3 Herd of Models // arXiv preprint. — 2024. — Vol. arXiv:2407.21783. — Access mode: https://arxiv.org/abs/2407.21783.

[2] Angouri J., Locher M. A. Theorising disagreement // Journal of Pragmatics. — 2012. — Vol. 44, No. 12. — P. 1549–1553. Access mode: https://doi.org/10.1016/j.pragma.2012.06.011.

[3] Benz A., Jasinskaja K. Questions Under Discussion: From Sentence to Discourse // Discourse Processes. — 2017. — Vol. 54, No. 3. — P. 177–186. Access mode: https://doi.org/10.1080/0163853x.2017.1316038.

[4] Bolander B., Locher M. A. Conflictual and consensual disagreement // In: Hoffmann C., Bublitz W. (Eds.). Pragmatics of Social Media. — De Gruyter Mouton, 2017. — P. 607–632. Access mode: https://doi.org/10.1515/9783110431070-022.

[5] Bousfield D. Impoliteness in Interaction. — John Benjamins Publishing Company, 2008.

[6] Brown P., Levinson S. C. Politeness: Some universals in language usage. — Cambridge: Cambridge University Press, 1987. — (Original work published 1978).

[7] Culpeper J. Towards an anatomy of impoliteness // Journal of Pragmatics. — 1996. — Vol. 25, No. 3. — P. 349–367. Access mode: https://doi.org/10.1016/0378-2166(95)00014-3.

[8] Ghosh Aniruddha, Li Guofu, Veale Tony, Rosso Paolo, Shutova Ekaterina, Barnden John, Reyes Antonio. SemEval-2015 Task 11: Sentiment Analysis of Figurative Language in Twitter // Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015). — Denver, Colorado, 2015. — P. 470–478.

[9] Gu Yuling, Fu Yao, Pyatkin Valentina, Magnusson Ian, Dalvi Mishra Bhavana, Clark Peter. Just-DREAM-about-it: Figurative Language Understanding with DREAM-FLUTE // Proceedings of the 3rd Workshop on Figurative Language Processing (FLP). — Abu Dhabi, United Arab Emirates (Hybrid): Association for Computational Linguistics, 2022. — P. 84–93.

[10] Kakava C. Opposition in modern Greek discourse: Cultural and contextual constraints // Journal of Pragmatics. — 2002. — Vol. 34, No. 10–11. — P. 1537–1568. Access mode: https://doi.org/10.1016/s0378-2166(02)00075-9.

[11] Kotthoff H. Disagreement and concession in disputes: On the context sensitivity of preference structures // Language in Society. — 1993. — Vol. 22, No. 2. — P. 193–216. Access mode: https://doi.org/10.1017/s0047404500017103.

[12] Langlotz A., Locher M. A. Ways of communicating emotional stance in online disagreements // Journal of Pragmatics. — 2012. — Vol. 44, No. 12. — P. 1591–1606. Access mode: https://doi.org/10.1016/j.pragma.2012.04.002.

[13] Lee C. J., Katz A. N. The differential role of ridicule in sarcasm and irony // Metaphor and Symbol. — 1998. — Vol. 13, No. 1. — P. 1–15.

[14] Leech G. N. Principles of pragmatics. — Longman, 1983.

[15] Liu Emmy, Cui Chenxuan, Zheng Kenneth, Neubig Graham. Testing the Ability of Language Models to Interpret Figurative Language // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Seattle, United States, 2022. — P. 4437–4452.

[16] Locher M. A. Power and politeness in action: Disagreement in oral communication. — Mouton de Gruyter, 2004.

[17] Min B., Ross H., Sulem E., Veyseh A. P. B., Nguyen T. H., Sainz O., Agirre E., Heintz I., Roth D. Recent advances in natural language processing via large pre-trained language models: A survey // ACM Computing Surveys. — 2023. — Vol. 56, No. 2. — P. 2. Access mode: https://doi.org/10.1145/3605943.

[18] Muntigl P., Turnbull W. Conversational structure and facework in arguing // Journal of Pragmatics. — 1998. — Vol. 29, No. 3. — P. 225–256. Access mode: https://doi.org/10.1016/s0378-2166(97)00048-9.

[19] Qi Chengwen, Li Bowen, Hui Binyuan, Wang Bailin, Li Jinyang, Wu Jinwang, Laili Yuanjun. An Investigation of LLMs' Inefficacy in Understanding Converse Relations // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Singapore: Association for Computational Linguistics, 2023. — P. 6932–6953.

[20] Reyes A., Rosso P., Veale T. A multidimensional approach for detecting irony in Twitter // Language Resources and Evaluation. — 2013. — Vol. 47, No. 1. — P. 239–268.

[21] Ruis L. et al. The Goldilocks of Pragmatic Understanding: Fine-tuning Strategy Matters for Implicature Resolution by LLMs // Advances in Neural Information Processing Systems. — 2024. — Vol. 36.

[22] Schiffrin D. Jewish argument as sociability // Language in Society. — 1984. — Vol. 13, No. 3. — P. 311–335. Access mode: https://doi.org/10.1017/s0047404500010526.

[23] Shum W., Lee C. (Im)politeness and disagreement in two Hong Kong internet discussion forums // Journal of Pragmatics. — 2013. — Vol. 50, No. 1. — P. 52–83. Access mode: https://doi.org/10.1016/j.pragma.2013.01.010.

[24] Sravanthi S., Doshi M., Tankala P., Murthy R., Dabre R., Bhattacharyya P. PUB: A pragmatics understanding benchmark for assessing LLMs' pragmatics capabilities // Findings of the Association for Computational Linguistics. — 2024. — P. 12075–12097. Access mode: https://doi.org/10.18653/v1/2024.findings-acl.719.

[25] Song Y., Wang G., Li S, & Lin B. Y. (2024). The Good, The Bad, and The Greedy: Evaluation of LLMs Should Not Ignore Non-Determinism. *arXiv preprint arXiv: 2407.10457*.
[26] Sykora M., Elayan S., Jackson T. W. A qualitative analysis of sarcasm, irony and related #hashtags on Twitter // Big Data & Society. — 2020. — Vol. 7, No. 2. Access mode: https://doi.org/10.1177/2053951720972735.
[27] Taguchi N. Pragmatic competence in Japanese as a second language: An introduction // In: Taguchi N. (Ed.). Pragmatic Competence. — Mouton de Gruyter, 2009. — P. 1–18.
[28] Tong Xiaoyu, Shutova Ekaterina, Lewis Martha. Recent advances in neural metaphor processing: A linguistic, cognitive, and social perspective // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. — Online: Association for Computational Linguistics, 2021. — P. 4673–4686.
[29] Uzelgun M. A., Mohammed D., Lewiński M., Castro P. Managing disagreement through yes, but… constructions: An argumentative analysis // Discourse Studies. — 2015. — Vol. 17, No. 4. — P. 467–484. Access mode: https://doi.org/10.1177/1461445615578965.
[30] Witek M. Irony as a speech action // Journal of Pragmatics. — 2022. — Vol. 190. — P. 76–90. Access mode: https://doi.org/10.1016/j.pragma.2022.01.010.
[31] Yao Binwei, Jiang Ming, Bobinac Tara, Yang Diyi, Hu Junjie. Benchmarking Machine Translation with Cultural Awareness // Findings of the Association for Computational Linguistics: EMNLP 2024. — Miami, United States, 2024. — pp. 13078–13096. Access mode:

6 Appendix A. Representative task examples

6.1 Direct disagreement

Example 1

A: Onlajn-obuchenie tak zhe e'ffektivno, kak tradicionnoe obuchenie v klasse, blagodarya svoej gibkosti. (Online learning is just as effective as traditional classroom learning because of the flexibility it offers.)

B: Ya tak ne dumayu. Otsutstvie ochnogo vzaimodejstviya uxudshaet process obucheniya. (I don't think so. The lack of face-to-face interaction hinders the learning experience.)

Example 2

A: Arenda doma luchshe, potomu chto ona pozvolyaet se'konomit' na remonte i nalogax na imushhestvo. (Renting a house is better because it saves so much money in repairs and property taxes.)

B: No vladenie domom e'konomit bol'she deneg so vremenem blagodarya nakopleniyu kapitala i vozmozhny'm nalogovy'm vy'chetam. (But, owning a house saves more money over time through equity buildup and potential tax deduction.)

Example 3

A: Mne kazhetsya, nado kupit' novyj noutbuk, etot uzhe ele rabotaet. (I think we need to buy a new laptop; this one is barely working.)

B: Nichego podobnogo, etot nout eshchyo paru let protyanet. (No way! This one's got a couple more years in it.)

Example 4

A: Ty smotrel novyj sezon seriala? (Have you seen the new season of the show?)

B: Net, poka eshche ne dobralsya do nego. (Not yet, haven't gotten around to it.)

6.2 Polite disagreement

Example 1 (With disagreement)

A: Tvoi reis vyletaet v 2 chasa dnya, ne opozdai (Your flight departs at 2 PM, don't be late).

B: Poleznaya informatsiya, no pokhozhe, chto on uzhe uletel (Useful information, but it seems it has already taken off)

Example 2 (With disagreement)

A: Ya s vami ne soglasen. Domashnie zhivotny'e prinosyat stol'ko radosti i uyuta v dom! Ne mogu predstavit' svoyu zhizn' bez svoego pushistogo druga. (I disagree. Pets bring us so much joy and comfort! I can't imagine life without my fluffy friend.)

B: Ya ne soglasen s vashim mneniem. Konechno, domashnie zhivotny'e mogut prinosit' radost', no e'to ne dlya vsex. U menya est' allergiya na sherst', poe'tomu ya predpochitayu ne imet' domashnix pitomcev. (I don't subscribe to your opinion. Pets can make people happier, sure, but they aren't for everyone. I'm allergic to fur so I wouldn't want to have a pet.)

Example 3 (With disagreement)

A: Poezd tol'ko v 5 vechera, ty uspevaesh'. (The train is only at 5 PM, you've got plenty of time.)

B: A ty uveren, chto 5 vechera, a ne v 7? (Are you sure it's 5 PM and not 7?)

Example 4 (Without disagreement)

A: Vot vam stakan vody, etu tabletku tak prosto ne proglotit' (Here's a glass of water; this pill isn't easy to swallow)

B: Aga, spasibo bol'shoe. Razlomit' zhe ee tozhe mozhno, da? (Oh, thanks a lot. I can break it in half too, right?)

Example 5 (Without disagreement)

A: Sovsem chto-to kust zasyhaet. Nado ego peresadit' v ten', navernoe... (This bush looks like it's drying out. Maybe we should move it to the shade?)

B: Togda, kak dumaesh', mozhet i elku peresadit'? (In that case, do you think we should move the spruce too?)

6.3 Interrogative Disagreement

Example 1 (With disagreement)

A: Nam nuzhno perestat' ispol'zovat' odnorazovye plastikovye izdeliya, oni zagryaznyayut okean (We need to stop using single-use plastic products; they pollute the ocean)

B: I chto ot etogo izmenitsya? (And what difference will that make?)

Example 2 (With disagreement)

A: Ya ne samy'j bol'shoj puteshestvennik, no kazhdoe puteshestvie prinosit mne neveroyatny'e vpechatleniya i novy'e znakomstva. (I don't travel that often but each time I do I get absolutely incredible experience and make new friends.)

B: Puteshestviya takzhe mogut by't' neveroyatno stressovy'mi i iznuritel'ny'mi. No pochemu by' ne provesti otpusk v uyute doma, e'konomya pri e'tom i vremya, i den'gi? (Travelling can also be incredibly stressful and tiring. Why wouldn't you spend your vacation in the comfort of your home, saving both time and money?)

Example 3 (With disagreement)

A: Luchshe vsekh sobak chipirovat', ne tol'ko bol'shih. (All dogs should be chipped, not just big ones.)

B: Ty s kakoj planety? Chto tebe chihuahua mozhet sdelat'? (What planet are you from? What harm could a chihuahua do to you?)

Example 4 (Without disagreement)

A: Ya vzyal nozhi i vilki, no kazhetsya chto-to zabyl… (I took the knives and forks, but I seem to have forgotten something...)

B: A kak zhe tarelki? (What about the plates?)

Example 5 (Without disagreement)

A: Ya 5 let prouchilsya v universitete i ochen' skuchayu po etomu vremeni. (I studied at university for five years and really miss that time.)

B: V kakom imenno? (Which one exactly?)

6.4 Repetitive Disagreement

Example 1 (With disagreement)

A: V takom sluchae stoit snachala razmorozit' kuritsu (In that case, you should defrost the chicken first)

B: V takom sluchae stoit snachala podumat', nuzhen li mne vash sovet. (In that case, you should first consider whether I need your advice)

Example 2 (With disagreement)

A: My vchera ne zakryvali dver', esli ya pravil'no pomnyu. (We didn't lock the door last night, if I remember correctly.)

B: Aga-aga, vizhu ya, kak vy ne zakryvali dver'. (Yeah, right, I can totally see how you didn't lock the door.)

Example 3 (Without disagreement)

A: Zavtra ozhidaetsya dozhdlivyy den', ne zabud' vzat' zont (Tomorrow is expected to be a rainy day, don't forget to take an umbrella)

B: Vzat' zont. (Take an umbrella)

Example 4 (Without disagreement)

A: Nado kupit' hleb ili ris na uzhin. (We need to buy either bread or rice for dinner.)

B: Tak hleb ili ris? (So which one—bread or rice?)

6.5 Referential disagreement

Example 1 (With disagreement)

A: Ideya universitetov kazhetsya slishkom uzhe ustarevshey (The idea of universities seems a bit too outdated)

B: Moy papa – professor, on smotrit na eto sovershenno inache. (My dad is a professor; he sees it completely differently).

Example 2 (With disagreement)

A: Plavanie — samoe nelepoe olimpijskoe sobytie. (Swimming is the most ridiculous Olympic event.)

B: Plavanie — eto zdorovo. Ya govoryu eto kak chelovek, kotoryj zanimalsya sportivnym plavaniem bolee 15 let. (Swimming is great. I say this as someone who did competitive swimming for over 15 years.)

Example 3 (Without disagreement)

A: Nichego ne uspel kupit' v podarok. Kak dumaesh', emu ponravitsya kniga? (I didn't manage to buy a gift. Do you think he'd like a book?)

B: Govoryat, chto kniga – luchshij podarok! (They say a book is the best gift!)

Example 4 (Without disagreement)

A: Ya kupil sol' i perec, no ne mogu najti v paketah. (I bought salt and pepper, but I can't find them in the bags.)

B: V cheke oni est', davaj iskat'. (They're on the receipt, so let's keep looking.)

6.6 Irrelevancy claim

Example 1 (Disagreement)

A: Ty pojdesh' na novuyu vystavku? (Are you going to the new exhibition?)

B: Net, eto slishkom daleko ot doma. (No, it's too far from home.)

Example 2 (Disagreement)

A: Kak ty dumaesh', nam stoit zavesti sobaku? (Do you think we should get a dog?)

B: Mozhet, koshku? (How about a cat?)

Example 3 (Refusal to engage with the topic)

A: Kstati, ya sobirayus' kupit' novye krossovki. (By the way, I'm going to buy new sneakers.)

B: I kak eto pomozhet otremontirovat' mashinu? (And how will that help fix the car?)

Example 4 (Refusal to engage with the topic)

A: Povyshenie minimal'noj zarplaty privedet k rostu bezraboticy. (Raising the minimum wage will lead to higher unemployment.)

B: Vopros bezraboticy ne imeet otnosheniya k obsuzhdeniyu. My govorim o prave rabotnikov na dostojnuyu oplatu truda. (Unemployment isn't the issue here. We're talking about workers' right to fair pay.)

6.7 Mitigated/Direct/Reprimand and Profanity

Example 1 (Mitigated disagreement)

A: Ochen' lyublyu Dzhuliyu Roberts! (I really love Julia Roberts!)

B: Slushaj, ne mogu tut soglasit'sya, mne ona, k sozhaleniyu, ni v odnom fil'me ne ponravilas'. (Honestly, I can't agree with you here; unfortunately, I haven't liked her in any movie)

Example 2 (Mitigated disagreement)

A: Davaj pouzhinaem v etom novom restorane, govoryat, tam vkusno. (Let's have dinner at that new restaurant, I've heard the food is good.)

B: V celom ya podderzhivayu, no mozhet, snachala pochitaem otzyvy? (I'm on board, but maybe we should check the reviews first?)

Example 3 (Direct disagreement)

A: Kakoe horoshee utro! (What a lovely morning!)

B: Net, nu kakoe zhe ono horoshee? Chetyre utra. (Lovely? How is it lovely? It's 4 a.m.)

Example 4 (Direct disagreement)

A: Nado obnovit' garderob, eti veshchi uzhe ne v mode. (I need to update my wardrobe, these clothes are out of fashion.)

B: Ne dumayu, zachem tratit' den'gi. (I don't think so, why spend money?)

Example 5 (Direct disagreement and Reprimand/Profanity)

A: Kakoye vsyo-taki zamechatel'noye utro! (What a wonderful morning it is!)

B: Slushay, idi ty k chertu. Utro u nego zamechatel'noye... (Listen, go to hell. Wonderful morning, he says...)

Example 6 (Direct disagreement and Reprimand/Profanity)

A: Ya vse v mashine ostavil. Hochesh' – idi sam posmotri. (I left everything in the car. If you want, go check yourself.)

B: Ty menya za duraka derzhish'? Slushaj, ded ty staryj, ya svoimi dvumya glazami videl, kak ty papku v shkaf polozhil! (Are you seriously that dumb? Listen, I saw with my own eyes how you put the folder in the cupboard!)

6.8 Ironic disagreement

Example 1 (Disagreement with irony)

A: Tebe ponravilsya novyy restoran? (Did you like the new restaurant?)

B: O, da, osobenno yego unikal'nyy servis s podachey kholodnoy yedy. (Oh, yes, especially its unique service of serving cold food.)

Example 2 (Disagreement with irony)

A: Mne vse ravno, chto anglijskij – yazyk mezhdunarodnogo obshcheniya, ya ne sobirayus' ego uchit'. (I don't care that English is the international language of communication, I'm not going to learn it.)

B: Chem bol'she takih prekrasnyh mnenij, tem men'she konkurenciya na rynke. Spasibo! (The more of these wonderful opinions, the less competition in the market. Thanks!)

Example 3 (Disagreement with irony)

A: Eta vecherinka byla prosto potryasayushchej, ne tak li? (This party was amazing, wasn't it?)

B: Absolyutno! Osobenno chast', gde vse utknulis' v svoi telefony. (Absolutely! Especially the part where everyone was glued to their phones.)

Example 4 (Disagreement without irony)

A: Vy ushli ot otveta. (You dodged the question.)

B: Ya? (Me?)

Example 5 (Disagreement without irony)

A: My zapuskaem novyj eko-frendli proekt. (We are launching a new eco-friendly project.)

B: Podobnaya iniciativa ni k chemu ne privedet. (Such an initiative will lead to nothing.)

6.9 Sarcastic disagreement

Example 1 (Disagreement with sarcasm)

A: A pochemu vy protiv? chto on takogo sdelal? (Why are you against? What did he do?)

B: Glupo etogo ne znat' =))) no ya podskazhu unikal'nyj sposob dlya prosveshcheniya: set' INTERNET + poiskovik = otvet na vopros. (It's silly not to know =))) but here's a unique way to educate yourself: the INTERNET + a search engine = your answer.)

Example 2 (Disagreement with sarcasm)

A: Pomoch' tebe s domashnej rabotoj po matematike? (Do you need help with your math homework?)

B: Uchitel' goda ob"yavilsya! (Here comes the Teacher of the Year!)

Example 3 (Disagreement with sarcasm)

A: Ya uveren, chto vegetarianstvo – luchshee, chto mozhno pridumat' dlya zdorov'ya. (I'm sure vegetarianism is the best thing for health.)

B: Kak zdorovo, chto ty tak verish' v eti chudesa. Tak derzhat'))) (How wonderful that you believe these miracles. Keep it up!)))

Example 4 (Disagreement without sarcasm)

A: Vy skazali, chto deficit vitamina D vedet k vospaleniyu desen. (You said that a vitamin D deficiency leads to gum inflammation.)

B: Gde ya takoe govoril..? (When did I say that...?)

Example 5 (Disagreement without sarcasm)

A: Vtoraya chast' "Vlastelina kolec" samaya luchshaya. Mneniya? (The second part of "The Lord of the Rings" is the best. Thoughts?)

B: Eeee, chto? Ty, navernoe, hotel skazat' "tret'ya". (Uhh, what? You probably meant "the third.")

7 Appendix B. Significance analysis

Dataset name	Assessor	Chi-Square Statistic	P-Value	Is Biased	Max Bias Category	Significance Level	Number of samples
Interrogative Disagreement	claude-3.5-sonnet	0.7619	0.3827	False	yes	0.0500	48
Interrogative Disagreement	gpt-4o	1.3545	0.2445	False	yes	0.0500	48
Interrogative Disagreement	llama-3-405b	0.3386	0.5606	False	yes	0.0500	48
Interrogative Disagreement	mistral-nemo	6.9725	0.0083	True	yes	0.0500	47
Interrogative Disagreement	gigachat	0.3386	0.5606	False	yes	0.0500	48
Interrogative Disagreement	yagpt-pro	14.3069	0.0002	True	yes	0.0500	48
Interrogative Disagreement	gemini-pro	16.5926	0.0000	True	yes	0.0500	48
Interrogative Disagreement	saiga-llama3-70b	0.3386	0.5606	False	yes	0.0500	48
Ironic Disagreement	claude-3.5-sonnet	17.8183	0.0000	True	yes	0.0500	51
Ironic Disagreement	gpt-4o	15.5217	0.0001	True	yes	0.0500	51
Ironic Disagreement	llama-3-405b	20.2733	0.0000	True	yes	0.0500	51
Ironic Disagreement	mistral-nemo	17.5750	0.0000	True	yes	0.0500	46
Ironic Disagreement	gigachat	0.0000	1.0000	False	yes	0.0500	51
Ironic Disagreement	yagpt-pro	9.7424	0.0018	True	yes	0.0500	50
Ironic Disagreement	gemini-pro	41.8929	0.0000	True	yes	0.0500	51
Ironic Disagreement	saiga-llama3-70b	17.8183	0.0000	True	yes	0.0500	51
Polite Disagreement	claude-3.5-sonnet	0.7212	0.3958	False	yes	0.0500	50
Polite Disagreement	gpt-4o	0.7212	0.3958	False	yes	0.0500	50
Polite Disagreement	llama-3-405b	0.0801	0.7771	False	no	0.0500	50
Polite Disagreement	mistral-nemo	13.8478	0.0002	True	yes	0.0500	49
Polite Disagreement	gigachat	5.1282	0.0235	True	no	0.0500	50
Polite Disagreement	yagpt-pro	13.5417	0.0002	True	yes	0.0500	50
Polite Disagreement	gemini-pro	8.0128	0.0046	True	yes	0.0500	50
Polite Disagreement	saiga-llama3-70b	0.7212	0.3958	False	yes	0.0500	50
Referential disagreement	claude-3.5-sonnet	0.3379	0.5610	False	yes	0.0500	49
Referential disagreement	gpt-4o	0.0845	0.7713	False	yes	0.0500	49
Referential disagreement	llama-3-405b	1.3517	0.2450	False	yes	0.0500	49
Referential disagreement	mistral-nemo	0.7603	0.3832	False	yes	0.0500	49
Referential disagreement	gigachat	24.4155	0.0000	True	no	0.0500	49
Referential disagreement	yagpt-pro	14.2776	0.0002	True	yes	0.0500	49
Referential disagreement	gemini-pro	0.3379	0.5610	False	yes	0.0500	49
Referential disagreement	saiga-llama3-70b	1.3517	0.2450	False	yes	0.0500	49
Repetition and Rewording	claude-3.5-sonnet	0.0821	0.7745	False	yes	0.0500	50
Repetition and Rewording	gpt-4o	0.0000	1.0000	False	yes	0.0500	50
Repetition and Rewording	llama-3-405b	0.7389	0.3900	False	yes	0.0500	50
Repetition and Rewording	mistral-nemo	0.3386	0.5606	False	yes	0.0500	48
Repetition and Rewording	gigachat	5.2545	0.0219	True	no	0.0500	50
Repetition and Rewording	yagpt-pro	3.0000	0.0833	False	yes	0.0500	49
Repetition and Rewording	gemini-pro	1.3136	0.2517	False	yes	0.0500	50
Repetition and Rewording	saiga-llama3-70b	5.2545	0.0219	True	no	0.0500	50
Sarcastic Disagreement	claude-3.5-sonnet	2.8986	0.0887	False	yes	0.0500	50
Sarcastic Disagreement	gpt-4o	0.0805	0.7766	False	yes	0.0500	50
Sarcastic Disagreement	llama-3-405b	0.7246	0.3946	False	yes	0.0500	50
Sarcastic Disagreement	mistral-nemo	35.5072	0.0000	True	yes	0.0500	50
Sarcastic Disagreement	gigachat	8.0515	0.0045	True	yes	0.0500	50
Sarcastic Disagreement	yagpt-pro	0.0889	0.7655	False	yes	0.0500	45
Sarcastic Disagreement	gemini-pro	20.6119	0.0000	True	yes	0.0500	50
Sarcastic Disagreement	saiga-llama3-70b	2.0129	0.1560	False	yes	0.0500	50

Table 3. Significance analysis of bias for the models

Dataset name	Assessor	Chi-Square Statistic	P-Value	Is Biased	Max Bias Category	Significance Level	Number of samples
Interrogative Disagreement	claude-3.5-sonnet	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	gpt-4o	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	llama-3-405b	0.0000	1.0000	False	yes	0.0500	16
Interrogative Disagreement	mistral-nemo	5.4545	0.0195	True	yes	0.0500	15
Interrogative Disagreement	gigachat	1.3333	0.2482	False	yes	0.0500	16
Interrogative Disagreement	yagpt-pro	3.0000	0.0833	False	yes	0.0500	16
Interrogative Disagreement	gemini-pro	3.0000	0.0833	False	yes	0.0500	16
Interrogative Disagreement	saiga-llama3-70b	0.0000	1.0000	False	yes	0.0500	16
Ironic Disagreement	claude-3.5-sonnet	4.7009	0.0301	True	yes	0.0500	22
Ironic Disagreement	gpt-4o	6.7692	0.0093	True	yes	0.0500	22
Ironic Disagreement	llama-3-405b	9.2137	0.0024	True	yes	0.0500	22
Ironic Disagreement	mistral-nemo	12.4444	0.0004	True	yes	0.0500	21
Ironic Disagreement	gigachat	0.1880	0.6646	False	yes	0.0500	22
Ironic Disagreement	yagpt-pro	4.7009	0.0301	True	yes	0.0500	22
Ironic Disagreement	gemini-pro	15.2308	0.0001	True	yes	0.0500	22
Ironic Disagreement	saiga-llama3-70b	6.7692	0.0093	True	yes	0.0500	22
Polite Disagreement	claude-3.5-sonnet	0.4706	0.4927	False	yes	0.0500	34
Polite Disagreement	gpt-4o	1.0588	0.3035	False	yes	0.0500	34
Polite Disagreement	llama-3-405b	0.1176	0.7316	False	no	0.0500	34
Polite Disagreement	mistral-nemo	12.1324	0.0005	True	yes	0.0500	33
Polite Disagreement	gigachat	1.0588	0.3035	False	no	0.0500	34
Polite Disagreement	yagpt-pro	7.5294	0.0061	True	yes	0.0500	34
Polite Disagreement	gemini-pro	4.2353	0.0396	True	yes	0.0500	34
Polite Disagreement	saiga-llama3-70b	1.0588	0.3035	False	yes	0.0500	34
Referential disagreement	claude-3.5-sonnet	0.0000	1.0000	False	yes	0.0500	23
Referential disagreement	gpt-4o	0.0000	1.0000	False	yes	0.0500	23
Referential disagreement	llama-3-405b	0.1825	0.6692	False	yes	0.0500	23
Referential disagreement	mistral-nemo	2.9206	0.0875	False	yes	0.0500	23
Referential disagreement	gigachat	8.9444	0.0028	True	no	0.0500	23
Referential disagreement	yagpt-pro	8.9444	0.0028	True	yes	0.0500	23
Referential disagreement	gemini-pro	0.7302	0.3928	False	yes	0.0500	23
Referential disagreement	saiga-llama3-70b	0.1825	0.6692	False	yes	0.0500	23
Repetition and Rewording	claude-3.5-sonnet	0.2017	0.6534	False	yes	0.0500	24
Repetition and Rewording	gpt-4o	0.0000	1.0000	False	yes	0.0500	24
Repetition and Rewording	llama-3-405b	3.2269	0.0724	False	yes	0.0500	24
Repetition and Rewording	mistral-nemo	0.8214	0.3648	False	yes	0.0500	23
Repetition and Rewording	gigachat	3.2269	0.0724	False	yes	0.0500	24
Repetition and Rewording	yagpt-pro	0.8067	0.3691	False	yes	0.0500	24
Repetition and Rewording	gemini-pro	0.2017	0.6534	False	yes	0.0500	24
Repetition and Rewording	saiga-llama3-70b	7.2605	0.0070	True	no	0.0500	24
Sarcastic Disagreement	claude-3.5-sonnet	0.8333	0.3613	False	yes	0.0500	20
Sarcastic Disagreement	gpt-4o	0.0000	1.0000	False	yes	0.0500	20
Sarcastic Disagreement	llama-3-405b	0.2083	0.6481	False	yes	0.0500	20
Sarcastic Disagreement	mistral-nemo	13.3333	0.0003	True	yes	0.0500	20
Sarcastic Disagreement	gigachat	1.8750	0.1709	False	yes	0.0500	20
Sarcastic Disagreement	yagpt-pro	0.2250	0.6353	False	no	0.0500	18
Sarcastic Disagreement	gemini-pro	10.2083	0.0014	True	yes	0.0500	20
Sarcastic Disagreement	saiga-llama3-70b	0.2083	0.6481	False	yes	0.0500	20

Table 4. Significance analysis of bias for the models in samples including negative particles

Appendix C. Model generation parameters

Model	Temperature	Intended generation strategy
Claude-3.5 Sonnet	0.0	Greedy
GPT-4o	0.0	Greedy
Llama-3 405B	0.0	Greedy
Mistral Nemo	0.0	Greedy
Gigachat Pro	0.0	Greedy
Yagpt Pro	0.0	Greedy
Gemini Pro	0.0	Greedy
Saiga Llama3 70B	0.0	Greedy

An Approach to Information Extraction from Texts of a Limited Subject Domain Based on a Chain of Large Language Models

Sidorova E. A.
A.P. Ershov Institute of Informatics Systems; Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
lsidorova@iis.nsk.su

Ivanov A. I.
Novosibirsk State University, Novosibirsk, Russia
a.ivanov15@alumni.nsu.ru

Ilina D. V.
A.P. Ershov Institute of Informatics Systems; Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
dviljina@gmail.com

Ovchinnikova K. A.
Novosibirsk State University, Novosibirsk, Russia
k.ovchinnikova2@alumni.nsu.ru

Osmushkin N. M.
Novosibirsk State University, Novosibirsk, Russia
n.osmushkin@g.nsu.ru

Sery A. S.
A.P. Ershov Institute of Informatics Systems; Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
alexey.seryj@iis.nsk.su

Abstract

The paper considers the approach for extracting the information from texts of limited domains of knowledge based on a chain of neural language models. The task is represented in the form of three subtasks solved sequentially: (1) term extraction and classification; (2) coreference resolution; (3) extraction of relations of entities named with the terms. The dataset was based on texts on computational linguistics from the Habr forum. In the markup for term classification and relation extraction, 17 classes of terms and 51 relations were used in accordance with the ontology of computational linguistics. Prompt chain-based methods were used to apply LLMs, where each next query to the LLM is based on the results of the previous step. Six types of prompt templates were developed: for extracting, classifying, verifying terms, extracting coreferential relations, relations specified by the ontology, and a specialized template for relations linking entities of the same class. Sentence-BERT, GPT-4 and Mistral-based models were used at different steps of the study; a comparison with the SFT approach (ruRoBERTa) was made; hybrid approaches that have shown the best results were also developed. For term extraction and classification, F1=0.77 was obtained, for coreference resolution—F1=0.897, and for relation extraction—F1=0.847.

Keywords: information extraction; term extraction; relation extraction, natural language processing; machine learning; large language models; ontology of domain of knowledge; prompt chaining

DOI: 10.28995/2075-7182-2025-23-361-373

Подход к извлечению информации из текстов ограниченной предметной области на основе цепочки больших языковых моделей

Сидорова Е. А.
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия
lsidorova@iis.nsk.su

Иванов А. И.
Новосибирский государственный университет, Новосибирск, Россия
a.ivanov15@alumni.nsu.ru

Ильина Д. В.
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия
dviljina@gmail.com

Овчинникова К. А.
Новосибирский государственный университет, Новосибирск, Россия
k.ovchinnikova2@alumni.nsu.ru

Осьмушкин Н. М.
Новосибирский государственный университет, Новосибирск, Россия
n.osmushkin@g.nsu.ru

Серый А. С.
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия
alexey.seryj@iis.nsk.su

Аннотация

В статье рассматривается подход к извлечению информации из текстов ограниченной предметной области на основе цепочки нейросетевых языковых моделей. Задача представлена в виде трех подзадач, решаемых последовательно: (1) извлечение и классификация терминов; (2) разрешение кореференции среди терминов, найденных на первом этапе; (3) извлечение отношений между сущностями, обозначаемыми терминами. В качестве материала были взяты тексты по компьютерной лингвистике с форума Habr. Разметка текстов для классификации терминов и извлечения отношений проводилась в соответствии с типами сущностей и отношений, заданных онтологией. Всего при разметке датасетов использовалось 17 классов терминов и 51 отношение, относящихся к предметной области компьютерной лингвистики. Для применения больших генеративных языковых моделей (LLM) использовались методы на основе цепочки промптов, когда каждый следующий запрос к LLM генерируется на основе результатов предыдущего этапа. Было разработано 6 видов шаблонов промптов: для извлечения, классификации, верификации терминов, извлечения кореферентных отношений и отношений, заданных онтологией предметной области, а также специализированный шаблон для отношений, связывающий сущности одного класса. На разных этапах исследования применялись языковые модели Sentence-BERT, GPT-4 и модели на основе архитектуры Mistral; проведено сравнение с подходом на основе обучения (модель ruRoBERTa) и разработаны гибридные подходы, показавшие лучшие результаты. Достигнуты значения F_1=0.77 для извлечения и классификации терминов, F_1=0.897 для разрешения кореференции и F_1=0.847 для извлечения отношений.

Ключевые слова: извлечение информации; извлечение терминов; извлечение отношений; обработка естественного языка; машинное обучение; большие языковые модели; онтология предметной области; цепочки инструкций

1 Introduction

Information extraction (IE) focuses on extracting structured knowledge from natural language texts. Solutions are applied across various domains, including the analysis and monitoring of news, social networks, and scientific publications, as well as in automatic database population, and the construction of knowledge graphs and ontologies.

Ontologies and knowledge graphs are not only intended for human perception, but can also be used in information retrieval systems, monitoring, learning, etc. Subject domain ontologies play a special role when dealing with domain-specific texts, providing structures for representing information and, as a rule, offering the possibility to store the obtained information as a knowledge graph. They can also serve as a basis for generating prompts for LLMs in automatic text processing (NLP) systems.

In the paper, ontology is a formal representation of knowledge — whether abstract or specific — within a particular subject area. It is developed based on a description of objects, facts, and relationships and is designed for a versatile application across various tasks [1]. A knowledge graph refers to a semantic network that stores information about objects and the relationships between them [2].

Recently, Large Language Models (LLMs) have shown great promise in text understanding and generation. There has been a lot of work on the application of LLMs to IE tasks based both on discriminative and generative paradigms [3]. The main disadvantage of LLM of the first type is the need to prepare a representative dataset, which is not always possible in limited knowledge domains. A second-type LLM requires the selection and systematic refinement of instructions (prompts) in natural language, which also requires the involvement of experts.

Our study was motivated by the following factors:

- the models in few-shot and zero-shot settings still have a huge performance gap behind the SFT (Supervised Fine-Tuning);
- models perform significantly worse on limited subject areas (e.g., the GPT-4 performs 83.48 vs. 58.44 F-measure on the news corpus compared to the medical one [3]);
- despite the similarity of LLM-based architectures, the results vary greatly depending on the methods used, including design prompts;
- the results on the Russian-language dataset may differ significantly from those for the English-language dataset.

The aim of our study is to develop an LLM chain-based approach for solving the IE task from Russian-language texts with limited topics. For this, two research questions were formulated.

Q1. What level of quality can be achieved in IE from Russian-language texts within a subject area using zero-shot and few-shot methods based on generative LLM?

Q2. Can we combine the classic and LLM-based approaches for solving the IE task?

Experimental studies were carried out on a corpus of computational linguistics texts collected from the Habr forum (habr.com).

2 Related Work

The application of generative LLM for IE tasks can be called generative information extraction, since structural information is generated rather than extracted from text. To utilize an LLM, it is essential to formulate a query in natural language (a prompt), consisting of an instruction — a task for the model — and a text fragment — a context from which the model should extract information and generate the response.

Several approaches to generative IE can be distinguished.

- Decomposed prompting is a step-by-step, or specialized, approach when the solution is modeled sequentially using a chain of specialized models (Prompt Chaining) [4, 5]. The main subtasks are identified, and for each subtask a different prompt is developed, which can take into account the results obtained at the previous steps.
- A universal approach that assumes a unified prompt structure [6]. Chain-of-Thought Prompting (CoT) technique can be used to encode the information extraction scheme.
- QA prompt approach — a dialog approach in which an IE task is transformed into a multi-pass question-and-answer system [7].

The universal approach provides a simpler and faster solution by relying on a single query to a single LLM, avoiding the error accumulation associated with chain-of-solvers or dialog-based approaches. The dialog approach addresses the problem through multiple queries to a single LLM, implementing either a sequential chain or a reasoning tree. Both of them use prompts constructed from templates and previously extracted data, allowing for iterative corrections between the stages. The specialized approach assigns each task to a dedicated, fine-tuned model, enabling integration of various methods and the creation of hybrid solvers. This introduces a potential error-correction mechanism, enhancing the robustness and adaptability of the solution for complex reasoning tasks.

Prompts are task-specific instructions to guide the behavior of a model [8]. Advanced techniques of prompt engineering such as CoT, self-verification, self-consistency, meta-prompting, and generated

knowledge, significantly enhance model performance. In addition, it is noted that the use of several examples (few-shot prompting) also improves model performance compared to zero-shot or one-shot prompting.

According to recent studies, the performance of using LLMs for IE remains suboptimal [9]. Consequently, it is worth splitting IE to the multiple subtasks, considering them separately (Figure 1).

1. Named Entity Recognition (NER) involves extracting the boundaries of subject domain terms and classifying them according to the subject domain ontology. Terms are the basic concepts of the subject area necessary to describe its main phenomena, processes and events and constitute the terminological core of the ontology.

2. Coreference Resolution (CR) — establishing coreference relations between the terms found in the first step and clustering them to identify all mentions of each subject domain entity in a given text.

3. Relation Extraction (RE) — predicting semantic relationships between the entities found in the first two steps, and classifying them according to the relation types specified by the ontology.

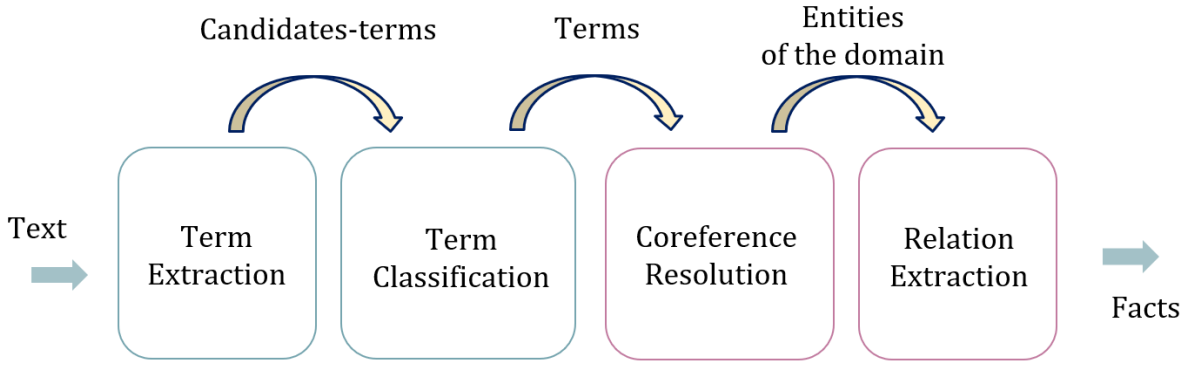


Figure 1: Main steps in information extraction

The application of LLMs to NER has been extensively studied, as this task is integral to many automatic text processing applications [10, 11]. However, solutions based on discriminative models and Supervised Fine-Tuning (SFT) still maintain an advantage over generative models, particularly in specialized knowledge domains. For instance, quality scores on the GENIA benchmark (biomedical data) reached 80–82% F_1-measure, whereas the best-performing generative model achieved 64.4% [12]. Furthermore, there is considerable variation in the quality of different generative models, highlighting the instability of these solutions.

Despite this, the generative-based approaches are still being actively explored [13]. In [14], two methods for generating prompts for the RE task are proposed: (a) based on knowledge of the possible types of relations between pairs of retrieved entities, and (b) using Question-Answering (QA) techniques. The quality of solutions to the relation classification task varies considerably, both depending on the dataset and across different models.

The CR task is also more efficiently addressed using SFT [15]. Although, typically additional entity information is not used, the authors of [16] prove that semantic information about entities enhances model predictions.

Competency questions (CQ) [17], which are essential to many ontology development methodologies, describe and constrain the domain of knowledge represented in the ontology. They define what the ontology should be able to answer. CQs can be generated automatically [18] and used as supplementary knowledge when composing prompts [19].

3 Preparing the data

Data preparation in IE tasks implies the creation of a representative corpus of texts of the subject domain, which serves as a basis for verification of results at all stages of the solving of the task. For this study, we collected a corpus of texts from the Russian-language website Habr related to the field of computational linguistics (CL) for the last 10 years, with a volume of about 1.5 thousand texts. The selection of necessary articles was based on the collected list of hubs. Based on this corpus, the subject area was analysed and datasets were created.

To capture the information important for a given subject domain (SD), formalise the structure of this information and all the data needed to retrieve it, a methodology was used to develop a subject domain ontology based on a base ontology whose concepts are specialised to more precise SD concepts and relations. In this paper, a top-level ontology on Computer Linguistics (CL) was used, which was obtained by instantiating the basic ontology of Scientific Knowledge [20]. The created ontology contains 17 classes and 111 relations. A fragment of this ontology is shown in Figure 2.

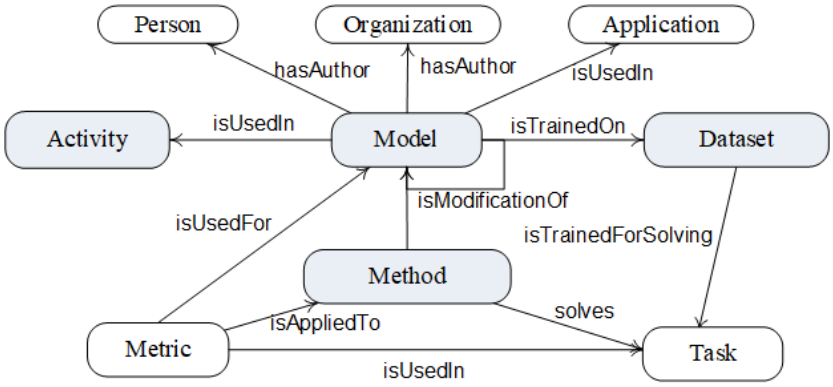


Figure 2: A fragment of the ontology on computational linguistics

To apply machine learning techniques to the IE tasks, we manually created a dataset by annotating terms, term classes, and the semantic relationships between terms. For the experimental studies, relations that were rarely mentioned in the texts were excluded from consideration. A total of 17 classes and 59 relations were used to mark up the datasets (the number of unique relation class names is 20; however, in the developed approach, the same-name relationships between different entity types are considered different, which allows for better analysis and evaluation of results).

To facilitate manual term annotation, the texts were initially processed using the SpaCy and a subject-specific dictionary. This dictionary was generated automatically from a top-level ontology, a thesaurus, and a web-portal on computational linguistics [20], and included 2640 words and word combinations related to the ontology classes.

The annotated relations were named based on the corresponding properties from the ontology, using the template <Subj.Property.Obj>, for example ***Model.isTraindOn.Dataset***. The format used in the CoNLL competition was used to annotate coreferential relations [15].

This resulted in the creation of two datasets.

a) The annotated corpus SentIE_CL contained 1088 informative sentences, in which 3136 terms and 1517 relations were annotated using the BIO format. The SentIE_CL dataset was used to train models using SFT.

b) Another dataset (DocIE_CL) is in json format and includes the markup of 4 full articles. The dataset containing 91 sentences, 606 terms, 232 semantic relations and 64 coreferences, was used for evaluation.

The created dataset and a complete list of the considered classes of terms and relations is presented at GitHub repository[1].

4 Approach to Information Extraction

To populate an ontology and build a knowledge graph, it is necessary to extract facts about entities of the subject domain defined in the ontology. The process of fact extraction included two steps: term extraction and relation extraction (Figure 1).

Based on the retrieved terms and their classes, entities (objects of a certain class) were created. Several terms may refer to the same entity, so before extracting the relations defined by the ontology, it was necessary to resolve coreference between the extracted terms (ideally also anaphora for pronouns). Next, the relations between terms were identified, and the attributes and relations (DataType Property and Object Property) between entities were established based on the relations type. It should be noted that

[1] https://github.com/Inscriptor/approach-to-IE-based-on-a-chain-of-LLMs.

for all the tasks considered, precision was prioritized, as the ultimate goal was to construct an accurate knowledge graph.

4.1 Comparative study of the approaches to Term Extraction

For the term extraction task, several LLM-based and prompt engineering approaches have been investigated and compared with the SFT-based approach, and two hybrid approaches have been considered. These approaches are an extension of the methods proposed by the authors earlier [20, 21].

The following instructions were developed for term extraction, classification and verification (Figure 3).

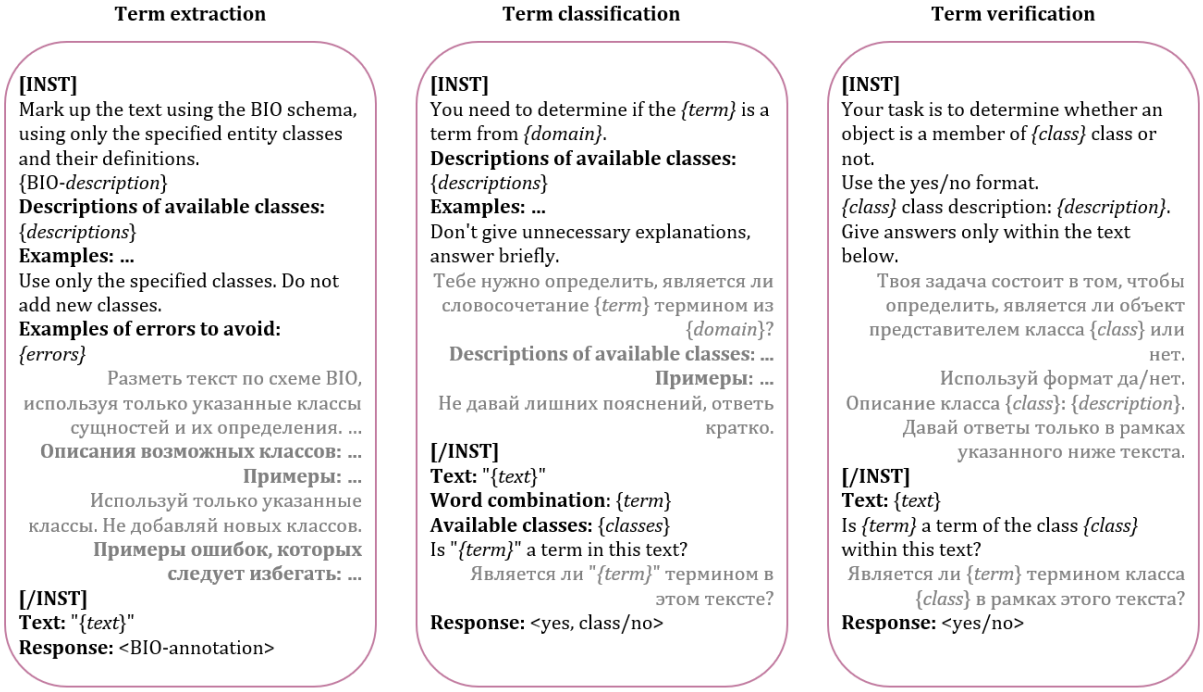


Figure 3: Instruction templates for term extraction: translations are black, originals (in Russian) are grey

The ruRoBERTa model [22] was employed in the SFT-based approaches. Additionally, three Mistral-based models — Mistral (Mistral-7B-Instruct-v0.3), Nemo (Mistral-Nemo-Instruct-2407), and the Mixtral (Mixtral-8x7b-Instruct) [23] — along with GPT-4 [24], were utilized for prompt engineering. To obtain embeddings, the Sentence-BERT model (paraphrase-multilingual-mpnet-base-v2) [25] was applied.

The following approaches were considered.

1. **ruRoBERTa (identify) + ruRoBERTa (classify) —** training the roBERTa model on the SentIE CL data first to extract terms, then to classify them in a given context [20].
2. **GPT4 (extract)** and **Mixtral (extract)** — instruction-based term extraction.
3. **Mixtral (extract) + Mixtral (verify)** — extracting terms using instruction designed for extraction and then verifying using an instruction designed for verification.
4. **ruRoBERTa (identify) + ruRoBERTa (classify) + Mistral-based LLM (verify)** — applying the first model and verifying terms afterwards.
5. **LSP + Sentence-BERT + Vocabulary + Mistral-based LLM (classify) + Mistral-based LLM (verify)** — a hybrid method that integrates linguistic and neural network models; linguistic models are used to generate hypotheses and LLM-based approaches to test and refine them (Figure 4).

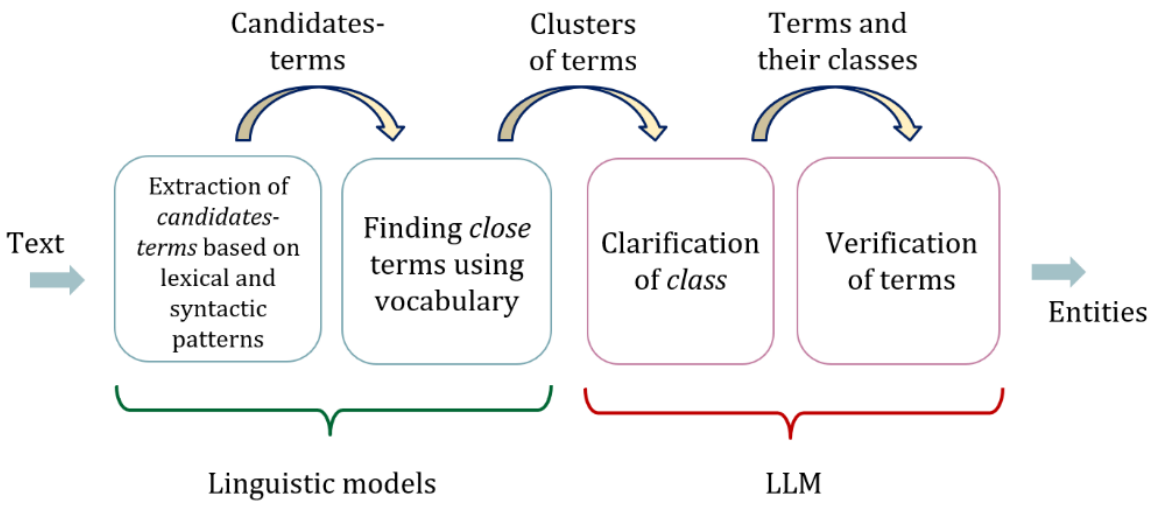


Figure 4: Hybrid approach to term extraction

The proposed hybrid approach for term extraction includes four components:

1. Lexico-syntactic patterns **(LSP)** allows the extraction of name groups based on 16 patterns.
2. **Vocabulary + Sentence-BERT** — subject dictionary and the method of comparison of embeddings obtained with the Sentence-BERT model allows to find for selected name groups close dictionary terms and on this basis to determine whether the new term belongs to the subject domain and to build hypotheses about its class (up to 6 classes).
3. **Mistral-based LLM (classify)** — predicting for each extracted term the most likely class from a set of predefined classes.
4. **Mistral-based LLM (verify)** — instruction-based verification of a term and its class.

Figure 5 shows examples of input and output data in the classification and verification stages of the suggested approach.

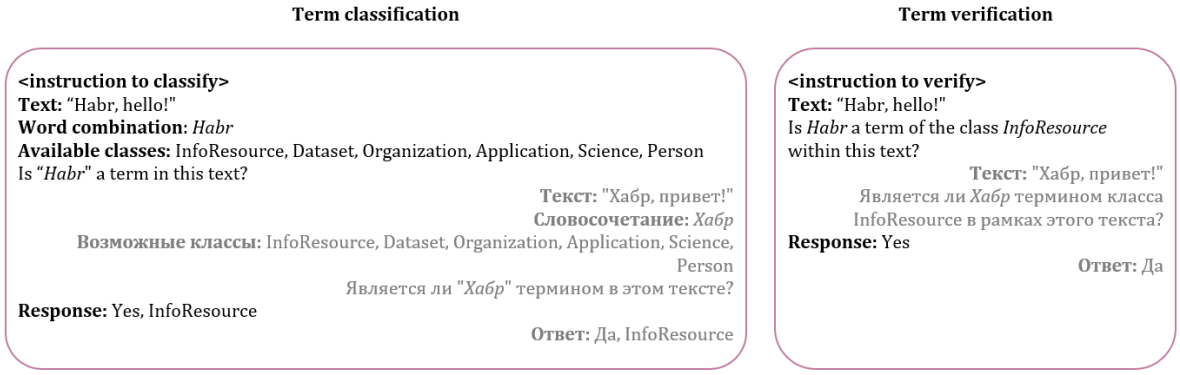


Figure 5: Example of input and output data in the classification and verification stages

In the example above, six available classes were proposed for the term *Habr* based on comparison with dictionary terms. At the first step, the model identified one class — ***InfoResource*** — using the prompt for classification. In the second step, the model performed a binary check of this result using the prompt for verification.

The obtained results for the above approaches are presented in Table 1.

Method	Recall	Precision	F_1
ruRoBERTa (identify) + ruRoBERTa (classify)	**0.81**	0.71	0.73
GPT4 (extract)	0.53	0.61	0.56
Mixtral (extract)	0.71	0.11	0.19
Mixtral (extract) + Mixtral (verify)	0.63	0.17	0.27
LSP + Vocabulary + Sentence-BERT + Mistral (classify) + Mistral (verify)	0.41	0.24	0.3
LSP + Vocabulary + Sentence-BERT + Nemo (classify) + Nemo (verify)	0.41	0.3	0.34
LSP + Vocabulary + Sentence-BERT + Mixtral (classify) + Mixtral (verify)	**0.78**	**0.71**	**0.74**
ruRoBERTa (identify) + ruRoBERTa (classify) + Mistral (verify)	0.44	0.35	0.39
ruRoBERTa (identify) + ruRoBERTa (classify) + Nemo (verify)	0.58	0.32	0.41
ruRoBERTa (identify) + ruRoBERTa (classify) + Mixtral (verify)	0.73	**0.89**	**0.77**

Table 1: Results of the term extraction and classifying

It can be concluded that adopting a comprehensive approach improved recall by leveraging templates and sufficiently large models, such as Mixtral. In contrast, the use of smaller, lightweight models has led to a decline in performance, primarily due to their lower classification quality. Using a combination of ruRoBERTa models for term extraction, ruRoBERTa for classification and LLMs for verification enabled the achievement of sufficiently high precision.

