ONLINE: ISSN 2075-7182

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной международной конференции «Диалог» (2023)

Выпуск 22

Computational Linguistics and Intellectual Technologies

Papers from the Annual International Conference “Dialogue” (2023)

Issue 22

Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной международной конференции «Диалог». Вып. 22. 2023. С. I–602.

Сборник включает 54 доклада международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог 2023», 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.

Предисловие

22-й выпуск ежегодника «Компьютерная лингвистика и интеллектуальные технологии» содержит избранные материалы 29-й международной онлайн-конференции «Диалог». В 2023 году для публикации в основном томе сборника редколлегией были отобраны 54 доклада из 120, поданных на конференцию. 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 (Intelligent Document Processing):** классификация, Name Entity & Relation Extraction, саммаризация, генерация, анализ тональности, Argumentation Mining, Propaganda & Fake News Detection, etc., мультимодальные подходы (совместное использование моделей NLP и Computer Vision);
- **Глубокое обучение в компьютерной лингвистике:** методики применения нейронных сетей в исследованиях, содержательная интерпретация;
- **Компьютерные лингвистические ресурсы:** 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;
- Лингвистический анализ текста: морфология, синтаксис, семантика (модели анализа);
- Компьютерная лексикография;
- **Полевая компьютерная лингвистика:** применение методов NLP для малоресурсных языков.

В соответствии с традициями «Диалога», конференции по компьютерной лингвистике с почти полувековой историей, 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 этом году состоялись 4 соревнования:

- **RuCoCo:** Соревнование по разрешению кореференции;
- **RuSentNE:** Соревнование по анализу тональности к именованным сущностям в новостных текстах;
- **RECEIPT-AVQA:** Соревнование по генерации ответов на вопросы к изображениям;
- **SEMarkup:** Соревнование по автоматической семантической разметке.

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 www.dialog-21.ru, 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

Азарова Ирина Владимировна
Андрианов Андрей Иванович
Антонова Александра Александровна
Баранов Анатолий Николаевич
Беликов Владимир Иванович
Богданова-Бегларян Наталья Викторовна
Богуславский Игорь Михайлович
Бурцев Михаил Сергеевич
Васильев Виталий Геннадьевич
Гусев Илья Олегович
Добров Борис Викторович
Добровольский Владимир Андреевич
Добровольский Дмитрий Олегович
Жарков Андрей Алехандрович
Зализняк Анна Андреевна
Захаров Леонид Михайлович
Золотухин Денис Денисович
Иванов Владимир Владимирович
Ивойлова Александра Михайловна
Ильвовский Дмитрий Алексеевич
Инденбом Евгений Михайлович
Инькова Ольга Юрьевна
Иомдин Леонид Лейбович
Киосе Мария Ивановна
Клышинский Эдуард Станиславович
Клячко Елена Леонидовна
Князев Сергей Владимирович
Кобозева Ирина Михайловна
Козеренко Елена Борисовна
Копотев Михаил Вячеславович
Коротаев Николай Алексеевич
Котельников Евгений Вячеславович
Котов Артемий Александрович
Куратов Юрий Михайлович
Кутузов Андрей Борисович
Лапошина Антонина Николаевна
Левонтина Ирина Борисовна
Лобанов Борис Мефодьевич
Логинов Василий Васильевич
Лукашевич Наталья Валентиновна
Малафеев Алексей Юрьевич
Митрофанова Ольга Александровна
Мичурина Мария Александровна
Недолужко Анна
Никишина Ирина Юрьевна
Орлов Евгений Анатольевич
Пазельская Анна Германовна
Переверзева Светлана Игоревна
Петрова Мария Владимировна
Подлесская Вера Исааковна
Рыгаев Иван Петрович
Селегей Владимир Павлович
Слюсарь Наталия Анатольевна
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
Шамардина Татьяна Вячеславовна
Шаров Сергей Александрович
Янко Татьяна Евгеньевна

Contents[1]

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

Receipt-AVQA-2023 Challenge

Artur Begaev
Budapest, Hungary
artur.begaev@aol.com

Evgeny Orlov
Budapest, Hungary
eugene.a.orlov@gmail.com

Abstract

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task. Receipt AVQA is comprised of 21835 questions in English over 1957 receipt images. The receipts contain a lot of numbers, which means discrete reasoning capability is required to answer the questions. The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition. We hope that the published dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

Keywords: VQA, computer vision, multimodal dataset
DOI: 10.28995/2075-7182-2023-22-1-11

Соревнование Receipt-AVQA-2023

Артур Бегаев
Будапешт, Венгрия
artur.begaev@aol.com

Евгений Орлов
Будапешт, Венгрия
eugene.a.orlov@gmail.com

Аннотация

В данной работе мы представляем новый датасет для задачи VQA, названный Receipt AVQA, и результаты проведенного соревнования RECEIPT-AVQA-2023. Датасет Receipt AVQA состоит из 21835 вопросов на английском языке к 1957 изображениям товарных чеков. Товарные чеки содержат большое количество числовой информации, что требует опредленной степени дискретного мышления для ответа на вопросы. Сопровождающее датасет соревнование привлекло 4 команды, которые смогли улучшить результаты по сравнению с базовой экстрактивной VQA моделью. Мы надеемся, что опубликованный датасет и многообещающие результаты участников соревнования вдохновят дальнейшие исследования в области автоматического понимания изображений документов в сценариях, где требуется дискретное мышление.