When analyzing the results, the following conclusions were made.

1. In the absence of verification, all methods extracted redundant terms, for which no corresponding ontology class existed.

2. A classifier may predict the term class differently depending on the context. In particular, some terms belonging to the ***Model*** class were incorrectly assigned to the ***Application*** class.

3. Challenges in identifying the complete phrase that is a term led to errors in extracting terms belonging to the ***Activity***, ***Dataset*** and ***Object*** classes. For instance, terms, such as *project model* instead of *project*, or *continuing experiments with optimizations* instead of *experiments* were assigned to the ***Activity*** category.

4.2 Coreference Resolution

Currently, the coreference resolution problem for the Russian language lacks robust solutions achieving performance comparable to the NER task [26], especially in restricted subject domains. As a result, significant limitations were introduced for the purposes of this study.

- Coreferential relations were sought only for the terms found (pronouns and name groups not related to subject domain terms were not considered).
- Relations were established only between terms of the same class (there was no deep hierarchy of classes in the subject domain).
- Coreference relations were established only within a single paragraph.

Coreference resolution was carried out using an LLM with an instruction designed for coreferential relation extraction. The model was provided with an instruction, the context in which a potential term-antecedent appeared, and additional auxiliary guidelines. Based on the model's responses, terms were clustered according to their coreferential relationships. Each resulting cluster corresponded to a distinct entity.

The results for some of the most representative classes of terms (considering only cases where the test dataset contained at least 30 occurrences of coreferential relations for each class) are in Figure 6. The mean F_1-measure for terms of all classes was 89.7%.

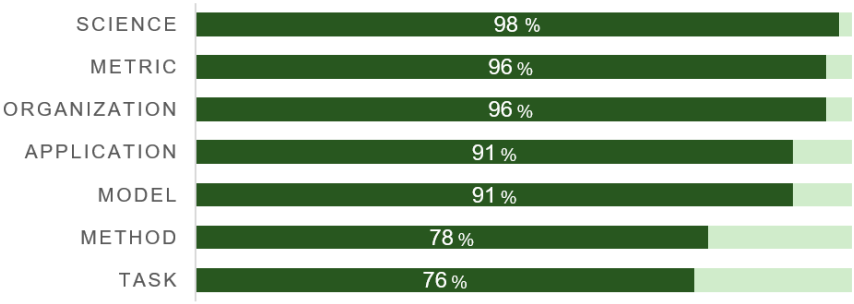


Figure 6: Coreference resolution results

4.3 Special Cases Study in Relation Extraction

The relation extraction involves identifying relations between the extracted entity terms. The set of possible relations is defined within the ontology, meaning that a relation between two terms is only possible if it exists between their respective classes in the ontology. Therefore, this task can be framed as a task of classifying the relation between a given pair of terms that are obtained in the previous steps and meet this requirement.

For each pair of terms, an instruction was generated, which included a text fragment (either a sentence or a paragraph) containing them, descriptions of the term classes, an explanation of the relation type to be tested, and examples of desired model responses (Figure 7).

Extraction of coreferential relations

[INST]
Your task is to determine whether or not the terms are names for the same entity in this text.
Answer only yes or no to the question below.
Твоя задача состоит в том, чтобы определять, являются термины названиями одной и той же сущности в этом тексте или нет.
Ответь только да или нет на вопрос ниже.
[/INST]
Text: {*text*}
Term {*class1*}: {*term1*}
Term {*class2*}: {*term2*}
Are {*term1*} and {*term2*} names of the same entity in this text?
Являются {*term1*} и {*term2*} названиями одной и той же сущности в этом тексте?
Response: <yes/no>

Relation extraction

[INST]
Your task is to determine whether a relationship exists between a pair of terms that belong to classes {*class1*} and {*class2*}.
The following questions may help you:
{*questions_list*}
Examples: {*examples_list*}
If a relationship exists, choose a relation for a pair of terms from the relation set {*relations_list*}. If the relation does not exist, write none.
Твоя задача определить, существует ли отношение между парой терминов, которые относятся к классам {*class1*} и {*class2*}.
Тебе могут помочь следующие вопросы: ...
Примеры: ...
Если отношение существует, выбери отношение для пары терминов из набора отношений {*relations_list*}. Если отношения нет, то напиши none.
[/INST]
Text: {*text*}
Term {*class1*}: {*term1*}
Term {*class2*}: {*term2*}
Is there an appropriate relationship between the terms {*term1*} and {*term2*} in this text?
Есть ли подходящее отношение между терминами {*term1*} и {*term2*} в этом тексте?
Response: <yes, relation/none>

Special relation extraction

[INST]
<...>
The terms should be arranged according to the meaning of the relation.
Термины должны быть упорядочены по смыслу отношения.
<...>
If a suitable relation exists, write the terms in the correct order so that it matches the direction of the relation from {*relation_names*}.
Если подходящее отношение есть, то запиши термины в правильном порядке, чтобы он соответствовал направлению отношения из {*relation_names*}.
[/INST]
Text: {*text*}
Term {*class1*}: {*term1*}
Term {*class2*}: {*term2*}
Is there an appropriate relation between the terms {*term1*} and {*term2*} in this text?
Есть ли подходящее отношение между терминами {*term1*} и {*term2*} в этом тексте?
Response: <yes, relation/no>

Figure 7: Instruction templates for relation extraction

Figure 8 shows the examples of the input and output data.

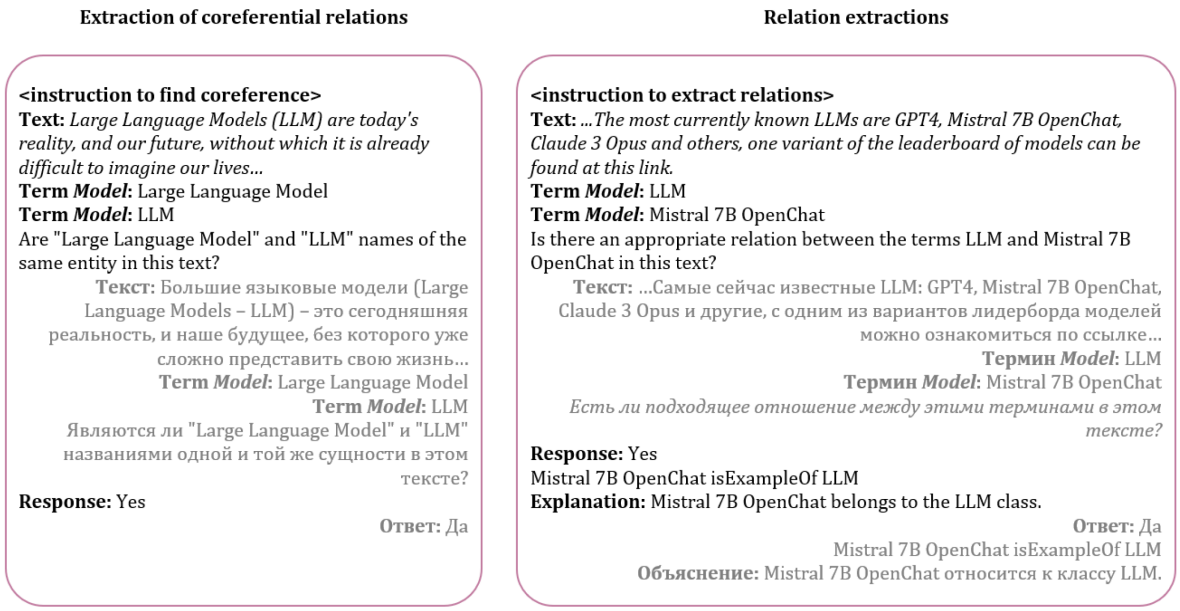


Figure 8: Examples of input and output data for the extraction of relations

In the example, the model receives as input the text fragment (sentence or paragraph) and two terms belonging to the class ***Model***. In the first case, the model calculates whether these terms are names of the same entity and gives a binary answer. In the second case, it answers the question whether there is a relationship between the given terms. The list of possible relations {relation_list} is computed when generating a prompt based on an ontology (a list of possible relations between given classes). In this case:

relation_list = {isModificationOf, isExampleOf, isPartOf, isAlternativeNameFor}

The model response includes three components: a binary response about the presence/absence of a relation, an indication of the type of relation, and an explanation of the response (this component was added later and performed better than the non-explanation response).

The average F1-measure was 84.7% (relations were extracted only within one sentence). The detailed results for all base types of relations are presented in [21].

A further refinement of the approach represented in the paper involves addressing cases that the model struggled to handle effectively. The primary complications arose from asymmetric relations between entities of the same class, such as inclusion relation (***Method.isPartOf.Method***) or example relations (***Method.isExampleOf.Method***). To improve the model's handling of these cases, an instruction for extracting asymmetric relations was developed, resulting in an increase in the precision. The change of recall and precision for the ***Model.isExampleOf.Model*** relation are detailed in Table 2.

Model.isExampleOf.Model	Recall	Precision
Before introducing the specialized instruction	0.50	0.28
After introducing the specialized instruction	0.21	**0.35**

Table 2: Results of relation extraction with standard and specialized instructions

The results presented in the table lead to a conclusion that while the precision increases significantly the recall decreases, which is in line with the expected, considering the ultimate goal of the ontology population, that was stated above. Therefore, the approach focused on improving the precision was considered justified and reasonable.

During the analysis, it was decided to subdivide the ***Model.isExampleOf.Model*** relation into two new types: hypo-hyperonymy and meronymy.

The ***Model.isExampleOf.Model*** relation is a hypo-hyperonymic relation that defines the relation between a model and the class of models to which it belongs.

The ***Model.isPartOf.Model*** relation is a meronymic relation that describes the part-to-whole relationship in situations where a model is part of another model in a literal sense, such as in the sentence, "The model created included a CRF layer" ("Созданная модель включала CRF-слой"). Once the relations were separated, their descriptions were also changed.

Splitting the ***Model.isExampleOf.Model*** relation also improved the extraction precision. The values of the metrics obtained after splitting the relation are shown in Table 3.

Relation	Recall	Precision	F_1
Model.isExampleOf.Model[2] *(before splitting into two relations)*	0.21	0.35	0.26
Model.isExampleOf.Model *(after splitting into two relations)*	**0.42**	**0.94**	**0.58**
Model.isPartOf.Model *(new relation)*	**0.33**	**0.60**	**0.43**
Model.isExampleOf.Model (Nemo) *(after splitting into two relations)*	0.12	0.34	0.17
Model.isPartOf.Model (Nemo) *(new relation)*	0.17	0.31	0.22

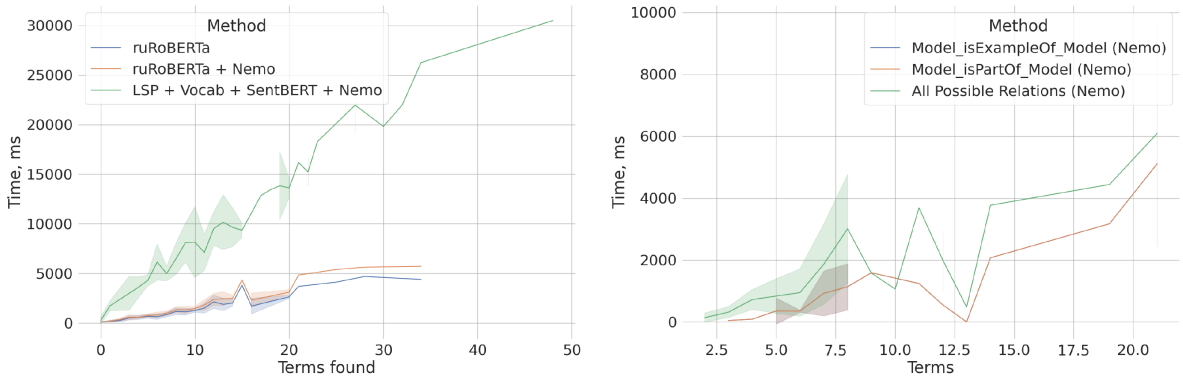
Table 3: Results of the term extraction and classifying

The values of the precision and recall metrics for the ***Model.isExampleOf.Model*** relation became higher. In addition, the values of the metrics for the ***Model.isPartOf.Model*** relation also seem to be quite high. Also, smaller models demonstrated significantly lower performance compared to the larger Mixtral model in the relation prediction task.

Figure 7 shows examples of real data sent with the prompts and received in response.

4.4 Evaluating the performance of the models

The approaches proposed by the authors are based on the application of transformer neural network models. Their computational complexity in comparison with recurrent and convolutional models is investigated in the original article [27].



a) Term Extraction b) Relation Extraction

Figure 9: Average text processing time

Figure 9 shows the average processing time per text for term extraction and the average time for predicting relations within a given set of terms. The average text volume is 383 characters or 52 words (the texts of the articles are divided into paragraphs).

The graphs are presented for local methods, i.e. methods that do not use external network resources, which showed the best results. The Mixtral model was called via the HuggingFace API, so most of the

[2] Unless otherwise specified, it is assumed that the Mixtral model was used.

time was spent on exchanging messages over the network. For this reason, the graphs for methods using this model are not shown in the figure. It is worth noting that the context length of the ruRoBERTa model is limited to 512 tokens, and instructions for large language models assume a simple response within a few tokens. This limits the potential number of terms that can be present in the text, and also simplifies the interaction with large language models, whose performance is highly dependent on the number of tokens that need to be generated in response to a query.

The term extraction time is shown as a function of the number of terms predicted by the initial model. This value affects both the text processing speed and the number of queries sent to the LLM. The graphs are presented for methods corresponding to methods **ruRoBERTa (identify) + ruRoBERTa (classify)**, **ruRoBERTa (identify) + ruRoBERTa (classify) + Nemo (verify)** и **LSP + Vocabulary + Sentence-BERT + Nemo (classify) + Nemo (verify)** in Table 1. As can be seen, the latter tends to predict a greater number of terms on average and, as a consequence, is a bit slower than methods based on the ruRoBERTa model.

The prediction time of relations on the example of relations ***Method.isExampleOf.Method***, ***Method.isPartOf.Method*** and all possible relations (see Table 2) is shown as a function of the number of terms of the corresponding classes in the given text, since a query to the LLM is formed for each pair of such terms. Coreference resolution was not considered an independent stage, since ***isAlternativeNameOf*** is considered a relation type. Therefore, the time required for coreference resolution is accordingly included in the total time for extracting all relations. As can be seen from Figure 9b, the plots for the relations ***Method.isExampleOf.Method***, ***Method.isPartOf.Method*** are almost identical.

Computations were performed on the following hardware configuration: AMD Ryzen 9 7950X, 32 GB RAM, NVidia GeForce RTX 4090 24 GB VRAM.

5 Conclusion

In this paper, an LLM-based information extraction approach (prompt chaining) was proposed and showed the following results: term extraction and classification (F_1=0.77, with the hybrid approach), simplified coreference resolution (F_1=0.89) and relation extraction (F_1=0.847). A new hybrid approach for term extraction integrating linguistic and neural network methods was proposed.

Although some of the problems were not solved, the research demonstrates that the use of LLM and few-shot prompting techniques can achieve results comparable to SFT methods. It is also noted that the involvement of an expert consistently improved the quality of the obtained solutions at every stage of the research: from the implementation of templates for assembling multi-word terms and the development of definitions for term classes to the development of instructions for extracting special relations.

The development of the approach can consist of applying data fusion techniques that show consistently good results on the NER task [28], as well as applying various techniques for automatic prompt improvement.

References

[1] Loukachevitch Natalia V., Dobrov Boris V. (2015), Developing Linguistic Ontologies in Broad Domains [Proektirovanie lingvisticheskikh ontologiy dlya informatsionnykh sistem v shirokikh predmetnykh oblastyakh], Ontology of Designing [Ontologiya Proektirovaniya], Vol. 5, no. 1(15), pp. 47–69.

[2] Ehrlinger L, Wöß W. Towards a Definition of Knowledge Graphs. Joint Proc. of the Posters and Demos Track of 12th Int. Conf. on Semantic Systems (SEMANTiCS2016) and 1st Int. Workshop on Semantic Change & Evolving Semantics (SuCCESS16). — Leipzig, Germany. — 2016. — P. 13–16.

[3] Derong Xu, Wei Chen et al. Large Language Models for Generative Information Extraction: A Survey // Frontiers of Computer Science. — 2024. — Vol. 18, no. 6. — Article no. 186357. — https://doi.org/10.1007/s11704-024-40555-y.

[4] Shichao Sun, Ruifeng Yuan, Ziqiang Cao, Wenjie Li, Pengfei Liu. Prompt Chaining or Stepwise Prompt? Refinement in Text Summarization // Findings of the Association for Computational Linguistics ACL 2024. — Bangkok, Thailand and virtual meeting. — 2024. — P. 7551–7558. — https://doi.org/10.18653/v1/2024.findings-acl.449.

[5] Khot Tushar, Trivedi Harsh et al. Decomposed Prompting: a Modular Approach for Solving Complex Tasks // Computing Research Repository. — 2022. — Vol. ArXiv:2210.02406. — version 2. Access mode: https://arxiv.org/abs/2210.02406.

[6] Yaojie Lu, Qing Liu et al. Unified Structure Generation for Universal Information Extraction. // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland. — 2022. — P. 5755–5772. — https://doi.org/10.18653/v1/2022.acl-long.395.

[7] Xiang Wei, Xingyu Cui et al. Zero-Shot Information Extraction via Chatting with ChatGPT // Computing Research Repository. — 2023. — Vol. arXiv:2302.10205. — version 2. Access mode: https://arxiv.org/abs/2302.10205.
[8] Pengfei Liu, Weizhe Yuan et al. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing // ACM Computing Surveys. — 2023. — Vol. 55, iss. 9. — P. 1–35. — https://doi.org/10.1145/3560815.
[9] Ridong Han, Tao Peng et al. Is Information Extraction Solved by ChatGPT? An Analysis of Performance, Evaluation Criteria, Robustness and Errors // Computing Research Repository. — 2023. — Vol. arXiv:2305.14450. — version 2. Access mode: https://arxiv.org/abs/2305.14450v2.
[10] Ashok D, Lipton Z C. PromptNER: Prompting for Named Entity Recognition // Computing Research Repository. — 2023. — Vol. arXiv:2305.15444. — ersion 2. Access mode: https://arxiv.org/abs/2305.15444.
[11] Tingyu Xie, Qi Li, Jian Zhang, Yan Zhang, Zuozhu Liu, Hongwei Wang. Empirical Study of Zero-Shot NER with ChatGPT // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. Singapore. — 2023. — P. 7935–7956. — https://doi.org/10.18653/v1/2023.emnlp-main.493.
[12] Wang S, Sun X, Li X, Ouyang R, Wu F, Zhang T, Li J, Wang G. GPT-NER: Named Entity Recognition via Large Language Models // Computing Research Repository. — 2023. — Vol. arXiv:2304.10428. — version 4. Access mode: https://arxiv.org/abs/2304.10428.
[13] Zhen Wan, Fei Cheng et al. GPT-RE: In-context Learning for Relation Extraction using Large Language Models // Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. — Singapore. — 2023. — P. 3534–3547. — https://doi.org/10.18653/v1/2023.emnlp-main.214.
[14] Kai Zhang, Gutierrez Bernal Jimenez, Yu Su. Aligning Instruction Tasks Unlocks Large Language Models as Zero-Shot Relation Extractors // Findings of the Association for Computational Linguistics: ACL 2023. — Toronto, Canada. — 2023. — P. 794–812. — https://doi.org/10.18653/v1/2023.findings-acl.50.
[15] Coreference Resolution on CoNLL 2012 // Paper with Code Repository. — Access mode: https://paperswithcode.com/sota/coreference-resolution-on-conll-2012.
[16] Khosla Sopan, Rose Carolyn. Using Type Information to Improve Entity Coreference Resolution // Proceedings of the First Workshop on Computational Approaches to Discourse. November 20, 2020. — Online. — 2020. — P. 20–31. — https://doi.org/10.18653/v1/2020.codi-1.3.
[17] Wiśniewski Dawid, Potoniec Jedrzej, Ławrynowicz Agnieszka, Keet Catharina Maria. Analysis of Ontology Competency Questions and their formalizations in SPARQL-OWL // Journal of Web Semantics. — 2019. — Vol. 29. — P. 105098. — https://doi.org/10.1016/j.websem.2019.100534.
[18] Pan Xueli, van Ossenbruggen Jacco, de Boer Victor, Zhisheng Huang. A RAG Approach for Generating Competency Questions in Ontology Engineering // Computing Research Repository. — 2024. — Vol. arXiv:2409.08820. — version 1. Access mode: https://arxiv.org/abs/2409.08820.
[19] Jiacheng Liu, Alisa Liu et al. Generated Knowledge Prompting for Commonsense Reasoning // Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Dublin, Ireland. — 2022. — P. 3154–3169. — https://doi.org/10.18653/v1/2022.acl-long.225.
[20] Ovchinnikova K., Ivanov A., Sidorova E. Automation of the construction of the terminological core of ontology in computer linguistics based on a corpus of texts [In Russian] // System Informatics. — 2023. — No. 23. — P. 13–32.
[21] Sidorova E., Ivanov A., Ovchinnikova K. Information extraction from texts based on ontology and large language models [In Russian] // Ontology of designing. 2025. — Vol. 15, no. 1. — P. 114–129. — https://doi.org/10.18287/2223-9537-2025-15-1-114-129.
[22] Zmitrovich D., Abramov A., Kalmykov A., Tikhonova M., Taktasheva E., Astafurov D., Baushenko M., Snegirev A., Shavrina T., Markov S., Mikhailov V., Fenogenova A. A Family of Pretrained Transformer Language Models for Russian // Computing Research Repository. — 2023. — Vol. arXiv:2309.10931. — version 4. Access mode: https://arxiv.org/abs/2309.10931.
[23] Mistral AI // Hugging Face Repository. — Access mode: https://huggingface.co/mistrala.
[24] GPT4-Turbo and GPT-4 // OpenAI. — Access mode: https://platform.openai.com/docs/models#gpt-4-turbo-and-gpt-4.
[25] paraphrase-multilingual-mpnet-base-v2 // Hugging Face Repository. — 2021. — Access mode: https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2.
[26] Dobrovolskii Vladimir, Michurina Mariia, Ivoylova Alexandra. RuCoCo: a new Russian corpus with coreference annotation // Computing Research Repository. — 2022. — Vol. arXiv:2206.04925. — version 1. Access mode: https://arxiv.org/abs/2206.04925.
[27] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I. Attention is all you need // Advances in neural information processing systems. — 2017. — No. 30. — P. 5998–6008. — https://doi.org/10.48550/arXiv.1706.03762.
[28] Hu X, Jiang Y, Liu A, Huang Z, Xie P, Huang F, Wen L, Philip S Y. Entity-to-Text based Data Augmentation for various Named Entity Recognition Tasks // Findings of the Association for Computational Linguistics: ACL 2023. — Toronto, Canada. — 2023. — P. 9072–9087. — https://doi.org/10.18653/v1/2023.findings-acl.578.

LORuGEC: the Linguistically Oriented Rule-annotated corpus for Grammatical Error Correction of Russian

Alexey Sorokin
MSU Institute for AI; Yandex
Lomonosov avenue, 27, corp. 1
a.sorokin at iai.msu.ru

Regina Nasyrova
MSU Institute for AI
r.nasyrova at iai.msu.ru

Abstract

We release LORuGEC – the first rule-annotated corpus for Russian grammatical error correction. The sentences in it are accompanied with the grammar rules governing their spelling. In total, we collected 48 rules with 348 sentences for validation and 612 for testing. LORuGEC appears to be challenging for open-source LLMs: the best F0.5-score is achieved by Qwen2.5-7B using two-stage finetuning and is only 50%. The closed YandexGPT4 Pro model achieves the score of 75%. Using a rule-informed retriever for fewshot example selection, we improve these scores up to 57% for Qwen and 81% for YandexGPT4 Pro.

Keywords: Grammatical Error Correction, Rule Annotation, Large Language Models, Fewshot Learning, Demonstration Selection.

DOI: 10.28995/2075-7182-2025-23-374-383

LORuGEC: лингвистически ориентированный корпус для задачи исправления грамматических ошибок в русском языке

Аннотация

Работа посвящена созданию корпуса LORuGEC – первого корпуса для исправления грамматических ошибок, в котором каждой ошибке сопоставлено правило, регулирующее корректное написание. Всего собрано 48 правил, корпус состоит из 348 валидационных примеров и 612 тестировочных. Из открытых языковых моделей среднего размера наилучшие результаты показывает модель Qwen2.5-7B, достигающая F0.5-меры в 50% при использовании двухступенчатого дообучения. Среди закрытых моделей лучше всех себя показывает модель YandexGPT4 Pro, чья F-мера при использовании fewshot равна 75%. При этом если использовать правиловую разметку для обучения ретривера, подбирающего демонстрационные примеры, то качество этих моделей можно улучшить до 57 и 81%, соответственно.

Ключевые слова: исправление грамматических ошибок, правиловая разметка, большие языковые модели, фьюшот-обучение, подбор демонстраций.

1 Introduction

Since the first works on Grammatical Error Correction (GEC), its primary application was for second language learning. When studying a foreign language, people tend to make multiple errors. That is why most of the corpora for GEC is based on foreign learners' texts or contain a mix of second (L2) and first (L1) language data. For example, in the case of English only the LOCNESS Corpus (Bryant et al., 2019) is based on both sources, while NUCLE(Dahlmeier et al., 2013) and Cambridge English Write&Improve Corpus (W&I)(Bryant et al., 2019) include only L2 data. The same holds for Russian, where both RULEC-GEC(Rozovskaya and Roth, 2019) and RU-Lang8(Trinh and Rozovskaya, 2021) consist of L2 and heritage data and only the recent GERA(Sorokin and Nasyrova, 2024) is based on native school texts.

As observed multiple times, L2 and L1 texts differ by the error distribution(Bryant et al., 2019; Flachs et al., 2020; Sorokin and Nasyrova, 2024). However, another factor that affects the complexity of grammatical errors is the source of the data. Most of the time, free-form essays serve as source texts for

GEC corpora. People tend to select expressions they are more confident in, reducing the risk of making grammatical errors, so some complex constructions may be underrepresented in GEC corpora. Thus, the models trained on such data have limited ability to collect complex errors, reducing their educational usefulness.

This observation was verified empirically by training large language models (LLMs) on GEC task using existing Russian corpora, such as GERA and RU-Lang8. We found that after such training, LLMs mostly improve precision, the change in recall is either less notable or even negative. Generally speaking, LLMs become more strict and less creative, which indicates that they excel in fixing "familiar" types of errors, which they were trained on, but mostly refrain from correcting other types – performing even worse than its basic version applied in a zero-shot mode.

Due to these considerations, our initial goal was to study the ability of large language models to correct complex grammatical errors. The current work primarily describes the first and the main part of this study – data collection. Our approach is case-oriented: we form a list of complex rules, using grammar handbooks as sources of data and then ask the annotators to collect sentence examples whose spelling is guided by these rules.

Eventually, we apply several models and approaches to the test sample of our data. As we expected, fine-tuning on existing corpora occurs to be suboptimal, even when the validation part of the collected data is included in the training data. In contrast, the best results were achieved by the few-shot approach. We also briefly discuss the algorithm of few-shot example selection that heavily affects the results. We hope that our data will become useful both for NLP and educational purposes. We make it freely available[1].

2 LORuGEC

We aimed at creating a Russian GEC dataset that would be more challenging for large language models (LLMs) and more linguistically-oriented, than existing corpora. LLMs are trained on large amounts of text data and therefore excel in extracting word cooccurrence patterns. Extracting such patterns is helpful in grammar-related tasks; for example, a model easily learns that the conjunction *что* requires a comma before it.

However, to implement some of rules of the Russian language, it is not enough to have profound knowledge of tokens and their co-occurrence, as they also require deeper understanding of semantics. Some of the challenges that models face are with common particles that may be written in one word or separately with the next token, depending on their meaning in a context (see Example 1), or commas that may be omitted in the sentences with several clauses (where commas, as a general rule, must be), if they have a common semantic component, for example, expressing place or time (Example 2).

(1) a. Он пошел **не** смотря вниз.

Он пошел не смотря вниз
He went not looking down
'He went without looking down.'

b. Он пошел **не**смотря на предупреждение.

Он пошел несмотря на предупреждение
He went not looking at warning
'He went despite the warning.'

(2) a. Они заполняли форму, и им приходило уведомление.

Они заполняли форму и им приходило уведомление
They filled form and to them came notification
'They filled the form, and a notification came to them.'

b. **Ранее** они заполняли форму и им приходило уведомление.

[1] https://github.com/ReginaNasyrova/LORuGEC

Ранее *они* *заполняли* *форму* *и* *им* *приходило* *уведомление*
Earlier they filled form and to them came notification

'Earlier they filled the form and [earlier] a notification came to them.'

2.1 Data

Rules of Russian grammar as well as examples to them were selected manually from the following grammar reference books, their electronic versions and educational websites (see more details on data extraction in the section below):

- High school Unified State Exam preparation books: (Berezina and Borisov, 2017) (Simakova, 2016)
- Academic handbook on spelling and punctuation: (Valgina et al., 2009), `http://orthographia.ru/`
- Handbook on the contemporary Russian language: (Valgina et al., 2002), `https://pedlib.ru/Books/6/0262/`
- Handbook on spelling and stylistics: (Rozental', 1997), `https://rosental-book.ru/`
- Dictionary of Russian collocations: (Kochneva, 1983)
- Educational web-sources: `https://orfogrammka.ru/`, `https://gramota.ru/biblioteka/spravochniki/`, `http://old-rozental.ru/`, `https://grammatika-rus.ru/`, `https://licey.net/free/4-russkii_yazyk/`, `https://www.yaklass.ru/p/russky-yazik/`

2.2 Collection and Annotation

Data were extracted and annotated by three bachelor students with a linguistic background, who are also Russian native speakers.[2] The annotators were given the following instruction:

- Select a source/sources of rules (see the list of chosen educational handbooks and websites in the section above), then choose several rules from different grammar sections: punctuation, spelling, grammar[3] and semantics. The authors of the paper also manually created a seed list of challenging rules, that the annotators were using as examples. During the annotation process, the selection of rules was additionally controlled by the paper authors.
- Find or construct 15 examples for each of the selected rules. As there is no available information on large language models' training data, to reduce the risk of compromising the dataset, several precautions were taken:
  - Preferably, choose sentences from different sources.
  - Avoid using quotations from fiction.
  - Refrain from selecting commonplace examples.
- Corrupt a fragment of each sentence which has to do with the rule it was selected for. If there may be a number of ways to make a mistake in a rule, it should be taken into account, while transforming the sentences for this rule. For instance, in Russian converbial clauses in the middle of the sentence must be marked with commas on both sides – Example 3 – so there are at least three ways of making a mistake: by overlooking the first comma, the second one or both.

  (3) Дети, гуляя по парку, ели мороженое.

  Дети *гуляя* *по* *парку* *ели* *мороженое*
  Kids walking around park ate ice cream

  'Kids ate ice cream, [while] walking around the park.'

  Our goal is to include diverse examples of errors for each rule, since it would more precisely reflect the set of possible mistakes in a text.
- For each rule test the YandexGPT3 Pro[4] model on the constructed sentences. If the model is imperfect in correcting the sentences, then try to generate several other sentences that belong to the same rule and present challenges to the model.

[2] The students earned credit hours as a result.

[3] The word *grammar* is polysemous, in GEC all kinds of errors in a text, except for the factual ones, are considered to be *grammatical errors*. Yet there are also specifically *grammatical errors*, which have to do with grammatical categories, e.g. wrong choice of number.

[4] `https://yandex.cloud/ru/docs/foundation-models/concepts/yandexgpt/models`

Since the source data creation requires targeted corruption of correct sentences, there is little chance for ambiguous corrections. In addition, a random subset of the annotated sentences was analyzed by the principal annotator, who found no discrepancies with the annotators‘ decisions.

2.3 Data Format

The dataset consists of rules, their definitions, information on their complexity for the YandexGPT Base model, pairs of corresponding tokenized[5] grammatical and ungrammatical sentences (see Table 1). There is some additional information, representing grammar sections which rules pertain to, sources of rules as well as indication of the subset for each sentence (validation or test, see more in the next section). There are few sentences in the dataset that do not contain any errors (see column *Correct source sentences* in Table 2), because it is also crucial to verify if models are prone to hypercorrection. These sentences are also marked with metadata.

The rule	**Did the base model have difficulties with the rule?**	**Initial sentence**	**Correct sentence**
Запятая перед союзом “как”: 2 [случай] Commas before the conjunction *kak*: second case	No	Иванова , как художника , я совсем не знаю . I don’t know Ivanov at all , as an artist.	Иванова как художника я совсем не знаю . I don’t know Ivanov at all as an artist.

Table 1: An example of rule from the dataset. Some metadata as well as other sentences for this rule were omitted for illustrative purposes.

We also present our data in .M2, which is a conventional GEC format. According to the .M2-standard, the source text is denoted with `S`, while the corresponding edits are prefixed with `A`. Each edit consists of the error span, error type, correction, if the edit is optional or required, additional remarks and annotator ID, yet we do not make use of error types:

S Иванова , как художника , я совсем не знаю .
A 1 2|||None||||||REQUIRED|||-NONE-|||0
A 4 5|||None||||||REQUIRED|||-NONE-|||0

2.4 Rules Description and Statistics

We gathered 48 rules from 4 grammar sections. The majority of them represent punctuation and spelling:

- **Grammar**
  - 1 Incorrect expression of government
  - 2 Declension of cardinal numerals
  - 3 Declension of numerals *poltora*, *poltory*, *poltorasta*
  - 4 Agreement between the participle and the word it defines
- **Punctuation**
  - 5 Commas in idiomatic expressions
  - 6 Commas between homogeneous subordinate clauses
  - 7 Commas between subordinate and main clauses
  - 8 Commas between the two conjunctions
  - 9-11 Commas before the conjunction *kak*: 3 instances
  - 12 Sentences with homogeneous parts
  - 13 Converbs after conjunctions

[5]We made use of NLTK Tokenizer: `https://www.nltk.org/api/nltk.tokenize.html`.

14 Clauses related to the personal pronoun
15 Clauses that are distant from the word they define
16 Punctuation in meaningful (indecomposable) expressions
17 Linking words and constructions
18 Recurring conjunctions
19 Dashes in sentences with no conjunctions
20 Dashes between the subject and the predicate
21 Dashes in case of appositions

- **Semantics**
  22 Collocations
  23 Pleonasms
- **Spelling**
  24 *n* and *nn* in the suffixes of adjectives
  25 Vowels in the suffixes of participles
  26 Noun suffixes *on'k*, *en'k*
  27 Suffixes *ic*, *ec* in neuter nouns
  28 Suffixes *ek*, *ik*
  29 Adjective suffixes *insk*, *ensk*
  30 Prefixes *pre* and *pri*
  31 *y* and *i* after prefixes
  32 Vowels after *c*
  33 Vowels after sibilants
  34 Separating letters[6]
  35 Hyphens as part of written equivalents of complex words
  36 Joint, separate or hyphenated spelling of adverbs
  37 Compound adjectives
  38 Particle *taki*
  39 *zato*
  40 *ottogo*
  41 *prichyom* and *pritom*
  42 *takzhe*
  43 *chtoby*
  44 *pol-*
  45 *ne* with verbs
  46 *ne* with adjectives
  47 *ne* with participles
  48 *ne* with nouns

Our research during the annotation showed that 29 out of 48 collected rules were challenging for the YandexGPT3 Pro. As may be observed on the Figure 1, the largest percentages of collected complex rules occur among punctuation and semantics. This partly proves our hypothesis that rules which require the understanding of semantics pose a more serious challenge to LLMs.

We collected 960 pairs of sentences, which were split into validation and test subsets so that for each rule at least 9 sentences or approximately two thirds of collected sentences would be allocated to the test partition (see Figure 2). Consequently, the size of the test subset is twice as large as the size of the validation one (see Table 2). Additionaly, unlike the latter, only the test subset includes initially correct sentences (for hypercorrection considerations). In both samples, however, two thirds of the sentences come from complex rules.

[6] ь, ъ are implied here, as they do not have literal ways of transliteration

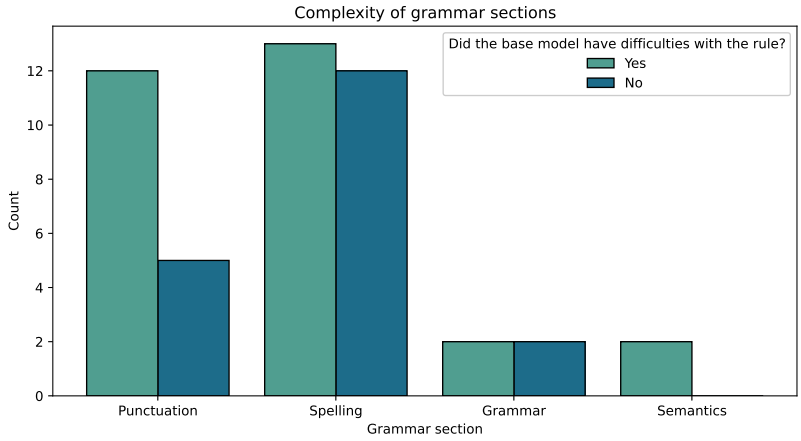


Figure 1: Complexity of different grammar sections is expressed by the number of complex for the YandexGPT3 Pro model rules. We considered the rule to be difficult if the model failed to correct some of its sentences (see 2.2).

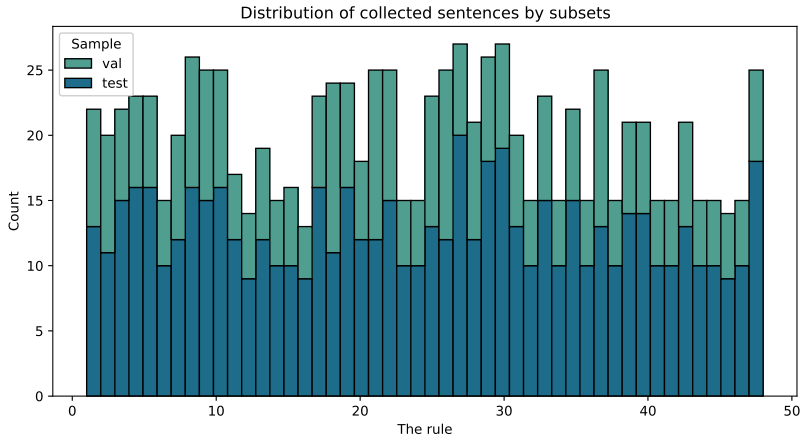


Figure 2: Distribution of sentences for each rule among validation and test samples.

Sample	**Sentences**	**Correct source sentences**	**Sentences for complex rules (%)**	**Tokens**
Validation	348	0	250 (71.84)	5,579
Test	612	31	419 (68.46)	10,131

Table 2: Statistics on the validation and test samples of LORuGEC.

2.5 Comparison to other corpora

Here we compare our corpus to existing corpora for Grammatical Error Correction of Russian: RULEC-GEC(Rozovskaya and Roth, 2019), RU-Lang8(Trinh and Rozovskaya, 2021) and GERA(Sorokin and Nasyrova, 2024). It differs from them in several aspects:

- To the best of our knowledge, that is the only GEC corpus where all the errors are matched with corresponding grammar rules.
- Our corpus is purposely created for evaluation purposes, not for training. Therefore, it has no

Sample	Sentences	Tokens
RULEC-GEC	12,480	206,258
RU-Lang8	4,412	54,741
GERA	6,681	119,068
LORuGEC	960	15,710

Table 3: Statistics on the validation and test samples of our dataset.

training subset and is much smaller than other corpora. On the other hand, almost all sentences of our corpus contain errors and are supposed to be challenging in contrast to other GEC data.

- Since corpus examples were created via corruption, for the vast majority of mistakes there is only one possible correction, increasing the trustworthiness of evaluation scores.

3 Model evaluation

Our preliminary studies demonstrated that large language models (LLMs) achieve state-of-the-art (SOTA) or near-SOTA results on the existing Russian GEC corpora, outperforming other methods. Therefore we restricted our attention to decoder-based LLMs. In the first series of our experiments we evaluate several open-source models as well as the closed YandexGPT model[7]. We prefer YandexGPT to GPT4 and other API-based large models since Russian was the main language during its training. Between the open-source models we select the multilingual Qwen2.5-3B Instruct[8], Qwen2.5-7B Instruct[9] (Yang et al., 2024) and the T-Lite 7B model[10], which is also based on Qwen. We selected these models among other variants as during preliminary experiments they showed a decent ability to correct grammatical errors in zero-shot mode, outperforming other open-source models, such as LLaMA or Mistral. Since our goal is to investigate the ability of LLMs to perform error correction task without finetuning, we selected the instruction-based versions of Qwen models. Additionally, these models show state-of-the-art performance on other Russian GEC datasets, outperforming other approaches.

In the first series of experiments we report the results of 0-shot, 1-shot and 5-shot runs. The demonstrations for fewshot are selected at random. We also evaluate finetuned versions of open-source models. Since the validation part of our corpus is rather small, we compare two variants of finetuning:

1. Train the models on the concatenation of available Russian GEC corpora: RULEC-GEC, RuLang8 and GERA.
2. Further tune the model on the validation part of the LORuGEC corpus.

As it is commonly done, we score the tokenized model outputs with M2scorer(Dahlmeier et al., 2013) and report precision, recall and F0.5 score, using F0.5 as the main metric.

The first result of our work is the difference between closed-source and open-source models. A partial explanation is the larger size of YandexGPT Pro model, however, the Lite model also clearly outperforms the open-source models. We have two possible explanations: first, many examples are taken from the school textbooks that likely were in the training data of Yandex models. Second, open-source LLMs are aligned on "creative" instruction-following tasks that require the model to rewrite the input text significantly. This makes them prone to overcorrection and hallucination that explains their moderate precision in comparison to recall.

Concerning open-source LLMs, we also observe a clear difference between the behaviour of the basic and finetuned models. The finetuned models follow the pattern of traditional GEC models based on smaller LLMs or Transformer networks as their precision is much higher than recall. Conversely, the

[7] `https://yandex.cloud/ru/docs/foundation-models/concepts/yandexgpt/models`, assessed 20th January, 2025.

[8] `https://huggingface.co/Qwen/Qwen2.5-3B-Instruct`

[9] `https://huggingface.co/Qwen/Qwen2.5-7B-Instruct`

[10] `https://huggingface.co/t-tech/T-lite-it-1.0qwen`

Model	0-shot	1-shot	5-shot	FT	FT+LORuGEC
Qwen-3B	29.4/33.9/30.2	29.9/31.1/30.1	34.8/33.1/34.4	45.1/18.6/35.1	48.4/32.1/44.0
Qwen-7B	38.6/39.8/38.9	43.5/38.4/42.4	46.7/39.5/45.0	50.8/18.4/37.6	54.2/39.8/50.5
T-Lite	31.9/48.5/34.3	37.9/43.7/38.9	41.8/44.5/42.3	54.1/22.4/42.1	61.0/46.5/57.4
YaGPT4 Lite	64.0/68.5/64.8	68.3/70.8/68.8	67.4/69.8/67.9	NA	NA
YaGPT4 Pro	68.6/72.8/69.4	72.5/71.0/72.2	75.2/73.6/74.9	NA	NA

Table 4: Comparison of different LLMs on the test set in zero-shot, fewshot and finetuning (FT) mode. FT+LORuGEC denotes two-stage funing with initial training on 3 basic Russian GEC corpora and further finetuning on the validation set of LORuGEC. We report precision, recall and F0.5 score (the main metric), separated by slashes.

pretrained models without finetuning tend to overcorrect, not only correcting grammar, but also trying to improve sentence fluency or make it more "standardized". This results in decent recall but poor precision. We suppose the alignment procedure of modern instruction-tuned LLMs to be the reason, since traditional alignment datasets contain a significant fraction of text editing tasks that require more extensive rewriting, than GEC. Additionally, the T-Lite model often fails to follow the prompt precisely, adding superfluous explanations or comments, but these format violation errors are nearly avoided with fewshot demonstrations.

To unveil the potential of fewshot learning and the usefulness of additional linguistic information, we perform a second series of experiments. We try to expose the models with the most relevant fewshot information as possible. The simplest solution could be to select the demonstrations from the same rule subset as the sentence under consideration. However, for arbitrary sentences their rule labels are not available in practice. Since the rule set is open and no corpus can cover all the grammatical rules of the language, training a rule classifier is also not a complete solution.

Our approach is to equip the demonstration selection algorithm with a similarity model. Common sentence embedding models mostly reflect semantic similarity that is irrelevant for our task. We hypothesize that grammatical similarity might be reflected by an encoder model trained on grammar-related task. To verify this hypothesis, we train an analogue of GECTOR model(Omelianchuk et al., 2020). This approach reduces grammatical error correction to sequence labeling, with labels encoding elementary edits operations. Thus, similar states of GECTOR encoder correspond to similar edit operations and, consequently, to similar grammatical rules. We train the GECTOR model on Russian grammatical error data, using the three available corpora (RULEC-GEC, RU-Lang8 and GERA) as well as up to 1M sentences with synthetic errors. During preliminary experiments we found that a single vector representation for a sentence is not enough and decided to represent a sentence by up to 3 states, corresponding to the most probable error positions.

We further tune GECTOR embedder on the task of rule classification using contrastive learning. To elaborate, for every source sentence in the validation set we search for its closest neighbour in embedding space and compel this neighbour to belong to the same class. Formally, for every embedding h we minimize the conventional triplet loss:

$$L(h, h^+, h^-) = \max(\rho(h, h^+) - \rho(h, h^-) + d, 0),$$

where h^+ is the closest example belonging to the same rule class and h^- is the closest neighbour from another class. Finetuning is performed on the validation set of our corpus.

The comparison of different embedding strategies is provided in Table 5.

We observe that GECTOR-based demonstration selection consistently outperforms the random one. This shows that encoder-based GEC models actually encode information about grammatical rules in its hidden states even when training on external GEC data. Tuning the encoder on in-domain data further

Model	k	random	GECTOR	GECTOR+ft
Qwen2.5-7B	1	43.5/38.4/42.4	47.6/42.9/46.6	**49.1/46.5/48.5**
	5	46.7/39.5/45.0	53.3/48.1/52.1	**58.9/52.9/57.6**
YandexGPT4 Lite	1	68.3/70.8/68.8	71.0/72.8/71.3	**73.4/74.4/73.6**
	5	67.4/69.8/67.9	70.9/71.6/71.0	**76.3/73.1/75.6**
YandexGPT4Pro	1	72.5/71.0/72.2	74.3/75.9/74.6	**78.5/76.9/78.2**
	5	75.2/73.6/74.9	80.4/75.4/79.3	**82.5/77.8/81.5**

Table 5: Comparison of random (*random*), embedder-based(*GECTOR*) and finetuned embedder-based (*GECTOR+ft*) for several LLMs. Best results across selection methods are in bold. We report precision, recall and F0.5 score (the main metric), separated by slashes.

improves the quality of demonstrations. Note that contrastive training requires rule labels, thus rule type annotation of our corpus is useful not only for linguistic, but also for practical purposes.

4 Conclusion

We created a linguistically-oriented evaluation corpus for Grammatical Error Correction of Russian. It appears to be challenging to current open-source models both in zero-shot mode or after finetuning on other Russian GEC corpora. However, the closed YandexGPT Pro4 model yields much higher scores, achieving the F0.5 score of 69% in zero-shot mode and 74% with 5-shot.

Since our corpus is additionally equipped with rule type information, we also show the utility of this annotation by training an encoder to assign similar vectors for examples with analogous mistakes. Using the trained encoder to select similar examples, we improve the quality of 5-shot error correction up to 81%. We hope that our study will shed additional light on the role of linguistic information in grammatical error correction and provide further insight for investigating in-context learning methods.

5 Acknowledgements

We thank Yandex and especially Marina Kosheleva for providing a grant for access to YandexGPT.

References

Svetlana Berezina and Nikolaj Borisov. 2017. *Russkij yazyk v sxemax i tablicax*. Eksmo, Moskva.

Christopher Bryant, Mariano Felice, Øistein E Andersen, and Ted Briscoe. 2019. // *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, P 52–75.

Daniel Dahlmeier, Hwee Tou Ng, and Siew Mei Wu. 2013. Building a large annotated corpus of learner English: The NUS corpus of learner English. // Joel Tetreault, Jill Burstein, and Claudia Leacock, *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, P 22–31, Atlanta, Georgia, June. Association for Computational Linguistics.

Simon Flachs, Ophélie Lacroix, Helen Yannakoudakis, Marek Rei, and Anders Søgaard. 2020. Grammatical error correction in low error density domains: A new benchmark and analyses. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8467–8478.

Elena Kochneva. 1983. *Slovar' sochetaemosti slov russkogo yazyka*. Russkij yazyk, Moskva.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. // Jill Burstein, Ekaterina Kochmar, Claudia Leacock, Nitin Madnani, Ildikó Pilán, Helen Yannakoudakis, and Torsten Zesch, *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, P 163–170, Seattle, WA, USA → Online, July. Association for Computational Linguistics.

Ditmar Rozental'. 1997. *Spravochnik po pravopisaniyu i stilistike*. Komplekt, SPB.

Alla Rozovskaya and Dan Roth. 2019. Grammar error correction in morphologically rich languages: The case of Russian. *Transactions of the Association for Computational Linguistics*, 7:1–17.

Elena Simakova. 2016. *Russkij yazyk: Novyj polnyj spravochnik dlya podgotovki k EGE'*. AST:Astrel', Moskva.

Alexey Sorokin and Regina Nasyrova. 2024. GERA: a corpus of Russian school texts annotated for Grammatical Error Correction. *// Proceedings of The 12th International Conference on Analysis of Images, Social Networks and Texts, to appear*.

Viet Anh Trinh and Alla Rozovskaya. 2021. New dataset and strong baselines for the grammatical error correction of Russian. *// Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 4103–4111.

Nina Valgina, Ditmar Rozental', and Margarita Fomina. 2002. *Sovremennyj russkij yazyk: Uchebnik*. Logos, Moskva.

Nina Valgina, Nataliya Es'kova, Ol'ga Ivanova, Svetlana Kuz'mina, Vladimir Lopatin, and Lyudmila Chel'cova. 2009. *Pravila russkoj orfografii i punktuacii. Polnyj akademicheskij spravochnik*. AST, Moskva.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Gradual Acceptability Judgments with LLMs: Evidence from Agreement Variation in Russian

Kseniia Studenikina
Lomonosov Moscow State University,
Moscow, Russia
xeanst@gmail.com

Ekaterina Lyutikova
Lomonosov Moscow State University,
Moscow, Russia
lyutikova2008@gmail.com

Anastasia Gerasimova
Lomonosov Moscow State University
Moscow, Russia
anastasiagerasimova432@gmail.com

Abstract

This study examines the LLMs' performance on gradual acceptability judgments task. Previously, the linguistic competence of LLMs was evaluated using binary acceptability scales, which contradicts the theoretical concept of acceptability. We present a new benchmark KVaS (***K**orpus **Va**riativnogo **S**oglasovanija* 'Corpus of Variable Agreement') derived from syntactic experiments on variable agreement in Russian. Our dataset contains multiple phenomena of agreement variation, ideal for modeling diverse acceptability levels, and compiles 7013 sentences rated by native speakers on a 1–7 Likert scale. We evaluated two LLMs, mainly Russian-trained GigaChat-Pro and multilingual Mistral Large, comparing their capability to treat acceptability as a scale to the reference human scores from KVaS. We used prompting providing benchmark sentences in two modes: zero-shot mode included only instructions while a few-shot mode added training sentences and their scores. The results show that GigaChat-Pro underperformed compared to Mistral Large. GigaChat-Pro improved significantly in a few-shot mode while Mistral Large exhibited more stable behavior. The case study shows that Mistral can detect nearly all significant contrasts in an experiment, whereas GigaChat performed near-randomly. The corpus may be useful for ranking LLMs, fine-tuning, and enhancing Russian text generation quality.

Keywords: linguistic competence, syntax, agreement, LLM, benchmark
DOI: 10.28995/2075-7182-2025-23-384-392

Градуальная оценка приемлемости с помощью больших языковых моделей: данные русского вариативного согласования

Студеникина К. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
xeanst@gmail.com

Лютикова Е. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
lyutikova2008@gmail.com

Герасимова А. А.
МГУ им. М. В. Ломоносова,
Москва, Россия
anastasiagerasimova432@gmail.com

Аннотация

В данном исследовании рассматривается способность больших языковых моделей (БЯМ) оценивать приемлемость предложений по градуальной шкале. Ранее способность БЯМ оценивать приемлемость языковых выражений выявлялась в бинарных задачах. Мы представляем новый бенчмарк, данными для которого послужили результаты синтаксических экспериментов, посвященных вариативному согласованию в русском языке.

Датасет включает 7013 предложений с оценками по шкале от 1 до 7, в которых представлены феномены вариативного согласования, соответствующие различным уровням приемлемости. С использованием датасета мы провели тестирование двух БЯМ на задаче градуальной оценки приемлемости в двух режимах диалога. Режим zero-shot включал только инструкцию, в режиме few-shot были добавлены тренировочные предложения и их оценки. Результаты показывают, что результат работы модели GigaChat-Pro, обучавшейся преимущественно на русскоязычных данных, зависит от режима тестирования: качество повышается в режиме few-shot. Качество мультиязычной модели Mistral Large выше и не зависит от типа инструкции. Mistral выявляет почти все значимые контрасты в одном из рассмотренных экспериментов, тогда как ответы модели GigaChat близки к случайным. Представленный корпус может быть использован не только для ранжирования БЯМ, но и для дообучения и улучшения качества русскоязычной генерации.

Ключевые слова: языковая способность, синтаксис, согласование, большие языковые модели, бенчмарк

1 Introduction

Modern large language models (LLMs) exhibit near-human proficiency in natural language processing. They are effectively used to understand and generate texts during interactions with user. The performance of LLMs is assessed through benchmarks that primarily focus on evaluating the semantic and pragmatic capabilities of an LLM, e.g. possessing information correctly, building causal relationships. For example, the SuperGLUE benchmark (Wang et al. 2019) includes tasks such as question answering and recognizing textual entailment, while the MMLU benchmark (Hendrycks et al. 2020) contains multiple choice questions from 57 subject areas.

Besides semantic adequacy, LLMs' linguistic behavior should be approximated to human behaviour based on formal linguistic criteria. Native speakers can not only produce correct linguistic expressions but also distinguish them from incorrect ones. Ideally, similar competencies are expected from LLMs. Prior assessments of LLM linguistic competence were limited to binary modes, such as the CoLA benchmark's binary classification of acceptable vs. unacceptable sentences (Warstadt et al. 2019) or the BLiMP's selection of the more acceptable sentence from a minimal pair (Warstadt et al. 2020). Meanwhile, theoretical linguistics views sentence acceptability as a gradual concept involving two aspects of a linguistic expression: (i) its grammatical well-formedness, which entails adherence to the grammatical constraints; (ii) its processing, which is subject to usage factors like word frequency, syntactic complexity, etc. (Chomsky 1965; see Sprouse 2007; Schütze, Sprouse 2014; Lau, Clark, Lappin 2017 a.o.). The existence of marginal acceptability ratings suggests that treating acceptability as a scale is a key feature of language competence.

This study aims to determine whether LLMs can assess sentence acceptability on a gradual scale. To achieve this goal, we make use of a well-studied phenomenon of Russian morphosyntax — the variable agreement. Agreement rules, being rather straightforward and unequivocal in standard cases, allow for multiple alternative strategies with non-canonical controllers, such as numeral or coordinated phrase. This phenomenon enables us to analyse the contribution of various factors to the choice of a specific agreement strategy and compare their impact in both language models and human speakers.

The rest of the paper is organized as follows. Section 2 presents our benchmark. Section 3 discusses the experiment with LLMs and evaluates its results. Section 4 concludes.

2 Corpus of Variable Agreement in Russian

We created a new benchmark called KVaS (***K**orpus **Va**riativnogo **S**oglasovanija* 'Corpus of Variable Agreement') designed to test language models on the gradual acceptability judgment task in Russian. In standard agreement, the target's grammatical form is unambiguously determined by the controller features (person, number and gender), which makes it possible to distinguish grammatical and ungrammatical variants (1). Variable agreement occurs if the calculation of the target features becomes ambiguous, which happens in case of several potential controllers (constructions with a postpositive relative clauses (2), coordinated subjects (3)) or within a non-canonical controller (quantitative noun phrases (4), quantified subjects (5)).

(1) *Marina* ***poliva-et*** / ****poliva-em*** *cvety* *po subbotam.*
Marina water-3SG / water-1PL flowers on Saturdays
'On Saturdays Marina waters the flowers.'

(2) *Vse, kto* ***zajdet/*** ***zajdut*** *v* *buhgalteriyu,*
all who come-3SG come-3PL in accounting_department
poluchat *zarplatu.*
receive-3PL salary
'Everyone who enters the accounting department will receive a salary.'

(3) *Petya i ya* ***id-em /*** ***?id-u /*** ***?id-ut /*** ****id-et*** *domoj.*
Petya and I go-1PL go -1SG go-3PL go-3SG home
'Petya and I are going home.'

(4) *Na l'du* ***svalil-as'/*** ***?svalil-is' /*** ****svalil-os'*** *ujma lyud-ej.*
On ice fell-SG.F fell-PL fell-SG.N heap people-GEN
'A lot of people fell on the ice.'

(5) *Dvoe iz nas* ***pridut /*** ***?pridet /*** ****pridem*** *v* *gosti.*
Two from us.GEN come-3PL come-3SG come-1PL in guests
'Two of us are coming to visit.'

Data for the corpus was gathered from the experimental studies conducted during the practical course on experimental syntax[1] in 2022-2024 published in the Database of Agreement Variation[2]. Each study employed a factorial design, wherein two or more independent variables (factors) were manipulated to examine their effect on a dependent variable, with one factor always corresponding to agreement pattern and the dependent variable being acceptability rating. Other factors in the experiments described different features of sentences that might influence the choice of agreement, e.g. order, semantics of the controller and target. By combining the factors, researchers investigated interactions between them, i.e. whether acceptability of an agreement pattern changes depending on the grammatical environment.

All the source studies applied the method of 1–7 Likert scale acceptability judgment task, which is a standard tool in the field of experimental syntax providing a quantifiable way to compare acceptability of different sentence structures. Unlike simpler methods, such as using asterisks (*) or question marks (?) commonly employed in generative syntax to indicate degrees of grammaticality, the Likert scale offers a more nuanced and systematic approach. An uneven number of points on the scale allows for non-binary judgments, which is especially important for detecting fine-grained effects of syntactic manipulations. To reduce the impact of lexical content on acceptability, stimuli were arranged in lists such that each sentence (lexicalization) appeared in each list under one condition (a specific combination of factor values). An example of lexicalization from the experiment on constructions with a postpositive relative clause is given in (2) and (6). The experiment examined agreement variation that emerges if there is a conflict in the number feature between the head (*vse / vse sotrudniki* 'all.PL / all.PL employees') and the relative pronoun (*kto* 'who.SG') and comprised two factors: agreement pattern and head structure.

(6) *Vse sotrudniki, kto* ***zajdet/*** ***zajdut*** *v* *buhgalteriyu,*
all employees who come-3SG come-3PL in accounting_department
poluchat *zarplatu.*
receive-3PL salary
'All employees who enters the accounting department will receive a salary.'

Alongside the stimuli, each experiment included filler sentences without any factors: grammatical ones represented fully correct sentences, while ungrammatical ones contained agreement errors. The two types of fillers are used as the threshold of (un)grammaticality when analysing ratings in experiments (Gerasimova 2023). Fillers can also serve as a baseline to demonstrate the model's ability to score non-variable agreement.

[1] The course is taught by Dr. Anastasia Gerasimova and Prof. Dr. Habil. Ekaterina Lyutikova at the Department of Theoretical and Applied Linguistics, Philological Faculty, Lomonosov Moscow State University. See more information at the website of Moscow Experimental Syntax Group, URL: https://expsynt.com/.
[2] URL: https://expsynt.com/table.html/.

The benchmark KVaS is publicly available in the GitHub repository[3]. It contains 7013 sentences from 15 experimental studies (from 66 to 137 participants each, mean number is 96, standard deviation is 20). For all experiments, outliers were excluded based on criteria from (Gerasimova 2021), from 5 to 15 participants in each study. The distribution of contexts in the corpus is shown in Figure 1. The apparent disparity in the number of examples stems from the fact that different contexts may exhibit varying number of relevant factors influencing agreement.

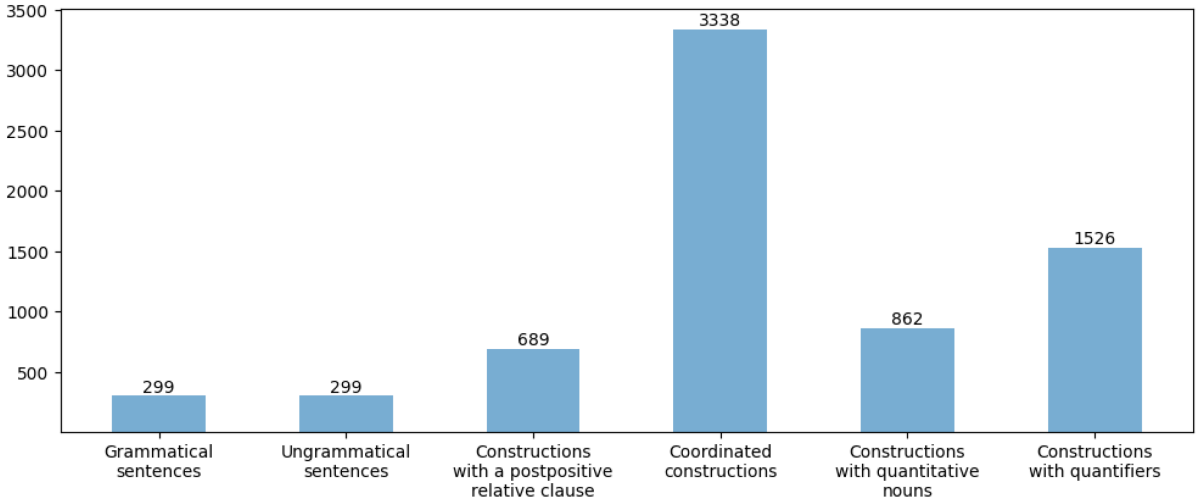


Figure 1. Distribution of contexts in the benchmark KvaS

The dataset includes the following information: the experiment code (author and year), the sentence, the average score for all participants, the type of construction / its grammaticality, the presence of variable agreement, the lexicalization number. A fragment of the corpus is shown in Table 1.

Experiment	**Sentence**	**response**	**subtype**	**type**	**lexi-caliza-tion**
Davidjuk_2023	*Ya i Maksim progulyaet poslednij urok.* 'Me and Maxim are skipping the last lesson.'	2.56	coordina-tion	stimu-lus	32
Belova_2023	*Shestero iz nas spravyatsya s zamenoj lampochek.* 'Six of us can handle replacing the light bulbs.'	5.27	quantifier	stimu-lus	11
Krainova_2023	*Eti million marok hranilis' v al'bomah.* 'These million stamps were stored in albums.'	3.88	quantita-tive_ nouns	stimu-lus	28
Dorofeeva_2023	*Kazhduyu mat', kto yavilas' na sobranie roditelej, rassprashivaet klassnyj rukovoditel'.* 'Every mother who attends the parents' meeting is questioned by the homeroom teacher.'	3.43	relative_ clauses	stimu-lus	34
Davidjuk_2023	*I Maksim, i ya progulyayut poslednij urok.* 'Both Maxim and I will skip the last lesson.'	2.70	coordina-tion	stimu-lus	32
Pasko_2024	*Na aukcione tret' skul'ptur pokupaesh' kollekcioner.* 'A collector buys a third of the sculptures at auction.'	1.81	bad	filler	505
Danilova_2023	*Klient poprosil vegetarianskoe ili postnoe menyu.* 'The client asked for a vegetarian or lean menu.'	6.03	good	filler	51

Table 1. Fragment of the benchmark KvaS

Additionally, each experiment has its own dedicated dataset with the markup from an experimental study. Table 2 shows such a dataset for T. Davidyuk' experiment (2024). This information complements the main corpus by specifying the type of conjuncts (*I* + proper name / proper name + *I*), the agreement

[3] URL: https://github.com/Xeanst/KVaS.

strategy (1st person singular / 3rd person singular / 1st person plural / 3rd person plural), the type of conjunction (*and*, *and..and*, *or*, *or..or*) and the word order (subject + predicate / predicate + subject).

sentence	**response**	**subtype**	**type**	**conjuncts**	**agreement**	**conjunction**	**order**
I ya, i Grisha sygraem shahmatnuyu partiyu. 'Grisha and I will play a chess game.'	4.50	coordination	stimul	I_noun	1pl	and_and	SV
Vlad i ya pokazhu korotkuyu dorogu. 'Vlad and I will show you a shortcut.'	2.44	coordination	stimul	noun_I	1sg	and	SV
Dima ili ya poslushayut eto golosovoe soobshchenie. 'Dima or I will listen to this voice message.'	4.88	coordination	stimul	noun_I	3pl	or	SV
Ili Mitya, ili ya sostavit etot dlinnyj spisok. 'Either Mitya or I make this long list.'	2.20	coordination	stimul	noun_I	3sg	or_or	SV
Lyova oformit etot delovoj dogovor. 'Leva will formalize this business agreement.'	6.44	good	filler	NaN	NaN	NaN	NaN

Table 2. Fragment of grammatical markup for the experiment by T. Davidyuk (2024)

The KvaS benchmark has both strengths and weaknesses compared to other acceptability judgments benchmarks. On the one hand, the existing corpora like CoLA and BLiMP cover a broader array of grammatical phenomena than KVaS, enabling the identification of which phenomena pose greater challenges for LLMs. On the other hand, they only allow for binary evaluations of language competence, which may seem overly simplistic and insufficiently nuanced. KVaS stands out as the first dataset for gradual acceptability judgments on a scale from 1 to 7. The dataset allows us to test how effectively language models trained on unlabelled texts are able to detect subtle morphosyntactic and semantic differences between sentences. Since each example in the KVaS benchmark was rated by many respondents, this markup is reliable and reflects the linguistic knowledge of Russian native speakers. Thus, KVaS may be used to examine whether LLMs' scoring consistency for different lexicalizations matches that of human evaluators.

Our corpus further enables ranking LLMs based on their level of linguistic competence, specifically in the domain of grammatical variation. In the next section, we will outline the experiment involving gradual acceptability judgments with various LLMs. The Likert-scale scores by LLMs will be compared with the reference human scores from KVaS.

3 Experiment on Gradual Acceptability Judgments with LLMs

This section describes the LLMs' evaluation on gradual acceptability judgments task using KVaS benchmark. We first present the method and the choice of models and then discuss the results.

3.1 Method

Previously, evaluation of LLMs linguistic competence mainly required either prior training on binary acceptability classification as seen with CoLA, or direct comparison of probability metrics as in BLiMP. In our study, we investigate LLMs' grammatical preferences by analysing direct responses to verbal instructions. A similar procedure was used with MMLU data on factual information (Hendrycks et al. 2021): the LLM was presented with multiple-choice questions and tasked with generating the correct answer's number. Since LLMs are specifically trained on instructional datasets, this approach aligns naturally with their capabilities. Our method not only tests the model's ability to predict that an unacceptable sentence is less likely than an acceptable one but also explores the models' concept of acceptability. Essentially, we replicate the experimental method used for human data labelling treating the model the same way.

We aimed to compare models trained primarily on Russian data with multilingual ones. The first models were expected to perform better in the gradual acceptability judgments, consistent with prior observations when contrasting the language competence of LLMs unprepared for Russian data versus those extensively trained on Russian texts (Grashchenkov et al. 2024). In this study, we tested two models: GigaChat-Pro[4] as a mainly "Russian-speaking" one and Mistral Large[5] as a multilingual one. Both were selected as top performers according to MERA benchmark (Fenogenova et al. 2024) at the time of the experiment (fall 2024) regardless of the number of parameters. The model choice was limited by API availability.

The final version of the instruction, yielding optimal results, is given in example (7). Two approaches were tested: (i) zero-shot – the prompt contains only the instructions and the target sentence, (ii) few-shot – in addition, the prompt contains two training examples: sentences and their respective scores. An example of a training pair is given in (8).

(7) *Tebe nuzhno ocenit' predlozhenie po shkale ot 1 do 7. Esli predlozhenie zvuchit horosho, tak mozhno skazat', postav' emu vysokuyu ocenku (6 ili 7). Esli predlozhenie zvuchit ploho, "ne po-russki", tak skazat' nel'zya, postav' emu nizkuyu ocenku (1 ili 2). Nekotorye predlozheniya mogut kazat'sya ne ochen' horoshimi, no v principe dopustimymi. Takim predlozheniyam postav' srednyuyu ocenku (ot 3 do 5).*
Oceni predlozhenie po shkale ot 1 do 7: "{ _}". Otvet' odnoj cifroj, nichego ne dobavlyaya.
'You need to grade the sentence on a scale from 1 to 7. If the sentence sounds good, you can say so, give it a high score (6 or 7). If the sentence sounds bad, "not in Russian", it is impossible to say so, give it a low score (1 or 2). Some sentences may not seem very good, but they are acceptable in fact. Give these sentences an average score (from 3 to 5).
Grade the sentence on a scale from 1 to 7: "{Target sentence}". Answer with one digit, without adding anything.'

(8) *Naprimer, predlozhenie "Pered paroj Yaroslav nadeli i pidzhak, i zhiletku." soderzhit oshibku. Emu stoit postavit' ocenku 1 ili 2. Predlozhenie "V aprele Marina poseyala semena i pomidora, i tykvy." yavlyaetsya vpolne estestvennym. Emu mozhno postavit' ocenku 6 ili 7.*
'For example, the sentence "Before the class, Yaroslav wore$_{PL}$ both a jacket and a vest." contains an error. It should be graded with 1 or 2. The sentence "In April, Marina sowed$_{SG}$ the seeds of both tomatoes and pumpkins." is quite natural. It can be given a score of 6 or 7.'

3.2 Results

Figure 2 shows the average human and LMMs' scores of sentences grouped by context type. Human scores for grammatical sentences fall within the upper range of the scale (6 out of 7), while ungrammatical sentences receive scores in the lower range (2 out of 7). GigaChat scores obtained in the zero-shot mode do not reflect this pattern. However, its scores generated in the few-shot mode are close to human judgments, indicating that the inclusion of training sentences enhances GigaChat performance. Conversely, there is no difference between the modes for Mistral. In both zero-shot and few-shot modes, grammatical fillers obtain high scores while the ungrammatical ones are rated low. The average scores for stimulus sentences – both for human and LLMs – lie in the middle of the scale (3-4 out of 7).

[4] URL: https://developers.sber.ru/docs/ru/gigachat/models.
[5] URL: https://docs.mistral.ai/getting-started/models/models_overview/.

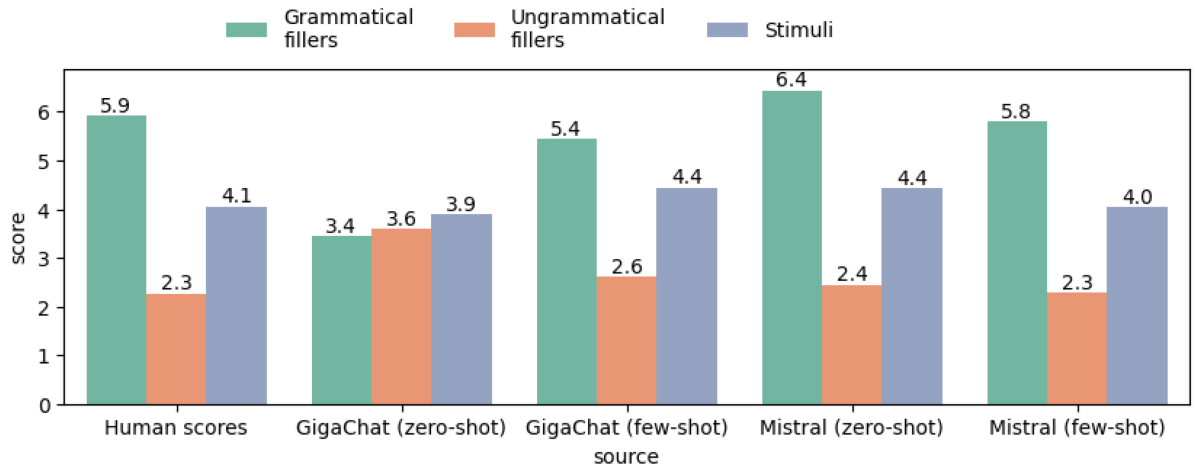


Figure 2. Average acceptability scores for human and LLMs

For a more detailed analysis of stimuli scores, we calculated the mean absolute error (MAE). The results are shown in Figure 3. The highest error for grammatical and ungrammatical fillers is obtained by GigaChat in zero-shot mode. Mistral's few-shot mode exhibits the lowest error for filler sentences. The error on stimuli for GigaChat in few-shot is slightly greater than for Mistral.

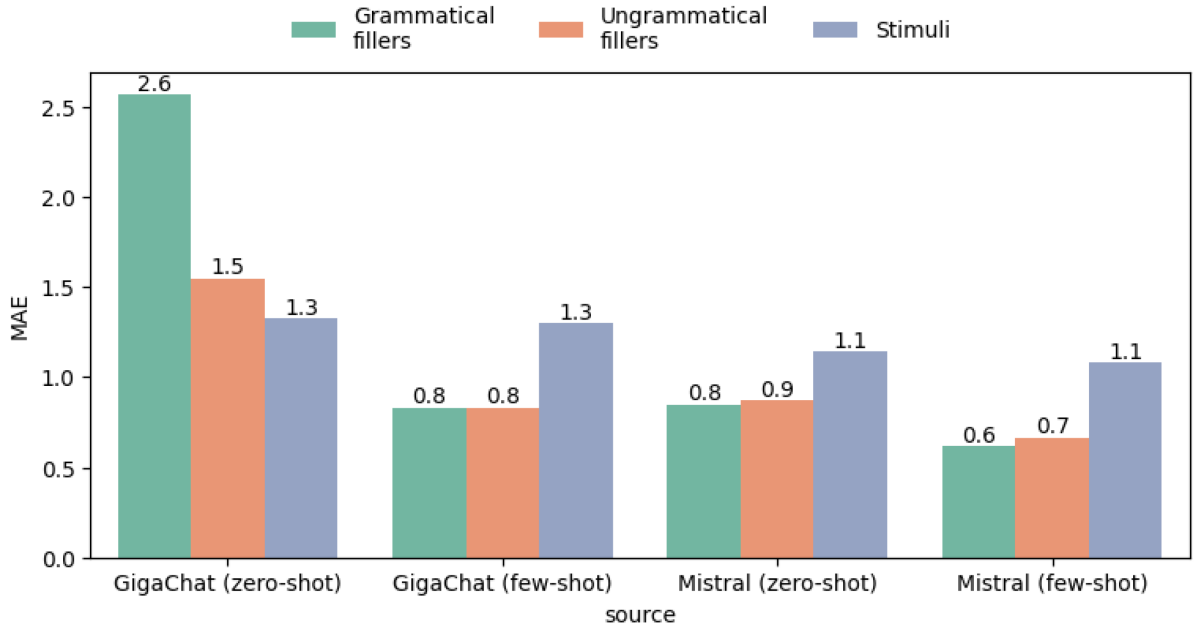


Figure 3. Mean absolute error (MAE)

MAE provides a quantitative estimate of the discrepancy between LLMs' predictions and human scores. However, it fails to indicate whether LLMs can capture relevant contrasts between experimental conditions as effectively as humans. To estimate this, we need to calculate if the conditions that showed significant differences in human experiments also exhibit significant differences in LLM scores, and conversely, if those insignificant for humans remain insignificant for LLMs. The ultimate metric is the percentage of matching pairwise comparisons between human scores and LLMs' responses.

As an illustration, we evaluated an experiment on constructions with a postpositive relative clause conducted by A. Golovnina (2022). The relevant experimental factors are agreement pattern (singular / plural verb in the embedded clause) and presence of the noun within the head (with noun / no noun) which gives four experimental conditions, see examples (2), (6) above. Figure 4 demonstrates mean values for conditions and grammatical and ungrammatical fillers. We employed the Mann-Whitney

U test and the Student's t-test to compute the significance tests[6]. The percentage of matches between human scores and LLM is presented in Table 3. The results show that most significant contrasts are captured by Mistral. The zero-shot mode performs even better than the few-shot one. On the contrary, GigaChat captures only half of the relevant contrasts, though the few-shot mode improves performance. Future research should involve conducting similar evaluations for other experiments in the KVaS.

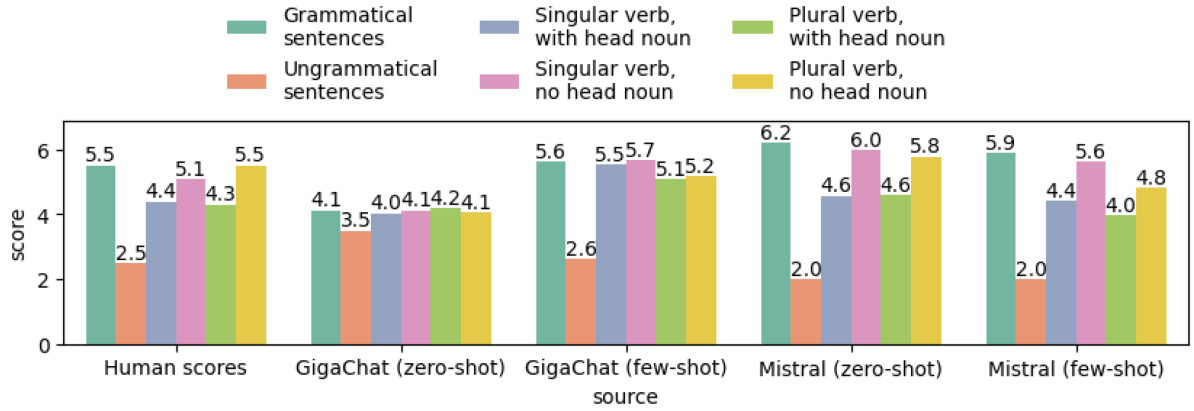


Figure 4. Mean values by condition from A. Golovnina's experiment (2022)

	Percentage of matching significant contrasts calculated with	
	Mann-Whitney U test	Student's t-test
GigaChat (zero-shot)	40%	40%
GigaChat (few-shot)	53.3%	53.3%
Mistral (zero-shot)	86.7%	86.7%
Mistral (few-shot)	73.3%	80%

Table 3. Percentage of matching significant contrasts between human and LLMs' scores

To conclude, GigaChat's performance varies depending on the selected mode, with the few-shot mode yielding superior results. Mistral's results are less dependent on the selected mode. Contrary to our hypothesis, the quantity of Russian data used during training did not affect the model performance. Multilingual Mistral shows higher proficiency than predominantly Russian-trained GigaChat. This suggests that LLMs' linguistic competence may correlate with their common-sense knowledge and rankings based on the tasks in the MERA benchmark (Fenogenova et al. 2024) or that the training data lacked a sufficient amount of variable grammatical phenomena.

4 Conclusion

This paper presents the first benchmark for evaluating LLMs on gradual acceptability judgments. Unlike binary acceptability classification, this task enables a fine-grained assessment of LLMs linguistic competence. The presented dataset comprises the results of syntactic acceptability judgment experiments on agreement variation in Russian. The stimulus sentences featuring agreement variation occupy the middle part of the acceptability scale, making them perfect material for testing gradual acceptability. The fillers presenting standard agreement serve as a baseline, occupying either high or low positions on the scale.

LLMs testing using the benchmark involved prompting. Two models – GigaChat Pro and Mistral Large – have been tested in two modes: zero-shot (instructions only) and few-shot (instructions plus training sentences). The mean absolute error (MAE) was calculated to compare the model predictions against human responses. The results revealed that GigaChat trained mainly on Russian data yielded lower quality compared to multilingual Mistral. GigaChat proved relatively unstable, with performance varying significantly between zero-shot and few-shot modes. Mistral demonstrated lower MAE value and maintained results regardless of the chosen mode. Importantly, Mistral preserved the majority of

[6] We used both parametric and non-parametric since it is unclear whether the data from LLMs is interval or ordinal.

human contrasts in the single experiment we examined, although these are preliminary conclusions as it is necessary to test all the experiments in the corpus. Our study indicates that the level of LLM linguistic competence as measured by grammatical phenomena of variation does not correlate with the amount of Russian data during training.

Acknowledgements

The work of Ekaterina A. Lyutikova and Anastasia A. Gerasimova, who collected the corpus and developed the metrics, was carried out with the support of MSU Program of Development, Project No 23-SCH02-10 “Linguistic Competence of Natural Language Speakers and Neural Network Models”. The work of Kseniia A. Studenikina, who conducted models’ evaluation, was supported by Non-commercial Foundation for Support of Science and Education “INTELLECT”.

References

[1] Chomsky N. Aspects of the theory of syntax. — Cambridge, MA : MIT Press, 1965.
[2] Davidjuk T. Agreement with disjoined subjects in Russian. — Argumentum, 2024. — Vol. 20. — P. 322–335.
[3] Fenogenova A., Chervyakov A., Martynov N., Kozlova A., Tikhonova M., Akhmetgareeva A., Emelyanov A., Shevelev D., Lebedev P., Sinev L., Isaeva U., Kolomeytseva K., Moskovskiy D., Goncharova E., Savushkin N., Mikhailova P., Minaeva A., Dimitrov D., Panchenko A., Markov S. MERA: A Comprehensive LLM Evaluation in Russian // Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics. — Bangkok, Thailand, 2024. — P. 9920–9948.
[4] Gerasimova A.A. (2021), The training materials of the workshop on experimental syntax. Selection of respondents. [Uchebnye materialy praktikuma po eksperimental'nomu sintaksisu. Otbor respondentov]. URL: https://agerasimova.com/wp-content/uploads/Gerasimova_Practice_Outliers.pdf
[5] Gerasimova A.A. (2023), Quantitative Methods of Investigating Grammar (A Case Study of Agreement Variation in Russian). PhD Thesis. Lomonosov Moscow State University. Moscow, 2023.
[6] Grashchenkov P.V., Pasko L.I., Studenikina K.A., Tikhomirov M.M. (2024). Russian parametric corpus RuParam [Parametricheskij korpus russkogo yazyka RuParam], Scientific and Technical Journal of Information Technologies, Mechanics and Optics [Nauchno-tekhnicheskij vestnik informacionnyh tekhnologij, mekhaniki i optiki.], Vol. 24, № 6, pp. 991–998.
[7] Golovnina A. S. (2022). Predikativnoje soglasovanije v otnositel’nyh klauzah s sojuznym slovom kto i tip mestoimennoj veršiny [Predicate agreement in relative clauses with the relative pronoun kto and the pronominal head type]. The student conference “Experimental studies of language”, Moscow, Russia.
[8] 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.
[9] 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.
[10] Likert R. A technique for the measurement of attitudes. — Archives of Psychology, 1932. — Vol. 22, № 140). — P. 5–55.
[11] Schütze C., Sprouse J. Judgment data // Research methods in linguistics / editors : D. Sharma, R. Podesva. – Cambridge : Cambridge University Press, 2014. – P. 27–50.
[12] Sprouse J. Continuous acceptability, categorical grammaticality, and experimental syntax // Biolinguistics. – 2007. – Vol. 1. – P. 123–134.
[13] 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.
[14] 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.
[15] Warstadt A., Singh A., Bowman S. R. Neural Network Acceptability Judgments. — Transactions of the Association for Computational Linguistics, 2019. — Vol. 7. — P. 625–641.
[16] 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.

Видовые пары глагольных идиом в русском языке: условия и механизмы аспектуальной редукции

Татевосов С. Г.
МГУ имени М. В. Ломоносова,
Москва, Россия
tatevosov@gmail.com

Киселева К. Л.
ИРЯ им. В. В. Виноградова РАН,
Москва, Россия
xkisseleva@gmail.com

Аннотация

В работе описывается аспектуальное поведение глагольных идиом в русском языке. Задача исследования — определить, какую информацию о структуре и интерпретации аспектуальной системы можно получить на фразеологическом материале. Исследование показывает, что аспектуальная редуцированность наблюдается в парах «непроизводный имперфектив — префиксальный перфектив», но не в парах «перфектив — вторичный имперфектив». Согласно нашей гипотезе, она возникает с использованием двух механизмов. Редукция к перфективу происходит в случае включения в состав идиомы лексического префикса, задающего результирующее состояние. Редукция к имперфективу происходит благодаря акциональному сдвигу, в результате которого идиома описывает ситуацию, не способную развиваться к кульминации.

Ключевые слова: идиомы, вид, префиксация, аспектуальная редукция
DOI: 10.28995/2075-7182-2025-23-393-406

Aspectual pairs in Russia idioms: conditions on and mechanisms of aspectual reduction

Tatevosov S. G.
Lomonosov Moscow State University,
Moscow, Russia
tatevosov@gmail.com

Kisseleva X. L.
Vinogradov Russian Language Institute
of the Russian Academy of Sciences,
Moscow, Russia
xkisseleva@gmail.com

Abstract

This article explores the aspectual behavior of VP idioms in Russian. The goal of the study is to find out how evidence from idiomatic material can inform a theory of the structure and interpretation of the aspectual system. The study shows that systematic gaps are observed in aspectual pairs consisting of a non-derived imperfective and a prefixed perfective verb, but not in pairs of the form "perfective — secondary imperfective". According to our hypothesis, such gaps come about due to two distinct mechanisms. First, what we call reduction to the perfective, i.e., the absence of the imperfective variant of an idiom, obtains if the idiom includes a lexical prefix. Reduction to the imperfective, in contrast, is brought about by a shift in the eventuality type which result in there only being non-culminating eventualities in the extension of the idiom.

Keywords: idioms, aspect, prefixation, aspectual reduction

1 Введение

Исследования идиоматики познавательны во многих отношениях. Они позволяют увидеть механизмы метафоризации и концептуализации разнообразных языковых значений и выявить возможности их некомпозиционального или не полностью композиционального взаимодействия. Они также позволяют получить информацию о грамматической структуре выражений, в составе которых используются.

Дж. Катц и П. Постал в [1], одной из первых работ, посвященных структуре идиом, обсуждают примеры (1)–(2):

(1) John kicked the bucket.
a. ‘Джон умер.’
b. ‘Джон пнул ведро.’

(2) ˜ The bucket was kicked by John[1].

В пассивном предложении (2) возможно только буквальное прочтение ‘пнуть ведро’, но не идиоматическое ‘умереть’, несмотря на то, что в (1)–(2) представлен идентичный лексический материал. Это говорит о том, что у идиоматизации есть не только лексические, но и структурные условия. В (1) они, по-видимому, выполнены, а в (2) нет.

Выявление этих условий среди прочего позволяет ограничить возможный анализ пассива. Дж. Катц и П. Постал, опираясь на тогдашнюю синтаксическую теорию, отклоняют анализ (2) (только) как результата трансформации (1): трансформации сохраняют семантику, и если значение ‘умереть’ представлено в (1), в результате пассивной трансформации оно не должно исчезнуть.

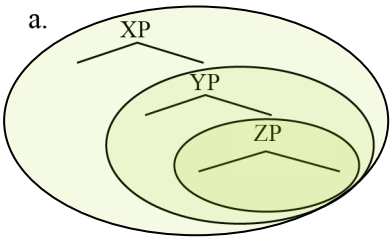
В этой работе мы излагаем предварительные наблюдения относительно случаев, похожих на (1)–(2), однако не в залогово-диатезной, а в аспектуальной сфере. С аспектуальной точки зрения идиомы изучались относительно немного (например, [2], [3], [4], [5]). Наша задача — определить, какую информацию о грамматической структуре русского языка могут дать **аспектуально редуцированные VP-идиомы**, например, *пить горькую* или *съесть собаку*. Такие идиомы содержат материал глагольной группы, в первую очередь сам глагол и прямое дополнение, а их аспектуальная редуцированность проявляется в отсутствии у них видовых коррелятов вида ˜*выпить горькую* и ˜*есть собаку*.

Дальнейшее изложение организовано следующим образом. В разделе 2 приводятся минимальные сведения о структурных ограничениях на образование идиом. В разделе 3 представлены основные эмпирические обобщения о русских VP-идиомах, у которых отсутствует ожидаемый видовой коррелят. В разделе 4 обсуждаются и критически оцениваются возможные варианты анализа этих обобщений. Материал исследования — корпус идиом, представленный в [7], из которого были отобраны VP-идиомы. Для подмножества этих идиом, описанного в разделе 3, осуществлялся поиск видовых коррелятов по корпусам (НКРЯ и интернет-ресурсы).

2 Структурные ограничения на образование идиом

Чем и в какой степени ограничено образование идиом? Важнейшая идея, которая обсуждается в литературе в последние десятилетия ([2], [7], [8], [9], [10], [11], [12], [13], [14]), состоит в том, что подвергаться идиоматизации могут только монотонно возрастающие сегменты структуры, например синтаксические составляющие все большего и большего размера, что иллюстрируется в (3a) и описывается в (3b):

(3) a.



b. Если XP представляет собой элемент идиомы, то составляющие YZ и ZP — элементы той же идиомы.

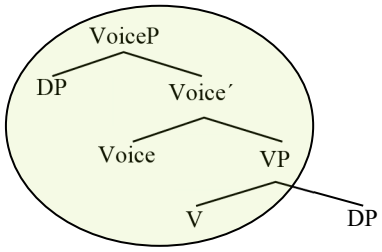
[1] Здесь и далее, помимо обычных помет «*», «?», «??», отражающих степень (не)приемлемости языкового выражения, мы используем помету «˜». Она сопровождает выражения, имеющие только буквальное значение и лишенные ожидаемого идиоматического прочтения.

Эта идея верно предсказывает целый ряд свойств идиом.

Например, если идиома включает внешний аргумент предиката, она также должна включать внутренний аргумент. В языках представлено огромное количество идиом, состоящих из глагола и его пациенса / темы (*склеить ласты*, *kick the bucket* ‘букв. пнуть ведро’, *casser sa pipe* ‘букв. сломать свою трубку’). Меньшее, но также внушительное количество идиом, иногда называемых речевыми формулами, устойчивыми коллокациями или аналогично, содержит двухвалетный глагол и оба его актанта (*гора родила мышь*, *shit hit the fan* ‘дерьмо попало в вентилятор’, *le jeu n'en vaut pas la chandelle* ‘игра не стоит свеч’). Однако полностью или почти полностью отсутствуют идиомы, состоящие из глагола и подлежащего, где позиция дополнения остается свободной для заполнения неидиоматическим материалом.

Монотонность дает этому ограничению принципиальное объяснение. Пусть в (3a) ZP = VP, XP = VoiceP, внутренний аргумент расположен внутри VP, а внешний — в спецификаторе VoiceP ([15] и последующая литература). Тогда недопустимость идиом, включающих внешний, но не внутренний аргумент, объясняется тем, что идиоматический материал не образует составляющую, как показано в (4).

(4)



(3) также предсказывает видо-временные ограничения на идиомы. В частности, если верно, что элемент, ответственный за временную интерпретацию («T»), расположен в структуре выше, чем элемент, определяющий аспектуальную («Asp»), можно ожидать следующее отношение между идиоматизацией времени и вида:

(5) Идиоматизация времени и вида
Если временная информация выступает частью идиоматического значения, то же самое происходит и с видовой информацией.

Пусть в (3a) YP = AspP, а XP = TP. Если время — составная часть идиомы, это означает, что идиоматизации подверглась TP. Но в таком случае в состав идиомы входит и AspP, составляющая TP, вместе со всей своей аспектуальной информацией.

(5) предсказывает возможность трех типов идиом — с фиксированными временем и видом, с нефиксированными временем и видом, с фиксированным видом и нефиксированным временем.

Примером первой возможности можно признать идиому *без поллитра не разберешься*, составными частями которой выступают морфологическое настоящее время и совершенный вид. По-видимому, у этой идиомы нельзя изменить ни видовую, ни временную характеристику, ср. *˜без поллитра не разобрался* и *˜без поллитра не разбирается*.

Пошла писать губерния — пример идиомы, в состав которой входит видовое, но не временное значение. Хотя у глагола *пойти* нет видовой пары в строгом смысле, он допускает ингрессивно-непредельный акциональный коррелят *идти* (*пойти в магазин — идти в магазин*). Такого коррелята нет у рассматриваемой идиомы (**шла/идет писать губерния*). Временная интерпретация идиомы, однако, не фиксирована, ср. (6) с временной референцией к будущему[2]:

[2] Идиома *пошла писать губерния* состоит из подлежащего, глагола и целевого инфинитивного оборота. Пациенс «писания» не входит в ее состав, однако и не образует открытую позицию, в которую можно внедрить прямое дополнение без потери идиоматичности. Соответственно, идиома не является контрпримером к обобщению в (4). Отметим, что для *писать* вне идиом, как и для многих других глаголов способа, синтаксическая реализация внутреннего аргумента факультативна. В составе идиомы, очевидно, используется тот вариант глагольной группы, который представляет собой дескрипцию деятельности агенса и где прямое дополнение отсутствует. Мы признательны рецензенту «Диалога», который рекомендовал нам сделать это уточнение.

(6) *Тоже нельзя, он проболтается Анке, та — Марии Алексеевне, Мария Алексеевна вызовет Клаву,… и* ***пойдет писать губерния****!* [poisk.ngonb.ru]

В отличие от *пойти* в *пошла писать губерния*, глагол *уйти* в идиоме *душа ушла в пятки* / *сердце ушло в пятки* имеет видовой коррелят *уходить*, который широко используется в составе идиомы, особенно в хабитуальных контекстах:

(7) *Бывает легче написать, чем позвонить — сердце* ***уходит в пятки****, пока слышишь гудки.* [vk.com]

Более того, для этой идиомы спорадически фиксируется и вариант с *идти*, как в (8), не отмеченный для идиомы *пошла писать губерния*:

(8) *Сюрпризов не люблю. Даже от приятных сперва* ***душа идет в пятки****. Пока в себя приду, уже и радоваться неохота.* [otvet.mail.ru]

Это третий из типов, которые предсказывает (5): ни видовая, ни временная информация не входят в семантику идиомы. Как и следовало ожидать исходя из (5), не засвидетельствован четвертый тип: такие идиомы, у которых временная интерпретация фиксирована, а вид может варьировать.

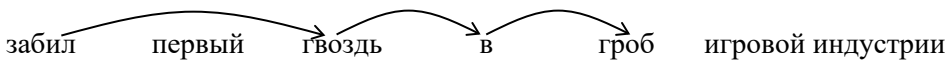
Идея структурной монотонности в (3a-b) как общий принцип, ограничивающий состав идиоматических выражений, выглядит эмпирически привлекательно. Тем не менее в том виде, в котором он сформулирован в (3b), запрет оказывается слишком сильным. Имеется значительное количество единиц, в которых он нарушается, и внутри идиомы оказываются возможны неидиоматические компоненты.

У многих единиц, содержащих именные и предложные группы, в состав идиомы не попадают разнообразные компоненты именной группы — посессоры, прилагательные, кванторные слова и т. п. *Забить гвоздь (какой-л.) в гроб (чего-л./чей-л.)* — пример идиомы, у которой *гвоздь* допускает адъективную модификацию, а *гроб* имеет открытую позицию посессора. Несколько примеров показано в (9a-c).

(9) a. *Кто, по словам разработчика, забил* ***первый гвоздь в гроб игровой индустрии****?* [kanobu.ru]
b. *Либерман забил* ***последний гвоздь в «гроб» коалиции****: Израиль ждут выборы.* [az.sputniknews.ru]
c. *Ты снова повел себя не как мужчина, а как милый, хороший парень и тем самым* ***забил жирный гвоздь в гроб ваших ЛТР****.* [forum.pickup.ru]

Ввиду многочисленных данных типа (9a-c) в литературе предлагалось, сохранив общую идею монотонности, сделать ограничения более мягкими. У. О'Грейди в [9] предположил, что идиоматическое значение создается системой синтаксических вершин, связанных друг с другом отношением зависимости. От структуры, которая внешне напоминает дерево зависимых, требуется, чтоб эти вершины составляли связную цепочку [9: 284], как в (10). Не требуется, однако, чтобы цепочка включала все вершины[3].

(10)



Б. Брюнинг [16] модифицирует эту идею, заменяя отношение зависимости отношением селекции. Не входя в подробности и не обсуждая предложенные в литературе альтернативы (например, [12], которая опирается на минималистский синтаксис), мы будем считать, что всё сказанное ниже совместимо с любой эмпирически адекватной теорией, описывающей ограничения на идиоматизацию.

[3] В таком виде ограничение оказывается слишком слабым. Оно допускает цепочку вершин, содержащую только глагол и подлежащее, и ошибочно предсказывает допустимость соответствующих идиом. Поэтому У. О'Грейди [9: 293-294] приходится вводить дополнительный принцип, опирающийся на иерархию семантически ролей: агенс может быть частью идиомы, только если пациенс — часть той же идиомы.

3 Аспектуально редуцированные VP-идиомы

3.1 VP-идиомы в русском языке

VP-идиомами мы называем идиомы, состоящие из глагола и материала (чаще всего прямого дополнения), размещающегося внутри глагольной группы; это такие идиомы, как *убить время* или *открыть Америку*. Кроме прямого дополнения, частью таких идиом могут быть пространственные и направительные адъюнкты (*давить на массу, водить за нос*), обстоятельства образа действия (*есть глазами, сидеть сложа руки*) и тому подобные элементы. В последнем случае прямое дополнение может быть частью идиомы, как в *пустить козла в огород*, или представлять собой переменную (*есть кого-л./что-л. глазами*). Кроме прямого дополнения, в VP-идиомах в качестве переменной систематически допускаются разнообразные косвенные аргументы и аргументы аппликативного типа (адресат, цель, бенефактив, малефактив и т. п.), ср. *открыть кому-л. глаза*, *смотреть на что-л. сквозь пальцы*, *играть у кого-л. на нервах*).

Важный для VP-идиом параметр варьирования — возможность модификации дополнения, обсуждаемый на русском материале, в частности, в [17] и [18], а, например, на французском — в [19].

Контраст между *нести чушь* и *заморить червячка* иллюстрируется в (11a-b).

(11) a. *нести эту/любую/всякую/какую-нибудь/несусветную/полную чушь*
b. *заморить ?этого/??всякого/??какого-нибудь/?злого/?голодного червячка*

Хотя запрет на модификацию в (11b), вероятно, не категоричен, и носители ограниченно допускают перечисленные комбинации, варианты в (11b) явно проигрывают вариантам в (11a) и в оценке приемлемости и в корпусной частотности[4].

3.2 Аспектуальная редуцированность

Рабочее определение аспектуальной редуцированности дается в (12):

(12) Аспектуальная редуцированность
Глагол V проявляет аспектуальную редуцированность в составе идиомы *I*, если в качестве элемента *I* он имеет иной состав аспектуальных возможностей, чем вне *I*.

Принимая деривационную трактовку русского вида, условимся без дальнейшего обсуждения, что аспектуальные возможности глагола описываются так, как в (13):

(13) Аспектуальные возможности глагола V
a. Набор аспектуальных дериватов, полученных из V в результате префиксации и вторичной имперфективации
b. Набор глаголов, по отношению к которым V выступает аспектуальным дериватом.

Например, аспектуальные возможности глагола *прочитать* включают *читать*, *прочитывать*, *допрочитать* и прочие не очень частотные, но довольно многочисленные дериваты с супралексическими префиксами.

В этой работе из всех аспектуальных возможностей мы рассмотрим лишь основные — когда глагол образует предельную видовую пару в терминах Е. В. Падучевой [20] или входит в состав видовой тройки. В [21] этим случаям соответствуют сильные предельные акциональные группы. Они иллюстрируются в (14)–(16):

(14) Видовая пара, состоящая из первичного имперфектива и префиксального перфектива
писать — написать, петь — спеть

[4] Например, в НКРЯ на 36 вхождений *нести чушь*, где *чушь* модифицирована прилагательным, приходится 50 случаев, когда *чушь* ничем не модифицирована. Для *заморить червячка* аналогичное соотношение — 4 к 44.

(15) Видовая пара, состоящая из непроизводного или префиксального перфектива и вторичного имперфектива
открыть — открывать, записать — записывать, решить — решать

(16) Видовая тройка
есть — съесть — съедать, сверлить — просверлить — просверливать

Существует традиция рассматривать в качестве видовой пары к перфективному элементу тройки только вторичный имперфектив (*просверлить — просверливать*). Мы, однако, будем иметь в виду все три — учитывая, что отношения между *сверлить* и *просверлить* неотличимы от (14), а между *просверлить* и *просверливать* — от (15).

3.3 Аспектуальные возможности идиом

Первое наблюдение, которое позволяет сделать наш материал, показано в (17):

(17) Общее устройство аспектуальной системы у идиом
Среди множества VP-идиом представлены те же возможности (14)–(16), что и за его пределами.

Большинство VP-идиом не показывает аспектуальных ограничений. За счет этого в идиоматической подсистеме засвидетельствованы все варианты, которые наблюдаются у обычных перфективных и имперфективных глаголов. (18), например, иллюстрирует пару *ломать копья — сломать копья*, состоящую из непроизводного имперфектива и приставочного деривата, а (19) — пару из приставочного перфектива и вторичного имперфектива *открыть глаза — открывать глаза*.

(18) a. *Ученые долго **ломали копья**, пытаясь определить, как частое использование мобильников влияет на здоровье.* [kp.ru]
b. ***Немало копий было сломано** в дискуссиях о том, чем отличается пиджак от блейзера.* [Приложение Яндекс.Ритм]

(19) a. *Пловец Колесников признался, что Олимпиада **открыла** ему **глаза** на многие вещи.* [matchtv.ru]
b. *Она неизменно ставила рядом с человеком всё многообразие природы и **открывала** ему **глаза** на «трепет жизни», на мудрую целесообразность, величие и гармонию.* [arhlib.ru]

Пример (20) показывает аспектуальную тройку *ворошить/разворошить/разворашивать (улей / осиное гнездо)*:

(20) a. *Журналист своими статьями **ворошил улей**, наглядно показывая, как советская государственная система истощала природные и человеческие ресурсы страны.* [dzen.ru]
b. *Товарищ из филармонии **разворошил улей**, и, конечно, всё, на что он только намекнул, — чистая правда.* [bloknot-voronezh.ru]
c. *Европа же двумя руками и ногами вляпывается в заигрывание с неонацистами, **разворашивая** такой **улей** — что мало никому не покажется.* [ng.ru]

Таким образом, первый вывод, к которому нас подводят данные идиоматики, состоит в том, что в идиоматической подсистеме нет глобальных пустот по сравнению с обычным для русского глагола набором аспектуальных возможностей.

В следующем разделе мы изложим более локальные наблюдения, относящиеся к идиомам с редуцированной аспектуальной парадигмой.

4 Типы аспектуальной редукции

Идиомы с редуцированной аспектуальной парадигмой, как показывают наши данные, не очень многочисленны. Более того, их наличие зависит от того, к какой группе из (14)–(16) относятся соответствующие глаголы. Эта зависимость описывается обобщением ниже:

(21) Обобщение об аспектуальной редуцированности
Аспектуальная редуцированность в идиомах наблюдается, только если вне идиом соответствующая глагольная основа образует видовые корреляты типа (14), состоящие из непроизводного имперфектива и префиксального перфектива.

В соответствии с (21) имеются два случая аспектуальной редуцированности в (22)–(23):

(22) Редукция к непроизводному имперфективу
a. Идиома представлена только в форме непроизводного имперфектива и не допускает приставочного перфективного варианта, который имеется у соответствующего глагола вне идиомы
b. *пить / ??/*выпить горькую, есть / ??/?съесть глазами, есть / ˜съесть землю*

(23) Редукция к перфективу
a. Идиома представлена только в перфективной форме и не допускает непроизводного имперфективного варианта, который имеется у соответствующего глагола вне идиомы
b. *съесть / ˜есть собаку, съесть / ˜есть пуд соли*

Согласно (21), аспектуальной редукции не наблюдается в тех идиомах, где глагол входит в пару вида (15): «префиксальный перфектив — вторичный имперфектив».

Идиомы с вторичными имперфективами, как и их неидиоматичные аналоги, всегда имеют перфективный видовой коррелят (непроизводный или префиксальный), от которого они образованы (*открывать Америку / открыть Америку, убивать время / убить время*). И наоборот, если вне идиомы перфективный глагол образует пару с вторичным имперфективом, эта возможность доступна и после идиоматизации. Вторично-имперфективные корреляты обнаруживаются даже у таких на первый взгляд типично перфективных идиом, как *дать дуба* и *двинуть кони*. В соответствующих примерах преобладают хабитуальные употребления, как в (24a), но регулярно встречаются и актуально-длительные, как в (24b):

(24) a. *Ставишь так себе на будильник голос давно умершей бабушки и с утра заслышав, что тебя зовет умершая бабушка,* ***двигаешь кони*** *с перепугу.* [habr.ru]
b. *Бомж лежал под сосной и* ***давал дуба****. Она заметила его не сразу, испугалась и остановилась, не зная, что предпринять.* [online-biblioteka.com].

Если глагол входит в видовую тройку, как, например, *есть — съесть — съедать*, то аспектуальная редукция затрагивает только отношение между первыми двумя ее элементами. *Есть собаку*, по-видимому, невозможно, в отличие от *съедать собаку*:

(25) *Многие родители на этом* ***съедают собаку****, ведь кажется, что ребенок всё таки понимает, но просто не хочет*. [Блог Виктории Токаревой]

Имея факты в (21)–(23), мы должны ответить на три вопроса:

(26) Вопросы в связи с (21)–(23)
a. Чем объясняется обобщение об аспектуальной редуцированности?
b. Как устроены аспектуально редуцированные идиомы?
c. Имеет ли описание и анализ таких идиом последствия для анализа русской аспектуальной системы?

Не беремся утверждать, что знаем окончательные ответы на эти вопросы. В разделе 5, мы, однако, постараемся существенно сузить диапазон разумных возможностей.

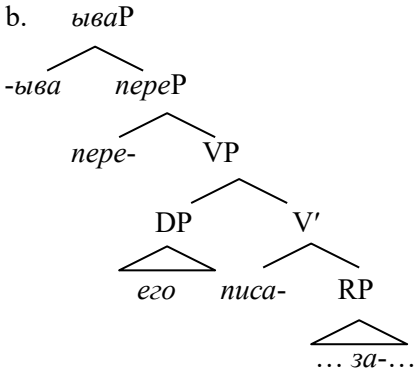
5 Идиомы как окно в структуру глагольной группы

5.1 Вторичный имперфектив

Начнем с обобщения в (21). Мы опираемся на представления о структуре русской глагольной группы и ее расширенных проекций, которые укладываются в синтаксически ориентированные варианты анализа глагольной морфологии. При таком анализе морфологические элементы, составляющие глагольную основу, рассматриваются как синтаксические вершины. (27b) показывает в качестве иллюстрации структуру предложения (27a), предполагаемую в [22].

(27) a. *Это видео мне далось не легко. Я его четыре раза* ***перезаписывал****.* [rutube.ru]

b.



По поводу обобщения (21) мы вслед за [23] делаем самое минималистичное из возможных допущений. Мы полагаем, что оно указывает на то, как материал, образующий расширенную проекцию глагольной группы, включая вторичный имперфектив, соотносится со стандартным доменом VP-идиоматизации. Вторичный имперфектив не участвует в образовании VP-идиом, а значит, находится за пределами этого домена.

(28) $[_{XP}$ … *-ыва-* … $\boxed{[_{VP}\ \textit{откры- Америку}\ ]}\ ]$

Ровно такую же структуру имеет любая вторично-имперфективная конфигурация, например, *открывать дверь* как неидиоматичный аналог *открывать Америку* или вторичный имперфектив в (27b), где позиция показателя *-ыва-* обосновывается независимыми соображениями.

Гипотеза (28) предсказывает, что возможности вторичной имперфективации перфективных глаголов в идиоматичных и неидиоматичных глагольных группах идентичны. Любая такая глагольная группа может либо присоединить показатель вторичного имперфектива (*открывать дверь/Америку*), либо продолжить деривацию без его участия (*открыть дверь/Америку*).

Приняв (28) как правдоподобную гипотезу, описывающую аспектуально полноценный вторичный имперфектив, рассмотрим две неправдоподобные, которые относятся к случаям аспектуальной редуцированности.

5.2 Две неправдоподобные возможности

Как возникает аспектуальная редуцированность, описываемая в (22)–(23)?

Вот первая возможность. Если идиома «не изменяется по виду», это означает, что в ее состав включена видовая семантика. Следовательно, идиома содержит в себе весь материал предикации вплоть до той позиции включительно, где в структуру входит аспектуальный оператор. Вариант такого описания показан в (29):

(29) [… $\boxed{[_{AspP}$ … PFV … $[_{VP}$ *съе- собаку*]]]]

(29) предполагает, что видовая семантика возникает в функциональной зоне клаузы — как и у других грамматических показателей. Обычно вершина, которая несет видовую интерпретацию, обозначается как Asp.

С (29), однако, есть проблема: Asp расположен слишком высоко — выше, чем позиция, где в структуру вводится внешний аргумент. В конфигурации такого типа последний обязан быть частью идиомы по соображениям монотонности:

(30) [… [$_{\text{AspP}}$ … PFV … [$_{\text{VoiceP}}$ … DP_{EXT} … [$_{\text{VP}}$ *съе- собаку*]]]]

VP-идиома такую структуру иметь не может. (Естественно, это верно не только для *съесть собаку*, но и для любых случаев аспектуальной редукции.)

Кроме того, если составляющая такого значительного размера свободно подвергается идиоматизации, мы утрачиваем объяснение отсутствия идиом с вторичным имперфективом, представленное в (28). Если идиомы могут включать в себя целую AspP, становится непонятно, почему мы не встречаем AspP, «захватившие» вторичный имперфектив, в качестве частного случая[5].

(29), таким образом, можно отклонить как неправдоподобную на эмпирических основаниях.

Есть другая возможность, вытекающая из того, что аспектуальная редукция — не единственное явление, для которого засвидетельствовано вариативное поведение сходных на первый взгляд идиом.

Невозможность пассива в английском языке, которая иллюстрируется выше в (1)–(2), характеризует не все VP-идиомы, а только их часть. Имеется обширный класс единиц, для которых пассив допустим, например, *spill the beans* ‘проболтаться, выдать секрет’:

(31) The beans were spilled (by Lisa).
$^{\text{OK}}$‘Секрет был выдан.’