Ключевые слова: VQA, компьютерное зрение, мультимодальный датасет

1 Introduction

Receipt understanding is an important problem, which has to be solved in many applications. For example, customers want to analyze the prices of positions and total paid money, extract information about quantities of products, and how individuals could adjust their budget as so to purchase the needed amount of goods. In addition, most people would like to ask questions about these properties without the usage of any APIs or complicated computational programs. There are no such existing datasets, which cover the issue of answering the natural language questions over a receipt. In this paper, we propose such a dataset based on SROIE (Huang et al., 2019) and CORD (Park et al., 2019) receipt datasets, which is expected to cover and reveal problems with existing solutions to the question answering tasks.

VQA (Antol et al., 2015) is quite a novel task in the machine learning domain. In order to resolve such an issue a solution should combine approaches from both computer vision (process an image) and

natural language processing domains (process a question). However, most of the existing datasets focus on real-life photos; documents in general, such as invoices, industry documents, food and nutrition-related collections; and screenshots (Patadia et al., 2021). These datasets provide a markup of layouts for images or bounds of the objects and relations between the objects in an image. Meanwhile, the questions in these datasets are mostly formulated in an extractive manner, meaning that required answers are already presented on an image.

There is a special subset of such datasets – TextVQA (Singh et al., 2019). The markup for this subset also contains the recognized text (OCR Tokens) from the scene on an image. TextVQA datasets require some reasoning about the placing and the nature of OCR tokens. However, the extraction of such answers doesn't involve complex mathematical reasoning or calculations.

In this challenge, we present a new dataset: Receipt AVQA Dataset, comprising 21837 questions over 1957 images. By introducing this dataset we want to stress an important problem in the TextVQA task: extracting and calculating answers from the data presented on the image of a receipt. This task is not so trivial, because most of the current state-of-the-art methods for the TextVQA problem mostly focus on the extraction of the answers without any calculations using extracted tokens.

VQA tasks commonly contain questions about the objects in an image or the relations between them. In the TextVQA datasets questions about text tokens are offered. Solutions are required to extract the requested tokens from an image following the rules defined in a question. Our dataset also offers a new challenge – a solving model has to make calculations and aggregations over the extracted text tokens when requested in a question. We used the receipts from two datasets, in which the scales of numbers are different. Expected solutions should take into account this variation in the scaling and produce a required numerical answer.

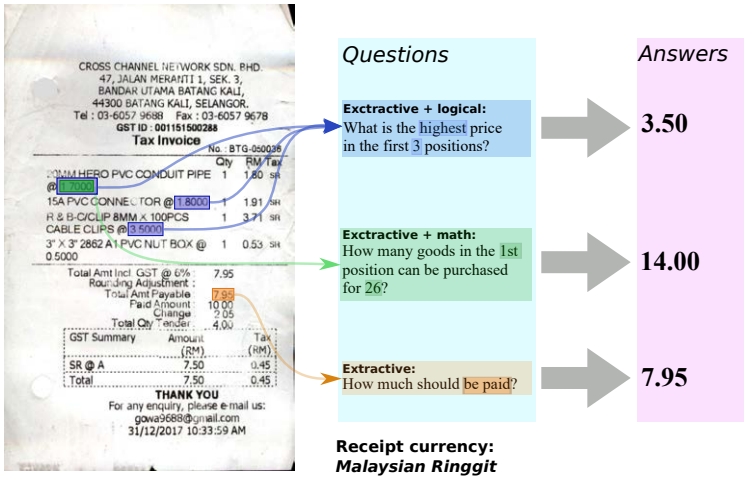


Figure 1: An example of a receipt with questions from the Receipt AVQA dataset

Moreover, our dataset introduces several types of questions like mathematical, logical, and finding the ratio between some values. The complete scheme for one receipt and some questions for that is represented in Figure 1.

The main contributions of this work can be summarized as follows:

- We introduce Receipt AVQA, a dataset of 1957 receipt images, over which we have defined 21837 questions and answers (§3) [1];
- We introduce a baseline solution for the shared task (§5);
- We conduct an analysis of the received submissions for both sub-tasks (§6) and discuss potential research directions (§7);
- We set up the shared task environment, which remains open for community submissions to facilitate future research in the area (§4.2)[2].

[1] https://github.com/dialogue-evaluation/Receipt-AVQA-2023
[2] https://codalab.lisn.upsaclay.fr/competitions/11087

2 Related Work

ST-VQA (Biten et al., 2019) and TextVQA (Singh et al., 2019) datasets extend VQA over natural images to a new direction where understanding scene text on the images is necessary to answer the questions.

OCR-VQA (Mishra et al., 2019) introduces a task similar to ST-VQA and TextVQA, but instead of natural images, images of book covers are used. Template questions are generated from book metadata such as author name, title, and other information.

Related to OCR-VQA, DVQA (Kafle et al., 2018), FigureQA (Kahou et al., 2017), and LEAF-QA (Chaudhry et al., 2019) are VQA datasets that operate on various chart images, InfographicVQA (Mathew et al., 2021) introduces VQA dataset of infographics images.

DocVQA (Mathew et al., 2020) dataset shifts the image domain to documents completely. DocVQA is a VQA dataset that is comprised of the document images of industry/business documents, and questions requiring understanding document elements such as text passages, forms, and tables. Similarly to most aforementioned VQA datasets, DocVQA is focused on *extractive questions*, where answers can always be extracted verbatim from a text on the images.

To our best knowledge, there are few VQA datasets operating on documents that are focused on the *abstract questions*, where answers cannot be directly extracted from text in the images or questions.

FiqureQA is comprised of abstract questions, but an answer to any question in the dataset is limited to yes/no.

InfographicVQA contains some questions that require certain discrete operations resulting in numerical non-extractive answers. These discrete operations are limited to counting and sorting and are employed only in 20% of questions in the dataset.

VisualMRC (Tanaka et al., 2021) dataset is built for the abstract question answering. The dataset employs the screenshots of web pages and questions that don't involve operating on numericals for answers.

Most similar to Receipt-AVQA dataset is TAT-DQA (Zhu et al., 2021) dataset. TAT-DQA comprises the high-quality images of financial reports. In order to answer the questions in the dataset a wide range of operations on numericals is required. The distinct feature of Receipt-AVQA stems from the document type employed. Receipt images on average contain fewer text tokens than financial reports but a higher quantity and relative share of numerical tokens on the image.

In Table 1 we present a high-level summary of Document VQA datasets related to ours.

Dataset	Images	Synthetic Images	Template Questions	# Images	# Questions	Answer type
DocVQA	Industry documents	No	No	12K	50K	Ex
FigureQA	Charts	Yes	Yes	120K	1.5M	Y/N
InfographicVQA	Infographics	No	No	5.4K	30K	Ex, Nm
VisualMRC	Webpage screenshots	No	No	10K	30K	Ab
TAT-DQA	Financial reports	No	No	3K	16.5K	Nm, Ex
Receipt-AVQA	Receipts	No	Yes	2K	21.8K	Nm, Ex

Table 1: Summary of Document VQA datasets. Answer type abbreviations are: Extractive: Ex, Abstractive: Ab, Yes/No: Y/N, and Numerical (the answer is numerical and not extracted from image or question; but derived): Nm.

3 Dataset and Shared Task

In this section, we present the definition of the RECEIPT-AVQA-2023 task, the construction of the Receipt AVQA dataset, and the statistical analysis of the dataset.

3.1 Task Definition

The challenge has a focus on question answering for receipts. Two tracks are offered:

1. VQA Track - question answering over the images of receipts.

2. QA Track - in this track in addition to receipt images the solutions can use the ground-truth text tokens with their corresponding coordinates extracted from the images (essentially, error-less OCR output).

The competition is formulated as a regression task. Unlike the vast majority of VQA tasks where the answer is a string, each answer is a float number here, and we use a metric operating on numbers to score submitted solutions.

To come up with a correct answer, the VQA model needs not only to recognize and extract tokens from the receipt image but to apply a number of operations (e.g. sorting, counting, arithmetic operations) over it. As a result, discrete reasoning capability is required from a potential solution.

All the questions are in English and can be divided into several types. Answering each question can be done independently for each type, however, we expected from participants to build a solution to answer questions in an end-to-end manner.

In our challenge we propose the following types of questions:

- Amount – finding and extracting required values, sometimes with some aggregation: "How much should be paid?", "What is the average price of a position?"
- Count – counting or extracting the number of elements of some type (eg. positions, a changed amount of goods): "How many positions were bought?", "How many goods are in the 1st position?"
- Ratio – finding a ratio between required values: "What is net total and total amount ratio?", "What share of cash was returned as change?"

To make our task easier, we offer a list of operations for each question. Explicit formulas are not provided. Types of proposed operations are the following: division, sorting, subtraction, summation, counting, and multiplication. A question can contain zero or several operations.

The values of prices are scaled differently because we made our dataset based on SROIE and CORD. These datasets contain receipts from different countries. We introduced the currencies of receipts as follows: "Malaysian ringgit", "Indonesian rupiah".

We split our dataset into train, validation, and test subsets in the following manner:

- Train: 16611 questions over 1537 images
- Validation: 2302 questions over 210 images
- Test: 2924 questions over 210 images

The splits in Receipt AVQA are consistent with the splits in SROIE and CORD datasets.