В [7], а затем в [8] выдвигается предположение, что синтаксическое различие между *kick the bucket* в (2) и *spill the beans* в (31) можно свести к семантическому. *Spill the beans* характеризуется тем, что и глаголу и прямому дополнению приписывается нестандартная интерпретация, которая отсутствует у них за пределами идиомы: *spill* ‘рассыпать, пролить’ ⇝ ‘разгласить, выдать’, *beans* ‘бобы’ ⇝ ‘секреты, тайна’. Однако соединение этих двух элементов осуществляется стандартно — как у любого глагола с валентностью на пациенс/тему: ‘выдать’ + ‘тайна’ → ‘выдать тайну’. В этом смысле *spill the beans* представляет собой разложимую (decomposable, [12]) идиому. Это меньшая степень некомпозициональности, чем у *kick the bucket*, которая неразложима: невозможно предложить правдоподобный семантический сценарий, позволяющий получить ‘умереть’, соединив трансформировавшееся значение *kick* с трансформировавшимся значением *bucket*. Разложимые идиомы располагают большими синтаксическими возможностями по сравнению с неразложимыми (см. анализ в [12]).

Последующие исследования в этом направлении установили, что диапазон возможностей — не бинарная, а скалярная характеристика. В [11], в частности, показано, что в английском языке имеется импликативная иерархия, описывающая доступность идиом в трех типах конфигураций: в номинализациях (*spilling the beans*), структурах с инкорпорацией объекта (*bean-spilling*) и в пассивных предложениях типа (31). Сюда же можно добавить возможность для дополнения выступать в VP-идиомах мишенью релятивизации, обсуждаемую в работах Е. А. Лютиковой. На русском материале аналогичные наблюдения и обобщения были сделаны, в частности, в [25], [26] и [27].

Возникает соблазн попытаться свести наличие и отсутствие аспектуальной редуцированности к аналогичным соображениям. Различие состоит не в том, что одни идиомы включают в себя Asp, а другие нет, а в том, что разные идиомы некомпозициональны в разной степени, и это имеет последствия для их грамматической дистрибуции, в том числе видовой. Как и в случае с пассивом, разложимые идиомы показывают полный набор возможностей, а неразложимые — редуцированный.

[5] Рецензент «Диалога» задает справедливый вопрос о том, как соотносятся XP в (28), в котором размещается вторичный имперфектив, и Asp в (29). Здесь возможны два предположения. Первое: вторичный имперфектив — это одна из реализаций Asp, то есть XP и AspP — одно и то же. Второе: XP расположена внутри AspP. Аргументы в пользу второго анализа обсуждаются, в частности, в [24]. И в том и в другом случае объяснения для отсутствия идиоматичности вторичного имперфектива в (28) и для аспектуальной редукции в (29) несовместимы.

Мы, однако, не рассматриваем и эту линию анализа как правдоподобную. *Съесть собаку* и *съесть пуд соли* аспектуально редуцированы, однако с точки зрения синтаксических свойств, диагностирующих разложимость, попадают в разные классы. В корпусах без труда находятся примеры типа (32a-c), где показаны номинализация, пассив и относительное предложение, производные от *съесть пуд соли*:

(32) a. *Речь ведь шла о двух его самых близких друзьях, с которыми, как говорится,* ***не один пуд соли был съеден****.* [sinonim.org]
b. *Ключевые моменты экспедиций: <…>* ***съедение пуда*** *соли с тибетцами и любовь к Тибету навсегда!* [rgo.ru]
c. *Горький* ***пуд соли, который мы съели вместе****, переплавился в бесценный опыт.* [sevastopol.su]

Аналогичные примеры для *съесть собаку* с идиоматичным значением не обнаруживаются — ни пассив *˜(на этом) была съедена собака*, ни относительные предложения *˜собака, которую он (на этом) съел*, ни номинализация *˜съедение собаки (на этом)*. Во всех таких структурах представлено лишь буквальное значение, ср. (33) с номинализацией.

(33) *˜В Копейске мужчина пойдет под суд за* ***съедение собаки****. Ранее обвиняемый отбывал наказание за убийство.* [1obl.ru]

Имея в виду эти данные, мы приходим к неизбежному выводу. Различие между разложимыми и неразложимыми идиомами для русского языка реально (хотя вопрос о том, по каким параметрам и насколько оно отличается от описанных для английского языка случаев, нуждается в дополнительном изучении). Но аспектуальные ограничения идиом оно не предсказывает, поскольку идиомы с таким ограничениями могут находиться в разных классах с точки зрения разложимости.

5.3 Идиоматизация и префиксация

Отклонив две возможности, относящиеся ко всему классу редуцированных идиом, рассмотрим два случая редуцированности в (22)–(23) раздельно.

Мы установили, что видовая редуцированность не коррелирует с синтаксической дистрибуцией и предполагаемой разложимостью, а также не может объясняться через включение в состав идиомы элемента Asp, который размещается в функциональной зоне клаузы и отвечает за аспектуальную интерпретацию. Остается предположить, что идиома включает в себя не Asp, а лишь морфосинтаксические компоненты глагола, расположенные внутри глагольной группы и влияющие на аспектуальную интерпретацию, в частности, лексические префиксы[6]. Аспектуальная интерпретация затем однозначно вычисляется по общим правилам[7].

С таким предположением мы получаем объяснение для идиом, редуцированных к перфективу, без всяких дополнительных допущений. Однако с идиомами, редуцированными к непроизводному имперфективу, возникают затруднения.

Начнем со *съесть собаку* и аналогичных единиц. Предположим, что идиоматизации подвергается в точности глагольная группа — глагол и его зависимые. Предположим также, следуя [28], [29], [30], [31], что лексический префикс размещается внутри глагольной группы. Для конкретности примем, что префикс расположен, как в (27b) — в результативной составляющей RP, ответственной за внедрение в событийную структуру результирующего состояния [28], однако это не принципиальное допущение.

[6] О супралексических префиксах в составе идиом см. [23].
[7] Правила известны: непроизводная основа типа *ед-* → IPFV, последний шаг морфосинтаксической деривации — префиксация → PFV, последний шаг деривации — вторичный имперфектив → IPFV. См. [24], где предлагается формальная реализация этой системы.

Тогда редукция к перфективу — это включение проекции префикса в состав идиомы:

(34) [… [$_{VP}$ … [$_{RP}$…*съ-* …] *ед- собаку*]]

Аспектуальная редуцированность *съесть собаку,* то есть невозможность ˜*есть собаку*, следует из (34) потому, что глагольная основа и дополнение образуют идиому только вместе с префиксом. Никакой элемент нельзя удалить из идиомы, не потеряв идиоматического значения. Идиома возможна только до тех пор, пока все ее компоненты, в том числе префикс, находятся на месте.

Таким образом, редукция к перфективу связана с включением лексического префикса в состав идиомы.

Следуя этой логике, можно попытаться построить симметричное рассуждение для идиом, редуцированных к непроизводному имперфективу, таких, как *пить горькую*.

Будем по-прежнему считать, что идиоматизации подвергается глагольная группа, состоящая в этом случае только из глагола и прямого дополнения, как в (35). (36) показывает, почему можно ожидать, что выражение *выпить горькую* невозможно.

(35) [… [$_{VP}$ … *пи- горькую*]]

(36) ˜[… [$_{VP}$ …] [$_{RP}$… *вы* …] *пи- горькую*]]

Построение префиксального перфективного глагола *выпить* предполагает внедрение в глагольную группу префикса *вы-* — материала, не входящего в состав идиомы. Тогда в результате идиоматизации глагол и дополнение перестают образовывать составляющую. Это делает невозможным **выпить горькую* потому же, почему не бывает идиом, состоящих из глагола и подлежащего, но не включающих прямое дополнение, как в (4). Структура (36) с точностью до синтаксических категорий выглядит как структура в (4).

С этим рассуждением, однако, есть неразрешимая проблема: оно распространяется на любые VP-идиомы, состоящие из непроизводного имперфектива и прямого дополнения. Если (36) — невозможная конфигурация, образование видового коррелята с помощью префикса должно быть полностью запрещено — не только для *пить горькую*, но и, например, для идиом в (37).

(37) *играть на нервах, петь с чужого голоса, прятать голову в песок, смотреть сквозь пальцы, строить замки на песке, читать между строк*

Ни одна из них, однако, не показывает редукции к непроизводному имперфективу. Несколько иллюстраций приводится в (38):

(38) a. *Китайский онлайн-магазин* ***сыграл на нервах*** *42-летней скрипачки Антонины.* [staryiy.livejournal.com]
b. *Вы хотите* ***спеть с чужого голоса*** *всякую хрень гаражного разлива? Уже слышал, не надо.* [drive2.ru]
c. *Это просто Нина* ***построила замки на песке****, намечтав, что Никита на ней женится, предварительно разведясь с женой.* [www.kino-teatr.ru]
d. *Политолог* ***прочитал между строк*** *истинную цель санкций США.* [africa.tsargrad.tv]

Наши предварительные наблюдения указывают на то, что перфективные идиомы типа (38a-d) проигрывают своим имперфективным аналогам в частотности. Например, частотность *смотреть сквозь пальцы* в НКРЯ составляет 0,61 ipm против 0,03 ipm для *посмотреть сквозь пальцы*. Более того, соотношение этих частотностей существенно отличается от соотношения *смотреть* и *посмотреть* вне идиом — соответственно 649,01 ipm и 278,83 ipm. Очевидно, эти числа отражают тот факт, что образование перфективного коррелята для идиомы более затруднительно, чем для обычного глагола *смотреть*. Тем не менее, примеры типа (38a-d) ясно показывают, что речь в данном случае идет не о грамматическом запрете, а лишь о количественной асимметрии.

Если принять (36) как правдоподобное объяснение, система предсказывает ровно две возможности. Первая — это идиомы, редуцированные к перфективу, если в состав идиомы входит

префикс, как в (34); вторая — идиомы, редуцированные к непроизводному имперфективу, если префикс отсутствует, как в (35). Наиболее распространенный класс идиом — те, у которых нет аспектуальной редуцированности — полностью исключается.

Из этих наблюдений с неизбежностью следует вывод: невозможность *выпить горькую* нельзя объяснить тем, что внутри глагольной группы появляется неидиоматический материал, а именно, префикс. Если бы форма *выпить горькую* была недоступна из-за этого, не существовало бы и примеров (38a-d). Префикс оказывается элементом, который ведет себя так же, как посессоры и прилагательные в именных группах в примерах типа (9a-c): располагаясь внутри идиоматической VP, но не будучи компонентами идиомы, они тем не менее не разрушают идиоматическое значение. Это заставляет нас отклонить объяснение в духе (36) для идиом, редуцированных к непроизводному имперфективу.

5.4 Редукция к непроизводному имперфективу

Как в таком случае возможна редукция к непроизводному имперфективу? Мы выдвигаем следующую гипотезу:

(39) Редукция к имперфективу через изменение акциональной категории
Редукция к непроизводному имперфективу возникает в результате появления у идиомы акциональной характеристики, отличной от исходного глагола. А именно: идиома представляет собой дескрипцию деятельности / процесса, не предполагающего кульминации.

Идея, стоящая за (39), заключается в следующем. Роль лексического, в частности, так называемого чистовидового префикса типа *вы-* в *выпить*, — внедрение результирующего состояния в событийную структуру [21]. Результирующее состояние — это новое состояние одного из участников описываемой ситуации, возникающее после ее кульминации. Поэтому наличие кульминирующих событий в экстенсионале непроизводного имперфективного глагола типа *пить*, *читать* и т. п. — необходимое условие для присоединения лексических префиксов. Если глагол не описывает таких событий, т. е. представляет собой, по меткому выражению Ю. С. Маслова, глагол бесперспективного протекания, он не образует и перфективных коррелятов с лексическими префиксами, описывающими результирующее состояние. «Перфективные партнеры» таких глаголов образуются почти исключительно с помощью супралексических префиксов, — таких, например, как делимитативный *по-* или инхоативный *за-*, ср. *бегать*, *побегать*, *забегать* и т. п.

Мы предполагаем, что в идиоме *пить горькую* происходит акциональный сдвиг, уводящий ее от исходного класса глагола *пить*. В результате она начинает описывать процесс, не имеющий естественного предела и не допускающий кульминации. Примерно это же значение выражает в русском языке просторечный глагол *бухать*. Из-за этого *пить горькую* не в состоянии соединиться с префиксом *вы-*, который вводит результирующее состояние для кульминирующих процессов питья.

Как и любая единица, описывающая процессы, не направленные к кульминации, *пить горькую* соединяется с супралексическими префиксами, регулярно образуя инхоативы и делимитативы, ср. (40a-b), и реже — более экзотические дериваты наподобие эксессивного *пере-* в (40c).

(40) a. *Поплакал,* ***попил горькую*** *да и отстроил избу заново. Стал работать механизатором.* [dzen.ru]
b. *Отец их когда-то принадлежал к купеческому сословию, но после смерти жены* ***запил горькую****, дела сначала пошатнулись, а затем и вовсе развалились.* [litres.ru]
c. *Проблема даже не в том, что кто-то* ***перепил горькую****. Проблема в том, что народ массово трескает алкозаменители.* [сетевой комментарий к публикации об отравлении настойкой «Боярышник»]

В этом отношении *пить горькую* также обнаруживает параллелизм с *бухать* и его дериватами — *побухать, забухать, перебухать* и т. п.

Завершая этот раздел, отметим, что если семантически сценарий, обозначенный выше, реален, то редукция к непроизводному имперфективу и редукция к перфективу имеют принципиально

разную природу. Редукция к перфективу использует регулярный морфосинтаксический механизм идиоматизации — включение в состав идиомы конкретного лексического элемента, префикса. Префикс выполняет в составе идиомы ту же функцию, что и за ее пределами — вводит результирующее состояние, пусть даже его дескриптивные свойства отличны от тех, которые представлены в неидиоматическом контексте.

Редукция к имперфективу, напротив, — артефакт некомпозиционального семантического сдвига, сопровождающего формирование новой событийной дескрипции, в данном случае, сдвига в акциональной характеристике предиката.

6 Вместо заключения

Мы готовы дать предварительные ответы на вопросы, сформулированные в (26).

1. Аспектуальная редуцированность наблюдается в парах «непроизводный имперфектив — префиксальный перфектив», но не в парах «перфектив — вторичный имперфектив» потому, что вторичный имперфектив расположен за пределами глагольной группы, ограничивающей идиоматический домен у VP-идиом.

2. Аспектуальная редуцированность возникает с использованием двух механизмов. Редукция к перфективу означает включение лексического префикса, задающего результирующее состояние, в состав идиомы. Редукция к имперфективу происходит благодаря акциональному сдвигу, в результате которого идиома описывает ситуации, не способные развиваться к кульминации.

3. Материал идиом подкрепляет некоторые высказанные в литературе соображения о морфосинтаксической структуре глагольной группы, но не дает эмпирических оснований для радикального пересмотра представлений об устройстве русской аспектуальной системы.

Литература

[1] Katz Jerrold, Postal Paul. Semantic interpretation of idioms and sentences containing them // Quarterly Progress Report of the MIT Research Laboratory of Electronics, 1963. — Vol. 70. — P. 275–282.

[2] McGinnis Martha. On the systematic aspect of idioms // Linguistic Inquiry, 2002. — Vol. 33. — P. 665–672.

[3] Glasbey Sheila R. Aspectual composition in idioms // Recent advances in the syntax and semantics of Tense, Aspect and Modality. — Berlin, 2007. — P. 71–88.

[4] Kuznecov E.V. (2016), Semantic structure of idioms with fixed aspectual charactecteristics [Semanticheskaya struktura idiom s fiksirovannoj vidovoj harakteristikoj]. MA Thesis [Diplomnaya rabota]. Lomonosov Moscow State University.

[5] Klimek-Jankowska Dorota, Błaszczak Joanna. The interaction of idioms and aspect in Polish // Natural Language and Linguistic Theory, 2024. — Vol. 42. — P. 153–207.

[6] Baranov A. N., Dobrovol'skij D. O. (eds) (2007), Thesaurus of modern Russian idioms [Slovar'-tezaurus sovremennoj russkoj idiomatiki]. Moscow.

[7] Wasow Thomas, Nunberg, Geoffrey, Sag Ivan. Idioms: An interim report // Proceedings of the XIIIth International Congress of Linguists. Tokyo, Nippon Toshi Center, 1984. — P. 102–115.

[8] Nunberg Geoffrey, Sag Ivan, Wasow Thomas. Idioms // Language, 994. — Vol. 70. — P. 491–538.

[9] O'Grady William. The Syntax of Idioms // Natural Language & Linguistic Theory, 1998. — Vol. 16. — P. 279–312.

[10] Mateu Jaume, Espinal Teresa. Classes of idioms and their interpretation // Journal of Pragmatics, 2010. — Vol. 42. — P. 1397–1411.

[11] Schildmier Stone Megan. The difference between bucket-kicking and kicking the bucket: Understanding idiom flexibility. Ph.D. dissertation, University of Arizona, 2016.

[12] Nediger Will Alexander. Unifying structure-building in human language: The minimalist syntax of idioms. Ph.D. dissertation, University of Michigan, 2017.

[13] Siloni Tal, Horvath Julia, Klunover Hadar, Wexler Ken. Idiom storage and the lexicon // Journal of Linguistics, 2018. — Vol. 54. — P. 189–215.

[14] Wierzba Marta, Brown J. M. M., Fanselow Gisbert. Sources of variability in the syntactic flexibility of idioms // Glossa: a journal of general linguistics, 2023. — Vol. 8.

[15] Kratzer Angelika. Severing the external argument from its verb // Phrase structure and the lexicon. Johan Rooryck & Laurie Zaring (eds.). Dordrecht (Kluwer Academic Publishers), 1996. — P. 109–137.

[16] Bruening Benjamin. Ditransitives Asymmetries and a theory of idiom formation // Linguistic Inquiry, 2010. — Vol. 41.

[17] Dobrovol'skij D. O. (2007a), Lexicosyntactic variation in idioms: introducing a modifier into the structure of an idiom [Leksiko-sintaksicheskoe var'irovanie vo frazeologii: vvod opredeleniya v strukturu idiomy] // Russkij yazyk v nauchnom osveshchenii. — № 2. — P. 18–47.
[18] Dronov P. S. (2019), Introducing an adjective modifier into the structure of an idiom: on semantic motivation for lexicosyntactic modification of idioms [Vvod ad'ektivnogo opredeleniya v strukturu idiomy: o semanticheskoj obuslovlennosti leksiko-sintaksicheskih modifikacij idiom], Candidate of Sc. Dissertation, Vinogradov Russian Language Institute.
[19] Abeillé Anne. 1995. The flexibility of French idioms: A representation with lexicalized tree adjoining grammar // Martin Everaert, Erik-Jan van der Linden, André Schenk & Rob Schreuder (eds.), Idioms: Structural and psychological perspectives. Hillsdale, Lawrence Erlbaum, 1995. — P. 15–42.
[20] Paducheva E. V. (1996), Studies in semantics [Semanticheskie issledovaniya]. Moscow.
[21] Tatevosov S. G. (2016), Verb classes and actional typology [Glagol'nye klassy i tipologiya akcional'nosti]. Moscow, Yazyki slavyanskoj kul'tury, 2016.
[22] Tatevosov S. G., Kiseleva K. L. (2019), Even identification in circumfixal word formation [Identifikacija sobytij v prefiksal'no-postfiksal'nom slovoobrazovanii] // Moscow University Bulletin. Series 9. Philology. — №1. — P. 7–35.
[23] Tatevosov S. G., Kiseleva K. L. (2018), Idioms as a window into a TAM system [Idiomy kak okno v vidovremennuju sistemu] // La relation temps/aspect: approches typologique et contrastive. — Collection UL3 travaux et recherches. — Universite Charles-de-Gaulle — Lille 3, Lille (France). — P. 425–429.
[24] Tatevosov Sergei. On the Aspectual Architecture of Russian. Ms., Lomonosov Moscow State University, 2020.
[25] Melerovich A. M. (1980), Semantic structure of idiomatic items in modern Russian as a linguistic problem [Semanticheskaja struktura frazeologicheskih edinic v sovremennom russkom jazyke kak lingvisticheskaya problema], Doctor of Sc. Dissertation. Kostroma, 1980.
[26] Baranov A. N., Jushmanova S. I. (2000), Negation in idioms: semantic and syntactic constraints [Otricanie v idiomah: semantiko-sintaksicheskie ogranicheniya] // Voprosy yazkoznaniya. — №1 — P. 46–65.
[27] Dobrovol'skij D. O. (2007b), Passivization of idioms: on semantic motivations for syntactic transformations in idioms [Passivizaciya idiom (o semanticheskoj obuslovlennosti sintaksicheskih transformacij vo frazeologii)] // Voprosy yazykoznaniya. — № 5. — P. 39–61.
[28] Romanova Eugenia. Constructing Perfectivity in Russian. Ph.D. dissertation. Tromsø, 2006.
[29] Svenonius Peter. 2004. Russian prefixes are phrasal // Proceedings of Formal Descriptions of Slavic Languages 5. Frankfurt am Main, 2008. — P. 526–537.
[30] Tatevosov S. G. (2013), Multiple prefixation and its consequences [Mnozhestvennaya prefiksaciya i ee sledstviya] // Voprosy yazykoznaniya. — №3. — P. 42–89.
[31] Biskup Petr. Prepositions, case and verbal prefixes: The case of Slavic. Amsterdam, John Benjamins, 2019.
[32] Tatevosov S. G. (2024), A theory of lexical prefixation, argument structure and asymmetric entailment [Teoriya leksicheskoj prefiksacii, aktantnaya struktura i asimmetrichnoe sledovanie] // Moscow University Bulletin. Series: Philology. — № 4. — S. 23–44.

Clause Linkers for the Ruscon Database: Selecting Criteria and Statistical Evaluation

Svetlana Timoshenko
IITP, Russian Academy of Sciences
Moscow, Russia
timoshenko@iitp.ru

Natalia Serdobolskaya
Institute of Linguistics, Russian Academy of Sciences
Moscow, Russia
serdobolskaya@gmail.com

Irina Kobozeva
Lomonosov Moscow University / Institute of Linguistics, Russian Academy of Sciences
Moscow, Russia
kobozeva@list.ru

Abstract

The study is devoted to compiling a list of clause linkers, or connectives, for the database Ruscon (available at http://ruscon.belyaev.io/). It is aimed at differentiating between complex connectives (like *a ne to* 'else') and occasional combinations of connectives (*a vsledstvie etogo* 'and henceforth'). There is a large number of cases when phonetic, morphosyntactic and semantic properties of connectives do not permit to make this differentiation. In these cases we propose to use the MMI metric (Modified Mutual Information). It rests on the rule of probability multiplication, which allows to estimate the probability of independent events.

Keywords: connective, clause linker, Russian language, subordination, complex sentence, complex connective, multiword expression extraction, collocation, collocation extraction measure, mutual information

DOI: 10.28995/2075-7182-2025-23-407-415

Статистические метрики как критерии отбора составных коннекторов (на материале базы Рускон)

Тимошенко С. П.
ИППИ РАН
Москва
timoshenko@iitp.ru

Сердобольская Н. В.
Институт языкознания РАН
Москва
serdobolskaya@gmail.com

Кобозева И. М.
МГУ им. М. В. Ломоносова, Институт языкознания РАН
Москва
kobozeva@list.ru

Аннотация

Исследование направлено на определение словника составных коннекторов для базы данных Рускон (доступна на сайте http://ruscon.belyaev.io/). Цель работы — разграничить составные коннекторы (напр. *а не то*) и окказиональные сочетания (*а вследствие этого*). В тех случаях, когда фонетические, морфосинтаксические и семантические критерии не позволяют решить вопрос о включении составного коннектора в словник, мы предлагаем использовать метрику MMI (Modified Mutual Information). Она опирается на правило умножения вероятностей, позволяющее оценить вероятность независимых событий.

Ключевые слова: коннектор, союз, русский язык, сложное предложение, составной коннектор, извлечение словосочетаний, мера устойчивости словосочетаний, взаимная информация

1 Введение

Одним из базовых вопросов изучения полипредикации является выделение инвентаря союзов, или, шире – коннекторов – единиц, способных соединять клаузы в сложном предложении. В особенности, эта задача актуальна для языков с богатой литературной традицией, которые активно используют комплексы союзов и развивают единицы, промежуточные между самостоятельной клаузой и коннектором, такие как *в расчёте на то что*, *в надежде на то чтобы*. Зачастую бывает трудно разграничить самостоятельные коннекторы, которые должны быть включены в лексикографические источники, и окказиональные сочетания. Настоящая работа ставит цель выявить такие критерии для русских коннекторов на основе статистических метрик, применяемых для извлечения коллокаций.

Материалом исследования служат коннекторы базы данных Рускон (база данных, включающая синтаксическую и семантическую информацию по современным коннекторам русского языка, доступна на сайте http://ruscon.belyaev.io/). Инвентарь коннекторов создан на основе ряда источников, включая словари русского языка МАС (МАС, 1981 1984), БТС (БТС, 1998), (Морковкин (ред.), 2003), (Ефремова, 2004) и Академическую грамматику (Шведова and others, 1980). В отличие от каталога (Богданов and Рыжова, 1997), Рускон основывается на отредактированном словнике и, кроме непосредственно терминологической информации (например, частеречный статус коннектора), содержит данные о значении, употреблении и ряде синтаксических и семантических свойств каждого коннектора.

Часть этого исследования выполнена благодаря поддержке гранта РНФ номер 24-18-00988. Авторы выражают фонду глубокую признательность.

2 Внутриязыковые критерии выделения составных коннекторов

Традиционно союзы делятся на простые и составные, или неоднословные – «нецельноформленные соединения двух или более элементов, каждый из которых одновременно существует в языке и как отдельное слово» (Шведова and others, 1980, с. 716), например, *благодаря тому что*, *даром что*, *не то что*, *оттого что*. При выделении класса составных союзов возникает проблема разграничения единых союзов (*потому что*) и окказиональных сочетаний (*а из-за этого*) – условно говоря, единиц, которые содержатся в памяти как единое целое или порождаются в ходе речепроизводства. В частности, сочетания *так как*, *так что*, *отчего и*, *не то* вводятся в БТС в зоне производных слов под ярлыком «в значении союза»; в МАС *отчего и* отсутствует, а единицы *так как*, *так что* и *не то* признаются сложным союзом; в словаре (Морковкин (ред.), 2003) *так как* и *отчего и* являются отдельными входами, в отличие от *так что* и *не то*; в (Ефремова, 2004) все эти единицы, кроме *отчего и*, вводятся в качестве отдельных входов. Сводный каталог коннекторов и информация о их фиксации в словарях содержится в (Богданов and Рыжова, 1997). Как можно видеть, источники серьезно расходятся в мнениях по поводу трактовки конкретных сочетаний – как в плане их лексико-грамматического статуса, так и, шире, в плане включения их в базовый инвентарь. В Академической грамматике вводится термин «союзные соединения», включающие сочетания с конкретизаторами (слова, которые уточняют значение союза), ср. *потому – и потому – а потому*; *зато – но зато*; *иначе – или иначе* (Шведова and others, 1980, с. 714). Кроме того, вводится термин «аналог союза» для единиц «с квалифицирующими лексическими значениями, которые активно вовлекаются в сферу союзных средств»: *вдобавок*, *кстати*, *лишь*, *потом* и др. Тем самым, в обиход вводится значительное количество коннекторов, которые ранее не рассматривались в словарях, включая также конструкции с вариативным лексическим наполнением – *конечно… но*, *если… то*, *для того чтобы*; *может* (повтор)…. *а может*, *возможно… а может быть* («союзные соединения»).

Настоящая работа ставит целью составить словник коннекторов для базы данных Рускон. База состоит из ряда модулей. В одном из них представлена справочная информация о лексико-грамматических и семантических свойствах коннекторов согласно авторитетным источникам, указанным в разделе 1. Тем самым в отношении коннекторов мы следуем идее введения многокомпонентных единиц в инвентарь союзов, предложенной в монографии (Копотев and Стексова, 2016, с. 41), содержащей ценную информацию по рассматриваемой проблеме. База содержит разметку коннекторов русского языка с точки зрения их синтаксических и семантических свойств, с примерами употребления и ссылкой на их источники. Поскольку речь идет о коннекторах (единицах, способных связывать части сложного предложения), а не о союзах, авторы исходят из функционала данных единиц, а не их частеречного статуса. Такой подход позволяет рассматривать интегрально все релевантные единицы, включая союзы, союзные соединения и аналоги союзов. Тем самым, мы в целом следуем подходу (Шведова and others, 1980). Однако мы предлагаем корректировку списка единиц, представленных в данном источнике, поскольку, как показано выше, многие единицы потенциально могут быть расширены за счет «конкретизаторов», а выделенные конструкции допускают значительную вариативность частей (*если… то для того чтобы – если… то потому что –*

если… то оттого что). Такая вариативность учитывается в подходе О. Ю. Иньковой, где базовой единицей является речевая реализация коннектора, т.е. «та форма, в которой коннектор встречается в данном конкретном высказывании» (Инькова and Кружков, 2018, с. 171). Таким образом, любое сочетание включается в базовый инвентарь и описывается как обладающее собственными уникальными свойствами. Проблема, однако, состоит в том, что число речевых реализаций потенциально бесконечно (Инькова, 2019, с. 24), и в результате, невозможно дать подробное описание синтаксиса и семантики каждой единицы. Между тем, база данных Рускон включает более 30 параметров описания для каждого коннектора, что предполагает ограничение словника. В свете этого, необходимы объективные эксплицитные критерии выделения коннекторов, позволяющие принимать последовательные решения.

Критерии выделения составных коннекторов в типологических работах базируются на эвристике, что коннектор должен вести себя как единая словоформа (а не сочетание слов) с точки зрения различных уровней языка – фонетического, морфологического, синтаксического и семантического (Kortmann, 1997, p. 71-77). Данные критерии входят в более общий круг свойств грамматикализации – явления, при котором лексические единицы переходят в разряд грамматических, или шире, функциональных элементов (таких как союзы) (Hopper and Traugott, 2003; Kouteva et al., 2019). Приведем примеры применения критериев.

Фонетический критерий. Союзы *потому что* и *как будто* фонетически ведут себя как единая словоформа: в повседневной речи *потому что* произносится как [птуштъ, туштъ], *как будто* произносится как [кабут:ъ], союз *так как* — [так:ък], и все три союза имеют одно ударение и произносятся без паузы между частями, т.е. образуют одно фонетическое слово. Безусловно, далеко не все союзы характеризуются единым ударением и отсутствием паузы, поэтому данный критерий работает не как необходимый, но как достаточный для отнесения спорной единицы к союзам. В других случаях просодическое оформление может быть использовано в качестве эвристики за или против того или иного решения. [1]

Семантический критерий. Семантически у *так как* значение причины трудно вывести из семантики составляющих единиц, *так* и *как*; у союза *и то* значение уступки (*всю ночь работал, и то не успел*) не выводится из значений *и* и *то*. Таким образом, для данных единиц работает критерий семантической «неразложимости», или семантической некомпозициональности, который предполагает, что значение целого не может быть выведено из суммарного значения частей.

Данный критерий, однако, затруднительно применить к сочетаниям многофункциональных союзов и частиц *а*, *да*, *и*, *но*, *же*, *даже* и (реже) *-то* с имеющими ограниченный набор значений «конкретизаторами» (согласно (Шведова and others, 1980)): *а* / *и* / *но в результате*, *если и* / *же*, и т.п.). Зачастую значение всего комплекса является композициональным, однако происходит десемантизация союза/частицы, например, *и* в составе *и однако* может иметь только значение логической конъюнкции (но не сопоставления или временного следования). При этом *однако* выступает в своем обычном противительном значении «ненормального следствия». Такое ограничение на значение сочинительного союза может быть проинтерпретировано как десемантизация в рамках составного коннектора или же как ограничение на лексическую сочетаемость, стандартным образом возникающее при контактном употреблении лексем. От выбора интерпретации зависит решение о признании *и однако* единым коннектором.

Морфологический критерий предполагает, что коннектор (практически) не имеет способности к словоизменению и выступает в одной фиксированной форме (fossilization в теории грамматикализации); соответственно, его части, даже если они граматикализовались из знаменательной части речи, утрачивают способность к словоизменению. Например, существительное фиксируется в определенной падежной форме и не допускает изменения по числу в составе коннекторов *в тех целях чтобы – *в той цели чтобы*, но *с той целью чтобы – *с теми целями чтобы*; наречия часто закрепляются в форме сравнительной степени (при глаголах речи): *точнее говоря – *точно говоря*, *вернее сказать – *верно сказать*, *проще сказать – *просто сказать* и т.д. Заметим, одна-

[1] Интересную проблему представляют собой коннекторы, допускающие разрыв, напр. *потому... что, затем... чтобы* и др. Вопрос об их самостоятельности требует отдельного рассмотрения. При этом, неразрывный вариант таких единиц представляет собой одно фонетическое слово и, соответственно, включается в базу.

ко, что коннектор может сохранять способность различать некоторые формы, например *коротко говоря – короче говоря* (но не **коротко сказав*), *по правде говоря – по правде сказать* (но не **по правде сказав*) и т.п.

Синтаксический критерий. Единицы, входящие в состав союзов, не могут быть модифицированы, ср. *потому что – *по тому самому что*; союзы не разрываются[2]: *так как он пришел – *так он как пришел*, *так же как он пришел* (такое сочетание возможно, однако не имеет причинной семантики). Части коннектора не могут быть опущены, ср. *а то*; точнее, опущение частей меняет семантику отношения между клаузами, и следовательно единица *а то* трактуется как единый коннектор. Семантика повторяющихся союзов *или... или, либо... либо* не вполне идентична значению одиночного *или* и *либо*, т.к. обычно вводит непересекающиеся альтернативы; соответственно, такие сочетания признаются самостоятельными коннекторами. Особую проблема опущения представляют собой т.н. рамочные коннекторы, т.е. коннекторы, смысловая часть которых находится в зависимой предикации, а главная предикация содержит сочинительный союз, частицу или местоимение, отсылающее к ситуации в зависимой клаузе: *не только... но и; ещё; а также; не то чтобы... а; но; просто* (обычно список таких средств является открытым). Такие коннекторы ограниченно допускают опущение второй части, чаще всего при маркированном просодическом оформлении. В базе они включены в виде входов с вариативным квази-коррелятом. Трактовка рамочных коннекторов представляет собой особую проблему, которая находится за рамками настоящего исследования; ниже мы рассматриваем только случаи, когда многокомпонентной является основная часть коннектора (в составе зависимой клаузы).

Таким образом, части коннектора при грамматикализации обычно подвергаются изменениям, которые связаны с интеграцией частей коннектора в единое целое, неразложимое с точки зрения тех или иных уровней языка. Если же такие процессы не происходят, это обычно означает, что рассматриваемая единица является свободным сочетанием отдельных лексем.

В результате обработки данных источников база Рускон включала 700 составных коннекторов. Из них относительно 591 коннекторов решение о включении (или удалении из базы) было принято на основании внутриязыковых критериев; для остальных 109 данные критерии не позволяют принять объективное решение.

3 Статистический анализ коннекторов

3.1 Сравнение метрик, опирающихся на идею взаимной информации

Для решения вопроса о степени слитности 109 оставшихся кандидатов использовались статистические инструменты — меры оценки устойчивости словосочетаний. Эти меры позволяют ранжировать словосочетания и сравнивать их между собой. Они очень разнообразны: в диссертации (Evert, 2005) представлено около 30 метрик (этот обзор широко известен благодаря удачному воплощению в справочном сайте `collocations.de`), в статье (Su et al., 2024) сравнивается 16. Наиболее популярными метриками являются: хи-квадрат (Chi-square), t-тест (T-score), коэффициент Дайса (Dice's coefficient), отношение правдоподобия (Log-likelihood ratio), а также варианты оценки взаимной информации — точечная взаимная информация (Pointwise Mutual Information или PMI), кубическая взаимная информация (Cubic Mutual Information или MI^3) и другие. Все метрики рассчитываются на основе данных о частотности элементов в том или ином корпусе. В работе (Deng and Liu, 2022) сравнивается 7 различных метрик на материале разных типов устойчивых словосочетаний и разных корпусов. В среднем лучшие результаты показывают отношение правдоподобия и кубическая взаимная информация. Сравнению метрик на материале русского языка посвящена статья М. Хохловой (Khokhlova, 2018). Лингвистическим материалом в этой рабое выстпают двухсловные сочетания "прилагательное + существительное". Найденные в корпусе сочетания ранжируются с помощью 13 разных метрик, а потом верхние 100 позиций в каждом списке оцениваются экспертами и проверяются по словарям. В этом эксперименте лучшими метриками оказываются MI^3, t-тест, коэффициент Фишера и коэффициент Пуассона. Можно сказать, что MI^3 показывает хорошие результаты на материале разных языков.

[2]В настоящем исследовании мы не рассматриваем разрывные коннекторы, например *да и* (*да он и не хотел этого*), *если и, пока не* и проч.

Лингвистически содержательные интерпретации метрик расходятся: так, согласно результатам, изложенным в (Krenn, 2000), PMI и коэффициент Дайса лучше подходят для поиска высокочастотных коллокаций, а отношение правдоподобия — для низкочастотных. А в более современном учебнике (Копотев, 2014) утверждается противоположное: «MI лучше ищет довольно редкие коллокации узкой тематической области (например, термины, составные названия компаний или сочетания имени и фамилии), а t-score лучше справляется с высокочастотными общеязыковыми "эквивалентами слова" (сложные предлоги, вводные конструкции и т. п.)».

Особенность задачи выявления единых коннекторов состоит в том, что необходимо сравнивать между собой выражения разной длины, от двухсловных (*но и*, *а не*, *как будто*) до пятисловных (*в сравнении с тем как*, в *расчёте на то что*). Формула расчета отношения правдоподобия применима только к двухсловным словосочетаниям (Ramisch, 2014, p. 215). Выбор среди остальных представляет трудность, т. к. известные нам исследования, выполненные на материале русского языка, посвящены биграммам, как и упоминавшаяся выше работа (Khokhlova, 2018), их результаты не могут служить главным основанием при выборе метрики для нашей задачи. Мы начали экспериментировать с MI^3 как с одной из самых популярных и хорошо себя зарекомендовавших в разных задачах. Мы адаптировали ее для словосочетаний произвольной длины по той же логике, что и реализация PMI в программе mwetoolkit - см. формулу в работе (Ramisch et al., 2010). Результаты экспериментов с PMI и MI^3 показывают, что обе они не позволяют корректно сравнивать между собой словосочетания разной длины. Средние значения метрик для двухсловных словосочетаний достоверно ниже, чем для трехсловных, а для трехсловных - ниже, чем для четырехсловных, и т. д. Иными словами, любое более длинное словосочетание оказывается более устойчивым, чем более короткое. Поэтому мы предлагаем новую формулу, модификацию взаимной информации(далее - MMI, Modified Mutual Information) 1. Значения этой метрики для лингвистически однотипных двухсловных и трехсловных устойчивых выражениях оказываются близки (Тимошенко, 2024). В таблицах 1 и 2 приведены средние и медианные значения метрик для словосочетаний разной длины, полученные на материале 242 коннекторов. Программная реализация экспериментов и данные

	2 слова	3 слова	4 слова	5 слов
PMI	7.79	22.11	40.28	61.55
MI^3	22.05	29.9	38.68	53.68
MMI	18.5	19.47	19.95	19.06

Таблица 1: Средние значения метрик для словосочетаний разной длины

	2 слова	3 слова	4 слова	5 слов
PMI	7.61	21.91	40.14	61.12
MI^3	22.92	30.88	39.24	53.19
MMI	18.3	18.87	20.05	19.41

Таблица 2: Медианные значения метрик для словосочетаний разной длины

доступны по адресу: `https://github.com/therolinguist/mmi_and_other_measures`

3.2 Метрика MMI

$$MMI(collocation) = ln(frec_{collocation} \div \frac{w_1 \times ... \times w_i}{freq^i_{pattern}}) \quad (1)$$

где

$freq_{collocation}$ - наблюдаемая абсолютная частота словосочетания в корпусе;

$freq_{pattern}$ - наблюдаемая абсолютная частота морфологического паттерна, который ему соответствует;

$w_1 \ldots w_i$ - наблюдаемые абсолютные частоты словоформ, входящих в состав словосочетания, на "своей" позиции в паттерне.
Пример расчета метрики см. в Приложении А. В качестве источника количественных данных использовался основной корпус НКРЯ (НКРЯ, 2024).

Морфологический паттерн — это последовательность частеречных характеристик входящих в него слов. Для выражения *в сравнении с тем как* он выглядит так: "1-е слово — предлог, 2-е — существительное, 3-е — предлог, 4-е — местоименное наречие, 5-е слово — союз" ((PR)(S)(PR)(SPRO)(CONJ) на языке запросов Национального корпуса русского языка [(НКРЯ, 2024)). Однако если мы посмотрим на морфологическую разметку вхождений этого выражения в основном корпусе НКРЯ, то обнаружим, что из 12 вхождений только у 4 характеристики будут соответствовать этому паттерну. В 8 оставшихся случаях *как* будет считаться не союзом, а местоименным наречием. (Здесь и далее числа отражают состояние Основного корпуса на июль 2024 года. С тех пор корпус пополнился новыми текстами, поэтому многие числа изменились). В большей части основного корпуса разрешение морфологической омонимии выполнено автоматически с помощью программы rubic (Lyashevskaya et al., 2023). Тем не менее, мы не считаем подобные "биения" системы ошибками: как говорилось выше, коннекторы находятся в процессе грамматикализации, а этот процесс происходит постепенно. Непоследовательности в ответах системы, выбирающей предпочтительный морфологический разбор, могут отражать переходное состояние отдельных элементов выражения. Эта гипотеза представляется захватывающе интересной: дальнейшие исследования сулят статистические открытия в области грамматикализации. Чтобы компенсировать разночтения разметки, мы использовали в паттернах дизъюнкции: итоговый паттерн *в сравнении с тем как* допускает в качестве 5-го слова не только союз и местоименное наречие, но и частицу, вводное слово и наречие — (PR)(S)(PR)(SPRO)(ADV|PARENTH|ADVPRO|CONJ|PART). В общей сложности мы использовали 86 паттернов.

Чтобы на основе статистического показателя понять, какие спорные случаи следует отнести к единым коннекторам, необходимо выбрать пороговое значение. Для этого мы рассчитали значение метрики для 242 выражений, ранее отнесенных к классу единых коннекторов на основе типологических критериев (см. выше). В качестве порогового значения мы использовали нижнюю границу стандартного отклонения. Для двухсловных выражений она составляет 16.39, для трехсловных — 16.66, для четырехсловных — 17.9 и для пятисловных — 17.03. Было рассмотрено 109 спорных случаев. Для них также были рассчитаны значения метрики MMI. У 84 единиц они оказались больше порогового значения.

В нашей выборке нередки случаи, когда единицы большей длины включают в себя единицы меньшей. Например, трехсловные единицы *а тем более*, *и тем более*, *тем более что* включают в себя слова, совпадающие с самостоятельным коннектором *тем более*. Результаты поискового запроса к корпусу "словоформа **тем**, на расстоянии 1 — словоформа **более**" включают вхождения всех трех более длинных единиц. Поэтому корректный показатель абсолютной частотности вхождений *тем более* получается путем вычитания количества вхождений более длинных единиц из количества вхождений *тем более*. Такая поправка была сделана во всех случаях, когда более длинные единицы включают в себя элементы, соответствующие более коротким.

Предлагаемый подход к оценке устойчивости словосочетаний имеет свои ограничения. Строго говоря, он оценивает взаимную обусловленность словоформ в цепочке; семантика же словосочетаний никак не контролируется. Поэтому в случае омонимичных конструкций результат оказывается смазан. Рассмотрим коннектор *а также*:

(2) *Кроме того, что кандидаты являются тезками, у них есть еще две общие черты: оба они пришли в университет в 1966 году,* ***а также*** *оба являются докторами физико-математических наук.* [НКРЯ: Анастасия Гулина. Ректорство на день рождения // «Богатей» (Саратов), 2003.10.23]

Чтобы оценить частотность именно коннектора, нужно отобрать только те примеры, где справа от *а также* как минимум находится простое предложение, а в идеале — присоединяется. Иными

словами, нужна информация о синтаксической связи между вершиной клаузы и *а также*. Несмотря на то, что в основном корпусе в 2024 году появилась возможность накладывать синтаксические ограничения на запрашиваемые словосочетания, в данном случае она не может помочь нам отобрать вхождения потенциальных коннекторов. Причин две. Во-первых, сочинительные связи устанавливаются автоматически хуже других за счет своей вариативности. Во-вторых, даже если бы точность их определения была близка к 100%, принятый в основном корпусе формализм не позволяет различить сочинение предложений, как в примере (2) и сочинение сказуемых, как в примере 3:

(3) *Тем не менее, это очень страшная штука, которую хотелось бы, чтоб мы сами,* ***а также*** *наши политики стремились не допускать в будущем во благо следующего поколения.*[НКРЯ: Форум: 17 мгновений весны (2005-2010)]

В обоих случаях союз *а* будет зависеть от глагола (*назначил* и *зарабатывали* соответственно), а сами глаголы будут размечены как подчиненная клауза. Различить такие случаи может только человек. Принимая во внимание огромное количество примеров, можно выполнить экспертную оценку для ограниченной случайной выборки примеров, например, для 1000, а потом экстраполировать результат на абсолютные цифры. Т. е. если в корпусе нашлось 59995 вхождений *а также*, и в случайно отобранном их подмножестве окажется лишь 5% примеров, в которых действительно представлен коннектор, то можно рассчитывать MMI, принимая абсолютную частотность коннектора равной 3000. В данном исследовании мы таких поправок не делали.

4 Заключение

В ходе разработки базы данных русских коннекторов Рускон возникает необходимость отбора словника составных коннекторов как единиц лексикографического описания. На основании внутриязыковых критериев (фонетические, морфосинтаксические свойства коннектора и семантическая композициональность) не всегда удается разграничить устойчивые единицы (напр., *а и то*), которые являются единицами языка, и окказиональные сочетания, которые порождаются говорящим в процессе реализации высказывания (напр., *а вследствие этого*).

Мы предприняли попытку статистически оценить устойчивость составных коннекторов. Для этого мы использовали метрику MMI (модифицированную взаимную информацию). Ее главное достоинство состоит в том, что она позволяет сравнивать между собой словосочетания разной длины. Значения метрики были посчитаны для 351 выражения. В качестве источника количественных данных использовался основной корпус НКРЯ. 242 выражения из 351 были классифицированы как коннекторы на основе внутриязыковых критериев. Среди них были двухсловные, трехсловные, четырехсловные и пятисловные единицы. Для каждой длины мы рассчитали стандартное отклонение значений и приняли его нижнюю границу в качестве порога для классификации спорных случаев. На основе этого порогового значения мы классифицировали 109 кандидатов в коннекторы. В итоге 84 из них были включены в базу.

Главный недостаток предложенного способа классификации состоит в том, что оценивается устойчивость последовательности словоформ, а значение этой последовательности никак не проверяется. Между тем, в языке иногда сосуществуют коннекторы и омонимичные им свободные словосочетания (ср. *а также*). Корректный расчет значений статистических метрик в этих случаях требует предварительной ручной классификации примеров.

References

Yaochen Deng and Dilin Liu. 2022. A multi-dimensional comparison of the effectiveness and efficiency of association measures in collocation extraction. *International Journal of Corpus Linguistics*, 27(2):191–219.

Stefan Evert. 2005. *The statistics of word cooccurrences: word pairs and collocations*. Ph.D. thesis, Stuttgart University.

P. J. Hopper and E. C. Traugott. 2003. *Grammaticalization*. Cambridge University Press.

Maria Khokhlova. 2018. Similarity between the association measures: a case study of noun phrases. // *RASLAN*, P 21–27.

B. Kortmann. 1997. *Adverbial Subordination: A Typology and History ofAadverbial Subordinators Based on European Lnguages*. Mouton de Gruyter, Berlin; New York.

T. Kouteva, B. Heine, B. Hong, H. Long, H. Narrog, and S. Rhee. 2019. *World lexicon of grammaticalization*. Cambridge University Press.

Brigitte Krenn. 2000. Empirical implications on lexical association measures. // *Proceedings of The Ninth EURALEX International Congress*, P 359–371.

O.N. Lyashevskaya, I.A. Afanasev, S.A. Rebrikov, Y.A. Shishkina, E.A. Suleymanova, I.V. Trofimov, and N.A. Vlasova. 2023. Disambiguation in context in the russian national corpus: 20 yeas later. // *Proceedings of the International Conference "Dialogue*, volume 2023.

Carlos Ramisch, Aline Villavicencio, and Christian Boitet. 2010. mwetoolkit: A framework for multiword expression identification. // *LREC*, volume 10, P 662–669. Valletta.

C. Ramisch. 2014. *Multiword Expressions Acquisition: A Generic and Open Framework*. Theory and Applications of Natural Language Processing. Springer International Publishing.

Qi Su, Chen Gu, and Pengyuan Liu. 2024. Association measures for collocation extraction: Automatic evaluation on a large-scale corpus. *International Journal of Corpus Linguistics*, 29(1):59–86.

БТС. 1998. *Большой толковый словарь русского языка под ред. С. А. Кузнецова*. Норинт, СПб.

С. И. Богданов and Ю. В. Рыжова. 1997. *Русская служебная лексика. Сводные таблицы*. Издательство Санкт-Петербургского Государственного Университета.

Т.Ф. Ефремова. 2004. *Толковый словарь служебных частей речи русского языка*. Астрель, АСТ, Москва.

О.Ю. Инькова and М.Г. Кружков. 2018. Метод описания структуры неоднословных коннекторов в надкорпусных базах данных. *Системы и средства информатики*, 28(4):168–181.

О.Ю. Инькова. 2019. Структура коннекторов: лингвистические методы описания. // О.Ю. Инькова, *Структура коннекторов и методы ее описания*, P 5–47. ТОРУС ПРЕСС, Москва.

М.В. Копотев and Т.И. Стексова. 2016. *Исключение как правило: переходные единицы в грамматике и словаре*. Языки славянских культур: Рукописные памятники Древней Руси, Москва.

М. В. Копотев. 2014. *Введение в корпусную лингвистику: Учебное пособие для студентов филологических и лингвистических специальностей университетов. Корпусная лингвистика. Филология*. Animedia Company.

МАС. 1981-1984. *Словарь русского языка: в 4-х т*. Русский язык, Москва.

В.В. Морковкин (ред.). 2003. *Объяснительный словарь русского языка : структурные слова: предлоги, союзы, частицы, междометия, вводные слова, местоимения, числительные, связочные глаголы*. Астрель, АСТ.

НКРЯ. 2024. Национальный корпус русского языка. https://ruscorpora.ru.

С. П. Тимошенко. 2024. Оценка словосочетаний разной длины с помощью точечной взаимной информации. *Труды конференции ИТиС 2024*.

Н.Ю. Шведова et al. 1980. *Русская грамматика. В двух томах.* Наука, Москва.

Translation of references in Russian

References

Bogdanov S.I., Ryzhova YU.V. (1997), Russian functional words. Summary tables [Russkaya sluzhebnaya leksika. Svodnye tablicy], Saint-Petersburg: SPbGU.

BTS: Kuznetsov S.A. (ed.) (1998), A big explanatory dictionary of Russian [Bol'shoi tolkovyi slovar' russkogo yazyka], Saint-Petersburg: Norint.

Efremova T.F. (2004), An explanatory dictionary of functional words of the Russian language [Tolkovyi slovar' sluzhebnykh chastei rechi russkogo yazyka], Moscow: Astrel', AST.
In'kova O.Yu., Kruzhkov M.G. (2018), A method of description of the structure of multiword connectives in supracorpora databases [Metod opisaniya struktury neodnoslovnykh konnektorov v nadkorpusnykh bazakh dannykh], Systems of means of informatics [Sistemy i sredstva informatiki]. Vol. 28 № 4, pp. 168-181.
In'kova O.Yu. (2019), A structure of connectives and methods of its description [Struktura konnektorov i metody ee opisaniya], Moscow: Torus Press.
Kopotev M.V., Steksova T.I. (2016), Exception is a rule. Transmitional items in grammar and dictionary [Iskljuchenie kak pravilo. Perehodnye edinicy v grammatike i slovare]. Moscow: Jazyki slav'anskoj kul'tury.
MAS: Evgen'eva A.P. (ed.) (1981–1984), A dictionary of Russian [Slovar' russkogo yazyka]: four volumes, AN SSSR, In-t rus. yaz.; 2-e izd., ispr. i dop., Moscow: Russkii yazyk.
Morkovkin V.V. (ed.) (2003), An explanatory dictionary of the Russian language: functional words: prepositions, conjunctions, particles, interjections, parentheticals, pronouns, cardinals, copulas: approx. 1200 words [Ob"yasnitel'nyi slovar' russkogo yazyka : strukturnye slova: predlogi, soyuzy, chastitsy, mezhdometiya, vvodnye slova, mestoimeniya, chislitel'nye, svyazochnye glagoly: okolo 1200 edinits], Gos. in-t rus. yaz. im. A. S. Pushkina, izd. 2-e, ispr., Moscow: Astrel', AST.
NKRYA. 2024. Russian National Corpus. [Nacional'nyi korpus russkogo yazyka]
Timoshenko S. P. (2024), Pointwise mutual information as an evaluation metric for collocations of different length [Ocenka slovosochetanij raznoj dliny s pomoshh'yu tochechnoj vzaimnoj informacii], Proceedings of the ITAS Conference 2004 [Trudy konferencii ITiS 2024.]
Shvedova N.Yu. (ed.) (1980; 2005), Russian grammar [Russkaya grammatika], Moscow: Nauka.

Приложение А. Пример расчета метрики MMI

Коннектор: *тем более* (классифицирован как единый на основе внутриязыковых критериев)
Морфологический паттерн:

(SPRO) (ADV | PARENTH | ADVPRO | CONJ | PART)

Абсолютная наблюдаемая частотность коннектора составляет 35323 (результат запроса "словоформа **тем**, на расстоянии 1 от нее словоформа **более**").

После вычитания количества вхождений более длинных единиц *а тем более* (2360 вхождения), *и тем более* (2572 вхождения), *тем более что* (15649 вхождения) абсолютная наблюдаемая частотность составляет 14742.

Абсолютная наблюдаемая частотность словоформы *тем* на первой позиции паттерна составляет 143588 (результат запроса "словоформа **тем** с грамматической характеристикой "местоименное существительное" , на расстоянии 1 от нее слово, которое является либо наречием, либо вводным словом, либо местоименным наречием, либо союзом, либо частицей").

Абсолютная наблюдаемая частотность словоформы *более* на второй позиции паттерна составляет 53077 (результат запроса "первое слово — местоименное существительное, на расстоянии 1 от него словоформа **более**, с одной грамматической характеристикой из набора: наречие, вводное слово, местоименное наречие, союз, частица").

Абсолютная наблюдаемая частотность морфологического паттерна "первое слово — местоимен-ое существительное, на расстоянии 1 от него слово, которое является либо наречием, либо вводным словом, либо местоименным наречием, либо союзом, либо частицей" составляет 7099859.

$$MMI(\textit{тем более}) = ln(frec_{collocation} \div \frac{w_1 \times w_2}{freq^2_{pattern}}) = ln(14742 \div \frac{143588 \times 53077}{7099859^2}) = 18.39 \quad (4)$$

Structured Sentiment Analysis with Large Language Models: A Winning Solution for RuOpinionNE-2024

Aleksei Vatolin
FRC CSC RAS
Moscow, Russia
vatolinalex@gmail.com

Abstract

This paper presents the winning solution for the RuOpinionNE-2024 competition on structured sentiment analysis in Russian news texts. We propose a novel pipeline with large language models (LLMs) and adapter-based fine-tuning, demonstrating how modern LLMs can be effectively adapted to complex opinion tuple extraction tasks. Our method addresses three key challenges: (1) alignment of model predictions with original text spans through a fuzzy substring matching algorithm, (2) robustness to generation variability via multi-prediction aggregation strategies, and (3) efficient domain adaptation using QLoRA fine-tuning. The proposed approach achieved first place with a test F1 score of 0.405. Experimental results reveal that adapter-based fine-tuning of open-source 70B parameter models (Llama-3.3) surpasses prompt-engineered proprietary models like GPT-4o. Our analysis provides practical insights into adapting LLMs for structured information extraction in morphologically rich languages, showing that targeted fine-tuning with 4-bit quantization enables state-of-the-art performance without task-specific architectures.

Keywords: Structured sentiment analysis, Large language models, QLoRA fine-tuning, Prompt engineering
DOI: 10.28995/2075-7182-2025-23-416-434

Структурный анализ настроений с большими языковыми моделями: решение, выигравшее в соревновании RuOpinionNE-2024

Аннотация

Эта статья представляет победное решение для конкурса RuOpinionNE-2024 по структурированному сентимент-анализу текстов российских новостей. Мы предлагаем новый пайплайн с использованием больших языковых моделей (LLM) и адаптерного дообучения, демонстрируя, как современные LLM могут быть эффективно адаптированы для сложных задач извлечения мнений. Наш метод решает три ключевые задачи: (1) согласование предсказаний модели с исходными текстовыми фрагментами с помощью алгоритма нечеткого сопоставления подстрок, (2) устойчивость к вариативности генерации через стратегии агрегации нескольких предсказаний и (3) эффективная адаптация к домену с использованием дообучения QLoRA. Предложенный подход занял первое место с тестовым F1-результатом 0,405. Экспериментальные результаты показывают, что адаптерное дообучение моделей с открытым исходным кодом на 70 миллиардов параметров (Llama-3.3) превосходит результаты GPT-4o с подобранным промптом. Наш анализ предоставля-ет практические инсайты по адаптации LLM для извлечения структурированной информации в морфологически богатых языках, показывая, что целевое дообучение с 4-битной квантизацией позволяет достичь передовых результатов без использования архитектур, специфичных для задачи.

Ключевые слова: Структурный анализ настроений, Большие языковые модели, QLoRA дообучение, Промпт-инжениринг

1 Introduction

Sentiment analysis is a field within Natural Language Processing (NLP) that focuses on determining the emotional tone expressed in text. In recent years, this field has undergone significant advancements, evolving from simple tasks of overall sentiment classification (positive, negative, neutral) to more complex and fine-grained approaches.

One important direction is *Aspect-Based Sentiment Analysis* (ABSA). ABSA aims to identify opinions not at the text level, but concerning specific aspects (features, components) of the entity under consideration. Within ABSA, several subtasks are defined:

- **Aspect Term Extraction** (ATE): Identifying specific words or phrases in the text that indicate aspects.
- **Aspect Sentiment Classification** (ASC): Determining the polarity (positive, negative, neutral) of the expressed opinion towards each extracted aspect.
- **Opinion Term Extraction** (OTE): Identifying words or phrases that express the opinion about the aspect.
- **Aspect-Sentiment-Opinion Triplet Extraction:** A complex task that combines the extraction of aspects, opinion terms, and the determination of their mutual sentiment polarity.
- **Aspect-Category-Opinion-Sentiment Quadruple Extraction**: Extraction of the aspect, its corresponding category, the opinion expressed, and its sentiment polarity.

The development of ABSA has led to the emergence of an even more challenging task – *structured sentiment analysis*. This approach aims to extract complete *opinion tuples*, including not only the aspect and its sentiment but also the *opinion holder* and the *opinion expression*.

This paper presents a solution to the task of structured sentiment analysis of the Russian language within the framework of the RuOpinionNE-2024 competition (Loukachevitch et al., 2025).

2 Literature review

The field of sentiment analysis has seen significant advancements, evolving from general sentiment classification to more fine-grained approaches like Aspect-Based Sentiment Analysis (ABSA) and its subtasks. Early work focused on identifying overall sentiment polarity (Turney, 2002), while subsequent research delved into aspect-level sentiment (Tang et al., 2015). More recently, the focus has shifted towards extracting complete opinion tuples, encompassing aspect terms, opinion expressions, and sentiment polarity. This progression led to the introduction of tasks like Aspect-Based Sentiment Triplet Extraction (ASTE) (Peng et al., 2019) and the even more comprehensive Structured Sentiment Analysis, which aims to capture all sentiment graphs within a text (Barnes et al., 2022).

Current state-of-the-art approaches for these complex tasks often involve sophisticated neural network architectures, including sequence labeling models, graph-based methods, and encoder-decoder frameworks (Fei et al., 2021). Researchers have explored techniques like position-aware

tagging (Xu et al., 2020), multi-task learning, and the incorporation of external knowledge to improve performance (Zhao et al., 2020). The introduction of large language models (LLMs) has further pushed the boundaries of sentiment analysis, with studies investigating their zero-shot capabilities and fine-tuning strategies for tasks like ASTE (Scaria et al., 2024) and structured sentiment analysis (Zhang et al., 2021). Instruction learning paradigms, such as those used in InstructABSA, have shown promise in enhancing the reasoning abilities of LLMs and improving performance across various ABSA subtasks.

3 Methodology

3.1 Data Format

The task of structured sentiment analysis involves extracting opinion tuples from text, specifically in the format (Holder, Target, Polar_expression, Polarity). This task is part of the RuOpinionNE-2024 competition (Loukachevitch et al., 2025), which focuses on processing Russian language news articles. Each sentence in the dataset is annotated with potential opinion tuples, where:

- **Holder**: The entity expressing the opinion, which can be a person, organization, or marked as `NULL` for general opinions or `AUTHOR` for the author's opinion.
- **Target**: The entity or object towards which the opinion is directed, always a non-empty named entity.
- **Polar_expression**: The phrase indicating the sentiment, which must be a contiguous substring of the original text.
- **Polarity**: The sentiment label, either `NEG` for negative or `POS` for positive.

Each tuple includes text spans for the holder, target, and polar expression, defined by character offsets. The task requires precise extraction of these components, ensuring that non-contiguous entities are correctly identified and aligned with the original text.

The dataset is divided into three parts: training (2,556 sentences), validation (1,316 sentences), and test (803 sentences). Notably, 41% of the examples in both the training and validation sets contain no opinions. On average, there are 1.22 opinions per example, with a maximum of 23 opinions in the training set and 15 in the validation set.

3.2 Prompt-Based Extraction

Our approach to structured sentiment analysis leverages prompt engineering, following current trends in large language model (LLM) applications. The designed prompt contains a detailed description of the problem, including the task's objectives, constraints, and specific rules for extracting opinion tuples. To enhance the model's understanding, we employ few-shot learning by providing carefully selected examples that demonstrate correct solutions to the task (Brown et al., 2020).

To ensure reliable JSON output and minimize generation errors, we implement structured generation techniques. This approach guides the model's output to follow a specific format, reducing the likelihood of malformed responses (Willard and Louf, 2023). The prompt includes explicit instructions for the model to first reason about the task in a dedicated reasoning field before producing the final output. This two-step process allows the model to analyze the text and consider the extraction rules before committing to a specific answer.

The prompt design incorporates several key elements:

- Clear definitions of each tuple component (Holder, Target, Polar_expression, Polarity)
- Specific rules for handling edge cases (e.g., NULL holders, AUTHOR markers)
- Examples demonstrating cases of opinion extraction
- Instructions for handling sentences without opinions

This structured approach to prompt design (see Appendix A for the full prompt) significantly improves the model's ability to extract opinion tuples accurately while maintaining the required output format. The inclusion of reasoning steps also provides valuable insights into the model's decision-making process, which can be useful for error analysis and system improvement.

We evaluated multiple state-of-the-art models for the prompt-based approach, selecting architectures based on their performance in the MERA Russian-language benchmark (Fenogenova et al., 2024) and model scale diversity. We tested one proprietary model and several open-source models including Llama and Qwen, ranging in size from 7B to 70B parameters (Table 3.2).

Model Name	**Params**	**Hugging Face Hub**	**Citation**
Llama-3.2	1B	meta-llama/Llama-3.2-1B-Instruct	(Grattafiori et al., 2024)
Llama-3.1	8B	meta-llama/Llama-3.1-8B-Instruct	
Llama-3.3	70B	meta-llama/Llama-3.3-70B-Instruct	
GPT-4o	-	-	(OpenAI et al., 2024)
Qwen2.5	7B	Qwen/Qwen2.5-7B-Instruct	(Qwen, 2024)
Qwen2.5	32B	Qwen/Qwen2.5-32B-Instruct	
Qwen2.5	72B	Qwen/Qwen2.5-72B-Instruct	
Vikhr-Nemo	12B	Vikhrmodels/Vikhr-Nemo-12B-Instruct-R-21-09-24	(Nikolich et al., 2024)
RuadaptQwen2.5	32B	msu-rcc-lair/RuadaptQwen2.5-32B-Instruct	(Tikhomirov and Chernyshev, 2023)
T-pro-it-1.0	32B	t-tech/T-pro-it-1.0	(T-bank, 2024)

Table 1: Models Used in Experiments

3.3 Fuzzy Substring Matching Algorithm

The inherent limitations of language models in precisely reproducing exact text substrings necessitated the development of a specialized fuzzy matching algorithm. When extracting opinion components through prompt-based approaches, models frequently introduce subtle alterations

such as case variations, omitted punctuation, or minor lexical substitutions. These discrepancies render exact string matching ineffective for aligning model predictions with original text spans.

Our algorithm employs a two-stage approximate matching strategy that balances computational efficiency with matching accuracy. The first phase utilizes regular expressions to identify potential start and end points of text spans, treating sequences of letters and digits as words and punctuation marks as natural separators. This preprocessing step ensures robust handling of common model errors such as misplaced or omitted punctuation by defining candidate spans based on these word/punctuation boundaries.

The core matching mechanism iterates through all possible text spans derived from these boundaries. For each candidate span extracted from the source text, the algorithm computes a similarity score against the model-generated substring using the `fuzz.ratio` function from `thefuzz` library[1], which calculates a normalized score based on Levenshtein distance. It systematically compares the target substring against all potential spans within the source text, identifying the span with the highest similarity score.

Finally, this highest-scoring alignment is selected only if its similarity score meets a predefined minimum confidence threshold, empirically set to 70 based on validation experiments. This threshold filters out spurious matches where the model might have generated text fragments significantly divergent from or absent in the original source. If a suitable match is found, the algorithm returns the similarity score, the character offsets (start:end) of the best-matching span in the original text, and the matched text itself. Otherwise, it indicates that no satisfactory match was found.

3.4 Aggregation Strategies

To address the inherent stochasticity in large language model outputs and improve prediction robustness, we developed three aggregation strategies that process multiple model predictions. Based on validation set performance, we determined that using 5 predictions per input was optimal. This approach mitigates random fluctuations in individual predictions and substantially improves the quality of the prediction.

Our implementation provides three distinct aggregation approaches:

Most Common strategy selects the single most frequent prediction across all generations. This conservative approach works best when models consistently produce high-quality outputs, prioritizing precision over recall.

Most Common N strategy addresses the limitations of basic frequency selection by dynamically adapting to the number of opinions in the first prediction. Rather than using a fixed N, this method selects the top N most frequent predictions, where N is equal to the number of opinions detected in the initial model response. This approach solves the key problem of the basic "Most Common" method, which often collapses multiple legitimate opinions into a single highest-frequency candidate due to its strict maximal frequency requirement. By preserving multiple high-probability candidates, each of which received substantial model support, this strategy better handles cases with multiple valid opinions while maintaining frequency-based filtering of spurious predictions.

[1] `https://github.com/seatgeek/thefuzz`

Combine strategy acts as a union operator, aggregating all unique predictions while eliminating duplicates. This approach maximizes recall at the cost of potentially lower precision, which is particularly useful when dealing with texts containing multiple valid opinions.

3.5 Model adapter finetuning

While prompt-based approaches demonstrated initial feasibility, their limitations in handling the task's specific linguistic patterns and structural constraints motivated our transition to adapter-based fine-tuning. The QLoRA approach (Dettmers et al., 2023) enables efficient parameter updates through 4-bit quantized training, allowing adaptation of large language models to the structured sentiment analysis task without prohibitive computational costs.

All models underwent identical training conditions: 10 epochs with a batch size of 8 distributed across 8×NVIDIA A100 80GB GPUs. The final 'Llama-3.3-70B' model, for example, completed training in approximately 1.5 hours using this setup. While this configuration accelerated the process, such extensive computational resources are not strictly mandatory; a minimum of 48GB of VRAM is sufficient to fine-tune 70B parameter models using QLoRA, though requiring longer training times. The training format maintained similarity to the prompt-based approach in output structure, but with crucial modifications: models now received only the raw sentence text as input and were required to directly output a JSON-formatted list of opinion tuples. This format explicitly removed the dedicated reasoning field used in initial prompt-based approaches, forcing models to learn the extraction patterns directly rather than through explicit reasoning steps.

Model	Size	Single	Most Common	Most Common N	Combine
Llama-3.1	8B	0.123	0.133	0.133	0.167
RuadaptQwen2.5	32B	0.170	0.171	0.172	0.173
Vikhr-Nemo	12B	0.183	0.184	0.189	0.174
Qwen2.5	32B	0.192	0.197	0.195	0.186
Qwen2.5	72B	0.200	0.208	0.209	0.191
T-pro-it-1.0	32B	0.199	0.199	0.203	0.201
gpt-4o	-	0.199	0.204	0.203	0.212
Llama-3.3	70B	0.234	0.241	0.240	**0.252**

Table 2: Model performance with prompt-based approach

4 Results

Our experimental results demonstrate significant performance improvements through model fine-tuning and multiple prediction aggregation. The prompt-based approach (Table 3.5) achieved the best performance with Llama-3.3-70B using Combine aggregation (F1=0.252), while QLoRA fine-tuning (Tables 3-4) substantially boosted results, with the same model reaching F1=0.379 on validation and 0.405 on test data using the Combine strategy. Inference using this final model configuration ('Llama-3.3-70B' with 'Combine' aggregation) on the entire validation or test dataset took approximately 5 minutes on the 8×A100 GPU setup.

Three key findings emerge from the data:

Model	Size	Single	Most Common	Most Common N	Combine
Llama-3.2	1B	0.245	0.247	0.243	0.253
Qwen2.5	7B	0.304	0.305	0.309	0.311
Qwen2.5	32B	0.331	0.338	0.336	0.326
T-pro-it	32B	0.322	0.331	0.328	0.326
RuadaptQwen2.5	32B	0.323	0.324	0.327	0.328
Llama-3.1	8B	0.335	0.333	0.334	0.335
Vikhr-Nemo	12B	0.345	0.346	0.350	0.352
Qwen2.5	72B	0.364	0.359	0.362	0.356
Llama-3.3	70B	0.372	0.368	0.372	**0.379**

Table 3: Model performance with QLoRA adapter on validation dataset

Model	Size	Single	Most Common	Most Common N	Combine
Llama-3.2	1B	0.263	0.263	0.265	0.278
Qwen2.5	7B	0.315	0.320	0.329	0.319
Qwen2.5	32B	0.329	0.342	0.340	0.349
Llama-3.1	8B	0.335	0.337	0.342	0.349
T-pro-it	32B	0.350	0.360	0.348	0.357
Vikhr-Nemo	12B	0.385	0.405	0.380	0.381
Qwen2.5	72B	0.375	0.376	0.382	0.383
RuadaptQwen2.5	32B	0.375	0.373	0.378	0.384
Llama-3.3	70B	0.396	0.394	0.390	**0.405**

Table 4: Model performance with QLoRA adapter on test dataset

Team Number	Validation F1	Test F1
Team 1 (ours)	**0.38**	**0.41**
Team 2	0.17	0.35
Team 3	0.28	0.33
Team 4	0.21	0.28
Team 5	0.34	0.24
Team 6	-	0.24
Team 7	-	0.20
Team 8	-	0.11

Table 5: Public leaderboard from the RuOpinionNE-2024 competition showing team rankings based on Validation and Test F1 scores.

- **Aggregation Impact**: The Combine strategy consistently outperformed other methods in final test evaluations (Table 4), particularly for larger models. However, the Most Common N strategy showed better validation performance (Table 3), suggesting different generalization patterns between datasets.
- **Scale vs Specialization**: While model scale generally correlated with performance, specialized adaptations proved crucial - the 32B RuadaptQwen2.5 outperformed the base Qwen2.5-32B by 3.8% in prompt-based settings despite identical parameter counts.
- **Training Effectiveness**: QLoRA fine-tuning produced dramatic improvements, with Vikhr-Nemo 12B showing the largest gain of 88% increase from prompt-based (0.183) to fine-tuned (0.345) validation performance.

The public leaderboard results (Table 3.5) confirm our approach's effectiveness, with our solution achieving 0.41 test F1 versus 0.35 for the nearest competitor. Notably, our validation score (0.38) strongly predicted final test performance, demonstrating the robustness of our evaluation methodology.

The results reveal several patterns:

- Smaller fine-tuned models (Qwen2.5-7B) outperformed much larger prompt-based models (Qwen2.5-72B), highlighting the value of adapter-based fine-tuning.
- Proprietary GPT-4o underperformed open-source models after fine-tuning, indicating the importance of task-specific adaptation over general capability.

4.1 Impact of Fuzzy Substring Matching

To quantify the contribution of the fuzzy substring matching algorithm described in Section 3.3, we conducted an ablation study comparing its performance against a baseline exact substring search. The exact matching approach performs a simple, case-insensitive search for the model-generated substring within the original sentence text. If an exact match is found, its character offsets are used; otherwise, the predicted component is discarded. We re-evaluated the best-performing aggregation strategy (Combine) for both prompt-based and fine-tuned models using this exact matching method. The 'Impact of Exact Match' column in the following tables shows

the change in F1 score when the fuzzy matching algorithm is replaced with the exact matching baseline.

The results (Tables 6 and 7) quantify the substantial benefit derived from the fuzzy matching algorithm, particularly for the prompt-based approach. The 'Impact of Fuzzy Matching' column in Table 6 shows F1 score improvements ranging from +0.020 to +0.042 when fuzzy matching is employed instead of exact matching. This highlights that prompt-based models frequently generate text spans with minor deviations (e.g., punctuation, changes in word form or case) from the source text, which only the fuzzy matching algorithm can successfully align.

For the QLoRA fine-tuned models (Table 7), the F1 score gains from using fuzzy matching over exact matching were less pronounced but still consistently positive. The improvements ranged mostly between +0.001 and +0.007 F1. This suggests that fine-tuning trains models to generate more precise text spans that adhere closer to the original text, thus reducing the reliance on fuzzy matching. However, even our best-performing fine-tuned model, Llama-3.3-70B, benefited from a +0.007 F1 increase, confirming that fuzzy matching remains a valuable final alignment step to capture residual generation variations, even for highly adapted models.

Model	**Size**	**Exact Match F1**	**Impact of Fuzzy Matching ($\vec{\Delta}$F1)**
Llama-3.1	8B	0.132	+0.035
Vikhr-Nemo	12B	0.152	+0.022
RuadaptQwen2.5	32B	0.152	+0.021
Qwen2.5	72B	0.158	+0.028
T-pro-it	32B	0.159	+0.042
Qwen2.5	32B	0.166	+0.020
gpt-4o	-	0.187	+0.025
Llama-3.3	70B	0.187	+0.031

Table 6: Performance comparison using exact matching for the prompt-based approach (validation set).

Model	**Size**	**Exact Match F1**	**Impact of Fuzzy Matching ($\vec{\Delta}$F1)**
Llama-3.2	1B	0.248	+0.004
Qwen2.5	7B	0.316	+0.005
RuadaptQwen2.5	32B	0.322	+0.006
Qwen2.5	32B	0.325	+0.001
T-pro-it	32B	0.327	0.000
Llama-3.1	8B	0.333	+0.002
Vikhr-Nemo	12B	0.350	+0.002
Qwen2.5	72B	0.355	+0.001
Llama-3.3	70B	0.372	+0.007

Table 7: Performance comparison using exact matching for the QLoRA fine-tuned approach (validation set).

5 Challenges in Cross-Dataset Adaptation

An experiment involved training a QLoRA adapter on combined data from SemEval 2022 (Barnes et al., 2022) and RuOpinionNE-2024. Despite the structural similarity in the annotation format, this approach yielded substantially worse performance (F1=0.27-0.31) compared to RuOpinionNE-only training (F1=0.38), revealing critical challenges in cross-dataset adaptation.

The failed methodology consisted of:

- Merging 7 SemEval datasets (Basque, Catalan, English, Norwegian, Spanish) with Russian data
- Removing neutral sentiment examples (2.1% of total) to match RuOpinionNE's binary labels
- Maintaining identical training parameters (10 epochs, batch size 8) as successful single-dataset runs

Three primary failure factors emerged from our analysis:

- **Label Space Mismatch**: While neutral examples were removed, models struggled with residual three-way classification patterns in the multilingual data, particularly for near-neutral expressions.
- **Structural Divergence**: SemEval's allowance for multi-segment Polar_expressions (6.6% vs 0.6% in RuOpinionNE) introduced conflicting annotation patterns that degraded model performance on Russian texts.
- **Annotation Guideline Discrepancy**: Despite apparent structural similarity in the output format, the underlying annotation guidelines and the interpretation of what constitutes an "opinion" might differ between SemEval and RuOpinionNE. This semantic divergence in annotation intent could lead to the model learning conflicting patterns.
- **Domain Discrepancy**: The news-focused Russian dataset clashed with SemEval's mixed domains (product reviews, social media), preventing effective knowledge transfer.

The results underscore the importance of dataset homogeneity, with the combined training approach performing 27% worse than monolingual adaptation. Leaderboard simulations suggest this method would have ranked 6th (F1=0.24) versus our actual 1st place finish (F1=0.41), highlighting the critical impact of focused dataset selection.

6 Generalizability and Cross-Domain Application

Given that the model was fine-tuned on a dataset derived from Russian news texts, it is naturally suited for extracting opinions from similar news sources beyond the specific competition dataset. This suggests its potential utility in broader media monitoring and analysis tasks within the news domain.

To explore the model's applicability in a significantly different domain, we applied it to scientific texts. We utilized a publicly available dataset RuSciBench (Vatolin et al., 2024) comprising approximately 182 thousand abstracts of scientific articles from various disciplines, sourced from the Russian scientific library elibrary.ru. We processed these abstracts with our best-performing fine-tuned model (‘Llama-3.3-70B‘ with ‘Combine‘ aggregation) to identify expressed opinions.

Table 8 shows the percentage of abstracts within different GRNTI (State Rubricator of Scientific and Technical Information) categories where at least one opinion was detected.

GRNTI Category	% Abstracts with Opinions
A comprehensive study of individual countries and regions	58.06
Religion. Atheism	57.97
Literature. Literary criticism. Oral folk art	57.57
Art. Art history	56.01
Politics and political sciences	50.96
History. Historical sciences	48.70
Mass communication. Journalism. Media	42.49
Complex problems of social sciences	41.94
Patent business. Invention. Rationalization	41.67
Biotechnology	40.00
Philosophy	39.93
Social sciences in general	39.50
Demography	39.32
Culture. Culturology	38.52
Scientific	38.10
Rural and forestry	37.78
Nuclear equipment	37.68
Food industry	33.73
Medicine and healthcare	33.36
Fisheries. Aquaculture	32.49

Table 8: Percentage of scientific abstracts (elibrary.ru) containing at least one detected opinion, by GRNTI category (Top 20).

The results indicate a higher prevalence of detectable opinions in humanities and social sciences disciplines (e.g., regional studies, religion, literature, politics, history) compared to natural sciences and technical fields, although opinions are found across the board. This aligns with the expectation that humanities research often involves interpretation, evaluation, and argumentation, which are likely to contain sentiment-bearing expressions. While a model specifically trained on scientific discourse might achieve higher accuracy and capture domain-specific nuances more effectively, these results demonstrate that the news-trained model can still provide interesting insights even in a significantly different domain.

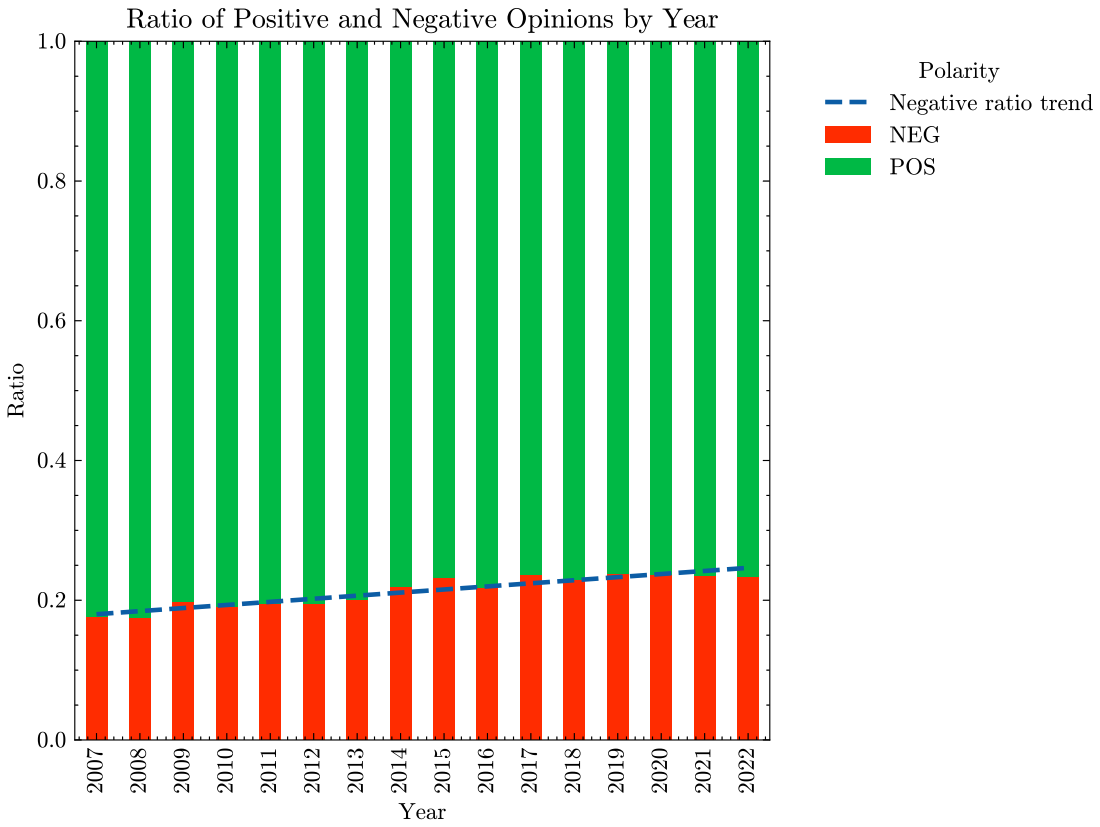


Figure 1: Negative sentiment ratio by year.

Further analysis of the extracted opinions revealed a slightly increasing trend in the proportion of negative opinions expressed in abstracts over the years covered by the dataset (see Figure 1). Additionally, examining the sentiment associated with specific targets (Figure 2) shows distinct patterns.

Beyond these expected extremes (like digital technology or terrorism), the analysis reveals more nuanced sentiment distributions for other entities within the context of scientific abstracts. For instance, terms central to the academic environment like "university" (83.9% positive) and "student" (73.2% positive) show predominantly positive sentiment, possibly related to research achievements or educational successes. However, the significant minority of negative mentions

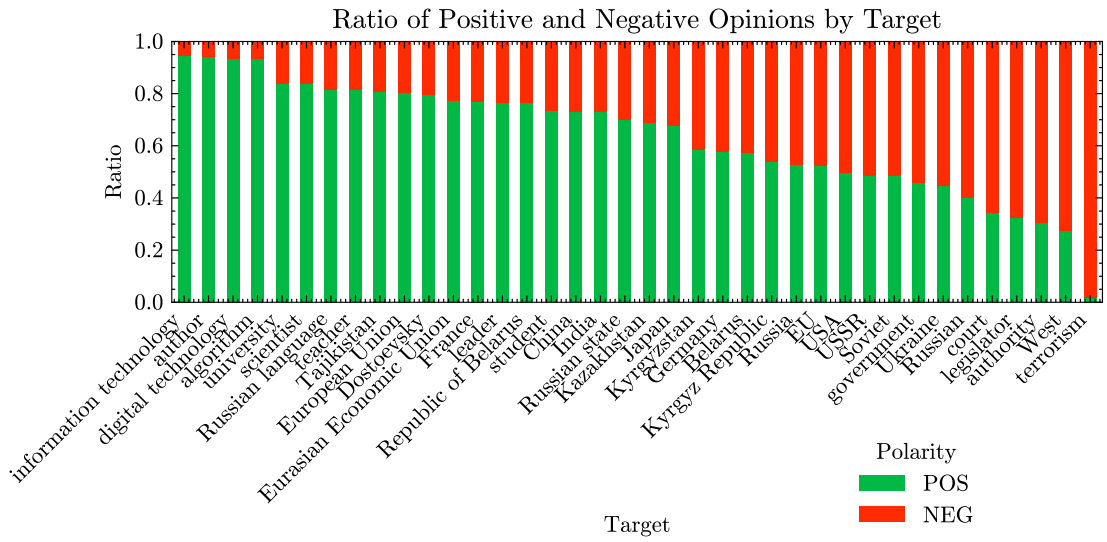


Figure 2: Most common opinion targets.

might stem from abstracts discussing challenges, critiques of educational systems, or studies focusing on student problems. Similarly, "scientist" (83.7% positive) follows this pattern.

Nations and political bodies often receive more mixed evaluations. For example, "Russia" (52.7% positive), "USA" (49.7% positive), and historical entities like "USSR" (48.5% positive) show near-even splits. This balance could reflect the objective and critical nature of academic discourse in fields like international relations, political science, or history, where abstracts might analyze policies, conflicts, or societal issues from multiple perspectives, leading to both positive and negative framing depending on the research focus. The comparatively lower positive sentiment towards "government" (45.8% positive) and higher negativity towards "West" (27.5% positive) might also indicate a tendency towards critical analysis of power structures and policies within certain academic disciplines prevalent in the dataset.

A closer look at the specific expressions associated with "Russia" provides further insight into this mixed picture.

Focusing specifically on "Russia" as an opinion target (Figure 3), the analysis of scientific abstracts up to the end of 2022 reveals a prevalence of terms associated with significant challenges. To generate this visualization, the extracted opinion expressions (polar expressions) associated with the target "Russia" were first lemmatized using pymorphy2 (Korobov, 2015) to group different word forms and then translated using Google Translate for representation in English. Negative expressions like "sanction", "corruption", and "problem" are most frequent, often linked to economic issues ("economic sanction", "lag behind") and security concerns ("threat", "information war", "threat of national security"). This reflects the intense focus within academic discourse during this period on Russia's geopolitical and socio-economic difficulties. However, positive terms also appear, highlighting perceived strengths and potential avenues for devel-

Figure 3: Word Cloud for Opinions Targeting "Russia" in Scientific Abstracts. Red color indicates negative sentiment, Green indicates positive.

opment. Words like "competitiveness", "advantage", "perspective", "cooperation", and "modernization" suggest discussions around national capabilities, strategic partnerships, and efforts towards improvement, such as "countering corruption" and "ensuring national security".

These exploratory applications suggest broader potential uses for the developed model. For instance, it could be integrated into specialized search engines to find articles mentioning specific entities (e.g., countries, organizations, individuals) in a positive or negative light. Furthermore, it could enable large-scale analytics, such as tracking sentiment trends towards certain topics or examining correlations in sentiment (e.g., identifying which countries are frequently mentioned negatively in articles where Russia, for example, is mentioned positively). These possibilities extend beyond the scope of this paper and represent promising avenues for future research.

7 Conclusion

Our work demonstrates that large language models can be effectively adapted for structured sentiment analysis in morphologically rich languages through targeted fine-tuning strategies. The combination of QLoRA-efficient adaptation and robust post-processing techniques enabled

state-of-the-art performance on the RuOpinionNE-2024 task, achieving a 0.405 F1 score without task-specific architectural modifications.

The results establish several important principles for applying LLMs to structured information extraction tasks. First, adapter-based fine-tuning with 4-bit quantization enables efficient specialization of billion-parameter models to complex annotation schemes. Second, open-source models can surpass proprietary alternatives when properly adapted to domain-specific requirements.

This research provides practical guidelines for adapting LLMs to opinion mining tasks in resource-rich language scenarios. Future work should investigate the generalizability of these techniques to other structured prediction tasks and low-resource language settings. The success of our approach suggests that continued development of efficient fine-tuning methods will further unlock the potential of large language models for complex information extraction challenges.

References

Jeremy Barnes, Laura Oberlaender, Enrica Troiano, Andrey Kutuzov, Jan Buchmann, Rodrigo Agerri, Lilja Øvrelid, and Erik Velldal. 2022. Semeval 2022 task 10: Structured sentiment analysis. // *International Workshop on Semantic Evaluation*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Ma teusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*, abs/2005.14165.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. 2021. Nonautoregressive encoder–decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34:5544–5556.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. // Lun-Wei Ku, Andre Martins, and Vivek Srikumar, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 9920–9948, Bangkok, Thailand, August. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and et al. 2024. The llama 3 herd of models, July.

Mikhail Korobov. 2015. Morphological analyzer and generator for russian and ukrainian languages. // Mikhail Yu. Khachay, Natalia Konstantinova, Alexander Panchenko, Dmitry I. Ignatov, and Valeri G. Labunets, *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, P 320–332. Springer International Publishing.

Natalia Loukachevitch, Natalia Tkachenko, Anna Lapanitsyna, Mikhail Tikhomirov, and Nicolay Rusnachenko. 2025. Ruopinionne-2024: Extraction of opinion tuples from russian news texts. // *Proceedings of International Conference on Computational Linguistics and Intelligent Technologies Dialogue 2025*.

Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: Constructing a state-of-the-art bilingual open-source instruction-following large language model for Russian. // *Proceedings of the 4rd Workshop on Multilingual Representation Learning (MRL) @ EMNLP-2024*. Association for Computational Linguistics.

OpenAI, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, Aj Ostrow, Akila Welihinda, Alan Hayes, and et al. 2024. Gpt-4o system card, October.

Haiyun Peng, Lu Xu, Lidong Bing, Fei Huang, Wei Lu, and Luo Si. 2019. Knowing what, how and why: A near complete solution for aspect-based sentiment analysis. // *AAAI Conference on Artificial Intelligence*.

Team Qwen. 2024. Qwen2.5: A party of foundation models, September.

Kevin Scaria, Himanshu Gupta, Siddharth Goyal, Saurabh Sawant, Swaroop Mishra, and Chitta Baral. 2024. InstructABSA: Instruction learning for aspect based sentiment analysis. // Kevin Duh, Helena Gomez, and Steven Bethard, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, P 720–736, Mexico City, Mexico, June. Association for Computational Linguistics.

T-bank. 2024. T-pro-it-1.0.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2015. Effective lstms for target-dependent sentiment classification. // *International Conference on Computational Linguistics*.

Mikhail Tikhomirov and Daniil Chernyshev. 2023. Impact of tokenization on llama russian adaptation. // *2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 163–168.

Peter D. Turney. 2002. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. // *Annual Meeting of the Association for Computational Linguistics*.

Aleksei Vatolin, Nikolai Gerasimenko, Anastasia Ianina, and Konstantin Vorotsov. 2024. Ruscibench: Open benchmark for russian and english scientific document representations. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 249–258*.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *ArXiv*, abs/2307.09702.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. // *Conference on Empirical Methods in Natural Language Processing*.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. // Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, P 504–510, Online, August. Association for Computational Linguistics.

He Zhao, Longtao Huang, Rong Zhang, Quan Lu, and Hui Xue. 2020. SpanMlt: A span-based multi-task learning framework for pair-wise aspect and opinion terms extraction. // Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 3239–3248, Online, July. Association for Computational Linguistics.

A Prompt for unsupervised approach

Инструкция по разметке данных для задачи определения мнений

Общие правила

Разметка выполняется на уровне предложений. Каждое предложение может содержать одну или несколько пар "мнение-отношение"(или не содержать их вовсе). Для каждой пары создается объект, включающий следующие ключи: source, target, polar_expression, polarity.

Ключи и их значения

1. **source (источник)**
   - **Определение**: Источник мнения — лицо или организация, выражающие мнение в предложении.
   - **Формат**: Строка или NULL, если источник не указан явно.
   - **Требования**:
     - Указывайте только конкретные именованные сущности (например, "Муртаза Рахимов "Единая Россия").
     - Если в предложении источник не указан явно или используется местоимение, но его можно вывести из контекста, источник отмечается как NULL.
     - Автор текста (AUTHOR) может быть источником, если мнение выражено в форме редакционного заявления, например, при оценке событий.
2. **target (цель)**
   - **Определение**: Цель мнения — лицо, организация или объект, на который направлено мнение.
   - **Формат**: Строка
   - **Требования**:
     - Указывайте только конкретные именованные сущности (например, "главу администрации "Монти").
     - Цель НЕ может быть отмечена как NULL! Если вы не можете найти цель, значит это мнение не подходит и его не нужно писать.
3. **polar_expression (эмоциональное выражение)**
   - **Определение**: Слово или фраза, выражающие отношение источника к цели.
   - **Формат**: Строка (например, "раскритиковал "восстановившая").
   - **Требования**:
     - Указывайте фразу в точности так, как она представлена в тексте.
     - Если эмоциональное выражение отсутствует, пара не создается.
     - Учитывайте фразы, выражающие скрытую оценку (например, "значительно ослабла").
     - Не может быть NULL!

4. **polarity (полярность)**
   - **Определение**: Тип эмоциональной окраски выражения.
   - **Допустимые значения**:
     - NEG — негативное мнение (например, "раскритиковал "уволил").
     - POS — позитивное мнение (например, "восстановившая "приглашен").
   - **Требования**:
     - Полярность определяется исходя из контекста предложения.
     - Если эмоциональное выражение нейтральное, пара не создается.

ВАЖНО! Требования ко полям source, target, polar_expression: Если значение не равно NULL или AUTHOR, то это должна быть подстрока исходного текста. Будь внимателен, подстрока должна состоять из подряд идущих слов и в точности совпадать с исходным текстом.

Примеры и уточнения

1. **Пример предложения с источником и целью:**
   Входной текст: Муртаза Рахимов подписал указ об отставке главы своей администрации
   Твой ответ:

```
{
   "opinions": [
      {
         "source": "Муртаза Рахимов",
         "target": "главы своей администрации",
         "polar_expression": "отставке",
         "polarity": "NEG"
      }
   ]
}
```

2. **Пример с** AUTHOR **как источником:**
   Входной текст: Правительство технократов проводило непопулярную экономическую политику.
   Твой ответ:

```
{
   "opinions": [
      {
         "source": "AUTHOR",
         "target": "правительство технократов",
         "polar_expression": "непопулярную экономическую политику",
         "polarity": "NEG"
      }
   ]
```

}

3. **Пример нейтрального предложения:**
   Входной текст: Указ о назначении нового главы своей администрации пока не подписан.
   Твой ответ:
   {"opinions": []}

Запрещенные значения

- source **и** target:
  - Не используйте местоимения (например, "он "она").
  - Не используйте обобщенные термины, если конкретная сущность не упомянута.
- polar_expression:
  - Не добавляйте субъективные интерпретации, отсутствующие в тексте.
- polarity:
  - Не отмечайте нейтральные фразы как NEG или POS.

Дополнительно

- Если вы сомневаетесь в полярности или источнике, добавляйте комментарии к разметке для последующего обсуждения.
- Проверяйте каждую размеченную пару на соответствие контексту предложения.
- Если в тексте нет мнений, на выходе должен быть пустой массив "{"opinions": []}".

Теперь твоя задача поррассуждать (рассуждения должны быть не больше 1000 слов) и разметить следующий текст:

ruSciFact: Open Benchmark for Verifying Scientific Facts in Russian

A. Vatolin
FRC CSC RAS
vatolinalex@gmail.com

N. Gerasimenko
MSU Institute for
Artificial Intelligence
nikgerasimenko@gmail.com

N. Loukachevitch
Lomonosov MSU
louk_nat@mail.ru

A. Ianina
Moscow Institute of
Physics and Technology
yanina.anastasia.mipt@gmail.com

K. Vorontsov
MSU Institute for
Artificial Intelligence
k.vorontsov@iai.msu.ru

Abstract

Against the backdrop of active LLM development, their tendency to hallucinate, as well as the growing volume of texts they generate, the validation of facts has become increasingly important and relevant. We propose ruSciFact[1], a new benchmark for fact-checking scientific claims in Russian. ruSciFact is structured as a NLI task, the goal is to verify whether a fact is confirmed by a given abstract. To generate facts, we used an 3-step pipeline based on LLaMA-405B, validating the resulting sentences with the help of assessors-terminologists. The ruSciFact dataset consists of 1128 pairs in the format <abstract, claim>, which we are releasing as open source together with the benchmark code. Additionally, we are open-sourcing the fact-generation pipeline[2], which facilitates the expansion of the dataset to specific scientific domains. We evaluated several popular language models on ruSciFact, including text embedders and generative models. The results show that this benchmark allows to effectively assess the fact-checking capabilities of LLMs in Russian.

Keywords: benchmark, large language model, embedder
DOI: 10.28995/2075-7182-2025-23-435-459

Аннотация

На фоне активного развития больших языковых моделей с их склонностью к галлюцинациям, а также роста количества сгенерированных с их помощью текстов, задача валидации фактов приобретает особую важность и актуальность. Мы представляем ruSciFact, новый бенчмарк для проверки научных утверждений на русском языке. ruSciFact имеет формат NLI: тестируется способность моделей проверить подтверждается ли утверждение аннотацией. Для генерации утверждений мы использовали 3-ступенчатый пайплайн на базе модели LLaMa-405B, затем валидируя полученные утверждения с помощью асессоров-терминологов. Всего ruSciFact содержит 1128 пар <аннотация, утверждение>, которые мы публикуем в открытом доступе вместе с кодом для оценки моделей на

[1] https://huggingface.co/collections/mlsa-iai-msu-lab/ruscifact-6803f71f2d3025988ebc489e
[2] https://github.com/mlsa-iai-msu-lab/ruscifact

бенчмарке. Также мы публикуем пайплайн генерации утверждений, который мы использовали, чтобы облегчить расширение датасета на дополнительные научные области. Мы протестировали на ruSciFact ряд современных языковых моделей, как текстовых эмбеддеров, так и генеративных LLM, и результаты показывают, что он позволяет эффективно оценить их способность к проверке научных фактов на русском языке.

Ключевые слова: бенчмарк, большие языковые модели, текстовый эмбеддер

1 Introduction

Assistance in evaluating the veracity of scientific claims becomes more and more important task in the modern research community due to widespread usage of Large Language Models in many spheres, including scientific investigation and paper writing processes. Such a task is called fact-checking – the verification of a statement against a corpus of documents that either support or refute the claim.

In this work we follow SciFact (Wadden et al., 2020), an expert-annotated dataset of $1,409$ scientific claims and related abstracts that either support or refute each claim. All the annotations in SciFact are provided with rationales (Lei et al., 2016), so each SUPPORTS / REFUTES decision has a reasoned justification. However, all the sentences presented there are written in English, making it impossible to use SciFact as a benchmark for fact-checking in other languages, thus limiting its usability. We propose a novel dataset and benchmark called ruSciFact, designed to evaluate the problem of scientific claim verification for texts written in Russian.

Another key distinction from SciFact is that, instead of relying on human annotators to reformulate naturally occurring claims in the papers, we used Large Language Models, making the data collection process entirely independent of human involvement. Specifically, we used a frontier-level open source language model LlaMa-405B (Touvron et al., 2023), which works across multiple languages including Russian. Using LLMs instead of human annotators enables the generation of larger datasets in less time.

We aimed to create a benchmark that is genuinely representative of the Russian scientific landscape. A direct translation of SciFact would inherently introduce biases toward international literature, thus potentially limiting the applicability and relevance of the benchmark to the Russian research community. Furthermore, translating specialized scientific claims and annotations involves complex terminology and domain-specific nuances. Automatic translation methods, such as using large language models (LLMs), do not yet reliably capture these nuances with perfect accuracy. Human translation by domain experts would mitigate this, but it requires substantial time and resources, making it impractical at scale.

Nevertheless, to support a more comprehensive evaluation of models, we still include a translated version of SciFact along with automated judgments by LLMs (prompts and evaluation results of wide range of LLMs can be found in Appendix A). This dual approach enables researchers to assess their systems across different dimensions: both in terms of language and scientific landscape, and also in terms of task formulation.

Our main contributions are as follows:

- a dataset of scientific claims with relevant abstracts;
- open-sourced code for the scientific claim generation pipeline, based on Llama-405B;
- open-sourced code for model evaluation using our proposed benchmark;
- the SciFact dataset translated to Russian.

2 Related Works

2.1 Scientific benchmarks

There are various benchmarks for assessing the quality of language models. Some are multi-purpose, covering a range of nearly unrelated tasks, while others focus on specific tasks, evaluating a model's ability to solve precise problems or work within a narrow domain. The first group of benchmarks include well-known GLUE (Wang, 2018) and SUPERGLUE (Wang et al., 2019) benchmarks. For evaluating embedders, there is the MTEB benchmark (Muennighoff et al., 2022), which spans 8 embedding tasks covering a total of 58 datasets and 112 languages.

A good example of benchmarks in the second group are those focused on scientific knowledge extraction and processing, such as SCIDOCS (Cohan et al., 2020). This benchmark consists of seven document-level tasks, ranging from citation prediction to document classification and recommendation of scientific articles. In our work, we follow its close relative, SciFact (Wadden et al., 2020), a dataset and benchmark for scientific claims and related abstracts that either support or refute the claims. SciFact consists of 1409 scientific claims verified against a corpus of 5183 abstracts. Each dataset entry includes a rationale providing an explanation for the final decision.

SciFact, in turn, is based on the Semantic Scholar Open Research Corpus (S2ORC) (Lo et al., 2019), a large corpus containing 81.1M English-language academic papers from `semanticscholar.org` spanning various academic disciplines. The original Semantic Scholar corpus (S2) consists of titles from scientific papers published in machine learning conferences and journals from 1985 to 2017, organized by year.

The SciFact public leaderboard is available online [3]. Based on SciFact, an AI fact-checking challenge called SciVER was developed. This shared task supports systems that should:

1. Take a scientific claim as input
2. Identify all relevant abstracts in a large corpus
3. Label them as Supporting or Refuting the claim
4. Select sentences as evidence for the label

SCIVER evaluation is performed at the abstract level (following (Thorne et al., 2018)) and sentence level (evaluating the correctness of the individual predicted evidence sentences).

All of the benchmarks discussed above lack non-English data, and we aim to bridge this gap for the Russian language. Despite being multilingual, even the MTEB benchmark (Muennighoff et al., 2022) lacks Russian-language tasks. A new benchmark for embedders designed specifically for the Russian language, called ruMTEB (Snegirev et al., 2024), has been introduced. Another example of a Russian benchmark is encodechka (Dale, 2022). However, neither of these contains a significant number of tasks focused specifically on scientific knowledge.

2.2 Automatic Generation of Claims for Fact-checking

Several approaches for the automatic generation of fact claims have been studied.

Pan et al. (Pan et al., 2021) proposed a method for automatically generating claims that can be supported, refuted, or unverifiable from evidence from Wikipedia. With this aim, the authors first employ a Question Generator to generate a question–answer pair for the evidence, then they

[3] `https://leaderboard.allenai.org/scifact/submissions/public`

convert the question-answer pair into a claim. The Question Generator and the QA-to-Claim model are off-the-shelf BART models trained on question-answering datasets.

The authors of (Wright et al., 2022) proposed both supervised and unsupervised methods for claim generation. For the supervised method, they trained the BART language model on a small set of <sentence, claim> pairs to directly generate claims (CLAIMGEN-BART). The unsupervised method, CLAIMGEN-ENTITY, involves identifying named entities in the target text fragment. For each named entity, the system generates a question about that entity, which can be answered from the sentence. The system then generates a claim from the question, as described in (Pan et al., 2021). Additionally, (Wright et al., 2022) proposed KBIN, a method for generating claims that contradict the text fragment. KBIN is designed for the biomedical domain. Entities from a given fragment are linked to concepts in the UMLS knowledge base. The initial concept is then replaced with a sibling concept (a sister concept) from UMLS. The closest concepts to the initial concept are selected based on pre-trained UMLS concept vectors. For each related concept, candidate claim variants are generated, with the best variants chosen according to the lowest perplexity and the highest probability of being a contradiction to the original claim.

Kozlova et al. (Kozlova et al., 2023) generate claims for a general fact-checking dataset in Russian. The authors used a trained paraphraser model to generate sentences for a text fragment. In addition, rule-based transformations were proposed as an alternative approach to synthetic data generation. For example, a randomly selected named entity in a statement was replaced with a different, randomly selected named entity.

Bussotti et al. (Bussotti et al., 2024) generate claims using both textual and tabular content. The authors experimented with the BART and LLaMa-2 models. To generate refuting claims, they used antonyms found by language models and resources such as WordNet and ConceptNet. The target entity in the initial claim is substituted with a related entity (e.g., antonym, sibling). Quantitative and human evaluations showed that the generated refuting examples demonstrate high variety, comparable to those created by humans.

3 Task Definition

Scientific fact-checking, or claim verification, focuses on finding evidence in the literature that SUPPORTS or REFUTES a given sentence (called a claim). According to (Wadden et al., 2020), "a scientific claim is an atomic verifiable statement expressing a finding". Validating such claims is a complex task that requires deep knowledge of specific scientific domains. This problem can be challenging even for human annotators, requiring specialists with a deep understanding of the topic. Therefore, the use of artificially accumulated knowledge systems, such as knowledge graphs or databases created using AI algorithms, can help address this issue. One of the most popular and effective examples of such approaches is Large Language Models (LLMs), whose prevalence in various applications is increasing daily. Using an LLM, one can automatically generate scientific claims and retrieve supporting abstracts without the involvement of human annotators.

4 ruSciFact dataset

4.1 Annotations dataset

We choose a publicly available dataset of Russian abstracts called ruSciBench as a starting point. It consists of 182000 pairs of titles and abstracts of scientific articles from Russian electronic library `elibrary.ru`. ruSciBench is the first benchmark for evaluating models on scientific texts in Russian. While all the papers in this dataset are written in Russian, they also include abstracts written in English.

4.2 Claims generation process

For the claim generation process, we used the largest model from the Llama family (Touvron et al., 2023). It has 405B parameters and is reported to perform on par with models like Gemini-1.5 (Team et al., 2024), GPT-4o (Achiam et al., 2023) and Claude-3.5 Sonnet for tasks such as MMLU Chat, instruction following, reasoning, code generation, long-context understanding (NIH/multi-needle task), and multilingual MGSM in a zero-shot scenario. However, Llama is an open-sourced model, unlike its proprietary counterparts from Google, OpenAI and Anthropic. We used the 8bit quantized variant of Llama-405B and ran it on 8 A100 with 80Gb Graphics RAM.

The generation pipeline can be described as follows (see Figure 1).

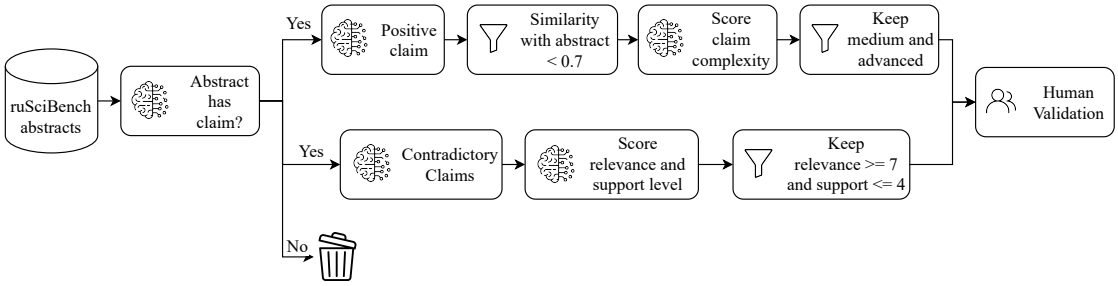


Figure 1: Overview of the data generation and validation pipeline.

4.2.1 Generation of positive examples (claims with evidence in the abstracts)

First, we filtered the abstracts in ruSciBench based on the following criteria. The eLibrary contains many abstracts that do not state meaningful claims. For instance, they may describe the conducted experiments or tell stories about the scientist, etc. A good example is the paper "COST-SENSITIVE LEARNING OF CLASSIFICATION KNOWLEDGE AND ITS APPLICATIONS IN ROBOTICS" by M.Tan[4].

Using a specific prompt (see Appendix B), we asked Llama-405B to generate a scientific claim based solely on the information in the abstract, or to specify that no claims were stated in the abstract. In the designed prompt, we listed the characteristics a potential claim should have and provided several examples of abstracts with corresponding claims. These characteristics include the following points:

[4] `https://www.elibrary.ru/item.asp?id=4996344`

- A fact must be related to the abstract and can be clearly supported by information from the abstract alone.
- A fact should not be too simple and must not contradict general knowledge or common sense.
- A fact should not contain uncertainty (e.g., a fact cannot include phrases like "in another study").

According to Llama-405B, only 42% of abstracts from eLibrary contain at least one scientific claim.

Second, after leaving out abstracts with no claims according to Llama, we continued the filtration process by eliminating facts that repeated information already present in the abstract. We performed this step using library **fuzzy** and its **partial_ratio** method to calculate how many words from the abstract and the fact overlapped. If the ratio was higher than 0.7, the claim was deleted.

Third, we filtered out simple facts. After manually annotating a small subset, we discovered that most of the generated facts were quite basic and could be derived from general knowledge (for example, "Russia is the biggest country territory-wise"). We created a prompt (see Appendix C) that asked the model to classify each fact into one of three classes: simple, medium, or advanced. Similar to the previous step, the prompt included a detailed description with examples. The distribution of fact complexity is the follows: most facts are of medium difficulty (61.4%), while simple and advanced facts are represented in roughly equal proportions (18.4% and 20.1% respectively).

4.2.2 Generation of negative examples (claims that contradict abstracts)

First, we selected abstracts that contained at least one fact (from the first step in the previous section). Then, we created a prompt (see Appendix E), that asks the model to generate a fact with the following features:

- A fact must contradict the abstract.
- A fact supporting the abstract cannot be obtained by removing the particle "not".
- A fact must be relevant to the abstract.
- A fact should not contain uncertainty (e.g., it cannot include phrases like "in another study").

Next, a fact should be derived from a chain of thought. It is a complex task for a language model. The model first changed truthful facts by adding particle "not" and then rephrased them so that the final claim would contain no negating expressions. The final fact was retrieved from such chains with another prompt.

The final step is to evaluate the fact's relevance and its contradiction to the abstract. We noticed that many facts neither relevant to the article nor contradict it. We designed an additional prompt (see Appendix F) that asks the model to grade each fact's relevance on a scale from 0 to 10 and indicate how much the abstract supports the fact. For ease of result interpretation, the model provides the response in JSON format. Only examples with a relevance score of 7 or higher and an abstract support level of 4 or lower were sent to the next step (labeling).