3.2 Dataset construction and verification

Original data from SROIE and CORD is not labeled for the VQA task: it doesn't contain any class labels for the fields on the receipt. We developed a method that allowed us to introduce the necessary labels in a semi-automatic way.

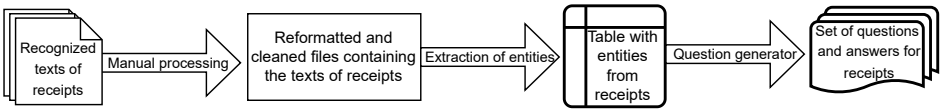


Figure 2: The scheme of data processing and question generation

We used recognized texts with bounding boxes made by authors of SROIE and CORD. In the first step all unnecessary data, like names of the shops, their addresses, and telephone numbers, was cleaned from the files. Then these files were used to extract the needed entities for our dataset: positions with prices and amounts, key-value pairs, containing, information about paid amounts, discount amounts and etc.

We used special heuristics for automatic data processing. After that, the information extracted for each receipt was reviewed and cleaned in order to contain valid data. SROIE and CORD were processed independently. So, a special table format was developed to merge the entities from both datasets. This table is intermediate and is not available for challengers.

The compilation of such a table made us available to generate the questions for receipts. We made up 13 types of questions for the positions in receipts and 9 types of questions for a total section of each

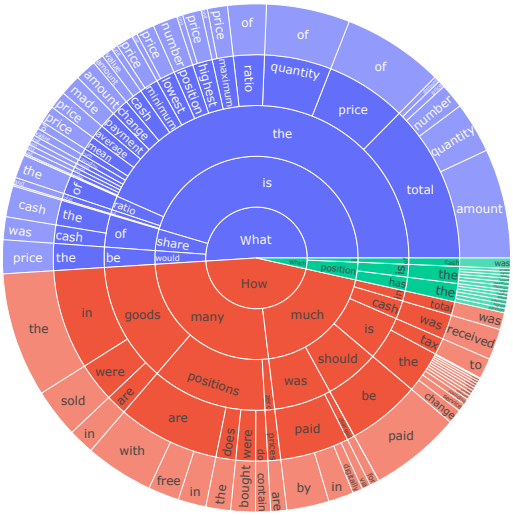


Figure 3: The word map of questions for the first five words.

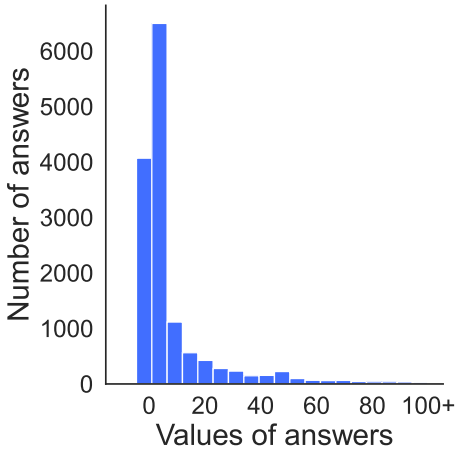


Figure 4: The distribution of answers lower than 100

receipt. For each kind of questions, several templates were prepared (from 3-5, depending on the type of question). Each template could be varied by several parameters: the numbers inside a question, the type of aggregation, and a type of value that is indented to be extracted. Some questions require calculations to get an answer. We obtained the answers explicitly by defining the needed formulas. However, applied formulas and functions are not provided in our dataset so the difficulty of provided tasks was increased.

Questions and answers have to be validated. Markup, entity extraction, and heuristics could introduce mistakes in the reasoning of getting the answers. Unfortunately, most part of the generated questions cannot be validated automatically. Still, the extractive questions for the total sections can be validated by calculating the prices of positions and summing them up. Other questions were validated by a human in several iterations.

Firstly, we prepared a large subset of our questions for the detection of mistakes in the extraction of answers. Answers and questions were validated manually by looking at a sample and checking the answers provided by the generative algorithm. This step revealed some issues with entity extraction and question generation. Some templates were reworked due to inappropriate or invalid formulations. After fixing the errors smaller subset was extracted. The same approach was applied to the new subset. We continued such iterations until there were no revealed mistakes in the dataset.

3.3 Dataset statistics and analysis

Hereby we provide some statistics over generated questions and answers. Challengers were awaited to do this by themselves in order to reveal hidden problems, which their models were intended to deal with.

In Figure 3 it can be seen that most of the questions start with the words "What" and "How" and the smallest part starts with the word "Which". There are a lot of questions that require getting some value by its extraction or calculation.

A hidden issue in our dataset was introduced – unbalanced data. Figure 4 shows the distribution of the values of answers lower than 100. Answers over 100 were omitted for more reasonable visualization. This distribution has a heavy tail – there are a lot of questions with an answer lower than 10 and many of them have an answer between 0 and 1. Extractive models should take into account this problem and resolve it properly.