4.3 Human validation process

The generated facts were annotated by two terminologists with experience in analyzing scientific texts across various domains. The annotators obtained pairs: (generated claim, initial abstract) and

should provide the pairs with the following labels: "confirms", "contradicts", or "problematic". For the last two labels, the annotators wrote explanations. Problematic facts could contain non-existent words or be unclear. Each pair received two annotations.

In some cases, discrepancies arose between the annotations. For instance, one annotator might have deemed a fact too general for the abstract, while another annotator might have considered the same general fact to be confirmed. Currently, only facts with consistent annotations are included in the dataset. The final dataset contains 1128 pairs. Statistics on the length of the texts are presented in Table 1, the class distribution is shown in Table 2, and the distribution across scientific categories is detailed in Table 3.

Text Type	Mean Words	25% Words	50% Words	75% Words
Claim	15	12	14	18
Abstract	165	86	144	234

Table 1: Word count statistics for claims and abstracts in the ruSciFact dataset.

Label	Count
confirms	758
contradicts	369

Table 2: Class distribution in the final ruSciFact dataset after human validation.

Examples of claims and abstracts from the final dataset can be found in Appendix G.

5 Experiments

5.1 Supervised classification task

The task for evaluating embedders was formulated as a supervised classification problem. The input to the model was the concatenation of the claim and abstract embeddings. The model's objective was to predict a binary class: whether the abstract confirms or contradicts the claim. A linear classification layer was added on top of the embedder model to perform this classification. All layers of the embedder model were trained (unfrozen) for 10 epochs. The dataset was split equally into training and testing sets. This split presented a challenge due to the relatively small size of the resulting training set.

We tested several well-known embedders using the ruSciFact benchmark and summarised the results in Table 4. The list of models includes:

- Different RuBERT models (Kuratov and Arkhipov, 2019): rubert-tiny, rubert-tiny-turbo and rubert-tiny2
- SciRus-tiny (Gerasimenko et al., 2024; Vatolin et al., 2024)
- Multilingual E5 (base and large models)
- BGE-M3 embeddings (Chen et al., 2024)
- Universal Sentence Encoder for Russian (USER)[5]
- Language-agnostic BERT sentence embeddings (LaBSE) (Feng et al., 2020)

[5](https://huggingface.co/deepvk/USER-base)

Scientific Category (GRNTI)	**Percentage (%)**
Medicine and healthcare	30.85
Physics	7.09
Biology	6.83
Chemistry	6.12
Rural and forestry	5.85
Mechanical engineering	3.99
Mechanics	3.46
Geology	3.28
Polygraphy. Reprography. Photokinotechnology	3.10
Mathematics	2.75
Popular education. Pedagogy	2.04
Construction. Architecture	1.95
Mining	1.68
Automation. Computing technique	1.68
Linguistics	1.51
Psychology	1.51
State and law. Legal sciences	1.33
Electronics. Radio engineering	1.24
Electrical engineering	1.15
Informatics	1.15
Metallurgy	1.06

Table 3: Distribution of scientific categories in the ruSciFact dataset (categories with >1% representation).

- MPNet (Song et al., 2020)
- XLM-RoBERTa (Conneau, 2019)
- General Text Embedding family of models (gte-Qwen2-7B-instruct) (Bai et al., 2023)

The best result on ruSciFact among the models presented is achieved by BGE-m3 (Chen et al., 2024). Despite having only 359M parameters, compared to the larger E5 models (560M) and the significantly larger Qwen (7.61B), BGE still delivers the best performance in this set of experiments. Furthermore, the size of the embedder does not always correlate with the score for the other models as well: RuBERT-tiny2 is larger than SciRus-tiny3.1, yet their F1-scores are 0.66 and 0.68, respectively. On the other hand, mE5 is nearly 5 times bigger than RuBERT-tiny2 (118M vs. 29M) and its embeddings size is bigger(384 vs. 312), but it has the same score. The main reason for this discrepancy is context length: for mE5 it's 4 times (512 vs. 2048) compared to RuBERT-tiny2.

Model name	Model Config			F1-Score
	Model size	Embedding size	Max context length	
rubert-tiny	12M	312	512	0,66
sci-rus-tiny	23M	312	1K	0,67
sci-rus-tiny3.1	23M	312	1K	0,68
rubert-tiny-turbo	29M	312	2K	0,75
rubert-tiny2	29M	312	2K	0,66
multilingual-e5-small	118M	384	512	0,66
USER-base	124M	768	512	0,69
LaBSE-en-ru	129M	768	512	0,84
multilingual-e5-base	278M	768	512	0,78
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,74
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,78
USER-bge-m3	359M	1024	8K	**0,85**
multilingual-e5-large	560M	1024	512	0,8
multilingual-e5-large-instruct	560M	1024	512	0,77
gte-Qwen2-7B-instruct	7.61B	3584	131K	0,82

Table 4: Results of supervised evaluation on ruSciFact benchmark

5.2 Generative LLMs

We also tested 10 proprietary and 7 open-source generative language models on our benchmark, ruSciFact (see Appendix D for the prompt used in the evaluation). For the closed-source LLMs, we focused on well-known, powerful models. Since technical details for these models are not disclosed, we provide only brief notes about each of them.

- Claude 3.5 Sonnet (claude-3-5-sonnet-20241022) and Claude 3.5 Haiku (claude-3-5-haiku-20241022)
- GPT-4o (gpt-4o-2024-11-20) and GPT-4o mini (gpt-4o-mini-2024-07-18)
- Gemini 1.5 Pro (gemini-1.5-pro-002)
- YandexGPT 4 Pro and YandexGPT 4 Lite
- GigaChat: We tested following model versions: GigaChat, GigaChat Pro and GigaChat Max. As these models lack explicit version identifiers, we specify the access date (05.01.2025).

As for open-source LLMs, we conducted experiments with the following models:

- T-pro-it-1.0 (T-bank, 2024b) and T-lite-it-1.0 (T-bank, 2024a)
- Qwen2.5-32B-Instruct, Qwen2.5-7B-Instruct (Bai et al., 2023), RuadaptQwen2.5-32B-Instruct (Tikhomirov and Chernyshev, 2023; Tikhomirov and Chernyshev, 2024).
- sainemo-remix-12b-gptq-8bit (Gusev, 2024)
- Vikhr-Nemo-12B-Instruct (VikhrModels, 2024).

Most of the generative LLMs, both closed and open-source, achieve F1-scores above 90% (see Table 5). The best results are obtained by Claude 3.5 Sonnet and GPT-4o among models developed internationally, and by T-pro among Russian-developed LLMs. Additionally, even relatively small models, such as T-lite and Qwen2.5-7B, demonstrate strong performance, scoring

Model name	Open weights	Model size	F1-Score
claude-3-5-sonnet-20241022	No	-	0,99
gpt-4o-2024-11-20	No	-	0,98
T-pro-it-1.0	Yes	32B	0,98
gemini-1.5-pro-002	No	-	0,97
Qwen2.5-32B-Instruct	Yes	32B	0,97
RuadaptQwen2.5-32B-instruct	Yes	32B	0,96
YandexGPT 4 Pro	No	-	0,95
gpt-4o-mini-2024-07-18	No	-	0,94
T-lite-it-1.0	Yes	7B	0,94
Qwen2.5-7B-Instruct	Yes	7B	0,93
claude-3-5-haiku-20241022	No	-	0,93
GigaChat Max	No	-	0,87
sainemo_remix_12b_gptq_8bit	Yes	12B	0,80
YandexGPT 4 Lite	No	-	0,79
GigaChat	No	-	0,77
Vikhr-Nemo-12B-Instruct-R-21-09-24	Yes	12B	0,75
GigaChat Pro	No	-	0,75

Table 5: Results of zero shot LLM evaluation on ruSciFact benchmark

over 93% on our ruSciFact benchmark. Overall, the fact-checking task does not appear to be particularly challenging for powerful models with large context windows and well-trained reasoning abilities across multiple languages.

5.3 Retrieval evaluation

The retrieval task was formulated following the methodology described in MTEB (Muennighoff et al., 2022). Dataset entry consists of a claim (query) and the set of all abstracts in the dataset (corpus). The goal is to retrieve the relevant abstract for each given claim. To achieve this, we used the evaluated model to embed all claims and all abstracts. Cosine similarity scores were computed between each claim embedding and all abstract embeddings. Based on these scores, abstracts were ranked for each claim. The primary metric used for evaluation in this setup is Mean Reciprocal Rank at k (MRR@k), specifically MRR@1, which measures whether the correct abstract is ranked first for a given claim.

Several retrieval models were tested using our proposed benchmark ruSciBench. The list of models includes:

- GritLM-7B (Muennighoff et al., 2024)
- SFR-Embeddings by Salesforce Research (SFR- Embedding-2-R) (Meng* et al., 2024)
- Multilingual E5 (small, base, large and large-instruct) (Wang et al., 2024)
- Qwen2-7B-instruct (Bai et al., 2023)
- Stella 1.5B (Alibaba, 2024)

Model name	Model Config			**MRR@1**
	Model size	Embedding size	Max context length	
GritLM-7B	7.24B	4096	32K	0,91
SFR-Embedding-2_R	7.11B	4096	32K	0,90
multilingual-e5-large-instruct	560M	1024	512	0,88
multilingual-e5-base	278M	768	512	0,80
multilingual-e5-large	560M	1024	512	0,78
gte-Qwen2-7B-instruct	7.61B	3584	131K	0,81
stella_en_1.5B_v5	1.54B	1024	131K	0,80
multilingual-e5-small	118M	384	512	0,70
USER-bge-m3	359M	1024	8K	0,76
sci-rus-tiny	23M	312	1K	0,57
sci-rus-tiny3.1	23M	312	1K	0,59
USER-base	124M	768	512	0,62
rubert-tiny-turbo	29M	312	2K	0,55
LaBSE-en-ru	129M	768	512	0,42
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,39
stella_en_400M_v5	435M	1024	8K	0,26
rubert-tiny2	29M	312	2K	0,15
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,12
rubert-tiny	12M	312	512	0,06

Table 6: Results of retrieval evaluation on ruSciFact benchmark

- BGE-M3 (Chen et al., 2024)
- SciRus models (Gerasimenko et al., 2024)
- RuBERT models (Kuratov and Arkhipov, 2019)
- Language-agnostic BERT sentence embeddings (LaBSE) (Feng et al., 2020)
- Multilingual MPNet (Song et al., 2020)
- XLM-RoBERTa (Conneau, 2019)
- Universal Sentence Encoder for Russian (USER)

The results are presented in Table 6. As shown, GritLM provides the best result (it is also the largest model tested in this experiment). Multilingual E5 models also achieve high results being significantly smaller than GritLM or SFR Embeddings.

6 Conclusion and Future Work

In this work, we introduced a new benchmark, ruSciFact, for evaluating fact-checking capabilities within the scientific domain in the Russian language. We open-sourced a dataset of scientific facts with rationales, generated using a multistep pipeline based on LLaMa-405B and validated by human assessors. In addition to the benchmark, we open-sourced the fact-generation pipeline, aiming to support research in LLM skill validation, benchmarking, scientific data processing, and automatic scientific knowledge extraction. ruSciFact is intended to encourage the development of language models designed for research-oriented tasks while providing a platform for comparing

Russian-speaking models.

Acknowledgements

The research is part of project #23-Ш05-21 SES MSU "Development of mathematical methods of machine learning for processing large-volume textual scientific information". We would like to thank eLibrary for providing the datasets.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alibaba. 2024. stella-en-1.5b-v5.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jean-Flavien Bussotti, Luca Ragazzi, Giacomo Frisoni, Gianluca Moro, and Paolo Papotti. 2024. Unknown claims: Generation of fact-checking training examples from unstructured and structured data. // *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, P 12105–12122.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*.

A Conneau. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

David Dale. 2022. Рейтинг русскоязычных энкодеров предложений, June. [Online; posted 12-June-2022].

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Nikolai Gerasimenko, Aleksei Vatolin, Anastasia Ianina, and Konstantin Vorotsov. 2024. Scirus: Tiny and powerful multilingual encoder for scientific texts. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 191–200*.

Ilya Gusev. 2024. sainemo-remix-12b-gptq-8bit.

Anastasia Kozlova, Denis Shevelev, and Alena Fenogenova. 2023. Fact-checking benchmark for the russian large language models. // *Proceedings of the International Conference "Dialogue*, volume 2023.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. *arXiv preprint arXiv:1606.04155*.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Dan S Weld. 2019. S2orc: The semantic scholar open research corpus. *arXiv preprint arXiv:1911.02782*.

Rui Meng*, Ye Liu*, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-2: Advanced text embedding with multi-stage training.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Liangming Pan, Wenhu Chen, Wenhan Xiong, Min-Yen Kan, and William Yang Wang. 2021. Zero-shot fact verification by claim generation. *// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, P 476–483.

Artem Snegirev, Maria Tikhonova, Anna Maksimova, Alena Fenogenova, and Alexander Abramov. 2024. The russian-focused embedders' exploration: rumteb benchmark and russian embedding model design. *arXiv preprint arXiv:2408.12503*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

T-bank. 2024a. T-lite-it-1.0.

T-bank. 2024b. T-pro-it-1.0.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

Mikhail Tikhomirov and Daniil Chernyshev. 2023. Impact of tokenization on llama russian adaptation. *// 2023 Ivannikov Ispras Open Conference (ISPRAS)*, P 163–168. IEEE.

Mikhail Tikhomirov and Daniil Chernyshev. 2024. Facilitating large language model russian adaptation with learned embedding propagation. *arXiv preprint arXiv:2412.21140*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: open and efficient foundation language models. arxiv. *arXiv preprint arXiv:2302.13971*.

Aleksei Vatolin, Nikolai Gerasimenko, Anastasia Ianina, and Konstantin Vorotsov. 2024. Ruscibench: Open benchmark for russian and english scientific document representations. *Doklady Mathematics, 2024, Vol. 110, No. 4, pp. 249–258*.

VikhrModels. 2024. Vikhr-nemo-12b-instruct-r.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *CoRR*, abs/1905.00537.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Alex Wang. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*.

Dustin Wright, David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Isabelle Augenstein, and Lucy Lu Wang. 2022. Generating scientific claims for zero-shot scientific fact checking. *// Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2448–2460.

A Scifact tranlation process and results

To adapt the original SciFact benchmark for Russian language evaluation, we undertook a multi-stage translation process for its English claims and abstracts. The initial translation was performed using the `gpt-4o-2024-11-20` large language model. Given the complexity of scientific terminology and the need for high fidelity, a rigorous quality control procedure was implemented based on the LLM-as-a-judge paradigm.

For quality assessment, we employed the `claude-3-5-haiku-20241022` model to act as an automated judge. This judge model evaluated each translated claim-abstract pair against the original English text, checking for semantic equivalence, grammatical correctness in Russian, and, crucially and the preservation of the core scientific meaning.

The translation and evaluation were conducted iteratively. After the initial translation pass by `gpt-4o`, all pairs were evaluated by the `claude-3-5-haiku` judge. Pairs flagged as containing translation errors (e.g., incorrect terminology, altered meaning, grammatical issues) were automatically collected. These problematic pairs were then fed back into the `gpt-4o` model for a second translation attempt. The re-translated pairs were subsequently re-evaluated by the judge model. This cycle of translation and evaluation was repeated for a total of three iterations.

Despite the iterative refinement, a small subset of examples remained flagged as incorrectly translated after the third iteration. These persistent cases, often involving particularly nuanced or complex scientific language, were manually reviewed and corrected by human experts to ensure the final dataset's accuracy and reliability. This combined approach of automated translation, iterative LLM-based judging, and final manual verification aimed to produce a high-quality Russian version of the SciFact dataset.

The evaluation of models on the translated SciFact dataset followed the same three setups as described for the ruSciFact benchmark: supervised classification, zero-shot LLM evaluation, and retrieval. The task formulations remain consistent with those outlined in the main body of the paper. The results for each setup are presented in Tables 7, 8, and 9.

Analysis of Results:

Comparing the results on the translated SciFact dataset (Tables 7, 8, 9) with those on the ruSciFact dataset (Tables 4, 5, 6), a noticeable decrease in performance is observed across all evaluation setups. This consistent trend suggests that the original SciFact data may be inherently more complex than ruSciFact, or it represents a different data distribution (e.g., varying scientific domains or writing styles) on which the evaluated models generally perform less effectively.

In the **supervised classification task** (Table 7), the highest F1-score achieved is 0.48 (by `multilingual-e5-large` and `gte-Qwen2-7B-instruct`), significantly lower than the 0.85 achieved on ruSciFact. This gap underscores the potential increased difficulty or distributional shift presented by the SciFact data, possibly amplified by translation nuances.

For the **zero-shot LLM evaluation** (Table 8), the performance drop is also substantial. While top models like `claude-3-5-sonnet-20241022` still perform relatively well (F1-score of 0.86), this is considerably lower than the near-perfect scores achieved by several models on ruSciFact. This difference highlights that even advanced LLMs find the translated SciFact version more challenging, likely due to the intrinsic properties of the data, although translation artifacts might also contribute.

The **retrieval task** (Table 9) shows the most dramatic performance decrease. The top MRR@1

Model name	Model Config			F1-Score
	Model size	Embedding size	Max context length	
rubert-tiny	12M	312	512	0,28
sci-rus-tiny	23M	312	1K	0,31
sci-rus-tiny3.1	23M	312	1K	0,33
rubert-tiny-turbo	29M	312	2K	0,23
rubert-tiny2	29M	312	2K	0,31
multilingual-e5-small	118M	384	512	0,28
USER-base	124M	768	512	0,27
LaBSE-en-ru	129M	768	512	0,40
multilingual-e5-base	278M	768	512	0,33
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0,27
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0,36
USER-bge-m3	359M	1024	8K	0,43
multilingual-e5-large	560M	1024	512	**0,48**
multilingual-e5-large-instruct	560M	1024	512	0,47
gte-Qwen2-7B-instruct	7.61B	3584	131K	**0,48**

Table 7: Results of supervised evaluation on translated SciFact benchmark

Model name	Open weights	Model size	F1-Score
claude-3-5-sonnet-20241022	No	-	0,86
T-pro-it-1.0	Yes	32B	0,81
gpt-4o-2024-11-20	No	-	0,81
gemini-1.5-pro-002	No	-	0,79
Qwen2.5-32B-Instruct	Yes	32B	0,77
GigaChat Max	No	-	0,75
gpt-4o-mini-2024-07-18	No	-	0,75
RuadaptQwen2.5-32B-instruct	Yes	32B	0,74
claude-3-5-haiku-20241022	No	-	0,73
YandexGPT 4 Pro	No	-	0,69
T-lite-it-1.0	Yes	7B	0,67
Qwen2.5-7B-Instruct	Yes	7B	0,64
GigaChat Pro	No	-	0,56
YandexGPT 4 Lite	No	-	0,53
Vikhr-Nemo-12B-Instruct-R-21-09-24	Yes	12B	0,49
sainemo_remix_12b_gptq_8bit	Yes	12B	0,48
GigaChat	No	-	0,47

Table 8: Results of zero shot LLM evaluation on translated SciFact benchmark

Model name	Model Config			MRR@1
	Model size	Embedding size	Max context length	
SFR-Embedding-2_R	7.11B	4096	32K	0.55
multilingual-e5-large-instruct	560M	1024	512	0.52
GritLM-7B	7.24B	4096	32K	0.52
multilingual-e5-large	560M	1024	512	0.52
gte-Qwen2-7B-instruct	7.61B	3584	131K	0.51
stella_en_1.5B_v5	1.54B	1024	131K	0.50
multilingual-e5-base	278M	768	512	0.49
USER-bge-m3	359M	1024	8K	0.49
multilingual-e5-small	118M	384	512	0.47
sci-rus-tiny3.1	23M	312	1K	0.41
sci-rus-tiny	23M	312	1K	0.39
USER-base	124M	768	512	0.37
LaBSE-en-ru	129M	768	512	0.37
rubert-tiny-turbo	29M	312	2K	0.36
paraphrase-multilingual-mpnet-base-v2	278M	768	512	0.32
stella_en_400M_v5	435M	1024	8K	0.23
rubert-tiny2	29M	312	2K	0.21
sn-xlm-roberta-base-snli-mnli-anli-xnli	278M	768	512	0.12
rubert-tiny	12M	312	512	0.09

Table 9: Results of retrieval evaluation on translated SciFact benchmark

score is 0.55 (`SFR-Embedding-2_R`), compared to 0.91 on ruSciFact. This indicates significant difficulty in accurately matching translated claims to translated abstracts based on semantic similarity, potentially stemming from the different nature of the SciFact content and how translation alters specific phrasing crucial for embedding models.

Overall, the translated SciFact benchmark serves as a valuable complementary evaluation tool. The observed lower scores across the board emphasize the challenges models face when dealing with potentially harder or differently distributed data, further complicated by cross-lingual transfer, underscoring the complexities of robust fact verification.

B Prompt for Positive Claim Generation

In the prompts presented in the appendices, the placeholder `{text}` is replaced with the actual text of the abstract during the generation process.

Вы ученый, который хорошо разбирается во всех областях науки. Ваша задача - записать один факт, который следует из аннотации к статье. Вы не можете скопировать текст из аннотации, вам нужно написать факт, который следует из аннотации, но прямо в ней не указан. Примечание: Убедитесь, что извлеченный факт точно выведен из содержания аннотации,

без добавления какой-либо дополнительной информации или интерпретации. При написании факта избегай ссылок на аннотацию (в приведенном текст, в данной работе, предложенный метод). Также при написании факта избегай неопределенности, например ”с определенным свойствами”, ”в некоторых состояниях”, ”определенной длины”. Если по аннотации невозможно написать точный факт, то напиши ”Аннотация не содержит факт”

Ниже приведены 2 примера фактов и аннотаций, на основе которых они были написаны.

Аннотация к статье: Ожидается, что снижение уровня гомоцистеина в сыворотке крови с помощью фолиевой кислоты снизит смертность от ишемической болезни сердца. Известно, что максимальное снижение уровня гомоцистеина достигается при приеме фолиевой кислоты в дозе 1 мг/сут, но эффект более низких доз (имеющих отношение к обогащению пищевых продуктов) неясен. МЕТОДЫ Мы рандомизировали 151 пациента с ишемической болезнью сердца на 1 из 5 доз фолиевой кислоты (0,2, 0,4, 0,6, 0,8 и 1,0 мг/сут) или плацебо. Первоначально, через 3 месяца приема добавок и через 3 месяца после прекращения приема фолиевой кислоты, были взяты образцы крови натощак для анализа на содержание гомоцистеина и фолиевой кислоты в сыворотке крови. РЕЗУЛЬТАТЫ: Средний уровень гомоцистеина в сыворотке крови снижался при увеличении дозы фолиевой кислоты до максимума при приеме 0,8 мг фолиевой кислоты в день, когда снижение уровня гомоцистеина (с поправкой на плацебо) составляло 2,7 мкмоль/л (23%), что аналогично известному эффекту приема фолиевой кислоты в дозах 1 мг/сут и выше. Чем выше был исходный уровень гомоцистеина в сыворотке крови человека, тем сильнее была реакция на фолиевую кислоту, но статистически значимое снижение наблюдалось независимо от исходного уровня. Уровень фолиевой кислоты в сыворотке крови повышался примерно линейно (5,5 нмоль/л на каждые 0,1 мг фолиевой кислоты). Индивидуальные колебания уровня гомоцистеина в сыворотке крови, измеренные в группе плацебо, были значительными по сравнению с эффектом приема фолиевой кислоты, что указывает на то, что мониторинг снижения уровня гомоцистеина у конкретного человека нецелесообразен. ВЫВОДЫ Для достижения максимального снижения уровня гомоцистеина в сыворотке крови во всем диапазоне уровней гомоцистеина в популяции, по-видимому, необходима доза фолиевой кислоты в размере 0,8 мг/сут. Нынешние уровни обогащения пищевых продуктов в США позволят достичь лишь небольшой доли достижимого снижения уровня гомоцистеина.

Факт из статьи: Дефицит витамина B9 снижает уровень гомоцистеина в крови.

Аннотация к статье: Для реакции выделения кислорода (OER) были разработаны известные на Земле катализаторы первого ряда (3d) на основе переходных металлов; однако они работают при потенциалах, значительно превышающих термодинамические требования. Теория функционала плотности предполагает, что не трехмерные металлы с высокой валентностью, такие как вольфрам, могут модулировать трехмерные оксиды металлов, обеспечивая почти оптимальную энергию адсорбции для предлагаемых промежуточных продуктов. Мы разработали метод синтеза при комнатной температуре для получения гелеобразных оксигидроксидных материалов с атомарно однородным распределением металлов. Эти гелеобразные оксигидроксиды FeCoW обладают самым низким перенапряжением (191 милливольт), зарегистрированным при 10 миллиамперах на квадратный сантиметр в щелочном электролите. Катализатор не проявляет признаков разложения после более чем 500 часов работы. Рентгеновское поглощение и компьютерные исследования показывают синергетическое взаимодействие между вольфрамом, железом и кобальтом в создании благоприятной локальной координационной среды и электронной структуры, которые повышают энергетику предложения.

Факт из статьи: Усовершенствованные катализаторы OER демонстрируют стабильную активность в течение нескольких сотен часов.
Если в аннотации не содержится фактов, например: В статье рассмотрено становление британо-японских отношений в период биполярности, отражена история формирования двусторонних отношений, а также сотрудничество в экономической, политической, научно-технической и социально-культурной сферах в указанный период. Главный акцент был сделан на рассмотрении зарождения отношений между Великобританией и Японией и анализа динамики их развития. то напиши ”Аннотация не содержит факт”.
Теперь ваша задача - записать этот факт в следующую аннотацию, точно следуя инструкциям
{text}
Не пиши никаких вводных слов (например ”из аннотации следует, что”), только факт. Напишите свой факт из этой аннотации:

C Prompt for Fact Complexity Classification

Инструкция по классификации фактов по научным статьям

Вы являетесь учёным, обладающим глубокими знаниями во всех областях науки. Ваша задача - определить сложность факта, изложенного в аннотации к научной статье. Факты могут быть классифицированы по трём уровням сложности: простой, средний и сложный. В отдельных случаях факт может

быть неопределённым.

Категории фактов:

1. Простой факт:

Факт очевиден большинству образованных людей и не требует дополнительных исследований или чтений для его подтверждения или опровержения. Такие факты известны из общих знаний.
Примеры простых фактов:

- Экономическая эффективность потребления может варьироваться в зависимости от личностных свойств потребителей.
- В России существует закон, регулирующий проведение медицинских научных исследований с участием человека и/или лабораторных животных.
- Нитраты могут вызывать токсические эффекты у животных.
- Российские вузы сокращают количество бюджетных мест.
- В Республике Алтай представлены многочисленные виды туризма и отдыха.
- У студентов вузов ценность внутреннего успеха выше, чем внешнего.
- Изменение жесткости элементов конструкции здания может быть вызвано разными факторами.
- Преступления, связанные с фальшивомонетничеством, совершались на территории Российской Федерации и Белгородской области.

2. Средний факт:

Факт достаточно сложен, большинству людей понадобится читать специализированные статьи или проводить запросы в интернете, чтобы понять, подтвердить или опровергнуть этот факт.
Примеры средних фактов:

- Поражение молочной железы эхинококком может быть излечено хирургическим путем.
- Татарстан стал более засушливым регионом за последние десятилетия.

3. Сложный факт:

Факт требует специфических знаний или экспертизы в данной научной области для его понимания.
Примеры сложных фактов:

- Прокатка СВС-продуктов в валках прокатного стана перед измельчением в шаровой мельнице увеличивает эффективность измельчения.
- У больных диабетическим макулярным отеком наблюдается повышенный уровень брадикинина в крови.
- Стентирование коронарных артерий не вызывает значимых изменений показателей глобальной и сегментарной продольной систолической деформации миокарда левого желудочка в первые 3 сут после процедуры.

4. Неопределённый факт:

Факт недостаточно ясен, неполон или содержит ссылки, требующие дополнительных разработок или исследований.
Примеры неопределённых фактов:

- Российские компании могут использовать разработанную шкалу для определения уровня развития ориентации на бренд.
- Предыдущие модели теплоусвоения вермикулита были неточными.

Ваша задача:

Внимательно изучите предоставленный факт и напишите одно из следующих определений сложности факта: "простой "средний "сложный"или "неопределенный". Напиши только одно слово, не объясняй причины такой классификации
Вот факт, для которого это нужно написать: {text}

D Prompt for zero shot LLM evaluation

Инструкция по разметке аннотаций и фактов
Цель
Цель этого задания - определить, насколько представленный факт соответствует содержанию аннотации научной статьи. Вы будете работать с парами "факт-аннотация"и назначать им один из трех предложенных классов.
Процедура разметки

1. **Внимательно прочитайте** указанный факт и соответствующую ему аннотацию из научной статьи.
2. **Определите соответствие** между фактом и содержанием аннотации, используя ниже приведенные классы.

Классы для разметки

1. **Аннотация подтверждает факт**: Выберите этот класс, если информация в аннотации подтверждает или прямо указывает на изложенный в ней факт. Выбирайте этот класс только если факт в точности соответствует аннотации, без необходимости "додумывать"что-либо.
2. **Аннотация не подтверждает или противоречит факту**: Аннотация опровергает факт: информация в аннотации прямо противоречит заявленному факту.

Важно

- Расценивайте смысл и контекст фактов и аннотаций, избегая слишком буквального истолкования.
- В ответе должно быть только название класса: ”Аннотация подтверждает факт”, ”Аннотация не подтверждает или противоречит факту”

E Prompt for Negative Claim Generation

Вы ученый, который хорошо разбирается во всех областях науки. Ваша задача - написать один факт, который был бы релевантен аннотации, но не следует из нее. Не используй отрицание, вместо этого напишите факт, который не подтверждается аннотацией, но релевантен ей! Вы не можете скопировать текст из аннотации, вам нужно написать факт, который не следует из аннотации. Примечание: Убедитесь, что извлеченный факт точно выведен из содержания аннотации, без добавления какой-либо дополнительной информации или интерпретации. При написании факта избегай ссылок на аннотацию (в приведенном текст, в данной работе, предложенный метод). Также при написании факта избегай неопределенности, например "с определенным свойствами "в некоторых состояниях "определенной длины". Если по аннотации невозможно написать точный факт, то напиши "Аннотация не содержит факт".
Ниже приведены 2 примера фактов и аннотаций, на основе которых они были написаны.

Пример 1

Аннотация к статье:
Статья посвящена актуальным проблемам установления уголовного запрета в сфере профессиональной медицинской деятельности. Проанализированы предложения Следственного комитета РФ о внесении изменений в действующий Уголовный Кодекс РФ, на основании социологического опроса и изучения уголовных дел, сделаны выводы о необходимости реформы уголовного закона. Основным является вывод о невозможности решения актуальных проблем в российском здравоохранении исключительно уголовно-правовыми средствами.
Факт из статьи: Актуальные проблемы в российском здравоохранении могут быть решены исключительно уголовно-правовыми средствами

Пример 2

Аннотация к статье:
Рассмотрены вопросы создания системы охраны территорий и объектов стратегического назначения. Предложены структура и способ построения такой системы, использующие методы теории решеток. Для обработки и анализа информации с датчиков физических величин и последующего принятия решений применяются решетки, построенные с помощью оператора замыкания.
Факт из статьи: Решетки могут быть построены без использования оператора замыкания

Пример аннотации без фактов

Если в аннотации не содержится фактов, например:
В статье рассмотрено становление британо-японских отношений в период биполярности, отражена история формирования двусторонних отношений, а

также сотрудничество в экономической, политической, научно-технической и социально-культурной сферах в указанный период. Главный акцент был сделан на рассмотрении зарождения отношений между Великобританией и Японией и анализа динамики их развития.
то напиши "Аннотация не содержит факт".

Пример неподходящего факта

Аннотация к статье: В статье обсуждаются вопросы взаимосвязи токсичности сточных вод и их химического состава. Для ряда гидрохимических показателей установлена достоверная связь между показателем токсичности, определявшимся с использованием методики, где в качестве тест-организма выступает P. Caudatum.
Факт из статьи: Для ряда гидрохимических показателей не существует достоверной связи между показателем токсичности, определявшимся с использованием методики, где в качестве тест-организма выступает P. Caudatum.
Факт не подходит, потому что можно удалить "не"и факт будет верным.

Не добавяй ничего к ответу, напиши только факт. Не пиши свои рассуждения, только факт! Теперь ваша задача - записать этот факт в следующую аннотацию, точно следуя инструкциям:

Аннотация к статье: {text}
Факт из статьи:

F Prompt for Relevance and Support Level Scoring

Описание Задачи

Вы ученый, который хорошо разбирается во всех областях науки.

Задача

Задача - оценить релевантность факта аннотации и насколько аннотация подтверждает факт.

Формат Вывода

На выходе нужно написать JSON.

Пример ответа:

```
{
   "relevance": "Релевантность факта аннотации",
   "support": "Насколько аннотация подтверждает факт"
}
```

Описание Полей

Поля relevance и support могут принимать значения от 0 до 10, где 0 — не релевантно (не подтверждает факт), 10 — максимально релевантно (подтверждает факт).

Входные Данные

Текст для анализа:
{text}

G Examples of Positive and Negative claims from ruSciFact

This section provides one positive (SUPPORTS) and one negative (REFUTES) example from the ruSciFact dataset.

Example 1

Claim: Пациенты с ишемической болезнью сердца, перенесшие чрескожные коронарные вмешательства, демонстрируют улучшение толерантности к физической нагрузке при использовании компьютеризированных систем поддержки врачебных решений.

Abstract: Цель. Изучить эффективность амбулаторных реабилитационно-профилактических программ у пациентов после чрескожных коронарных вмешательств (ЧКВ) с использованием компьютеризированной системы поддержки врачебных решений (СПКР), предназначенной для выбора режима контролируемых физических тренировок (КФТ) и предоставления полноценных рекомендаций по физической активности (ФА). Материал и методы. Исследование выполняли в течение 12 мес. с включением 194 пациентов (124 мужчины и 70 женщин, средний возраст 53,5) со стабильной формой ишемической болезни сердца (ИБС), перенесших ЧКВ (коронарную ангиопластику, коронарное стентирование). При выборе режима КФТ использовалась компьютеризированная СПКР. Традиционные врачебные решения анализировали по специально разработанной анкете. Результаты. Пациенты группы КФТ, продемонстрировали достоверное увеличение толерантности к физической нагрузке (ТФН) и средней продолжительности ФН, положительную динамику качества жизни (КЖ), высокий уровень приверженности лекарственной терапии на протяжении всего периода реабилитации. При формировании врачебных решений использовали, в среднем, 3 клинических признака. Наиболее типичные врачебные ошибки носили методологический характер. Заключение. Интегрирование реабилитационных программ с использованием СПВР в амбулаторных условиях у пациентов, перенесших ЧКВ, обеспечивает высокую эффективность реабилитационно-профилактических мероприятий и безопасность ФТ.

Label: SUPPORTS

Example 2

Claim: Порода карпа не влияет на содержание сухого вещества, жира, протеина у сеголетков карпа

Abstract: В статье приведены результаты сравнения биохимического состава тела сеголетков и годовиков некоторых коллекционных пород карпа, разводимых в спу «Изобелино»: немецкого, сарбоянского, отводок изобелинского карпа (столин XVIII, три прим, смесь чешуйчатая), выращенных одновременно, в одинаковых условиях и зимовавших совместно в одном пруду. Установлены породы, характеризующиеся повышенными уровнями содержания сухого вещества, жира, протеина у сеголетков карпа. В результате исследования биохимического состава тела сеголетков карпа разной породной принадлежности, выращенных в одинаковых условиях, проявляется тенденция к увеличению содержания сухого вещества, жира и протеина у коллекционных линий карпа белорусской селекции (изобелинский) по сравнению с породами зарубежной селекции (немецкий и сарбоянский), выращенными одновременно в одинаковых условиях. У годовиков коллекционных линий белорусской селекции отмечается тенденция к увеличению содержания сухого вещества, жира и протеина, снижению содержания влаги по сравнению с зимовавшими совместно коллекционными породами зарубежной селекции. В результате исследования изменения показателей, характеризующих биохимический состав тела, произошедших за зимний период, установлено, что отклонения биохимических показателей, особенно содержания сухого вещества и жира у пород зарубежной селекции значительно выше, чем у линий изобелинского карпа (белорусская селекция). Полученные данные свидетельствуют о большей приспособленности карпа коллекционных линий белорусской селекции к условиям зимовки в Беларуси, по сравнению с импортными породами (немецким и сарбоянским).

Label: REFUTES

Upstep in Russian echo-questions about expressions

Yanko T. E.
Institute of Linguistics, Moscow, Russia
tanya_yanko@list.ru

Abstract

The paper gives an account of Russian echo questions about expressions. The prosodic upstep model is viewed as a means of expressing the illocutionary force of the echo questions about expressions. Upstep can be construed as the fundamental frequency (F_0) contour aligned with the segment material long enough to implement a gradual rise from one step of an echo question to another up to the end of the speech act. It is proposed to consider a generalized step of the upstep as a rise of the F_0 on the stressed syllable of a phonetic word followed by high post-tonics. At the same time, it is shown that the prosody of the echo question varies, in particular, it depends on the length of the segment material of an echo question. In addition, the upstep model can as well be considered in another meaning: as a function that converts the F_0 contour of a *wh*-question into the F_0 contour of the corresponding echo question. Achieving average F_0 values significantly higher than the average F_0 of any other speech act for each particular speaker is viewed as the "prosodic target" of the upstep model.

Keywords: Russian language; upstep; prosody; echo questions about expressions
DOI: 10.28995/2075-7182-2025-23-460-470

Русский upstep: просодическая модель переспроса к плану выражения

Янко Т. Е.
Институт языкознания РАН, Москва, Россия
tanya_yanko@list.ru

Аннотация

Анализируется русский вопрос-переспрос к плану выражения. В качестве средства выражения переспроса рассматривается просодическая модель upstep. Она понимается как описание тонального контура переспроса, достаточно длинного для реализации последовательного подъема от одного шага переспроса к другому по мере продвижения к концу речевого акта. Просодической моделью шага переспроса предлагается считать подъем на ударном слоге шага переспроса с высокими ровными заударными. Кроме того, модель upstep может пониматься как функция, преобразующая частоту основного тона F_0 вопроса с вопросительным словом в соответствующий переспрос. Показано, что просодия переспроса к плану выражения вариативна. В частности, она зависит от объема сегментного материала переспроса, а также от стратегии говорящего, который при артикуляции переспроса к плану выражения может использовать скороговорку. Последний фактор «сглаживает» перепады частоты шага переспроса. Инвариантом — просодической целью — модели upstep предлагается считать достижение средних значений частоты переспроса, существенно превышающих среднюю частоту других типов речевых актов для каждого конкретного говорящего.

Ключевые слова: русский язык; переспрос к плану выражения; просодия; upstep

1 Вступительные замечания. Постановка задачи

Русский вопрос-переспрос характеризуется широким спектром прагматических значений по И.М. Кобозевой [9]. Такое понимание определяется семантикой русских слов *переспрос* и *переспросить* [9], а собственно выделение набора значений определяется через анализ вопросов, способных служить прямой речью при глаголе *переспросить*. Наличие в матричном

предложении глагола *переспросить* служит и одним из рабочих методов создания корпуса для анализа русского переспроса[1]. Другой метод, который мы называем экспертным, лежит в основе приписывания речевым актам пометы «переспрос» в Национальном корпусе русского языка (НКРЯ); о корпусе см. [13]. Переспрос по НКРЯ — это речевой акт, в котором аннотаторы корпуса усмотрели один из видов переспроса. Эти два метода дают материал для анализа переспроса. Учет контекста и средств выражения переспроса — вопросительных слов и просодических показателей — позволяют разделить массив переспросов на функциональные классы.

Приведем примеры некоторых типов переспроса: 1) переспрос при нерасслышанных словах, или переспрос к плану выражения (*Что-что-что вы говорите?* (Г. Александров и др., Волга-Волга, к/ф, НКРЯ[2], 1938), 2) переспрос непонятого смысла, или переспрос к плану содержания (<— *Ты откуда знаешь счет в банке?* > — *Какой счет?* (А. Рогожкин. Операция С Новым Годом, к/ф, НКРЯ, 1996); 3) переспрос при неудачном именовании или ошибочной идентификации (<— *А одна нога, которая в аварию попала, очень сексуальная.*> — *Нога сексуальная?* (А. Рогожкин. Операция С Новым Годом, к/ф, НКРЯ, 1996); 4) уточняющий переспрос (<— *Мы сбор проводим.*> — *Пионерский сбор*? (А. Салтыков и др., Друг мой, Колька!, к/ф, 1961, НКРЯ); 5) переспрос, направленный на выражение чувств (<— *Львы. Сто львов!*> — *Сколько?!* (Г. Александров и др., Цирк, к/ф, 1936, НКРЯ); 6) переспрос, подтверждающий прием информации (<— *И я, и Устинов, и вот она уже записались.*> — *Записались, значит?* (А. Салтыков и др., Друг мой, Колька!, к/ф, 1961, НКРЯ); 7) переспрос, задаваемый при обдумывании ответа (<— *…ну где же вы возьмете эти деньги?*> —*Я?* <*Понятия не имею.*> (А. Сурикова и др., Чокнутые, к/ф, 1991, НКРЯ).

И наконец, третий метод разработки массива для анализа, который мы используем в настоящей работе, сводится к поиску в звучащем мультимедийном подкорпусе НКРЯ примеров по ключевым словам *что-что-что*, *откуда-откуда*, *чего-чего*, *что вы говорите*, *что ты сказала* и др. Этот метод позволяет выделить из множества русских переспросов один подкласс, а именно — переспросы к плану выражения, они же металингвистические переспросы; ср. термины echo questions about expressions и metalinguistic echo questions в работе Y. Sudo [14].

Переспрос к плану выражения заслуживает отдельного рассмотрения в связи со следующими обстоятельствами. 1) Переспрос к плану выражения — это крупный класс речевых актов; 2) переспрос к плану выражения, несмотря на обязательное присутствие вопросительного слова, имеет уникальную просодию, которая отличается от просодии других вопросов с вопросительным словом. Эта просодия не получила в русистике (ср. о просодии переспроса [4: 400; 9: 244]), а предположительно — и в лингвистике вообще — окончательной интерпретации.

Причины невнимания к уникальной просодии переспроса мы видим в том, что переспрос — это, как правило, короткий речевой акт с минимальным сегментным материалом (*Что*?; *А*?). Это ограничивает доступ к репрезентативным данным. Соответственно, достаточный материал для анализа может быть получен только в результате поиска представительных (достаточно длинных в сегментном отношении) данных в больших звучащих корпусах. Для русского языка такую возможность предоставляет мультимедийный подкорпус НКРЯ. Массив обследованных данных составил 336 переспросов к плану выражения, из которых в результате поиска по ключевому слову *что-что* было получено 145 примеров, по ключевому слову *чего-чего* — 41 пример, *что вы сказали* — 28 примеров, *как-как* — 22, *что ты сказала* — 19, *что ты сказал* — 8, *что ты говоришь* и *что вы говорите* — по 16 примеров, *где-где* — 8, *откуда-откуда* — 11, *сколько-сколько* — 3. Уточним, что переспросами к плану выражения служат не все результаты поиска по приведенным здесь ключевым словам; «шум» отсеивался в соответствии с интроспекцией автора. В отсутствие редупликации, вообще говоря, характерной для переспроса к плану выражения (о редупликации в переспросе см. [1]), из массива переспросов, полученных в результате имеющейся в корпусе пометы «переспрос», например *Что*? и *Чего*?, было также использовано 19 примеров.

[1] Ср. пример переспроса с глаголом *переспросить*: *Какой-какой закон? — переспросил вездесущий юнга…* (В. В. Лебедев. Приключения капитана Селёдкина, 2015, НКРЯ).

[2] Здесь и далее примеры взяты из звучащего подкорпуса МУРКО Национального корпуса русского языка (НКРЯ).

Еще одну причину для объяснения того, что в лингвистической литературе отсутствует описание просодии переспроса, мы видим в том, что переспрос с сегментной формой типа *Что*? может относиться как к классу переспросов о форме, так и к классу переспросов о содержании. Это связано с тем, что, если слушающий не разобрал слов, он не понял и содержания. Соответственно, переспрос непонимания содержания может использоваться и при нераспознанных словах. Переспрос к содержанию — это немаркированный член оппозиции «переспрос к выражению vs. переспрос к содержанию». Между тем у переспроса к выражению и у переспроса к содержанию различная просодия. Обратимся к минимальной паре:

(1) *<— вы… воспитанники школы … имени Достоевского. — ШКИД!> — Что? Что-что? <—ШКИД!>* (Г. Полока и др., Республика ШКИД, к/ф, 1966, НКРЯ*).*

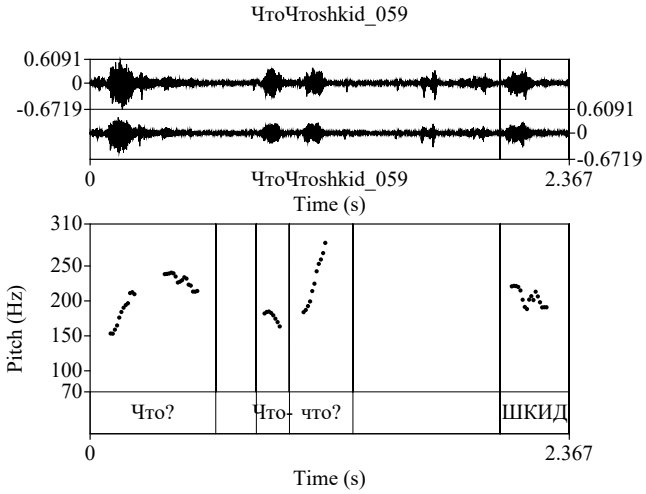


Рисунок 1: График изменения частоты (нижняя панель) в примере (1)

График изменения частоты F_0 в примере (1) говорит о том, что при переспросе *что-что*? достигается бóльшая частота, чем при первом переспросе *что*?. Эффект достижения высокой частоты усиливается т. н. отрицательным заносом (заходом частоты вниз перед повышением) на первом *что* в *что-что*. При этом первое *Что*? говорит об общем непонимании директором того, что сказал ученик, а также о замешательстве директора. Между тем второй переспрос *что-что*? возникает, когда директор, хоть и не сразу, осознает, что он не просто ничего не понял, а что он не опознал новое для него слово. И тогда он формулирует конкретный переспрос к форме.

Таким образом, переспросы с сегментной формой типа *что*? могут иметь различное интенсиональное значение и различную просодию. Это усложняет анализ просодической картины переспроса; ср. обсуждение переспроса у Е. А. Брызгуновой [4: 400], где переспрос к плану выражения (если судить по приведенному примеру[3]) получает описание в виде ИК-3, а переспрос к плану содержания (опять же если судить по примеру) — как ИК-6. Описание Е. А. Брызгуновой наиболее близко к тому, которое будет предложено ниже исходя из анализа корпусных данных.

Рассмотрим еще один пример, иллюстрирующий специфическую картину переспроса к плану выражения.

(2) *<Пожалуйста, билет…> Что-что вы сказали? <… Какой прием?>* (Л. Гайдай и др., Операция Ы и другие приключения Шурика, к/ф, 1965, НКРЯ)

[3] В [4: 400] приводится контекст, в котором на переспрос о форме указывает глагол *говорю*: *— Он живет в Московской области. —Где? —В Московской области, говорю.*

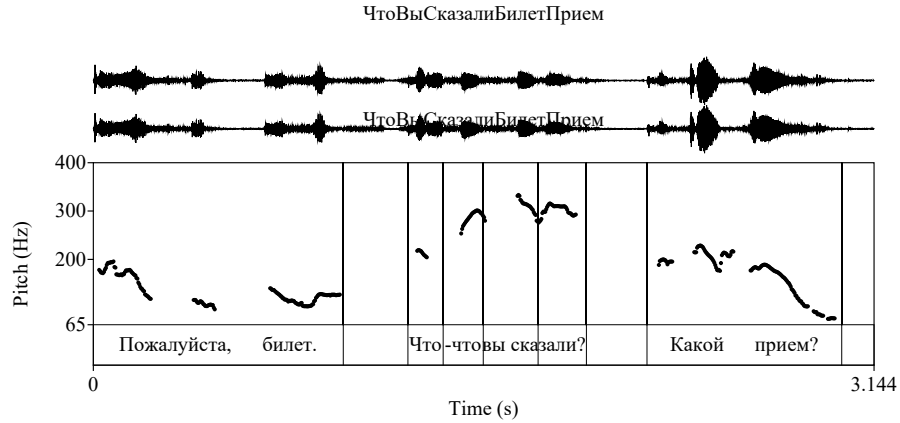


Рисунок 2: График изменения частоты в примере (2)

Перед нами сцена на экзамене. Средняя частота основного тона F_0, с которой преподаватель произносит переспрос *Что вы сказали*?, резко контрастирует со средней частотой двух других его высказываний: предложения взять билет и вопроса непонимания того, что имел в виду студент, настраивающий спрятанную рацию на прием информации от подсказывающего. Средняя частота переспроса к плану выражения примерно в два раза превышает среднюю частоту контекста.

Примеры (1) и (2) заставляют предположить, что просодическая стратегия переспроса к плану выражения состоит в достижении частоты, значимо превышающей среднюю частоту произнесения речевых актов других типов. Отметим также еще два важных свойства вопроса к плану выражения. Вопрос к плану выражения реализуется на сегментном материале вопроса с вопросительным словом и в переспросе с редупликацией вопросительного слова ударение приходится на последний редуплицированный член: *что-что-чт**о***.

Обратимся к механизмам, позволяющим в переспросе достичь высокой частоты. Предварительный анализ говорит о том, что просодические стратегии переспроса к плану выражения зависимы от структуры и объема сегментного материала. В частности, в длинных переспросах к плану выражения русским языком используется т. н. модель upstep Дж. Пьерхамберт [12: 74], при которой максимальная частота основного тона F_0 фонетического слова под номером n ниже минимальной F_0 фонетического слова под номером n+1 в линейной цепочке предложения.

Между тем понятие upstep используется в лингвистике в двух различных значениях. Первое представлено определением из диссертации [12], однако в применении к тоновым языкам upstep понимается как тональное преобразование, применяемое к языковым единицам, которые считаются исходными, и форму и значение которых upstep меняет. В этом случае upstep — это не просодия речевого акта в целом, а преобразование, которое применяется к тональной структуре исходной словоформы или другой единицы, частоту F_0 которой upstep повышает. При таком подходе upstep — это функция, подразумевающая существование некоего аргумента, выделение которого при анализе — это отдельная задача.

Раздел 2 ниже посвящен просодии переспроса при минимуме сегментного материала: *что*?; *чего*?. В разделе 3 — рассматривается длинный переспрос, например, *Что-что-что-чт**о** вы говорите, чуткое руководство*? (Г. Александров и др., Волга-Волга, к/ф, НКРЯ[4], 1938), и демонстрируется, что переспрос к плану выражения реализует upstep в понимании Дж. Пьерхамберт [12: 74]. Здесь же исследуется структура отдельного шага переспроса и пределы его вариативности. В разделе 4 upstep анализируется как функция, в которой в роли аргумента рассматривается вопрос с вопросительным словом. В заключении подводятся итоги анализа переспроса к плану выражения в русском языке.

Для верификации слуховых гипотез мы пользуемся системой анализа звучащей речи Praat [2].

[4] Здесь и далее примеры взяты из звучащего подкорпуса МУРКО Национального корпуса русского языка.

2 Просодия переспроса при минимуме сегментного материала

В переспросе типа *что-что* без заударных слогов реализуется одна из двух стратегий достижения высокой частоты. Это либо подъем тона с крутым набором частоты, как в примере (1) (второе вхождение словоформы *что*), либо немедленный относительно ровный высокий тон, как в примере (3):

(3) *Что-что-что? <Я умней?>* (М. Захаров, и др., Дом, который построил Свифт, к/ф, 1982, НКРЯ).

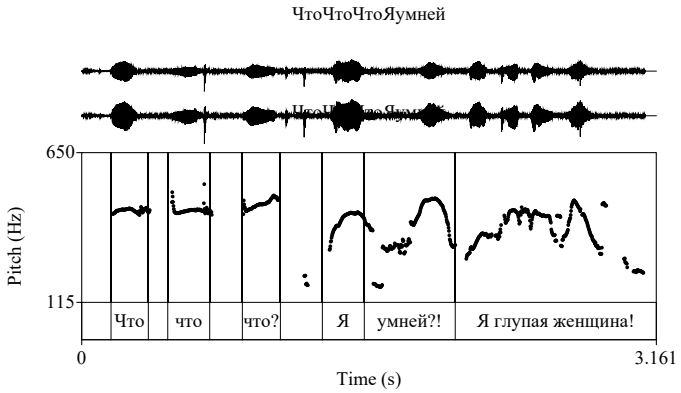


Рисунок 3: График изменения частоты в примере (3)

В переспросе же с наличествующими заударными слогами мы наблюдаем три вариации: 3) подъем на ударном слоге плюс высокий ровный или слабо нисходящий заударный слог по модели ИК-6 (Брызгунова [1: 98]), как в примере (4) (ср. также реализацию этой модели на высокой частоте в примере (2)); 4) подъем на ударном слоге плюс продолжающийся подъем частоты на заударном слоге, как в примере (5); 5) ровный тон на повышенной частоте, как в (6).

(4) *<— В пять тридцать устраивает?> — Во ско**о**лько?* (С. Герасимов. Журналист, к/ф, 1967, НКРЯ).

Рисунок 4: График изменения частоты в примере (4)

(5) *<— Ну тыща тридцать.> — Ско**о**лько?* (А. Райкин и др., Люди и манекены, к/ф, 1974, НКРЯ).

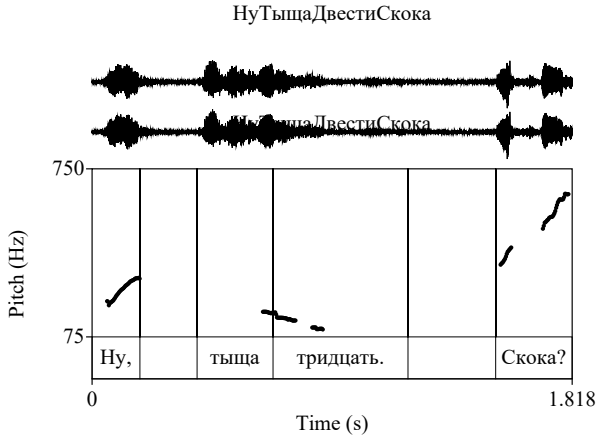


Рисунок 5: График изменения частоты в примере (5)

(6) <— *Ну и сколько?*> — *Сколько-ск**о**лько*? (Ф. Янковский, Г. Островский. В движении, к/ф, 2002, НКРЯ).

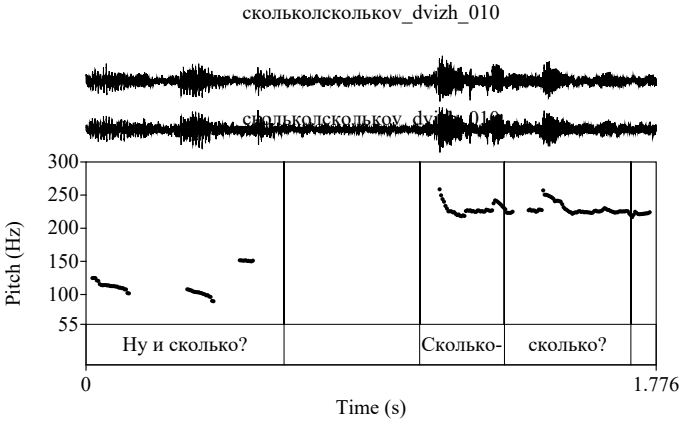


Рисунок 6: График изменения частоты в примере (6)

В примере (6) оба вопроса произнесены одним и тем же говорящим, что иллюстрирует различие частотного контура при простом вопросе и при переспросе к плану выражения. Средняя частота ударных слогов между первым и третьим вхождением *сколько* различается более, чем в два раза.

Если воспользоваться схематическим представлением изменения частоты в фонетическом слове, где тонкая линия символизирует безударный слог (заударный или предударный), а жирная — ударный, по аналогии с представлением, принятым в [7], то примеры (1)-(6) можно представить Таблицей 1.

Способы набора частоты в коротких переспросах				
1	2	3	4	5
*Что-что**о**?*	*Что-что-что**о**?*	*Во ск**о**лько?*	*Ск**о**лько?*	*(Сколько)-ск**о**лько?*
LL+H*%[5]	HH*+H%	L+H*-H%	L+*H-^H%[6]	H*+H-H%
Пример (1)	Пример (3)	Примеры (2) и (4)	Пример (5)	Пример (6)

Таблица 1: Схематическое представление способов набора частоты в коротких переспросах

Наша гипотеза состоит в том, что инвариантом короткого переспроса служит высокий тон H*-H%. Пределы вариативности при реализации короткого переспроса отражены в Таблице 1.

3 Просодия переспроса при репрезентативном сегментном материале. Длинный переспрос

Обратимся к переспросам с представительным сегментным материалом. Предварительно уточним терминологию. Традиционный для исторической акцентологии термин «фонетическое слово» (полноударное слово с присоединившимися к нему клитиками, см., например [17: 9]) мы предлагаем заменить здесь на термин «шаг переспроса», понимая под шагом словоформу, которая несет коммуникативно релевантный акцент, плюс присоединившиеся к ней словоформы, не несущие коммуникативно релевантного акцента. Так, в переспросе *Сколько-ск**о**лько*? (пример (6)) фонетических слов два, а коммуникативно релевантный акцент — только один: на втором *сколько*. Соответственно в этом переспросе только один шаг.

Обратимся к примеру (7).

(7) *Куд**а** уехал? Куд**а** ты, Гена, у**е**хал?* (Н. Михалков, В. Мережко. Родня, к/ф, 1981, НКРЯ).

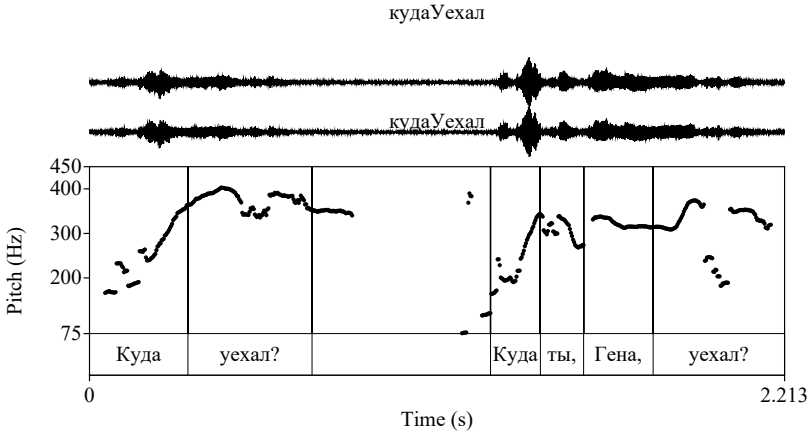


Рисунок 7: График изменения частоты в примере (7)

[5] Поясним, что в т. н. автосегментной нотации приняты следующие обозначения: L — низкий тон, H — высокий тон, L+H — восходящий тон, звездочка (*) — ударный слог, L+H-H — высокий тон сохраняется на заударном слоге, L+H* — подъем с ранним таймингом, L*+H — подъем с поздним таймингом, % — конец фонетического слова. В настоящее время просодические описания разнообразных языков используют именно эту нотацию с необходимыми дополнениями, см. пионерскую диссертацию [12].

[6] Знак «крышечка» (^) введен в [8] для описания немецкого акцента, основной характеристикой которого служит последовательное повышение F_0 на заударных слогах, начиная от точки, достигнутой в результате подъема на ударном слоге, т.е. так, как в примере (5) на *сколько*.

В примере (7) два переспроса: *Куда уехал?* и *Куда ты, Гена, уехал?*. Они задаются в ответ на телефонное заявление некоего коварного Гены, состоящее в том, что он не Гена и что Гена якобы уехал. Первый переспрос имеет подъем на ударном слоге словоформы *куда*. На словоформе *уехал* сохраняется высокий ровный тон по модели заударных в ИК-6. Мы предлагаем считать, что этот переспрос совершается в один шаг.

Второй переспрос имеет один подъем на *куда* и второй — на *уехал*. Соответственно, переспрос *Куда ты, Гена, уехал?* совершается в два шага, причем второй шаг, сохраняя форму кривой изменения частоты на первом шаге, мажорирует соответствующие значения первого шага. Можно также заметить существенно менее рельефные, чем первый и второй шаг, зачатки третьего шага на словоформе *Гена*. Вопрос о том, считать ли подъем на ударном слоге словоформы *Гена* с последующим ровным тоном на заударном слоге отдельным шагом переспроса, мы здесь оставляем открытым. Мы считаем, что вокатив *Гена*, может произноситься как с полноценным акцентом вокатива и формировать шаг переспроса, так и, находясь в срединной позиции и образуя с местоимением *ты* цепочку коммуникативно безударных словоформ, безакцентно.

Как указано в [4: 400], переспрос к плану выражения может произноситься скороговоркой. Наш анализ подтверждает это наблюдение. При этом скороговорка делает форму шага менее рельефной, сохраняя только общее направление значений частоты вверх на ударных слогах. Таким образом, в длинных речевых актах четкое выделение шагов переспроса прослеживается не всегда, обратимся, однако, к примеру (8).

(8) *Что-что-что-что**о**, вы говорите, стихийной*? (Г. Александров и др., Волга-Волга, к/ф, НКРЯ, 1938)

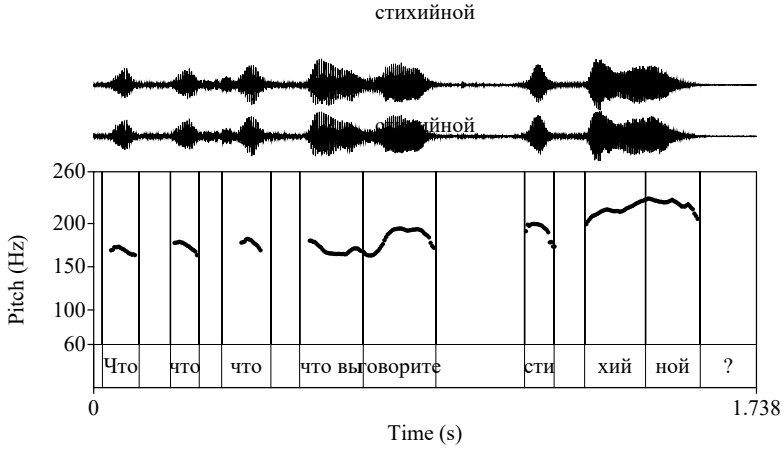


Рисунок 8: График изменения частоты в примере (8)

В (8) наблюдается последовательное возрастание частоты на редуплицированном *что* (ср. также пример (3)). *Что-что-что-что**о*** с акцентом на последнем вхождении *что* — это первый шаг переспроса. Далее, на клитике *вы* частота тона снижается, чтобы начать расти на акцентоносителе второго шага словоформе *говорите*. Третий шаг — реализуется на словоформе *стихийной* в виде подъема на ударном слоге с последующим высоким заударным слогом.

Итак, в длинных переспросах наблюдается последовательное возрастание частоты F_0 к концу речевого акта. Возрастание происходит пошагово, причем каждый шаг в зависимости от структуры сегментного материала реализует одну из фигур, зафиксированных в Таблице 1. При достаточном — максимальном — сегментном материале реализуется форма частотной кривой по модели ИК-6. В «образцовом» (т.е. без «сглаживания» значимых движений F_0) переспросе *Куда ты, Гена, уехал*? (пример (7)) артикулируется как минимум два развернутых шага переспроса в форме кривой интонационной конструкции ИК-6 каждый шаг. Как вариант реализации модели upstep встречается частотная кривая, сохраняющая высокий ровный уровень частоты к концу предложения, ср. пример (11) ниже.

В просодическую структуру шага переспроса включаются не имеющие собственного коммуникативно релевантного акцента фрагменты: безакцентные обращения и парентезы.

Пример (9) ниже демонстрирует «сглаженный» переход от одного шага к другому:

(9) *Что вы сказали, синьор*? *Что потайная*? (Л. Нечаев, И. Веткина. Приключения Буратино, к/ф, 1975, НКРЯ)

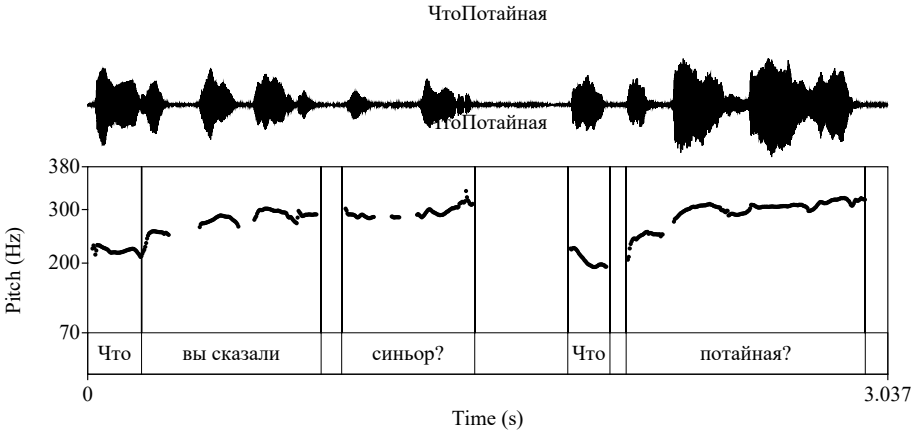


Рисунок 9: График изменения частоты в примере (9)

По поводу первого шага переспроса заметим, что при общей тенденции частоты переспроса к подъему движение на теоретически ровном высоком уровне при реализации на минимальном сегментном материале может нести и известное падение частоты: сравним движение F_0 на начальных *что* в примере (9), содержащем два переспроса. Аналогичная картина наблюдается также в примерах (6) и (8).

Итак, в русском переспросе к плану выражения, имеющем развернутый сегментный материал, представлена просодическая стратегия upstep: частота изменения основного тона последовательно повышается от одного шага к другому. В зависимости от сегментного материала каждого шага цельный контур переспроса может иметь различную тональную структуру. Для переспроса, иллюстрированного «образцовым» примером (7), мы предлагаем использовать следующую нотацию: L+¡H*-H%, где последовательное повышение тона от одного шага к другому в соответствии с автосегментной традицией обозначается перевернутым восклицательным знаком (¡). Обобщающую нотацию для различных вариантов переспроса, где каждый шаг может иметь иную, в частности, более «сглаженную» форму, чем прототипическая форма шага в (7), мы предлагаем формулировать как ¡H*-H%.

4 Upstep как акцентное преобразование

До настоящего момента стратегия upstep понималась здесь как способ достижения в предложении высокой частоты, что в свою очередь, рассматривалось в качестве средства выражения иллокутивной силы переспроса к плану выражения. Между тем в применении к языкам, где диагностируется upstep, в частности, к тоновым (ср., например, [5], язык нджем, группа банту, Камерун), но не только к ним (ср. работу [6] о кастильском испанском), upstep понимается как просодическое преобразование, повышающее частоту исходной единицы и служащее средством выражения дополнительного значения. В упомянутых здесь работах в качестве такого значения выступают определенные типы контраста, судя по описаниям, различные. При таком подходе областью upstep необязательно служит предложение в целом; upstep может выделяться и на уровне тональной структуры словоформы.

Подойдем к русскому материалу с этой же стороны и рассмотрим upstep как процедуру преобразования в речевой акт переспроса некоторой исходной структуры[7]. Исходной структурой, к которой применяется upstep, мы предлагаем считать соответствующий переспросу вопрос с вопросительным словом. Рассмотрим пару «вопрос с вопросительным словом (10) и переспрос (11)».

[7] О линейно-акцентных преобразованиях коммуникативных структур и о понятии коммуникативной парадигмы, в которой выделяется один (иногда два) — исходных члена, а другие члены парадигмы служат результатом линейно-акцентных преобразований, примененных к исходному члену, см. работы И. И. Ковтуновой [10] и Е. В. Падучевой [11].

(10) *Что вы сказ**а**ли*? (А. Коренев, Г. Садовников. Большая перемена, к/ф, 1973, НКРЯ).

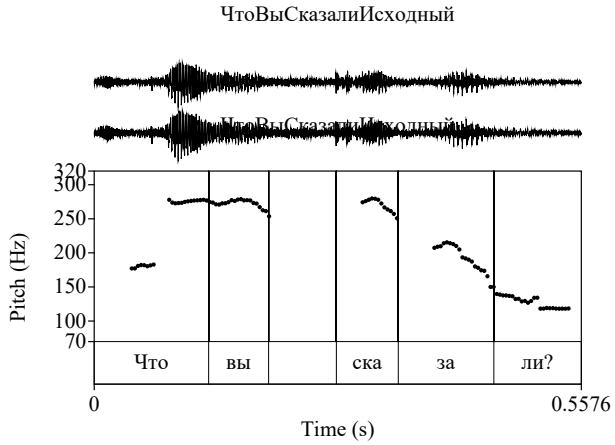


Рисунок 10: График изменения частоты в примере (10)

Прототипический, или т. н. нейтральный (т.е. такой, чья линейно-акцентная структура вносит минимальный вклад в семантическую структуру), вопрос с вопросительным словом в соответствии со своей просодической моделью [16] имеет подъем типа ИК-6 на начальном вопросительном слове и падение типа ИК-1 (или ИК-2) в исходе речевого акта. Эту модель иллюстрирует пример (10). Обратимся к переспросу, который мы предлагаем считать результатом преобразования upstep, примененного к вопросу типа (10).

(11) *Что вы сказ**а**ли*? (Н. Михалков, Р. Ибрагимбеков. Сибирский цирюльник, к/ф, 1998, НКРЯ)

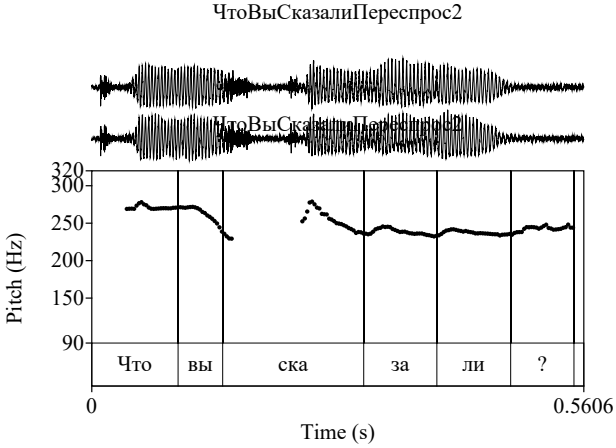


Рисунок 11: График изменения частоты в примере (11)

Сопоставление просодических структур (10) и (11) позволяет считать просодию переспроса результатом повышения значений частоты F_0 в соответствующих переспросу элементах вопроса с вопросительным словом. Отсчет значений, которые подвергаются повышению в результирующем предложении, мы предлагаем начинать со второго шага как исходного вопроса, так и соответствующего переспроса: начальный шаг вопросительного слова мы понимаем здесь как нулевой.

В настоящем подразделе понятие *upstep* рассматривается как функция, повышающая частоту вопроса с вопросительным словом в качестве исходного аргумента. В соответствии с гипотезой последовательное повышение частоты преобразует вопрос с вопросительным словом в соответствующий переспрос к плану выражения.

5 Заключение

Рассмотрен русский вопрос-переспрос к плану выражения. В качестве средства выражения переспроса предлагается рассматривать просодическую модель upstep, которая понимается как описание тонального контура переспроса, достаточно длинного для реализации последовательного подъема от одного шага переспроса к другому по мере продвижения к концу речевого акта. Моделью прототипического шага предлагается считать подъем на ударном слоге переспроса с высокими ровными заударными. Кроме того, модель upstep может также рассматриваться как функция, преобразующая F_0 вопроса с вопросительным словом в F_0 соответствующего переспроса. Показано, что просодия переспроса к плану выражения достаточно вариативна, в частности, она зависит от объема сегментного материала переспроса и стратегии говорящего, который при артикуляции переспроса к плану выражения может использовать скороговорку. Просодической целью модели upstep предлагается считать достижение средних значений частоты переспроса, существенно превышающих среднюю частоту других типов речевых актов для каждого конкретного говорящего.

References

[1] Apresjan V. Ju. Iomdin B. L. Russian interrogative pronouns as a lexicographic type. Russian Journal of Linguistics 26 (4). 2022. 1078–1113.

[2] Boersma P., Weenink D. Praat: Doing phonetics by computer. Version 6.4.04. Online: Praat: doing Phonetics by Computer (uva.nl), 2024 (accessed date: 15.11.2024).

[3] Bryzgunova E. A. Intonation [Intonatsija], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, 1982. P. 103-118.

[4] Bryzgunova E.A. Means of expressing the unknown in questions (interaction of vocabulary, context, and intonation) [Sredstva vyrazhenija neizvestnogo v voprose (vzaimodejstvie leksiki, konteksta i intonatsii)], Russian Grammar [Russkaja grammatika]. Vol. 2, Nauka, Moscow, 1982. P. 397-402.

[5] Chumbow B. S., Akumbu P. W. Upstep of low tone in Njyem and Register Tier Theory // Journal of West African Languages, 2007, 34 (1), pp.39-53.

[6] Fliessbach, J. Mirativity, obviousness, and reversal as instances of contrast on different levels of meaning: Evidence from Spanish intonation // Brysbaert J., K. Lahousse K., eds. On the Role of Contrast in Information Structure. Trends in Linguistics. Vol. 382. Berlin; Boston, 2024, pp. 101—130.

[7] Grønnum N. The Groundworks of Danish intonation. An introduction. 1992.

[8] Grice M., Baumann S., Benzmüller R. German Intonation in Autosegmental-Metrical Phonology // Sun-Ah Jun (ed.), Prosodic Typology. The Phonology of Intonation and Phrasing. Oxford: Oxford University Press. 2005. 55-83.

[9] Kobozeva, I.M. «Perespros» (≈ echo-question) as a peripheral subcategory of a communicative-grammatical category of interrogativity in Russian: semantics and formal properties [Perespros kak periferija kommunikativno-grammaticheskogo polja voprositel'nosti: semantika i sredstva vyrazhenija v russkom jazyke] // Voejkova M.D., Kazakovskaja V.V., eds. Problems of functional grammar. Reference to the speaker [Problemy funkcional'noj grammatiki. Otnoshenie k govorjashhemu v semantike grammaticheskih kategorij]. Moscow. 2020. Pp. 223—245.

[10] Kovtunova I. I. Modern Russian. Word order and theme-rheme division of a sentence [Sovremennyj russkij jazyk. Porjadok slov i aktual'noe chlenenije predlozhenija]. Prosveshchenije, Moscow, 1976. Pp. 132-194.

[11] Paducheva E.V. (1984) Communicative sentence structure and the concept of a communicative paradigm [Kommunikativnaja struktura predlozhenija i ponjatie kommunikativnoj paradigm], Scientific and technical information [Nauchno-tehnicheskaja informacija]. Ser. 2. N10. P. 25-31.

[12] Pierrehumbert J. B. The phonology and phonetics of English intonation (manuscript) Massachusetts Institute of Technology. 1980. Online: Pierrehumbert_PhD.pdf (ox.ac.uk).

[13] Savchuk, S.O., Arhangel'skij, T.A., Bonch-Osmolovskaja, A.A., Donina, O.V., Kuznecova, Ju.N., Ljashevskaja, O.N., Orehov, B.V. and Podrjadchikova, M.V. (2024) Russian National Corpus 2.0: New opportunities and development prospects [Nacional'nyj korpus russkogo jazyka 2.0: novye vozmozhnosti i perspektivy razvitija.]. Voprosy jazykoznanija [Topics in the Study of Language], 2. Pp. 7–34 (in Russ.).

[14] Sudo Y. Metalinguistic semantics for echo questions (manuscript). Massachusetts Institute of Technology. 2011. Online: Metalinguistic Semantics for Echo Questions (studylib.net)

[15] Yanko T. E. (2023) The prosody of the Russian question [Prosodicheskaja model' rechevogo akta voprosa] Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference 'Dialogue 2023' [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy Mezhdunarodnoj konferencii 'Dialog 2023']. Vyp. 22 (29), M., pp. 566–578.

[16] Zaliznjak A.A. Old Russian accent system. General introduction and dictionary. [Drevnerusskoe udarenie. Obshhie svedenija i slovar']. M. 2019.

Operator words versus MT systems and LLMs

Zimmerling A. V.
Pushkin State Russian Language Institute;
Institute of Linguistics, Russian Academy of Science

Baiuk A. M.
Russian State University
for the Humanities

Abstract

This paper investigates the meanings of two classes of Russian discourse words, defined by modal operators VER and AFF, and examines their translation equivalents in nine target languages using machine translation (MT) systems, large language models (LLMs), and human expert translations. VER (verification) indicates confirmation of a hypothesis, while AFF (affirmation) expresses a strengthened belief in a hypothesis. The study uses a set of 17 Russian discourse words and evaluates their translations in English, German, Danish, Swedish, Icelandic, Ukrainian, Bulgarian, Ossetic, and Arabic across 85 test sentences. The primary goal was to test the universality of the VER and AFF distinction, hypothesizing that these classes remain distinct across languages despite the lack of direct one-to-one translation equivalents. The study assumed that VER and AFF operators, corresponding to DE RE and DE DICTO attitudes respectively, differ fundamentally in semantics and distributional behavior. The study confirms the semantic and distributional independence of VER and AFF operators, supporting their universality. LLMs, despite not being specialized for MT tasks, showed remarkable adaptability and context awareness compared to traditional MT systems. The findings highlight the potential of LLMs in nuanced translation tasks and underscore the complexity of translating modal discourse words. Future work will explore custom models and further refine evaluation metrics for translation accuracy.

Keywords: discourse words, operators, machine translation systems, large language models
DOI: 10.28995/2075-7182-2025-23-471-480

Модальные операторы в автоматических системах перевода и больших языковых моделях

Циммерлинг А. В.
Государственный институт русского языка имени А. С. Пушкина; Институт языкознания РАН

Баюк А. М.
Российский государственный гуманитарный университет

Аннотация

Обсуждаются значения двух классов русских дискурсивных слов, определяемых модальными операторами VER и AFF, и рассматриваются их переводные эквиваленты на девяти целевых языках с использованием машинных переводчиков, больших языковых моделей и переводов, выполненных экспертами. Оператор VER (верификация) обозначает подтверждение гипотезы, в то время как AFF (аффирмация) выражает усиленную уверенность в гипотезе. В исследовании использован набор из 17 русских дискурсивных слов и оцениваются их переводы на английский, немецкий, датский, шведский, исландский, украинский, болгарский, осетинский и арабский языки на основе 85 тестовых предложений. Основной целью было проверить универсальность различия между VER и AFF на основе гипотезы о том, что различение этих классов сохраняется во всех языках, несмотря на отсутствие пословных эквивалентов. Предполагалось, что операторы VER и AFF, соответствующие отношениям DE RE и DE DICTO, фундаментально отличаются по семантике и дистрибутивным свойствам. Исследование подтверждает семантическую и дистрибутивную специфику операторов VER и AFF. Большие языковые модели продемонстрировали большую адаптивность и способность учитывать контекст по сравнению с традиционными системами машинного перевода. Результаты подчеркивают потенциал больших языковых моделей в решении задач перевода и акцентируют внимание на сложности перевода модальных дискурсивных слов. В будущих исследованиях планируется разработка новых моделей и доработка метрик оценки точности перевода.

Ключевые слова: дискурсивные слова, модальные операторы, машинный перевод, большие языковые модели

1 Introductory remarks

In this paper, we check the meanings of two groups of Russian discourse words and their translation equivalents provided by MT systems, LLMs and human experts. We selected 17 Russian discourse words and multiword expressions presumably containing two kinds of modal operators VER and AFF. Their values are defined as follows:

(i) VER (*p*): *X* confirms the hypothesis *p* about the outer world.

(ii) AFF (*p*): *X* states that his belief that *p* is strengthened.

The scenario VER entails that the speaker considered the hypotheses *p* and ~ *p* and verified *p* at the moment *t* [33]. The scenario AFF entails that the speaker was originally biased towards *p* at the moment *t-1* and is still biased towards *p* at the moment *t*, i.e. the moment of speech. Russian has a large amount of discourse words that have been discussed in detail [3; 10; 16], cf. also [24; 18: 122; 35] on the language-specific discourse word *pravda*, [11] on *razumeetsya*, *konechno*, *samo soboj*, *estestvenno* and [23] on the particle *taki*. For this study, we assumed that Russian has two classes of VER and AFF words that can have external correspondences in other languages. Class 1 contains 7 standard elements expressing VER. 6 of them always take over the phrasal accent, while *taki* is an enclitic.

(iii) Class 1 words: VER (*p*): *dejstvitel'no*, *real'no*, *(i) pravda*, *na samom dele*, *v samom dele*, *taki, i vpryam'*.

Class 2 contains 9 standard elements expressing AFF.

(iv) Class 2 words: AFF (*p*): *razumeetsya*, *konechno*, *samo soboj*, *estestvenno*, *ponyatno*, *yavno, ya tak i znal, tochno, opredelenno.*

The 17th item, Rus. *konkretno*, is a recent colloquial word that presumably patterns with Class 1. We however left the final verdict to the translators and MT systems depending on their ability to reconstruct VER or AFF equivalents in the target languages.

For the translation task, we offered 5 short Russian sentences containing a slot for a VER | AFF operator:

Fig. 1. Stimuli sentences (Russian)

1) Vasya VER | AFF durak 'Vasya is VER | AFF a fool'.
2) Vasya VER | AFF ne pridet 'Vasya VER | AFF won't come'.
3) Vasya VER | AFF ne prishel 'Vasya VER | AFF didn't come'
4) Vasya VER | AFF oshibsya 'Vasya VER | AFF made a mistake'.
5) Vasya VER | AFF ne sobiralsya prixodit' 'Vasya VER | AFF wasn't going to come'.

The number of the stimuli totals 17 x 5 = 85. The stimuli were translated into 9 languages — English, German, Danish, Swedish, Icelandic, Ukrainian, Bulgarian, Ossetic and Arabic — by 1) the human expert[1]; 2) MT systems (Google, Yandex and the pre-trained Google model); 3) LLMs (ChatGPT 4o and Gemini 1.5). The experts produced 765 target sentences (85 x 9) altogether. The same amount of target sentences was produced by each MT system and LLM. In addition, we checked the responses of MT systems and LLMs by giving them both Russian stimuli sentences with the bare operator word $pravda^{VER}$ and with the added proclitic *i* (*i* $pravda^{VER}$). The experts took the equivalence $pravda^{VER} \cong i\ pravda^{VER}$ for granted, but the MT systems and LLM reacted to these stimuli differently, so we got extra 5 target sentences from each MT system and LLM.

[1] The group of experts (6 women and 3 men, from 30 to 88 years) included one bilingual person and two L1 speakers of the target language with a near-native level of Russian. The rest were L1 speakers of Russian with a near-native or highly proficient level of the target L2 language. All experts but one had a linguistic background, were engaged in the teaching of foreign languages and had a translation experience.

The goal of our study is to check the hypothesis the distinction of Class 1 words containing VER versus Class 2 words containing AFF is universal despite possible asymmetries in the number of Class 1 and Class 2 elements across the world's languages and the lack of exact word-to-word equivalents of the tested Russian discourse words[2]. The translation was valued as semantically correct, if the source sentence containing VER or AFF was rendered by the target sentence with the same operator. The zero hypothesis was that Class 1 and Class 2 do not intersect in any language, so that no discourse word that is a standard part of the Class 1 lexicon patterns with to Class 2 in any of its uses, and vice versa.

2 Basic linguistic intuitions

Verification and affirmation are different meanings, despite they are occasionally conflated in the description of modals, cf. [21: 81; 19: 78; 2: 33] and impressionistic labels like 'enhanced indicativity' used for verification markers in [29: 299]. Informally, verification of *p* does not imply that *X* was originally biased towards *p,* and the fact that *X* preserved his belief that *p*, does not guarantee that *p* is true. More specific arguments for treating VER and AFF as independent operators come from the distribution of verification and affirmation markers in the world's languages. The meaning of verification can be encoded by the intonation [28: 82; 17], moreover, verification words normally get the phrasal accent, cf. *dejstvitel'no* 'really', *pravda* 'true that' that are always accented in Russian [27], while affirmative markers like Rus. *estestvenno* 'naturally', *razumeetsya* 'certainly' can be both accented and deaccented [36]. Further linguistic tests involve the asymmetry of VER and AFF words in the so-called indirect contexts, i.e. utterances about the beliefs of other people. Since VER words freely apply to all possible worlds given that *p* is true in some accessible possible world *W*, cf. *If X* ***really*** *were a good boy, he would buy his girl a gift*, they are licensed both in the counterfactual and in the indirect contexts, cf. (1a-b).

(1) Russian

a. Esli by Vasya **dejstvitel'no** opozdal, ego by oshtrafovali.

'If Vasya **really** came late, they would fine him.'

b. Petya (**dejstvitel'no**) dumaet, chto Vasya **dejstvitel'no** opozdal.

'Pete (**indeed**) believes that Vasya **indeed** came late.'

VER words normally combine with the irrealis markers and can be used in conditional clauses in the subjunctive mood, cf. [33] on the Old Russian verification particle TI_1 'indeed'. AFF words however do not combine with the irrealis markers and are blocked in the indirect contexts, since the speaker cannot project his own belief that *p* to other people, cf. (2a-b)[3].

[2] For modal particles in some of the target languages see [1; 15; 2].

[3] An anonymous reviewer points out that AFF items can be analyzed as embedded to a different depth, cf. Krifka's and Frey's accounts for German and English [14; 7]. This holds for the Russian AFF words as well, cf. the context provided by the reviewer, where *konechno* 'certainly' cannot be replaced by *estestvenno* 'naturally': *Ya*^{EGO} *eshche dumal, chto, konechno*^{EGO}*, ego i menya ravnyat' glupo* 'I^{EGO} thought that it was, of course^{EGO}, stupid to compare him and me', $^{??}$*dumal, chto, estestvenno*^{S}*, ego i menya ravnyat' glupo*. This contrast can prima facie be explained by the fact that *konechno* is an egocentric word projecting the point of view of the speaker / propositional subject [22: 311], while *estestvenno* and *naturally* introduce an assessment ('judgment', in Krifka's terms) shared by the speaker and the addressee. However, the classification of AFF words is orthogonal for our paper: we do not claim that Class 2 items containing AFF are absolute synonyms and replace each other in all root and embedded contexts. Neither do we claim this for Class 1 items containing VER. What we check, is the ability of translators to extract the operators VER and AFF from the corresponding word classes correctly.

(2) Russian

a. *Esli by Vasya **estestvenno** opozdal, ego by oshtrafovali.

Int. ‘If Vasya **naturally** came late, they would fine him.’

b. *Petya (**estestvenno**) dumaet, chto Vasya **estestvenno** opozdal.

Int. ‘Pete (**naturally)** believes that Vasya **naturally** came late.’

Class 1 items containing VER can be straightforwardly identified with *de re* modals, i.e. the attitude to describe the real or accessible possible world as it is. Class 2 items containing AFF can arguably be linked with *de dicto* modals, i.e. the attitude to interpret the text <on the model set of possible worlds> in the form as it is expressed. The proof of these claims is beyond the reach of this paper. We just assume here that the distinction of Class 1 versus Class 2 holds on the level of logical semantics and is universal[4]. In the following, we conventionally label Class 1 items ‘DE RE words’ and Class 2 items ‘DE DICTO words’[5]. Most Russian DE RE and DE DICTO words are homonymic to standard adverbs, nouns, cf. *pravda* ‘truth’, adjectives, cf. *ponyatno* ‘it is clear’ and even clauses, cf. *Ya tak i znal* ‘I knew it’. Therefore, the translation of DE RE and DE DICTO words amounts to a triple task: 1) recognizing the VER | AFF insertion as a discourse item in the position where it occurs in the source sentence; 2) extracting the correct operator from the discourse item; 3) providing a suitable equivalent from the set of DE RE or DE DICTO words in the target language. This is schematically shown in (v); the symbol W^{VER} stands for the set of all discourse elements containing VER, ‘SL’ stands for the source language, and ‘TL’ for the target language.

(v) $w_i \in \{_{SL}\, w_1, w_2 \ldots w_n\}^{DE\,RE} \Rightarrow W^{VER} \Rightarrow w_j \in \{_{TL}\, w_1, w_2 \ldots w_n\}^{DE\,RE}$

3 Linguistic models and the design of the experiment

In this section, we comment on the chosen linguistic model and its impact on the design of our experiment. We assume that verification is a universal meaning that is encoded lexically by the words containing VER, by prosody, or by both cues [27]. We also adopt the hypothesis that Russian, which is the source language in the experiment, has a class of VER words that have a different distribution compared to AFF words [36]. Provided that the meaning of VER is universal, this gives a ground to check whether the contrast between VER and AFF words is preserved in translation. Since operator semantics represents a more abstract layer of meaning than contextually-determined uses, our approach is complementary with the previous descriptions of Russian discourse words [3; 10; 11; 16] that focus on their cognitive profiles. Krifka’s layered model of speech acts distinguishes two kinds of abstract objects called ‘commitments’ and ‘judgments’ and identifies them with syntactic projections [13; 14]. This model describes the semantics-to-syntax interface but leaves no space for VER words[6] and does not suit our experiment.

An anonymous reviewer argues that artificial stimuli are not really valid as MT systems and LLMs have been pre-trained on rich sentences, but our stimuli are grammatically correct, and their shortness shouldn’t affect anything, besides, training data can’t contain only long (rich) ones. On the other side, prototype sentences with similar contexts may allow to evaluate the ability of neural models to distinguish the operator words themselves. The linguistic problem is that the set of our short lab sentences invites the translator to reconstruct the contexts, where these sentences are appropriate:

[4] The DE RE / DE DICTO distinction can be interpreted differently in logical studies. The analysis of verification predicates as DE RE modals is close to the approach dubbed ‘metaphysical’ by Michael Nelson [20].

[5] An anonymous reviewer objects that our categories “DE RE words” and “DE DICTO words” duplicate the labels VER and AFF. We prefer to withhold the double notation since the DE RE / DE DICTO terminology applies to sentences and propositional attitudes, while VER and AFF labels can apply to the sentence components and to the storage of operator words in the lexicon.

[6] This issue is acknowledged by Krifka himself, who admits that German VER words *tatsächlich* and *in der Tat* ‘indeed’, ‘in fact’ are not commitment modifiers but elements invoking ‘the contrast between the proposition and its negation, similar to cases of verum focus’ [14: 146].

otherwise a considerable part of the stimuli is redundant. Another anonymous reviewer suggests that we drop our 17 'lab stumps' and import a much richer collection of genuine examples from the parallel corpora. This is unfortunately not realistic since the diversity of the reconstructed contexts for our lab stimuli with 17 discourse words far exceeds the limits of the parallel corpora even for well-documented languages. The same reviewer asks what MT and LLM contribute compared to human translation in our experiment. To begin with, neural models are based on statistical language modeling and pre-trained on huge amounts of natural language data. Therefore, they may emulate human perception of operator words and cross-lingual connections between them. Finally, the quality of the MT and LLM translation of discourse words may reflect the gaps in the training data, and we check this in the experiment.

4 Translation issues

There are two extremes in translation studies. One group of scholars basing on monolingual dictionaries and lexicographical definitions capitalize the role of language-specific elements and their functional equivalents in the parallel corpora [30; 23; 5]. Discourse words are language-specific, since they have diverse correspondences in corpora, but such correspondences can be registered in super-corpus data bases [31; 32] and eventually added to monolingual dictionaries [6]. Other authors deny the existence of translation equivalents on the word-level and argue that discourse words lack their own meanings [19]. The last claim echoes the well-known theory by Victor V. Vinogradov that only content but not function words have lexical meanings. We are not going the mediate this dispute here, cf. the opposite statements on the need to account for function words in the lexicon [9; 34]. However, we would like to add an extra dimension to translation of operator words, namely, the distinction of *horizontal* and *vertical* translation. While horizontal translation involves an attempt to use one and the same target equivalent in all contexts, vertical translation is context-bound: a group of semantically close words tends to get similar or identical targets in the same context, but one source word gets different targets in different contexts. The vertical strategy for tackling Russian DE DICTO words is schematically shown in fig. 2: the symbols like '*b' stand for the slots where the translator failed to select the target predicted by the vertical strategy.

	Context 1	Context 2	Context 3	Context 4	Context 5
razumeetsya ^{AFF}	a	b	c	d	e
konechno^{AFF}	a	b	c	*c	e
samo soboj^{AFF}	a	b	c	d	*b
estestvenno^{AFF}	a	b	c	*c	*f
ponyatno^{AFF}	a	b	c	d	e

Table. 1: Translation grid for 5 Russian DE DICTO words (vertical translation)

The vertical strategy is based on the translator's capacity and will to look up or reconstruct the broad context, which is of double importance in the case of short lab sentences in our study, where neither the human experts nor the MT systems had access to prosody and coherent text fragments. We assume that this factor is relevant for the expected contrast between expert and MT translation as well as for the choice of the horizontal versus vertical strategy by MT systems and LLMs.

The recognition of VER and AFF brings about a more specific problem, notably, the overtness of the operator in the target language. Since verification can be encoded solely by intonation across the world's languages, we accepted the omission of the lexical VER operator ($w_{SL}{}^{VER} \Rightarrow \emptyset^{VER}$) as a legal move. Meanwhile, affirmation is always encoded lexically, therefore the dropping of the AFF operator is illegal. E. g., both Google MT and Yandex MT translate the source sentence with *Ya tak i znal*^{AFF} into German without the modal particle *ja* prescribed by our expert, therefore we considered the translation (3) a mistake.

(3) Russian ⇒ German

[14.2.] Rus. **Ya tak i znal**$^{\mathbf{AFF}}$, chto Vasya ne pridet ⇒ Ger. **Ich wusste Ø**, dass Vasya nicht kommen würde. (Google MT & Yandex MT).

5 5Technical aspects

In this work, several MT options were used: a) MT systems Google[7] and Yandex[8]; b) Large Language Models (LLMs) ChatGPT 4o [25; 4] and Gemini 1.5 [8]. We had also been planning to use a MT model based on Transformer architecture [26] (namely, T5 or Text-to-Text Transfer Transformer), but there are few pre-trained options with all the languages we need, and there were limits to our computational resources so we weren't able to train a custom model, although we may do so for the future research. We have conducted several experiments with a one-to-many languages MT model[9], but it translates from English to other languages, so it would be questionable to compare the results. It should be noticed, though, that the use of an open pre-trained MT model allows one to get token embeddings for further investigations, visualisation included.

Modern MT systems such as Google or Yandex Translate are also Transformer-based[10], but at least for Google Translate there are reports[11] that there are language pairs it doesn't translate directly, but with the use of an intermediate language (most often, English). We have chosen Google and Yandex Translate because of their popularity.

As for the LLMs, it is known that they weren't designed specifically for the task of MT, but are good at it, as research shows [12]; ChatGPT 4o is currently one of the leading LLMs in known benchmarks[12], whereas Gemini 1.5 is a model with a much smaller parameter size (200 billion vs 1 trillion), and this was one of the reasons we have chosen it, as we wanted to assess the gap in their performance.

We have conducted several experiments. For MT systems, we compiled lists of stimulus sentences and sequentially translated them into target languages. The LLMs work on different principles, so we needed to create instructions for them. For ChatGPT, we only had to describe the task (in Russian) as "Please translate these sentences into the following language", and provide it with the lists of sentences and the name of a language; the instruction then was repeated for all of the target languages. Gemini 1.5, however, couldn't cope with the same instruction as it didn't translate all of the sentences one by one and tried to elaborate on the given task instead of doing it, so we had to give it a sentence and a list of target languages at a time. Besides, we encountered a problem with Gemini 1.5 while trying to translate stimuli such as *Vasya ... durak* 'Vasya … is a fool': the model has rigid rule-based restrictions which make it plainly refuse to deal with any words considered offensive. That's why we had to replace such stimuli with *Vasya… umnyj* 'Vasya … is smart'.

In order to evaluate the results of the automatic MT, we used expert translations as a reference; but it must be noted that assessing the quality of translations is not straightforward, so a different translation can still be correct from the point of view of a speaker. We considered different sentences that were close semantically and correct grammatically as 'correct' ones. In the table 2, there are also two types of incorrect translations: 'incorrect' means that a model chose a wrong operator word (DE DICTO instead of DE RE and vice versa) but the sentence was grammatically well-formed. Sentences with a different meaning were considered translation errors ('error'). For example, a common issue is for the MT models to translate *samo soboj* 'naturally' as 'by himself' or 'alone', and even ChatGPT makes such mistakes. We also confirm that our chosen LLMs are almost incapable of translating into Ossetian (obviously it is drastically underrepresented in their training data), while Yandex and Google make lots of grammatical mistakes. Surprisingly, ChatGPT shows the worst quality in Ossetian for it hallucinates pseudo-Ossetian sentences with random words.

[7] translate.google.com
[8] translate.yandex.com
[9] https://huggingface.co/google/madlad400-3b-mt
[10] https://research.google/blog/recent-advances-in-google-translate/
[11] https://en.wikipedia.org/wiki/Google_Translate
[12] https://www.vellum.ai/llm-leaderboard

		Correct	Incorrect	Error	Total
	Expert	80	10	0	90
	Google	51	1	38	90
	Yandex	71	0	19	90
	ChatGPT	50	5	35	90
Arabic	***Gemini***	71	1	18	90
	Expert	80	10	0	90
	Google	77	9	4	90
	Yandex	69	9	12	90
	ChatGPT	90	0	0	90
Bulgarian	***Gemini***	82	8	0	90
	Expert I	85	5	0	90
	Expert II	80	10	0	90
	Google	80	10	0	90
	Yandex	77	13	0	90
	ChatGPT	85	5	0	90
Danish	***Gemini***	90	0	0	90
	Expert	82	8	0	90
	Google	90	0	0	90
	Yandex	85	5	0	90
	ChatGPT	85	5	0	90
English	***Gemini***	89	1	0	90
	Expert	86	4	0	90
	Google	77	8	5	90
	Yandex	65	11	14	90
	ChatGPT	80	5	5	90
German	***Gemini***	89	1	0	90
	Expert	90	0	0	90
	Google	80	6	4	90
	Yandex	67	7	16	90
	ChatGPT	90	0	0	90
Icelandic	***Gemini***	83	7	0	90
	Expert	80	10	0	90
	Google	0	0	90	90
	Yandex	0	0	90	90
	ChatGPT	0	0	90	90
Ossetian	***Gemini***	0	0	90	90
	Expert	73	17	0	90
	Google	81	9	0	90
	Yandex	77	13	0	90
	ChatGPT	85	5	0	90
Swedish	***Gemini***	90	0	0	90
	Expert	90	0	0	90
	Google	75	0	15	90
	Yandex	81	0	9	90
	ChatGPT	85	0	5	90
Ukrainian	***Gemini***	83	0	7	90

Table 2. The statistics of the results

What our results show from the 'technical' point of view, is that 1) Google Translate tends to make less grammatical mistakes than Yandex Translate, although the latter may choose incorrect translations (DE DICTO instead of DE RE and vice versa) less often; 2) ChatGPT translates closer to humans than other models; 3) LLMs surprisingly often choose translation variants close to those of the experts, but not the exact ones, as opposed to the MT systems. 4) Genetic relatedness of languages also matters: the automatic translations for Ukrainian are much closer to the expert ones than for Icelandic or German.

6 Discussion

The results of the study confirm the hypothesis that the contrast of DE RE versus DE DICTO words holds cross-linguistically, despite the class of DE RE words in the target languages is smaller than in Russian: no translator used 7 different DE RE words. The translators generally succeeded in extracting the corrector operator (VER vs AFF) from most DE RE and DE DICTO words, except for the colloquial *konkretno*. Most experts correctly identified the VER operator and placed *konkretno* in the DE RE class. Sporadic deviations, e.g., in the expert translation of *konkretno* via German *bestimmt*^{AFF} or Ossetic *bælwyrdæj*^{AFF} are probably explained by the fact that those experts who treated Rus. *konkretno*^{VER} as a close synonym of Rus. *opredelenno*^{AFF} do not have it in their active vocabulary. The failures of MT systems and LLMs to translate the stimuli with *konkretno*^{VER} can be explained by the insufficient training base, i.e. the lack of the modern input texts containing this discourse item. The same factor was apparently responsible for the mishaps with the stimuli containing the pejorative noun *durak* 'fool': the experts rejected the targets like Ger. *Narr*, *Dummkopf*, Ice. *fífl* as dated and weird. This issue however is marginal for the recognition of operators. One expert, a bilingual speaker of Russian and English, refused to provide English equivalents to *i vpryam'*^{VER} and *opredelenno*^{AFF} since he considered these words parasitic. A normative reaction of this kind was not an option for MT and LLMs, while other experts preferred to translate the stimuli even if they did not use them themselves and were not quite sure about their exact meanings.

Most translated versions show a compromise between the horizontal strategy, i. e. the preservation of the same target in Contexts 1 – 5, and the vertical strategy, i. e. the use of the same target standing for different stimuli in contexts of the same type. It is but noteworthy that the choice of the horizontal strategy indicates that the target language is relatively close to Russian regarding the size of DE RE and DE DICTO classes. E. g., our expert in Icelandic provided 5 lexical equivalents for 7 standard Russian DE RE words[13], 9 lexical equivalents for 9 standard Russian DE DICTO words and suggested a correspondence between Rus. *konkretno*^{VER} and Ice. *nefnilega*^{VER}. These moves both confirm the skill of the expert and indicate that Icelandic allows to differentiate nearly as many discourse words expressing the VER | AFF values as Russian. This level was not reached by the MT systems, e. g. Google Translate from Icelandic opted for the vertical translation of 7 Russian DE RE words providing 5 of the them with identical lines but systematically uses different targets in Contexts 1 & 4 versus Contexts 2, 3 & 5. However, the vertical strategy in the hands on an expert tells more about the expert's attitude to the experiment that about language data. E. g, our expert who approached the vertical strategy in their translations of Russian operator words into German looked for the most idiomatic variants and apparently considered the differentiation of 17 VER | AFF words in the test set of Russian a more futile task than the differentiation of Contexts 1 – 5 containing the VER | AFF insertion. This degree of freedom towards the source data is normally not encouraged in MT systems, which partly explains the gap between expert and MT translation[14]. One expert who consistently used the horizontal strategy for the AFF stimuli developed a special system of translating VER stimuli into English: he translated them lexically or by $\varnothing^{VER}$ but marked the licensing context by a special 'prefix'.

[13] The stimuli Rus. *dejstvitel'no*^{VER}, *na samom dele*^{VER} and *v samom dele*^{VER} were translated by our expert by identical lines. Meanwhile, the expert managed to preserve the subtle difference between Rus. *dejstvitel'no*^{VER} and *real'no*^{VER} by adding the particle *nú* in the target sentences corresponding to the stimuli with *real'no*^{VER}.

[14] There are hyperparameters like 'temperature' in most modern MT systems that allow for a certain degree of freedom, but typically production systems are deterministic by default in order to make their translations consistent and reliable. Also, emerging methods in the area of LLM development may allow and even stimulate variance in translations, as well as in text generation. This may well be seen in our data where Google and Yandex tend to give less variants than LLMs.

(4) Russian ⇒ English

[3.4.] Rus. Vasya **i pravda** $^{\text{VER}}$ oshibsya ⇒ Eng. [$_{\text{PREF}}$ **Yes, too bad**]. Vasya **did**$^{\text{VER}}$ make a mistake würde. (Expert translation).

[4.5.] Rus. Vasya **na samom dele** $^{\text{VER}}$ ne sobiralsya prixodit'. ⇒ Eng. [$_{\text{PREF}}$ **Come to think of it**]: Vasya ∅ $^{\text{VER}}$ didn't even mean to come. (Expert translation).

The 'prefixes' [$_{\text{PREF}}$ **Yes, too bad**] and [$_{\text{PREF}}$ **Come to think of it**] as such do not convey the meanings of VER or AFF but link the sentences containing VER to a broader context. It is difficult to say whether the corresponding method can be implemented in the MT translation of discourse words. MT depends strongly on the linear context and slight changes in the source sentence may produce seemingly unexplainable differences in the target one, e.g., 'pravda' vs 'i pravda'. There are also issues with the amount of data for the source and target languages: both MT systems and LLMs cope really well with resourced languages like English or German, but their performance deteriorates drastically for Arabic and is virtually null for Ossetian.

7 Conclusions and perspectives

This study confirms the semantic and distributional independence of VER and AFF operators, supporting their universality. LLMs, despite not being specialized for MT tasks, showed remarkable adaptability and context awareness compared to traditional MT systems. The findings highlight the potential of LLMs in nuanced translation tasks and underscore the complexity of translating modal discourse words. Future work will explore custom models and further refine evaluation metrics for translation accuracy.

References

[1] Akinina O. G. Particles from the Class INNA. Functions and Semantics [Chasticy, podobnye INNA. Funkcionirovanie i semantika]. Moscow, 2008.

[2] Averina A. V. German Particles [Chasticy v nemeckom yazyke]. Moscow, 2022.

[3] Baranov A. N., Plungian V. A., Rakhilina E. V. A Guide to Russian Discourse Words [Putevoditel' po diskursivnym slovam russkogo yazyka]. Moscow, 1998.

[4] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J. D., D. Prafulla, A. Neelakantan et al. "Language models are few-shot learners." // *Advances in neural information processing systems 33* (2020), 1877-1901.

[5] Dobrovol'skij D., Pöppel L. Discoursive Construction in the Russian-Swedish Dictionary Database: A Case Study of *v tom-to i N*. // T. Margalitadze, G. Meladze, I. Javakhishvili (eds.), *Proceedings of the XVII Euralex International Congress. Lexicography and Linguistic Diversity*. Tbilisi, 2016, 668 – 677.

[6] Dobrovol'skij D. O., Zaliznjak Anna A. The Parallel Corpus as an Instrument of Semantic Analysis: The German Modal Verb *SOLLEN* [Parallel'nyj korpus kak instrument semanticheskogo analiza: nemeckij modal'nyj glagol *sollen*] // *Proceedings of the International Conference "Corpus Linguistics 2021".* St-Petersburg, 2021, 209-218.

[7] Frey W. On the categorical status of different dependent clauses. In: J. M. Hartmann & A.Wöllstein (eds.): Propositionale Argumente im Sprachvergleich: Theorie und Empirie. Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues [Studien zur Deutschen Sprache]. Tübingen: Gunter Narr Verlag, 2023, 364 – 410. DOI: 10.24053/9783823394105

[8] Georgiev P., Ving Ian Lei, Burnell R., Libin Bai, Gulati A., Tanzer G. et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530* (2024).

[9] Haspelmath M. Types of Clitics in the World's Languages // *Linguistic Typology at the Crossroads 3-2* (2023), 1 – 59. DOI: 10.6092/issn.2785-0943/16057

[10] Kiseleva K. L., Paillard D. (eds.). Russian Discourse Words: Contextual Variation and Semantic Invariants [Diskursivnye slova russkogo yazyka: Kontekstnoe varjirovanie i semanticheskoe edinstvo]. Moscow, 2003.

[11] Kiseleva K.L. Konechno. Razumeetsya. Estestvenno. In Kiseleva K. L., Paillard D. (eds.) *Russian Discourse Words: Contextual Variation and Semantic Invariants* [Diskursivnye slova russkogo yazyka: Kontekstnoe varjirovanie i semanticheskoe edinstvo] Moscow, 2003.

[12] Kocmi T., Federmann C. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520* (2023).

[13] Krifka M. Modelling questions in commitment spaces. In M. Cordes (ed.), *Asking and answering: Rivalling approaches to interrogative methods*, 63–95. Tübingen: Narr Francke Attempto, 2021. DOI: 10.24053/9783823394808

[14] Krifka M. Layers of the clause: Propositions, judgements, commitments, acts. In: J.M. Hartmann & A.Wöllstein (eds.), *Propositionale Argumente im Sprachvergleich: Theorie und Empirie / Propositional Arguments in Cross-Linguistic Research: Theoretical and Empirical Issues* [Studien zur Deutschen Sprache]. Tübingen: Gunter Narr Verlag, 2023, 116 – 183. DOI: 10.24053/9783823394105
[15] Krylova E. V. The Communicative Functions of Danish Modal Particles [Kommunikativnye funkcii modal'nyx chastic v datskom yazyke]. Moscow, 2021.
[16] Levontina I. B. Particles of Speech [Chasticy rechi]. Moscow, 2022.
[17] Lohnstein H. Verum Focus, Sentence Mood, and Contrast. In C. Dimroth and S. Sudhoff (eds.), *The Grammatical Realization of Polarity Contrast: Theoretical, Empirical, and Typological Approaches*. [Linguistik Aktuell/Linguistics Today 249]. 2018, pp. 55–88.
[18] Mikhailov M., Cooper R. Corpus Linguistics for Translation and Contrastive Studies. London and New York: Routledge, 2016. DOI: 10.4324/9781315624570
[19] Mikhailov M. Mind the Source Data! Translation Equivalents and Translation Stimuli from Parallel Corpora. In Vincent X. Wang, Lily Lim, Defeng Li (eds.), *New Frontiers in Translation Studies*. Springer, 2021, 259 — 279. DOI: 10.1007/978-981-16-4918-9
[20] Nelson M. The De Re/De Dicto Distinction. In E.N.Zalta & U.Nodelman (eds.), *The Stanford Encyclopedia of Philosophy* (Fall 2024 Edition), URL = <https:// https://plato.stanford.edu/entries/prop-attitude-reports/dere.html>
[21] Paducheva E. V. Modality [Modal'nost']. In *Data for the Corpus Grammar of Russian. I. The Verb* [Materialy k korpusnoj grammatike russkogo yazyka. I. Glagol]. St.-Petersburg: Nestor-Istoriya, 2016, 19 – 94.
[22] Paducheva E. V. Semantic Studies. The Semantics of Aspect and Tense in Russian. The Semantics of the Narrative [Semanticheskie issledovanija. Semantika vida i vremeni v russkom yazyke; Semantika narrativa]. Moscow, 1996.
[23] Pailliard D. TAKI. In K. L. Kiseleva, D. Paillard (eds.) Russian discourse words. A contextual semantic description [Diskursivnye slova russkogo yazyka. Opyt kontekstno-semanchticheskogo opisaniya]. Moscow, 1998.
[24] Pailliard, Denis. PRAVDA as a discourse word. The guarantee and the perspective [PRAVDA kak diskursivnoe slovo: garant i tochka zrenija]. In Kiseleva K. L., Paillard D. (eds.). *Russian Discourse Words: Contextual Variation and Semantic Invariants* [Diskursivnye slova russkogo yazyka: Kontekstnoe varjirovanie i semanticheskoe edinstvo]. Moscow, 2003, PAGES.
[25] Radford A. Improving language understanding by generative pre-training. June 11, 2018. URL: https://openai.com/index/language-unsupervised/ [Accessed 20.01.2025]
[26] Vaswani A. Attention is all you need. // *Advances in Neural Information Processing Systems, 30*. 2017.
[27] Yanko T. E. The Communicative Status of Expressions Containing the Word PRAVDA [Kommunikativnyj status vyrazhenij so slovom PRAVDA]. In N. D. Arutyunova, N. K. Ryabceva (eds.), *Logical Analysis of the Language. Truth and Veracity in Culture and Language* [Logicheskij analiz yazyka. Istina i istinnost' v kul'ture i yazyke]. Moscow, 1995, 173-177.
[28] Yanko T. E. Communicative Strategies of Russian Speech [Kommunikativnye strategii russkoj rechi]. Moscow, 2001.
[29] Zaliznjak Andrej A. Studies of the Language of Birch Bark Letters [K izucheniju yazyka berestjanyx gramot]. In V. L. Janin, A. A. Zaliznjak. *Novgorodskie gramoty na bereste. Iz raskopok 1984 – 1989 godov* [[Novgorod Birch Bark Letters from the Excavations of 1984–1989]. Moscow, 1993, 191–319.
[30] Zaliznjak Anna A. Russian Language-Specific Words as an Object of Contrastive Analysis [Lingvospecificheskie edinicy russkogo yazyka v svete korpusnogo analiza] // *Computational linguistics and intellectual technologies. Iss. 14 (21).* 2015, 683 – 695.
[31] Zaliznjak Anna A. A Data-Base of Cross-Linguistic Equivalences as an Instrument of Linguistic Analysis [Baza dannyx mezh'yazykovyx ekvivalencij kak instrument lingvisticheskogo analiza] // *Computational linguistics and intellectual technologies. Iss. 15 (22).* 2016, 854 – 866.
[32] Zaliznjak Anna A., Loshchilova E. Y., Kruzhkov M. G., Durnovo A. A., Buntman N. V., Egorova A. Y., Zatsman I. M. The Library of Discourse Words Annotations in the Parallel Russian-French Texts, version 2.0 [Biblioteka annotacij diskursivnyx slov v russko-francuzskix parallel'nyx tekstax (versiya 2.0). Certificate of registration of the database RU 2022623727, 27.12.2022.
[33] Zimmerling A. Biased Questions and Hamblin Semantics // *Typology of Morphosyntactic Parameters*. 2023. Vol. 6. Iss. 2, 92-135. DOI: 10.37632/PI.2023.94.43.005
[34] Zimmerling A. Really: Syntactics without Semiotics? [Konkretno: sintaktika bez semiotiki?] // *Slovo. Ru: Baltic Accent*. 2023. № 3, 125-153. DOI: 10.5922/2225-5346-2023-3-9
[35] Zimmerling A. From God's Truth to Concession [Ot Bozhjej pravdy k ustupke]. In M. S. Milovanova (ed.), *Axiology and Cognitive Linguistics* [Aksiologiya i kognitivnaya lingvistika]. Moscow, 2024, 346 – 353. DOI: 10.5281/zenodo.15098320
[36] Zimmerling A.V., Yanko T. E. Verification and Truth. From Predicate to Connector [Verifikacija i pravda. Ot predikata k konnektoru]. In I. M. Kobozeva, N. V. Serdobolskaya, A. I. Kryukova (eds.), *Connectors in Modern Russian. Semantics, Syntax, Pragmatics, Diachrony* [Konnektory sovremennogo russkogo yazyka: semantika i sintaksis, pragmatika i diahroniya. Moscow, 2025.