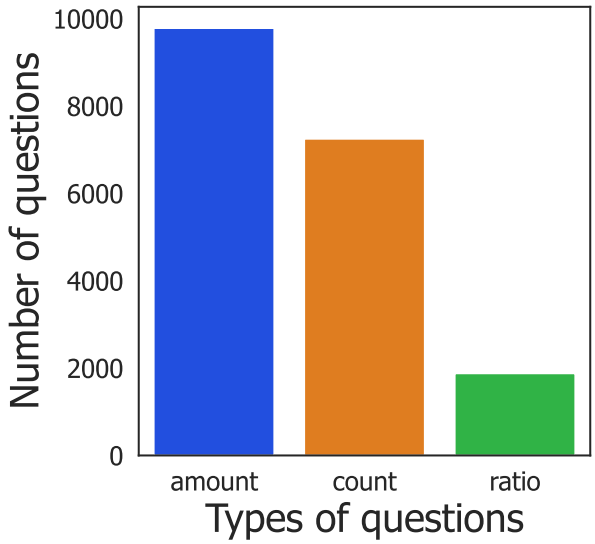


Figure 5: The distribution of questions by types

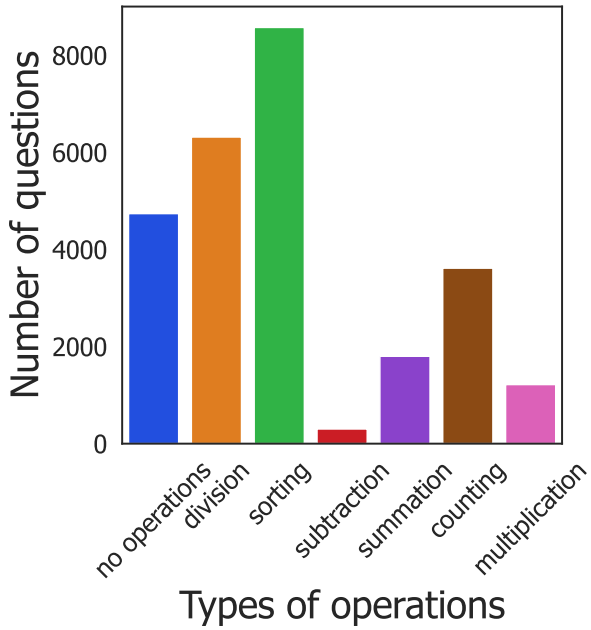


Figure 6: The distribution of questions by operations

Figure 5 provides an intuition on the number of questions from every proposed category. Most of the questions require answering counting questions and questions that ask for a specified amount of some kind.

4 Evaluation

In this section, we discuss our primary evaluation metric for the shared task and organization of the evaluation process.

4.1 Evaluation metric

Unlike traditional VQA tasks, where answers to questions are usually text, all answers in the Receipt AVQA dataset are real numbers. Consequently, both tracks in the RECEIPT-AVQA-2023 shared task are treated as a regression problems.

Mean absolute scaled error (MASE) (Hyndman and Athanasopoulos, 2013) metric is adopted as the primary metric for the shared task.

$$MASE = \frac{\frac{1}{N}\sum_{n=1}^{N}|QT_n - PR_n|}{\frac{1}{N}\sum_{n=1}^{N}|QT_n - \overline{QT_{train}}|}, \quad (1)$$

where N is a number of questions, QT_n is a real answer for the n-th question, PR_n is a predicted answer for the n-th question, and $\overline{QT_{train}}$ is an average of real answers on the train split of the dataset.

MASE metric was selected due to scale invariance, symmetry, and predictable behavior near 0.

One possible interpretation of the metric would be the ratio of the mean absolute error of a model to the mean absolute error of the baseline model, which is merely an average of the answers for the training dataset.

To get a better baseline model evaluation, questions are divided into 6 groups based on the receipt currency ("Indonesian rupiah", "Malaysian ringgit") and question type ("amount", "count", "ratio").

The 6 MASE values are then averaged, so the weights of all question types were equal.

In addition to MASE metric, we also use a more traditional accuracy metric, but with 10% leeway to account for possible rounding errors ($Acc_{\pm 10\%}$ metric).

4.2 Evaluation platform

We use the CodaLab (Pavao et al., 2022) competition platform to run the shared task.

Participation is allowed on either individual or team basis in both sub-tasks. The shared task consists of two stages: public and private testing. The first stage provides access to the public validation set and the leaderboard, allowing the participants to develop and improve their submissions during the competition. The second stage defines the final leaderboard ranking on the private test set, scoring up to twelve submissions selected by the participants. Participants are allowed to use any additional materials and pre-trained models, except for direct markup of the test set and looking for answers on the Internet.

5 Baseline

In this section, we describe the baselines we evaluated on the Receipt AVQA. These include heuristic baselines and upper bounds, and a document understanding model that was adopted as a baseline for the shared task.

5.1 Heuristics and Upper bounds

Heuristic baselines and upper bounds we evaluate are similar to the ones evaluated in other VQA benchmarks like DocVQA, and InfographicVQA.