Abstracts

USING RHETORICAL STRUCTURES TO ANALYZE ARGUMENTATION IN SCIENTIFIC COMMUNICATION TEXTS

Akhmadeeva I., **Kononenko I.**, **Sidorova E.**, **Shestakov V.**, A.P. Ershov Institute of Informatics Systems, Novosibirsk, Russia

This article explores the role of rhetorical structures in the argument mining task on the material of scientific Internet communication texts in Russian. Two approaches are proposed and studied for argumentative relation prediction: the first one constructs segment vector representations using a Graph Neural Network (GNN) based on rhetorical structure, and the second one uses multitask learning that combines argumentation extraction with rhetorical relations prediction tasks. With proposed approaches three models were implemented: two variations of the model using GNN and one model employing the multitask approach. These models were compared with a simple baseline using the Lonformer model on a dataset annotated with both argumentative and rhetorical structures. Argumentative annotating was performed manually by four experts. Existing resources and tools were used to obtain rhetorical markup. The conducted experiments showed that the approaches using additional rhetorical information improve the quality of argumentative relation prediction, particularly for long-distance relations. The best performance, with an F1 score of 72.32%, was achieved by a model incorporating GNN-enhanced statement representations.

RUTAR—A DATASET IN RUSSIAN FOR REASONING ABOUT TAXES

Alibekov A.[1,2], **Migal A.**[3,4], **Matenkov A.**[3], **Muryshev A.**[3], **Bolshakov V.**[3,5], **Kozachenko A.**[1], **Mukhtarova G.**[3], **Mikhaylovskiy N.**[3,2], [1]NTR Labs, Tomsk, Russia; [2]Higher IT School, Tomsk State University, Tomsk, Russia; [3]NTR Labs, Moscow, Russia; [4]HSE, Saint-Petersburg, Russia; [5]BMSTU, Moscow, Russia

In 2024, reasoning have emerged as a new frontier for artificial intelligence and computational linguistics. Reasoning models are typically evaluated either on STEM-related datasets, or on synthetic datasets. This ignores a huge area of human thought—namely, humanitarian. To bridge this gap partially, we present a new open dataset, RuTaR (Russian Tax Reasoning). The dataset consists of modestly modified content of 199 select Ministry of Finances of Russia and Russian Federal Tax Service letters that typically reason to answer some taxpayer question. Despite apparent simplicity of yes/no questions, both off-the-shelf Large Language Models (LLMs) and Retrieval Augmented Generation (RAG) systems struggle to achieve high results on the dataset, with top RAG system studied achieving 77% accuracy.

COBALD PARSER: JOINT MORPHOSYNTACTIC AND SEMANTIC ANNOTATION

Baiuk I.[1], **Baiuk A.**[2], **Petrova M.**[2,3], [1]MIPT, Russia, Moscow; [2]RSUH, Russia, Moscow; [3]Smart Data Expert, Russia, Moscow

Dependency parsing is a common task for modern NLP, and Universal Dependencies (De Marneffe et al., 2021) is widely acknowledged nowadays as a morphosyntactic annotation standard. Yet, its dependency relations are rather generalized, therefore, in order to take more syntactic details into account, the Enhanced UD standard was proposed. A newly developed CoBaLD annotation standard elaborates the E-UD principles by enriching it with the semantic level. It is aimed at structural simplicity and the compatibility with UD in all possible issues. Currently, there are several datasets annotated in CoBaLD standard, but until now, there has been no appropriate tool for automatic data parsing in CoBaLD format. In this paper, we present a neural-based joint parser capable of automatic annotation both in E-UD and in CoBaLD, including ellipsis restoration which is supposed by these standards. Additionally, we provide a qualitative analysis of automatic annotation errors.

EMBEDDED QUOTATION AS A FACTOR OF IDIOMATICITY

Baranov A. N., Vinogradov Russian Language Institute of the Russian Academy of Sciences

The report examines one of the types of idiomaticity—embedded quotation. This type is peripheral and is implemented in a small set of idioms, slightly more than twenty elements. Idioms with embedded citations include, in particular, the expressions po samoye ne baluysya (lit. to the very don't play) ≈ 'to the maximum extent possible', po samoye ne khochu (lit. to the very don't want) 'to the maximum extent possible', na otskrebis' (lit. on fuck off) ≈ 'to do something quickly and badly', prikazat' dolgo zhit' (lit. to order to live long) ≈ 'to die', derzhat'sya na chestnom slove (lit. to keep one's word of honor) ≈ 'be poorly connected to the main part of the object', pod chestnoye slovo (lit. under a word of honor) ≈ 'to do something to someone or allow him to do something under the guarantee of this person', za Khrista radi (lit. for Christ's sake) ≈ 'to beg', [i] pominay kak zvali (lit. [and] remember his name) ≈ 'disappear without a trace', vot tebe / te[i] na (lit. here's for you) ≈ 'expression of surprise at failure', za spasibo (lit. for thanks) ≈ 'doing something for free', za [prosto] tak (lit. for no reason) ≈ 'doing something for free' and some others.

Embedded quotation is a relevant factor of idiomaticity, leading to the emergence of new idioms in the language.

Embedded quotation as a factor of idiomaticity is directly related to the phenomena of quotation (citation) and generalized quotation (citation). The existence of generalized quotation in speech provides the very possibility of forming idioms, the idiomaticity of which is based on embedded quotation. This is a necessary intermediate link in the sequence: quotation—generalized quotation—embedded quotation.

COLLECTIVITY & DISTRIBUTIVITY

Boguslavsky I. M., IITP (Kharkevich Institute), FRC CSC, Moscow, Russia; Universidad Politécnica de Madrid, Madrid, Spain

The opposition 'separately' / 'together' manifests itself in different parts of the linguistic system—in the interpretation of noun phrases (distributive vs. collective), in syntax (200 roubles a piece, 100 kilometres per hour), in the lexicon (together, jointly, collectively, in aggregate, common, collective; apart, separate, apart, separately, each, etc.) in word formation (coauthor). We discuss these meanings in the context of developing a computational model of semantic analysis that highlights the following issues: 1) What information about the meaning of distributive/collective NPs should be available to the parser for reasoning? 2) What valency slots are characteristic of the meanings 'together' / 'separately' and how are they instantiated in syntactic constructions and lexical items?

RUSSIAN NATIONAL CORPUS 2.0: CORPUS PLATFORM, ANALYSIS TOOLS, NEURAL NETWORK MODELS OF DATA MARKUP

Bonch-Osmolovskaya A. A., Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Gladilin S. A.**, IITP (Kharkevich Institute), FRC CSC, **Kozerenko A. D.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Lyashevskaya O. N.**, HSE University, **Morozov D. A.**, NSU, **Kuznetzova Y. N.**, MSU, Institute of Linguistics of the Russian Academy of Sciences, **Makhova A. A.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, **Piskounova S. V.**, Independent Researcher, **Bujlova N. N.**, Lopukhin Federal Research And Clinical Center of Physical-chemical Medicine of Federal Medical Biological Agency, **Borodina D. G.**, St. Petersburg State University, **Vinogradova I. I.**, Prosveshchenie Publishers, **Sizov V. G.**, **Dyachenko P. V.**, **Kazennikov A. O.**, IITP (Kharkevich Institute), **Vlasova N. A.**, A.K. Ailamazyan Institute of Program Systems of the Russian Academy of Sciences, **Glazkova A. V.**, University of Tyumen, **Stolyarov S. S.**, **Garipov T. A.**, **Smal I. A.**, NSU, **Gubar'kova Ya. N.**, Yandex

The Russian National Corpus has existed for over 20 years and is a unique linguistic tool. However, the technical limitations of the software platform on which it was implemented significantly narrowed its development prospects. In 2020, work was launched on a comprehensive update of the RNC software platform, as a result of which the National Corpus switched to a new generation 2.0 platform. The implemented deep changes concerned both the development of functionality that meets modern approaches to corpus linguistics, and a fundamental restructuring of the platform architecture as a whole, from data preparation and indexing systems to the user interface. A separate area of development of the capabilities of the RNC was associated with the implementation of neural network models used for metadata tagging, disambiguation, word-formation markup, etc.

This article provides a short description of the new corpus platform as of 2024. The description includes key pa-rameters of changes in the architecture of the RNC platform and its user interface, descriptions of new corpus data analysis services and the specifics of their implementation, as well as a description of the experience of using neural network models for tasks related to corpus data markup.

The purpose of the article is to describe the technological layer of changes implemented in the National Corpus of the Russian Language as part of a large-scale update carried out in recent years.

LEXICAL DECISION TASK: MODALITY MATTERS

Burov D. A.[1], **Panich M. B.**[1], **Sadkovskii F. A.**[1,2], **Fedorova O. V.**[1], **Shevelev S. M.**[1], [1]Lomonosov Moscow State University, Moscow, Russia; [2]RAS Institute of Linguistics, Moscow, Russia

Lexical decision task is one of the most wide-spread methods used is psycholinguistic experiments, performed in either visual or auditory modality. The comparability of results acquired in the two modalities, however, has only recently received attention. In this paper, we present the results of two parallel experiments, one in the visual modality and one in the auditory modality, run on the same lexical material (words and pseudo-words). The reaction time distri-butions differ between the two modalities: mean reaction time is longer in the auditory modality. Mean reaction time for pseudo-words in significantly longer the mean reaction time for real words.

DERIVATION OF AKTIONSARTEN FROM (NON-)DIRECTIONAL VERBS OF MOTION: POSING RESTRICTIONS ON RESTRICTIONS

Chuikova O. Iu., Herzen State Pedagogical University of Russia, **Gorbova E. V.**, Independent researcher

The paper examines prefixal derivatives of Russian motion verbs that represent perfective Aktionsarten. The focus is on the statistical analysis of motion verbs as a specific subset of Aktionsarten characterized by their derivational features. Both sets (verbs of motion and Aktionsarten) are included in the database of Russian prefixed verbs developed by the authors based on the Dictionary of Russian language. One of the aims of the study is to verify the statement defended in previous works that the stem of a motion verb imposes restrictions on the possible Aktionsarten derived from it. This verification will be based on both dictionary data and the data from the Russian National Corpus and the Runet.

ENHANCING RAG AND KNOWLEDGE GRAPHS WITH DISCOURSE

Galitsky B., Knowledge Trail, San Jose, CA, USA, **Ilvovsky D.**, **Morkovkin A.**, HSE University, Moscow, Russia

We consider a number of Retrieval Augmented Generation (RAG) architectures to address a lack of specific information and hallucination issues of Large Language Models (LLM)—based question answering. We start with conformal prediction which acts on top of LLM and maintains a set of generations instead of a single one and attempts to find the best element of this set, which is assumed to be the "most average one". We then proceed to LLM self-reflection series of RAG architectures predicting the multi-hop question answering session before actual search for an answer. After that, we propose a mechanism for LLM to filter out answers inappropriate with respect to style. All these components need discourse-level analysis for more robust functioning. Knowledge graph (KG) and Abstract Meaning Representation (AMR)-based knowledge graph construction follow. We evaluate the contribution of all of these components to overall answer relevance and also zoom in on the role of discourse—based subsystem in each of these components. There is a substantial improvement of performance due to the four-component architecture introduced in this paper; the contribution of discourse-based subsystems is fairly modest.

CULTURAL EVALUATION OF LLMS IN RUSSIAN: CATCHPHRASES AND CULTURAL TYPES

Gromenko E.[1], **Kalacheva D.**[2], **Klokova K.**[2], **Krongauz M.**[2,3], **Moroz O.**[4], **Shulginov V.**[2], **Yudina T.**[2], [1]HSE, Moscow, Russia; [2]MIPT, Moscow, Russia; [3]RSUH, Moscow, Russia; [4]BHSAD, Moscow, Russia

This study addresses the gap in evaluating large language models' (LLMs) cultural awareness and alignment within the Russian sociocultural context by introducing a structured framework comprising 8 Cultural Types (e.g., Spiritual Practitioner, Soviet Intellectual) and 5 catchphrase groups (e.g., memes, proverbs). A 400-question evaluation dataset was developed to probe 10 multilingual LLMs, including GPT-4o, Claude 3.5 Sonnet, and Gemini 1.5 Pro, across fact-based cultural knowledge and nuanced linguacultural understanding in a zero-shot setting. Results show that closed-source models GPT-4o and Claude 3.5 Sonnet outperform other models, with one of the smallest models (Mistral NeMo 12B) achieving the lowest result. Performance disparities were noticed in separate evaluation on Cultural Type tasks and catchphrases. Model-specific skews emerged, with lower-ranked models showing inclination toward specific cultural types. Qualitative analysis revealed common errors, such as selecting synonymous but incorrect answers or failing to grasp culturally specific logic. The contribution outlines the limitations of LLMs in interpreting cultural context and lays the groundwork for further research in assessing the cultural-linguistic alignment of LLMs.

MULTI-SENSES CONNECTIVES: SOME PRECISIONS

Inkova O. Yu., Institute of Informatics Problems, FRC CSC RAS, Moscow, Russia; University of Geneva, Geneva, Switzerland

The article discusses multi-senses connectives and their annotation in text corpora. The author clarifies the concept, which is usually applied to three different phenomena: i) the uncertainty of annotators in choosing one of the meanings of a polysemic connector; ii) the possibility of establishing more than one relation, both explicit and implicit, between two fragments of text; iii) the "combination" of several meanings by a connective. The author then considers how these cases are annotated in the Penn Discourse Treebank and in the Supracorpora Database of Connectives, which use a multi-label annotation for discourse relations. The study also raises a number of theoretical questions: what information can we obtain from double labels, which relations are distributionally close, i.e. can be established in the same contexts, how to separate the contribution of the connective and of the context to the overall meaning of a se-quence of sentences, how legitimate is it to talk about the combination of meanings by the connective, how to annotate the features of the context of the connectives so that this data can be used for Natural Language Processing? The solution to these questions is important for both theoretical cognitive and applied research (in particular, for machine learning and machine translation).

AN ALGORITHM FOR GENRE IMBALANCE CORRECTION IN THE RUSSIAN SUBCORPUS OF THE GOOGLE BOOKS NGRAM CORPUS

Ivleva I., **Solovyev V.**, Laboratory of Linguistics and AI, Institute of Philology and Intercultural Communication, Kazan Federal University, Kazan, Russia

Recently, the latest version of the Google Books Ngram was shown to be imbalanced. This corpus being an important and widely used database, the imbalance can affect the results of research in different areas. In the paper, we present an algorithm for correcting the dynamics of word frequency in the Google Books Ngram corpus. The algorithm takes into account the discovered imbalance of the main literary styles—fiction, publicistics, and non-fiction. The rationale of the algorithm is given, as well as examples of correction.

COMPUTER TOOLS IN MENTAL DISORDERS DIAGNOSTICS BY ORAL SPEECH

Khomenko A.[1], **Komratova A.**[1], **Isakov D.**[1], **Balba D.**[1], **Shishkovskaya T.**[2,3], **Khudyakova M.**[1], [1]HSE University Nizhny Novgorod, Russia; [2]HSE University Moscow, Russia; [3]Mental Health Research Center Moscow, Russia

The integration of automated speech analysis in diagnosing mental health disorders is becoming increasingly significant in both clinical and computational linguistics. This study aims to construct linguistic profiles for individuals with neurocognitive and affective mental disorders. Using speech transcriptions and relevant to the study computational techniques like lexical clustering and stylostatistical analysis, this research looks for characteristics capable of distinguishing speech patterns indicative of various mental health conditions. A text corpus of oral speech from 136 people diagnosed with schizophrenia, schizotypal disorder, schizoaffective disorder, borderline personality disorder, other personality disorders, primary depressive episode, recurrent depressive disorder, bipolar affective disorder and 210 participants with no diagnosed diseases in the control group was used in the research. As a result of the study, it was proved that people with mental disorders display specific features in oral speech, that can be used in creation of an automatic mental disorders diagnostic model.

LOCAL STRUCTURE OF DESCRIPTIVE DISCOURSE IN THE SPEECH OF ADULT LEARNERS OF RUSSIAN AS A FOREIGN LANGUAGE

Kiose M.[1,2], **Rzheshevskaya A.**[1], **Petrov A.**[1,2], [1]Moscow State Linguistic University, Moscow, Russia; [2]Institute of Linguistics RAS, Moscow, Russia

This paper explores the local structure expressed in rhetorical, communicative and narrative relations and speech disfluencies in the descriptive discourse of adult learners of Russian as a foreign language. In the experiment, two learner groups with B1/B2 and C1/C2 levels of proficiency in Russian as a foreign language had to read two descriptive texts and relate their contents to an interested listener. The study contrasts the distribution of local discourse structure in stimuli texts (59 EDUs) and in the speech of experiment participants in two groups (629 EDUs). The results reveal that the speech in C1/C2 group manifests more frequent narrative relation Setting, communicative relation Metadiscourse, rhetorical relation Interpretation. Meanwhile, the speech of B1/B2 group demonstrates the predominance of communicative relation Agreement and rhetorical relation Joint. The most common type of speech disfluency of C1/C2 speakers is Self-correction, whereas for B1/B2 learners this is Repetition. Additionally, as opposed to the rhetorical structure of the stimuli, Joints are used far more frequently in the speech of both groups, which evidences of the problems in rhetorical coherence with both B1/B2 and C1/C2 learners.

CONTEXTS FOR OVERLAPPING TALK IN RUSSIAN CONVERSATIONS

Korotaev N. A., Institute of Linguistics RAS, Moscow, Russia; Russian State University for the Humanities, Moscow, Russia

The paper analyses instances of overlap in Russian triadic conversations. Based on the "Russian Pear Chats and Stories" corpus, I provide a multimodal treatment of 364 overlapping episodes. I mostly focus on two conversational parameters—discourse contexts for overlaps and respective epistemic statuses of the interlocutors. A vast majority of overlaps in the data can be attributed to one of the four basic discourse contexts: Sideline Turn, Turn Transition, Turn Interception, and Competitive Development. These contexts differ in the degree of cooperation vs. competition exhibited by the participants. I show that overlaps appearing in more competitive contexts tend to be longer and more often include truncated turns. Epistemic statuses of participants are fixed throughout the studied conversations. When statuses are considered equal, competitive overlaps are more frequent than when one participant has a higher epistemic status.

THE DEVELOPMENT OF A NATURAL LANGUAGE REASONING SYSTEM FOR A COMPANION ROBOT

Kotov A. A.[1,2], **Nosovets Z. A.**[1,2], **Filatov A. A.**[3], **Arinkin N. A.**[1,2], [1]NRC Kurchatov Institute, Moscow, Russia; [2]MSLU, Moscow, Russia; [3]Yandex.Technologies LLC, Moscow, Russia

The natural language inference system allows the robot to go from the meaning of the incoming text (from a visual or tactile stimulus) to the derived meaning—the inference. This system uses a rule-based parser, and pairs of semantic representations constructed by the parser are combined into a scenario. In this paper, we represent the robot's natural language inference space as a graph, where the robot can move from the premise of a scenario to the consequence, and from the consequence of one scenario to the premise of another. We involve annotators who propose derived sentences (semantic components) for a given premise and, in the annotator interface, immediately evaluate the proximity of the proposed sentence to the available scenarios. This procedure allows us to develop the graph of scenarios, to evaluate its connectivity and the absence of dead ends (deadlock vertices), as well as the adequacy of the analysis of incoming texts by scenarios within this graph. The graph contains 5,000 scenarios and approximately 22,000 nodes. We estimate that a graph consisting of 7,000 scenarios can be sufficient for modelling the mechanism of human natural language reasoning.

BUILDING RUSSIAN BENCHMARK FOR EVALUATIONOF INFORMATION RETRIEVAL MODELS

Kovalev G., **Tikhomirov M.**, **Kozhevnikov E.**, **Kornilov M.**, **Loukachevitch N.**, Lomonosov Moscow State University, Moscow, Russia

We introduce RusBEIR, a comprehensive benchmark designed for zero-shot evaluation of information retrieval (IR) models in the Russian language. Comprising 17 datasets from various domains, it integrates adapted, translated, and newly created datasets, enabling systematic comparison of lexical and neural models. Our study highlights the importance of preprocessing for lexical models in morphologically rich languages and confirms BM25 as a strong baseline for full-document retrieval. Neural models, such as mE5-large and BGE-M3, demonstrate superior performance on most datasets, but face challenges with long-document retrieval due to input size constraints. RusBEIR offers a unified, open-source framework that promotes research in Russian-language information retrieval. The benchmark is available for public use on GitHub.

GENERATING ENCYCLOPEDIC ARTICLES BASED ON A COLLECTION OF SCIENTIFIC PUBLICATIONS

Lee G. F., **Loukachevitch N. V.**, **Khokhlov A. R.**, Lomonosov Moscow State University

Generating texts that demand high factual accuracy and strict formatting, such as encyclopedic articles, presents numerous challenges: how and where to gather relevant information, how to structure it into a coherent and wellformatted text, and how to ensure that the compiled article does not contain factual errors. We propose a solution to these problems for the Russian language by developing a system for generating encyclopedic articles that extracts the most recent and relevant knowledge from scientific publications in the online library eLIBRARY.RU and structures it as a single context for input into a generative model. To evaluate both the impact of the extracted knowledge on the content of the final texts and the overall quality of generation, we considered several prompting strategies, some of which do not use the context found in publications, and compared these approaches using automatic metrics and human expert evaluation. We hope that the created framework will become a reliable reference material for scientists researching new and relevant topics in their field of expertise.

CAUSAL MODELS AND ADVERSARIAL TRAINING: SELECTING THE RIGHT PROPERTIES FOR ROBUST NON-TOPICAL TEXT CLASSIFICATION

Lepekhin M., Moscow Institute of Physics and Technology, Russia, **Sharoff S.**, University of Leeds, UK

The vast majority of datasets for non-topical classification of texts contain distribution shifts. In most cases, those are topical shifts. Their presence in the data forces the classifiers to fit topics-related features instead of focusing on those relevant for the target class. It causes a dramatic decrease in the accuracy of the trained models when the test data are taken from a different data source. To address this problem, we experiment with two techniques: causal models and adversarial domain adaptation. In our work, we apply CausalLM, Adversarial Domain Adaptation (ADA), and Energy-based ADA (EADA) for gender classification and compare the results. The results are novel for the non-topical classification task. We show that both causal and adversarial methods manage to make the model more resilient to the distribution shifts although it causes a decrease of accuracy when tested on the domain prevailing in the training dataset. Moreover, we describe the first attempt to reduce the impact of topical shifts in the task of non-topical classification with usage of causal methods. Besides, we provide a link to the GitHub repository with the code of our experiments to ensure their reproducibility: https://github.com/MikeLepekhin/CausalAndAdversarialMethods.

RUOPINIONNE-2024: EXTRACTION OF OPINION TUPLES FROM RUSSIAN NEWS TEXTS

Loukachevitch N. V., **Tkachenko N. M.**, **Lapanitsyna A. M.**, **Tikhomirov M. M.**, Lomonosov Moscow State University, Moscow, Russia, **Rusnachenko N. L.**, Bauman Moscow State Technical University, Moscow, Russia

In this paper, we introduce the Dialogue Evaluation shared task on extraction of structured opinions from Russian news texts. The task of the contest is to extract opinion tuples for a given sentence; the tuples are composed of a sentiment holder, its target, an expression and sentiment from the holder to the target. In total, the task received more than 100 submissions. The participants experimented mainly with large language models in zero-shot, fewshot and fine-tuning formats. The best result on the test set was obtained with fine-tuning of a large language model. We also compared 30 prompts and 11 open source language models with 3-32 billion parameters in the 1-shot and 10-shot settings and found the best models and prompts.

RUTERMEVAL-2024: CROSS-DOMAIN AUTOMATIC TERM EXTRACTION AND CLASSIFICATION IN RUSSIAN SCIENTIFIC TEXTS

Mamontova A., **Ischenko R.**, **Vorontsov K.**, MSU Institute for Artificial Intelligence

Automatic Term Extraction (ATE) is a critical NLP task for identifying domain-specific terms, which are essential for tasks like information retrieval, machine translation, and ontology construction. Cross-domain nested term extraction further complicates the task, as traditional methods often fail to handle hierarchical term structures and domain variability. This paper introduces both the CL-RuTerm3 dataset, a novel resource featuring nested term annotations across six domains (the main one is computational linguistics, also mathematics, medicine, economics, literature studies, and agrochemistry), and the RuTermEval-2024 competition, designed to evaluate term extraction systems on this data. The CL-RuTerm3 dataset, comprising 1270 abstracts and 15 full-text articles (over 165k tokens with over 37k annotated entities), is the largest of its kind for Russian scientific texts. Terms are classified into three categories based on lexical and domain specificity: specific terms, common terms, and nomens. The dataset's unique features, such as nested term markup and cross-domain coverage, enable more realistic evaluation of ATE systems. The paper concludes with an analysis of participant approaches in the RuTermEval-2024 competition, emphasizing the effectiveness of contrastive learning. This work aims to advance ATE research by providing a robust dataset and fostering discussions on term extraction methodologies.

BERT-LIKE MODELS FOR AUTOMATIC MORPHEME SEGMENTATION OF THE RUSSIAN LANGUAGE

Morozov D.[1,2], **Glazkova A.**[1,3], **Garipov T.**[1,2], [1]Russian National Corpus; [2]Novosibirsk State University Novosibirsk, Russia; [3]Russian National Corpus & University of Tyumen, Tyumen, Russia

Current approaches to automatic morpheme segmentation for the Russian language rely on machine learning, primarily neural network methods. Among the architectures presented, the best results have been achieved using convolutional neural networks and LSTM networks. However, the quality of automatic annotation is far from ideal, especially when dealing with roots that were not present in the training dataset. In this work, we present a new approach to morpheme segmentation based on fine-tuning BERT-like models. Through comparisons using two morpheme dictionaries with different segmentation paradigms, we demonstrated the superiority of our approach over previous ones, including when working with unfamiliar roots. The best result was achieved by fine-tuning the RuRoBERTa-large model: when working with random words, the share of completely correct segmentations increased from 88.5–90.8% to 92.5–93.5%, and when working with unfamiliar roots, it improved from 70.5–72.6% to 74.9–77.2%. Error analysis of the model showed that root nests not encountered in the training dataset can be distributed into two groups during testing: "recognizable", meaning those for which more than 90% of the words are correctly analyzed, and "unknown", meaning those for which the proportion of correct segmentations is less than 10%.

"OH, NO! NOT ON WEDNESDAY! ON THURSDAY!": INTERJECTION IN THE CONTEXT OF SPEECH DISFLUENCY

Podlesskaya V. I., Institute of Linguistics, RAS, Moscow, Russia

The subject of the study is a special class of uses of the Russian interjection OY, namely, uses in the context of speech disfluencies. The study is based on data from three corpora—the oral and the multimedia (MURCO) subcorpora of the Russian National Corpus and the pilot version of the corpus of oral personal stories "What I saw". The main types of speech disfluencies in the context of which speakers resort to using the interjection OY, the structure of constructions with OY, as well as the features of their phonetic—segmental and suprasegmental—implementation are investigated. From a quantitative point of view, a comparison of the use of OY in oral public and oral non-public speech is carried out. It is shown that although OY in oral non-public speech occurs, as could be expected, many times more often, the proportion of OY as a marker of a speech disfluency in relation to the total number of uses of OY is almost the same in public and non-public varieties of oral speech. These quantitative data suggest that the general distribution of the use of OY and the distribution of OY as a marker of speech disfluency are not regulated by the same factors. The general distribution is associated with the degree of spontaneity and publicity of speech: interjections as exclamatory discursive markers become more appropriate when the speaker is not constrained by disciplinary guidelines and can allow himself to express emotions and give assessments more freely. And the distribution of OY as a marker of speech disfluency is also associated with the density of speech disfluencies themselves. Of course, this density varies among individual speakers, but at the same time it largely depends on the general principles regulating the generation of unprepared discourse.

INTENSITY AND ITS MANIFESTATIONS IN SPEECH AND RECURRENT GESTURES IN SPONTANEOUS DIALOGUE

Prokofyeva O.[1], **Kiose M.**[1,2], **Leonteva A.**[1,2], **Smirnova E.**[1], [1]Moscow State Linguistic University, Moscow, Russia; [2]Institute of Linguistics RAS, Moscow, Russia

The present article explores the category of intensity in multimodal spontaneous dialogues in Russian. We regard intensity as a cognitive category expressed in notional, referential and sign meanings, i.e. in its manifestation degrees (high, medium, and low) and interaction of quantitative and qualitative meanings, referent types and types of lexical meaning. We study multimodal intensity patterns as revealed in the spontaneous dialogue in question, with the main focus on recurrent co-speech gestures, where the latter are attributed the stability of form and function. The research material consists of the recordings of about 3 hours featuring 20 participants, 1082 speech intensifiers, and 392 co-recurrent gestures of 9 groups. The results show, despite the susceptibility to individual preferences, that this is the notional meaning of intensity in speech that is more consistently revealed in recurrent gestures. The presenting recurrent gesture group appears to be most numerically frequent, nevertheless, these are enhancing and locating gestures that could help distinguish between pure quantity and merged quantity-quality cases. In general, exploring the use of gestures as mediated by intensity meanings allowed to specify their discourse functions attributed to intensity.

STRUCTURED SENTIMENT ANALYSIS USING FEW-SHOT PROMPTING OF AN ENSEMBLE OF LLMS

Rossyaykin P., Lomonosov Moscow State University, Moscow, Russia

This paper describes our participation in RuOpinionNE-2024 shared task (Loukachevitch et al., 2025). The objective of this task is to extract opinion tuples of the form <holder, target, polarity expression, polarity> from news texts in Russian. We approached this task with few-shot prompting of super large language models (LLMs). The quality of LLMs' predictions was improved in two ways. In the first stage we used prompts with examples which text embeddings were similar to that of the target text. In the second stage we augmented prompts with answers of LLMs from the previous stage, achieving the second-best F1 score in the competition in the post-evaluation stage. Our results show that the addition of answer suggestions to the prompt is particularly useful if they provide novel and variable information.

METHODS FOR RECOGNIZING NESTED TERMS

Rozhkov I. S., **Loukachevitch N. V.**, Lomonosov Moscow State University, Moscow, Russia

In this paper, we describe our participation in the RuTermEval competition devoted to extracting nested terms. We apply the Binder model, which was previously successfully applied to the recognition of nested named entities, to extract nested terms. We obtained the best results of term recognition in all three tracks of the RuTermEval competition. In addition, we study the new task of recognition of nested terms from flat training data annotated with terms without nestedness. We can conclude that several approaches we proposed in this work are viable enough to retrieve nested terms effectively without nested labeling of them.

FAROESE CORPUS DEVELOPMENT: STRATEGIES FOR LOW-RESOURCE LANGUAGES CORPORA

Satdarov K., **Kharlamova D.**, **Yakuboy A.**, **Lezina A.**, National Research University Higher School of Economics, Moscow, Russia

This paper presents the development of a comprehensive morphologically annotated corpus for Faroese, a low-resource Germanic language. We describe the creation of a large corpus of contemporary news texts automatically annotated using a custom-trained SpaCy model. The study demonstrates the effectiveness of creating linguistic resources for low-resource languages using minimal initial data. We trained a Transformer-based morphological parsing model on the small but high-quality OFT treebank using 5-fold cross-validation, achieving significant accuracy in morphological tagging and lemmatization. Manual evaluation confirms satisfactory performance of the automatic annotation, though certain challenges remain in distinguishing homonymous word forms across different parts of speech. This research provides a methodological framework for developing comprehensive linguistic resources for other low-resource languages with minimal initial data requirements.

COMPARING TRANSFORMER-BASED APPROACHES FOR TERM RECOGNITION IN RUSSIAN TEXTS

Semak V. V.[1], **Bolshakova E. I.**[1,2], [1]Lomonosov Moscow State University, Moscow, Russia; [2]HSE, Moscow, Russia

The paper describes an experimental comparative study of three transformer-based approaches applied for automatic term recognition (ATR) in Russian texts. The approaches include sequential labeling of word tokens in a given text, phrase classification with context sentence enclosing the phrase, and text span prediction by using vector representations obtained by contrastive learning. The BERT-based models were trained and evaluated on the data of RuTermEval-2024 competition for nested term identification and classification, which encompassed three tasks: binary term identification, term recognition with classification (into one of predefined types), and cross-domain term recognition. The experiments have shown that the span prediction models based on contrastive learning outperform the other models across all three RuTermEval tasks, but at the same time demonstrate the most significant decrease in quality in the cross-domain task.

"OKIE DOKIE, HERES THE NO-CAP TRUTH!": EVERYDAY RUSSIAN YOUTH SPEECH IN CORPUS REPRESENTATION (STRUCTURE AND APPLICATION OF THE ESC SOUND CORPUS)

Sherstinova T. Yu., **Melnik A. G.**, **Petrova I. A.**, **Azarevich K. I.**, **Melkozerova V. I.**, **Chepovetskaya S. V.**, Laboratory for Language Convergence, National Research University Higher School of Economics, St. Petersburg, Russia

The article focuses on the creation and potential applications of the Everyday Student Conversations Russian speech corpus (ESC corpus), referred to as KURS corpus in Russian. The corpus is being developed based on a modified methodology of the "one day of speech" recording approach, adopted from the ORD Corpus, with an emphasis on capturing contemporary youth speech. The article highlights the key aspects of the corpus creation process, with special attention to its structure, data collection and processing methodology, as well as the functionality of the corpus online demo version. The primary aim of the project is to explore linguistic changes in the youth environment, create a resource for scientific and applied research, and develop a big data collection of everyday speech for machine learning and advanced artificial intelligence systems. For linguists, the corpus offers unique opportunities to study sociolinguistic, pragmatic, and phonetic features of speech, making it a valuable tool for analyzing contemporary discourse.

EVALUATING THE PRAGMATIC COMPETENCE OF LARGE LANGUAGE MODELS IN DETECTING MITIGATED AND UNMITIGATED TYPES OF DISAGREEMENT

Shulginov V.[1,2], **Şimşek H. B.**[2], **Kudriashov S.**[1], **Randautsova R.**[1], **Shevela S. A.**[1,2], [1]HSE University, Moscow, Russia; [2]MIPT, Moscow, Russia

This study presents a framework for evaluating the effectiveness of language models (LLMs) in detecting disagreement across a wide range of pragmatic strategies, from mitigated forms to overt verbal aggression. Special attention is given to complex cases of implicit manifestations of irony and sarcasm, which pose significant challenges for both automated analysis and interpersonal communication. Experimental testing of LLMs was conducted in two types of tasks: binary classification for identifying disagreement and classification of specific strategies for its expression. The results showed that large multilingual models outperformed other models, especially in binary classification. However, models that focus primarily on the Russian language, such as GigaChat and YaGPT, tend to interpret irony and sarcasm more accurately and have a higher result density. Comparative analysis with human judgments revealed that, despite progress, the accuracy of sarcasm detection by LLMs still lags significantly behind human judgments. The results suggest a need for further optimization of LLMs to improve their pragmatic competence in real communicative situations.

AN APPROACH TO INFORMATION EXTRACTION FROM TEXTS OF A LIMITED SUBJECT DOMAIN BASED ON A CHAIN OF LARGE LANGUAGE MODELS

Sidorova E. A.[1], **Ivanov A. I.**[2], **Ilina D. V.**[1], **Ovchinnikova K. A.**[2], **Osmushkin N. M.**[2], **Sery A. S.**[1], [1]A.P. Ershov Institute of Informatics Systems; Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia; [2]Novosibirsk State University, Novosibirsk, Russia

The paper considers the approach for extracting the information from texts of limited domains of knowledge based on a chain of neural language models. The task is represented in the form of three subtasks solved sequentially: (1) term extraction and classification; (2) coreference resolution; (3) extraction of relations of entities named with the terms. The dataset was based on texts on computational linguistics from the Habr forum. In the markup for term classification and relation extraction, 17 classes of terms and 51 relations were used in accordance with the ontology of computational linguistics. Prompt chain-based methods were used to apply LLMs, where each next query to the LLM is based on the results of the previous step. Six types of prompt templates were developed: for extracting, classifying, verifying terms, extracting coreferential relations, relations specified by the ontology, and a specialized template for relations linking entities of the same class. Sentence-BERT, GPT-4 and Mistral-based models were used at different steps of the study; a comparison with the SFT approach (ruRoBERTa) was made; hybrid approaches that have shown the best results were also developed. For term extraction and classification, F1=0.77 was obtained, for coreference resolution—F1=0.897, and for relation extraction—F1=0.847.

LORUGEC: THE LINGUISTICALLY ORIENTED RULE-ANNOTATED CORPUS FOR GRAMMATICAL ERROR CORRECTION OF RUSSIAN

Sorokin A., MSU Institute for AI; Yandex, **Nasyrova R.**, MSU Institute for AI

We release LORuGEC—the first rule-annotated corpus for Russian grammatical error correction. The sentences in it are accompanied with the grammar rules governing their spelling. In total, we collected 48 rules with 348 sentences for validation and 612 for testing. LORuGEC appears to be challenging for open-source LLMs: the best F0.5-score is achieved by Qwen2.5-7B using two-stage finetuning and is only 50%. The closed YandexGPT4 Pro model achieves the score of 75%. Using a rule-informed retriever for fewshot example selection, we improve these scores up to 57% for Qwen and 81% for YandexGPT4 Pro.

GRADUAL ACCEPTABILITY JUDGMENTS WITH LLMS: EVIDENCE FROM AGREEMENT VARIATION IN RUSSIAN

Studenikina K. A., **Lyutikova E. A.**, **Gerasimova A. A.**, Lomonosov Moscow State University, Moscow, Russia

This study examines the LLMs' performance on gradual acceptability judgments task. Previously, the linguistic competence of LLMs was evaluated using binary acceptability scales, which contradicts the theoretical concept of acceptability. We present a new benchmark KVaS (Korpus Variativnogo Soglasovanija 'Corpus of Variable Agreement') derived from syntactic experiments on variable agreement in Russian. Our dataset contains multiple phenomena of agreement variation, ideal for modeling diverse acceptability levels, and compiles 7013 sentences rated by native speakers on a 1–7 Likert scale. We evaluated two LLMs, mainly Russian-trained GigaChat-Pro and multilingual Mistral Large, comparing their capability to treat acceptability as a scale to the reference human scores from KVaS. We used prompting providing benchmark sentences in two modes: zero-shot mode included only instructions while a few-shot mode added training sentences and their scores. The results show that GigaChat-Pro underperformed compared to Mistral Large. GigaChat-Pro improved significantly in a few-shot mode while Mistral Large exhibited more stable behavior. The case study shows that Mistral can detect nearly all significant contrasts in an experiment, whereas GigaChat performed near-randomly. The corpus may be useful for ranking LLMs, fine-tuning, and enhancing Russian text generation quality.

ASPECTUAL PAIRS IN RUSSIA IDIOMS: CONDITIONS ON AND MECHANISMS OF ASPECTUAL REDUCTION

Tatevosov S. G., Lomonosov Moscow State University, Moscow, Russia, **Kisseleva X. L.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This article explores the aspectual behavior of VP idioms in Russian. The goal of the study is to find out how evidence from idiomatic material can inform a theory of the structure and interpretation of the aspectual system. The study shows that systematic gaps are observed in aspectual pairs consisting of a non-derived imperfective and a pre-fixed perfective verb, but not in pairs of the form "perfective — secondary imperfective". According to our hypothesis, such gaps come about due to two distinct mechanisms. First, what we call reduction to the perfective, i.e., the absence of the imperfective variant of an idiom, obtains if the idiom includes a lexical prefix. Reduction to the imperfective, in contrast, is brought about by a shift in the eventuality type which result in there only being non-culminating eventualities in the extension of the idiom.

CLAUSE LINKERS FOR THE RUSCON DATABASE: SELECTING CRITERIA AND STATISTICAL EVALUATION

Timoshenko S. P.[1], **Serdobolskaya N. V.**[2], **Kobozeva I. M.**[3,2], [1]IITP, Russian Academy of Sciences, Moscow, Russia; [2]Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia; [3]Lomonosov Moscow University, Moscow, Russia

The study is devoted to compiling a list of clause linkers, or connectives, for the database Ruscon (available at http://ruscon.belyaev.io/). It is aimed at differentiating between complex connectives (like a ne to 'else') and occasional combinations of connectives (a vsledstvie etogo 'and henceforth'). There is a large number of cases when phonetic, morphosyntactic and semantic properties of connectives do not permit to make this differentiation. In these cases we propose to use the MMI metric (Modified Mutual Information). It rests on the rule of probability multiplication, which allows to estimate the probability of independent events.

STRUCTURED SENTIMENT ANALYSIS WITH LARGE LANGUAGE MODELS: A WINNING SOLUTION FOR RUOPINIONNE-2024

Vatolin A., FRC CSC RAS, Moscow, Russia

This paper presents the winning solution for the RuOpinionNE-2024 competition on structured sentiment analysis in Russian news texts. We propose a novel pipeline with large language models (LLMs) and adapter-based fine-tuning, demonstrating how modern LLMs can be effectively adapted to complex opinion tuple extraction tasks. Our method addresses three key challenges: (1) alignment of model predictions with original text spans through a fuzzy substring matching algorithm, (2) robustness to generation variability via multi-prediction aggregation strategies, and (3) efficient domain adaptation using QLoRA fine-tuning. The proposed approach achieved first place with a test F1 score of 0.405. Experimental results reveal that adapter-based fine-tuning of open-source 70B parameter models (Llama-3.3) surpasses prompt-engineered proprietary models like GPT-4o. Our analysis provides practical insights into adapting LLMs for structured information extraction in morphologically rich languages, showing that targeted fine-tuning with 4-bit quantization enables state-of-the-art performance without taskspecific architectures.

RUSCIFACT: OPEN BENCHMARK FOR VERIFYING SCIENTIFIC FACTS IN RUSSIAN

Vatolin A., FRC CSC RAS, Moscow, Russia, **Gerasimenko N.**, MSU Institute for Artificial Intelligence, Moscow, Russia, **Loukachevitch N.**, Lomonosov MSU, Moscow, Russia, **Ianina A.**, Moscow Institute of Physics and Technology, Moscow, Russia, **Vorontsov K.**, MSU Institute for Artificial Intelligence, Moscow, Russia

Against the backdrop of active LLM development, their tendency to hallucinate, as well as the growing volume of texts they generate, the validation of facts has become increasingly important and relevant. We propose ruSciFact, a new benchmark for fact-checking scientific claims in Russian. ruSciFact is structured as a NLI task, the goal is to verify whether a fact is confirmed by a given abstract. To generate facts, we used an 3-step pipeline based on LLaMA-405B, validating the resulting sentences with the help of assessors-terminologists. The ruSciFact dataset consists of 1128 pairs in the format <abstract, claim>, which we are releasing as open source together with the benchmark code. Additionally, we are open-sourcing the fact-generation pipeline, which facilitates the expansion of the dataset to specific scientific domains. We evaluated several popular language models on ruSciFact, including text embedders and generative models. The results show that this benchmark allows to effectively assess the fact-checking capabilities of LLMs in Russian.

UPSTEP IN RUSSIAN ECHO-QUESTIONS ABOUT EXPRESSIONS

Yanko T. E., Institute of Linguistics, Moscow, Russia

The paper gives an account of Russian echo questions about expressions. The prosodic upstep model is viewed as a means of expressing the illocutionary force of the echo questions about expressions. Upstep can be construed as the fundamental frequency (F_0) contour aligned with the segment material long enough to implement a gradual rise from one step of an echo question to another up to the end of the speech act. It is proposed to consider a generalized step of the upstep as a rise of the F_0 on the stressed syllable of a phonetic word followed by high post-tonics. At the same time, it is shown that the prosody of the echo question varies, in particular, it depends on the length of the segment material of an echo question. In addition, the upstep model can as well be considered in another meaning: as a function that converts the F_0 contour of a *wh*-question into the F_0 contour of the corresponding echo question. Achieving average F_0 values significantly higher than the average F_0 of any other speech act for each particular speaker is viewed as the "prosodic target" of the upstep model.

OPERATOR WORDS VERSUS MT SYSTEMS AND LLMS

Zimmerling A. V., Pushkin State Russian Language Institute; Institute of Linguistics, Russian Academy of Science,
Baiuk A. M., Russian State University for the Humanities

This paper investigates the meanings of two classes of Russian discourse words, defined by modal operators VER and AFF, and examines their translation equivalents in nine target languages using machine translation (MT) systems, large language models (LLMs), and human expert translations. VER (verification) indicates confirmation of a hypothesis, while AFF (affirmation) expresses a strengthened belief in a hypothesis. The study uses a set of 17 Russian discourse words and evaluates their translations in English, German, Danish, Swedish, Icelandic, Ukrainian, Bulgarian, Ossetic, and Arabic across 85 test sentences. The primary goal was to test the universality of the VER and AFF distinction, hypothesizing that these classes remain distinct across languages despite the lack of direct one-to-one translation equivalents. The study assumed that VER and AFF operators, corresponding to DE RE and DE DICTO attitudes respectively, differ fundamentally in semantics and distributional behavior. The study confirms the semantic and distributional independence of VER and AFF operators, supporting their universality. LLMs, despite not being specialized for MT tasks, showed remarkable adaptability and context awareness compared to traditional MT systems. The findings highlight the potential of LLMs in nuanced translation tasks and underscore the complexity of translating modal discourse words. Future work will explore custom models and further refine evaluation metrics for translation accuracy.

Авторский указатель

Азаревич К. И. 331
Алибеков А. 12
Аринкин Н. А. 183
Ахмадеева И. 1
Бальба Д. 147
Баранов А. Н. 38
Баюк А. 28
Баюк А. М. 471
Баюк И. 28
Богуславский И. М. 45
Большакова Е. И. 323
Большаков В. 12
Бонч-Осмоловская А. А. 58
Бородина Д. Г. 58
Буйлова Н. Н. 58
Буров Д. А. 74
Виноградова И. И. 58
Власова Н. А. 58
Воронцов К. 245
Гарипов Т. А. 59, 257
Герасимова А. А. 384
Гладилин С. А. 58
Глазкова А. 257
Глазкова А. В. 58
Горбова Е. В. 86
Громенко Е. 117
Губарькова Я. Н. 59
Дьяченко П. В. 58
Иванов А. И. 362
Ильина Д. В. 362
Инькова О. Ю. 129
Исаков Д. 147
Ищенко Р. 245
Казенников А. О. 58
Калачева Д. 117
Киосе М. 158, 281
Киселева К. Л. 393
Клокова К. 117
Кобозева И. М. 407
Ковалев Г. 195
Кожевников Е. 195
Козаченко А. 12
Козеренко А. Д. 58
Комратова А. 147
Кононенко И. 1
Корнилов М. 195
Коротаев Н. А. 169
Котов А. А. 183
Кронгауз М. 117
Кудряшов С. 345
Кузнецова Ю. Н. 58
Лапаницына А. М. 234
Лезина А. 312
Леонтьева А. 281
Ли Г. Ф. 206
Лукашевич Н. В. 195, 206, 234
Лютикова Е. А. 384
Ляшевская О. Н. 58
Мамонтова А. 245
Матенков А. 12
Махова А. А. 58
Мелкозерова В. И. 331
Мельник А. Г. 331
Мигаль А. 12
Михайловский Н. 12
Мороз О. 117
Морозов Д. А. 58, 257
Мурышев А. 12
Мухтарова Г. 12
Носовец З. А. 183
Овчинникова К. А. 362
Осьмушкин Н. М. 362
Панич М. Б. 74
Петров А. 158
Петрова И. А. 331
Петрова М. 28
Пискунова С. В. 58
Подлесская В. И. 267
Прокофьева О. 281
Рандовцова Р. 345
Ржешевская А. 158
Русначенко Н. Л. 234
Садковский Ф. А. 74
Сатдаров К. 312
Семак В. В. 323
Сердобольская Н. В. 407
Серый А. С. 362
Сидорова Е. А. 1, 362
Сизов В. Г. 58
Смаль И. А. 59
Смирнова Е. 281
Столяров С. С. 59
Студеникина К. А. 384
Татевосов С. Г. 393
Тимошенко С. П. 407
Тихомиров М. М. 195
Тихомиров М. М. 234
Ткаченко Н. М. 234
Федорова О. В. 74
Филатов А. А. 183
Харламова Д. 312
Хоменко А. 147
Хохлов А. Р. 206
Худякова М. 147
Циммерлинг А. В. 471
Чеповецкая С. В. 331
Чуйкова О. Ю. 86
Шевела С. А. 345
Шевелев С. М. 74
Шерстинова Т. Ю. 331
Шестаков В. 1
Шимшек Х. Б. 345
Шишковская Т. 147
Шульгинов В. 117, 345
Юдина Т. 117
Якубой А. 312
Янко Т. Е. 460

Author Index

Akhmadeeva I. 1
Alibekov A. 12
Arinkin N. A. 183
Azarevich K. I. 331
Baiuk A. M. 28, 471
Baiuk I. 28
Balba D. 147
Baranov A. N. 38
Boguslavsky I. M. 45
Bolshakova E. I. 323
Bolshakov V. 12
Bonch-Osmolovskaya A. A. 57
Borodina D. G. 57
Bujlova N. N. 57
Burov D. A. 74
Chepovetskaya S. V. 331
Chuikova O. Iu. 86
Dyachenko P. V. 57
Fedorova O. V. 74
Filatov A. A. 183
Galitsky B. 103
Garipov T. 257
Garipov T. A. 57
Gerasimenko N. 435
Gerasimova A. A. 384
Gladilin S. A. 57
Glazkova A. V. 57, 257
Gorbova E. V. 86
Gromenko E. 117
Gubar'kova Ya. N. 57
Ianina A. 435
Ilina D. V. 361
Ilvovsky D. 103
Inkova O. Yu. 129
Isakov D. 147
Ischenko R. 245
Ivanov A. I. 361
Ivleva I. 137
Kalacheva D. 117
Kazennikov A. O. 57
Kharlamova D. 312
Khokhlov A. R. 206
Khomenko A. 147
Khudyakova M. 147
Kiose M. 158, 281
Kisseleva X. L. 393
Klokova K. 117
Kobozeva I. M. 407
Komratova A. 147
Kononenko I. 1
Kornilov M. 195
Korotaev N. A. 169
Kotov A. A. 183
Kovalev G. 195
Kozachenko A. 12
Kozerenko A. D. 57
Kozhevnikov E. 195
Krongauz M. 117
Kudriashov S. 345
Kuznetzova Y. N. 57
Lapanitsyna A. M. 234
Lee G. F. 206
Leonteva A. 281
Lepekhin M. 224
Lezina A. 312
Loukachevitch N. V. .....
195, 206, 234, 299, 435
Lyashevskaya O. N. 57
Lyutikova E. A. 384
Makhova A. A. 57
Mamontova A. 245
Matenkov A. 12
Melkozerova V. I. 331
Melnik A. G. 331
Migal A. 12
Mikhaylovskiy N. 12
Morkovkin A. 103
Moroz O. 117
Morozov D. A. 57, 257
Mukhtarova G. 12
Muryshev A. 12
Nasyrova R. 374
Nosovets Z. A. 183
Osmushkin N. M. 361
Ovchinnikova K. A. 361
Panich M. B. 74
Petrov A. 158
Petrova I. A. 331
Petrova M. 28
Piskounova S. V. 57
Podlesskaya V. I. 267
Prokofyeva O. 281
Randautsova R. 345
Rossyaykin P. 291
Rozhkov I. S. 299
Rusnachenko N. L. 234
Rzheshevskaya A. 158
Sadkovskii F. A. 74
Satdarov K. 312
Semak V. V. 323
Serdobolskaya N. V. 407
Sery A. S. 361
Sharoff S. 224
Sherstinova T. Yu. 331
Shestakov V. 1
Shevela S. A. 345
Shevelev S. M. 74
Shishkovskaya T. 147
Shulginov V. 117, 345
Sidorova E. A. 1, 361
Şimşek H. B. 345
Sizov V. G. 57
Smal I. A. 57
Smirnova E. 281
Solovyev V. 137
Sorokin A. 374
Stolyarov S. S. 57
Studenikina K. A. 384
Tatevosov S. G. 393
Tikhomirov M. M. 195, 234
Timoshenko S. P. 407
Tkachenko N. M. 234
Vatolin A. 416, 435
Vinogradova I. I. 57
Vlasova N. A. 57
Vorontsov K. 245, 435
Yakuboy A. 312
Yanko T. E. 460
Yudina T. 117
Zimmerling A. V. 471

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 23, 2025

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**