Heuristic Baselines. The following heuristics were evaluated.

- *Random OCR number* measures performance when a random number from OCR results for the receipt image is picked as the answer;
- *Majority answer* measures performance when the most frequent answer in the train split is considered as the answer.

Upper Bounds. We also compute the following upper bounds:

- *Vocab UB* measures the upper bound on performance if the answer is predicted correctly, provided it is in the vocabulary of most common answers (> 1) of the train split;
- *OCR UB* measures the upper bound on performance if the answer is predicted correctly, provided it is one of the text tokens present on the corresponding receipt;
- *Vocab + OCR UB* measures the upper bound on performance if the answer satisfies either *Vocab UB* or *OCR UB*.

In the calculation of upper bounds, if the correct answer is not found within the defined scope, the mean of the corresponding question type answers on the train set is used instead.

The results of heuristic baselines and upper bounds are shown in Table 2.

Baseline	$MASE$ **val**	$MASE$ **test**	$Acc_{\pm 10\%}$ **val**	$Acc_{\pm 10\%}$ **test**
Random OCR number	1e11	3e11	6.17%	5.81%
Majority answer	0.8576	0.9662	18.64%	17.44%
Vocab UB	0.4420	0.5277	82.54%	81.67%
OCR UB	0.8216	0.8175	51.56%	47.02%
Vocab + OCR UB	0.4184	0.4812	88.44%	87.10%

Table 2: Results of heuristics and upper bounds.

5.2 VQA Baseline

It is not obvious whether SOTA methods for the TextVQA task, such as LayoutLMv3 (Huang et al., 2022), would be enough for solving the proposed tasks without any heavy modifications. We expect a model that can do some reasoning and calculations on extracted tokens. In this work, we want to present LayoutLMv3 as the baseline for the challengers to beat by introducing novel modules and methods for finding the relations between questions and texts from a receipt. We intentionally focus on the extraction questions in order to find out competitors who had beaten our solution.

LayoutLMv3 is a visual transformer that accepts bounding boxes, an image, and a tokenized text. We use pre-trained LayoutLMv3 from Hugging Face (Hug,) with LayoutLMv3TokenizerFast as a tokenizer. Still, it is not suitable to use for our dataset without some modifications.

Data preprocessing. LayoutLMv3 requires the recognized text from an image, we use PaddleOCR (Pad,) for the text extraction with default settings. Our task was defined as regressive, so it is expected to extract numbers, not strings, both from images and questions. Still, most of the TextVQA models treat answers as strings. Solutions for the TextVQA datasets extract the answers by using OCR tokens. We made a simple lookup algorithm for finding an answer in OCR tokens. This algorithm matches the decimal, the whole and fractional parts of a number. We treat an OCR token as matching to the answer when the absolute error between the token converted to a float and the answer itself equals zero. Perhaps, the OCR algorithm is not perfect and introduces some errors which our algorithms fail to resolve. Unresolved answers are encoded by a special token, which should be extracted by the model if the answer wasn't found.

Model training. A token prediction head was added on top of the LayoutLMv3. This head consists of 2 fully-connected layers with LeakyReLU between and Dropout before and after the first layer. The first layer outputs 768 features, and the second outputs 2 features - predicting is a token the answer or not. For the optimization, we used Adam with $learning\_rate = 0.00001$. We trained our model for 50 epochs, but the best result (by loss) was achieved on the 9th epoch. In order to train the model to select an appropriate token we use cross-entropy loss.

Answer extraction. The answer decoding is not as straightforward as could be expected. The tokenizer splits a string by punctuation marks and spaces, although we try to extract only one OCR token without finding spans. These factors introduce some complications to our algorithm for the extraction of answers. So, the output from our model represents a sequence from 2-dimensional vectors. Because we trained our model for the binary classification, we can extract the needed tokens after the tokenization by finding all positions in the sequence where the number in the last dimension is greater than 0.5. We concatenate these tokens to get a complete token representing the extracted answer. However, when the model fails to get the answer our post-processing outputs a special token. In order to get an answer to such questions we had pre-calculated means for each type of question and currency. When this unwanted situation occurs we put the mean defined by the type of question as the answer.

Results. Such a simple approach doesn't provide good results. Especially for the questions requiring calculations or aggregations.

$MASE$ **Total**	$MASE$ **Amount**	$MASE$ **Count**	$MASE$ **Ratio**	$Acc_{\pm 10\%}$
0.8786	0.8068	0.8291	1.0000	14.60%

Table 3: The values of metrics produced by our LayoutLMv3 model.

As it can be seen in Table 3 our model completely fails in the Ratio questions and doesn't provide precise results for the Amount and Count questions. This leads to the conclusion that a good model, which can properly solve our task, should inherit some architectural and methodical properties from the state-of-the-art TextVQA models. Thus, further modifications, such as computational trees over the extracted tokens or specific rules for token processing, are required.

6 Submitted solutions

The final phase of the Receipt-AVQA-2023 shared task attracted 5 participants. We provide brief descriptions of the 3 solutions, which outperformed the baseline for at least one track. We denote each team by their CodaLab user names. In case of multiple submissions from one team, we report only the best result. The scores of the teams are shown in Table 4.

surkov_evseev The solution of the team is based on two core pipelines, one for extraction and structuring of the information from receipt images and another for translating each question into a mathematical expression. To extract the textual information from an image fine-tuned PP-OCRv3 (Li et al., 2022) /

TrOCR (Li et al., 2021) pipeline is employed. The extracted text data is then tagged via finetuned BERT (Devlin et al., 2019) models to establish a standardized structure for the receipt. A finetuned T5 (Raffel et al., 2019) model is employed for parsing question text into a mathematical expression required for the answer calculation. The final answer is then derived based on the structured receipt data and the mathematical expression for the question. The team had to build auxiliary markup using custom rules to train their BERT and T5 models.

s231644 The team manually converted all questions types into mathematical expressions, and built a multimodal UDOP (Tang et al., 2022) model to predict sequence of operands and operands types for these mathematical expressions. If the output of the UDOP model is inconsistent with the mathematical expression for the question, a fallback T5 model is employed instead, which tries to answer the question directly without parsing it into a mathematical expression. The sequences in models operate on character level. The models require text tokens from the image as an input; these text tokens are extracted via custom OCR pipeline. The team had to annotate the dataset to create custom labels for the UDOP model.

daniyallaiev The participant created a separate solution for each question type in the dataset. For the *ratio* question type a finetuned multimodal LayoutLMv3 model outputs the numerator and the denominator tokens of the answer. For the *amount* question type a multimodal LayoutLMv3 model is trained to output the best suitable token for the answer. LayoutLMv3 models work in extractive fashion by trying to select the most appropriate tokens for the answer among all OCR'ed tokens available to the models as input. The pipeline for the *count* question type is different. Firstly, a BERT model is used to extract key tokens from the question text that are required to answer the question. If no token is found, a fallback ViT (Dosovitskiy et al., 2020) model is used to predict the answer directly. If the tokens are found, they are used as an extra input to another LayoutLMv3 model, which tries to output operand tokens required to answer the question. The participant had to use custom rules to annotate labels to finetune the BERT model which processes question text and to extract operand tokens for each question with the *count* type.

7 Results and discussion

Track	Solution	$MASE$	$Acc_{\pm 10\%}$
QA	surkov_evseev	**0.1164**	**91.45%**
VQA	s231644	**0.2165**	**90.08%**
VQA	surkov_evseev	0.2331	81.91%
VQA	fireee80	0.2652	86.15%
VQA	daniyallaiev	0.7874	25.31%
VQA	LayoutLMv3 baseline	0.8786	14.60%
VQA	poddiving	6892	10.53%

Table 4: The shared task results on the test dataset. The best results for each track are in **bold**.

We report the shared task results for both tracks in Table 4.

Our observations from the results table are the following.

- Only a single team hasn't managed to beat an extractive VQA baseline;
- The best models according to $MASE$ also perform the best according to the accuracy based metric $Acc_{\pm 10\%}$;
- The need to apply OCR to extract textual information from images puts significant pressure on the quality of question answering system as highlighted by the drop in metrics of the *surkov_evseev* solution going from the QA track to the VQA track.

The submitted solutions share certain commonalities:

- The performance of the solutions adapted to the task significantly exceeds the performance of a generic extractive VQA model, which hopefully indicates potential for further research in the area;
- The participants have to rely on a preliminary OCR step to explicitly extract text data; building an end-to-end OCR-free solution remains unattainable in practical setting;

- As expected, the participants turn to pretrained VQA models to improve solution performance; building a custom network architecture given the limited size of the dataset remains challenging;
- For practical reasons all the participants instead of doing arithmetic operations on numbers within neural network computations decided on predicting operands involved in calculation / mathematical expression of the calculation and doing the required operations in the post processing step; the ways on how operations on real numbers can be incorporated into the compute within neural networks remains under-explored.

One potential extension of the Receipt AVQA dataset would be the addition of explicit calculation steps for obtaining answers to questions. This should significantly decrease the need for custom annotation efforts when building the models using the dataset and facilitate creation of end-to-end models that require less post processing.

Another interesting direction of future work is expanding the number of templates used for QA generation in order to encourage building solutions that try to estimate required calculation steps automatically rather than rely on predefined rule-based formulas.

8 Conclusion

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task.

Receipt AVQA is comprised of 21835 questions over 1957 receipt images. The questions in the dataset are formulated in a way, that to answer them, VQA solution needs not only to recognize and extract tokens from the receipt image, but to apply a number of operations (e.g. sorting, counting, arithmetic operations) over it, thereby testing discrete reasoning capability of the solution.

The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition.

We hope that the dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

Acknowledgements

The authors gratefully thank Vasily Loginov for leading contribution in the Receipt AVQA dataset annotation and verification effort and valuable input in related discussions.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. *CoRR*, abs/1505.00468.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marcal Rusinol, Ernest Valveny, C.V. Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. *// Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October.

Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. 2019. Leaf-qa: Locate, encode & attend for figure question answering. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, P 3501–3510.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. Icdar2019 competition on scanned receipt ocr and information extraction. *// 2019 International Conference on Document Analysis and Recognition (ICDAR)*, P 1516–1520.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. *Proceedings of the 30th ACM International Conference on Multimedia*.

Hugging face – the ai community building the future. https://huggingface.co. Accessed: 2023-04-06.

Rob J Hyndman and George Athanasopoulos. 2013. Forecasting: principles and practice.

Kushal Kafle, Scott D. Cohen, Brian L. Price, and Christopher Kanan. 2018. Dvqa: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 5648–5656.

Samira Ebrahimi Kahou, Adam Atkinson, Vincent Michalski, Ákos Kádár, Adam Trischler, and Yoshua Bengio. 2017. Figureqa: An annotated figure dataset for visual reasoning. *ArXiv*, abs/1710.07300.

Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021. Trocr: Transformer-based optical character recognition with pre-trained models. *ArXiv*, abs/2109.10282.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoyue Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, Dianhai Yu, and Yanjun Ma. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *ArXiv*, abs/2206.03001.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. V. Jawahar. 2020. Docvqa: A dataset for vqa on document images. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, P 2199–2208.

Minesh Mathew, Viraj Bagal, Rubèn Pérez Tito, Dimosthenis Karatzas, Ernest Valveny, and C.V. Jawahar. 2021. Infographicvqa. *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, P 2582–2591.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. *// 2019 International Conference on Document Analysis and Recognition (ICDAR)*, P 947–952.

Paddleocr: Awesome multilingual ocr toolkits based on paddlepaddle. https://github.com/PaddlePaddle/PaddleOCR. Accessed: 2023-04-06.

Seunghyun Park, Seung Shin, Bado Lee, Junyeop Lee, Jaeheung Surh, Minjoon Seo, and Hwalsuk Lee. 2019. Cord: A consolidated receipt dataset for post-ocr parsing.

Devika Patadia, Shivam Kejriwal, Richa Shah, and Neha Katre. 2021. Review of vqa : Datasets and approaches. *// 2021 International Conference on Advances in Computing, Communication, and Control (ICAC3)*, P 1–6.

Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Xavier Baró, Hugo Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. 2022. Codalab competitions: An open source platform to organize scientific challenges. *Technical report*.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards VQA models that can read. *CoRR*, abs/1904.08920.

Ryota Tanaka, Kyosuke Nishida, and Sen Yoshida. 2021. Visualmrc: Machine reading comprehension on document images. *ArXiv*, abs/2101.11272.

Zineng Tang, Ziyi Yang, Guoxin Wang, Yuwei Fang, Yang Liu, Chenguang Zhu, Michael Zeng, Chao-Yue Zhang, and Mohit Bansal. 2022. Unifying vision, text, and layout for universal document processing. *ArXiv*, abs/2212.02623.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. *ArXiv*, abs/2105.07624.

Constructing a Semantic Corpus for Russian: SemOntoCor

Igor M. Boguslavsky
A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia;
Universidad Politécnica de Madrid, Madrid, Spain
bogus@iitp.ru

Vyacheslav G. Dikonov
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
sdiconov@mail.ru

Evgeniya S. Inshakova
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
e.s.inshakova@gmail.com

Leonid L. Iomdin
A. A. Kharkevich Institute for Information Transmission Problems
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
iomdin@gmail.com

Alexandre V. Lazursky
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
lazursky@mail.ru

Ivan P. Rygaev
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
irygaev@jent.org

Svetlana P. Timoshenko
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
timoshenko@iitp.ru

Tatyana I. Frolova
A. A. Kharkevich Institute for Information Transmission Problems,
B. Karetnyj 15, Moscow, 103051, Moscow, Russia
tfrolova@gmail.com

Abstract

The SemOntoCor project focuses on creating a semantic corpus of Russian based on linguistic and ontological resources. It is a satellite project with regard to a semantic parser (SemETAP) being developed, the latter aiming at producing semantic structures and drawing various types of inferences. SemETAP is used to annotate SemOntoCor in a semi-automatic mode, whereupon SemOntoCor, when reaching sufficient maturity, will help create new parsers and other semantic applications. SemOntoCor can be viewed as a further step in the development of SynTagRus with its several layers of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct layer of meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. The first phase of SemOntoCor consists in annotating a Russian translation of the novel “The Little Prince” by Antoine de Saint-Exupery (1532 sentences, 13120 tokens).

Keywords: semantic corpus; semantic parser; ontology; Saint-Exupery
DOI: 10.28995/2075-7182-2023-22-12-25

Разработка семантического корпуса русского языка: SemOntoCor

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

Диконов В. Г.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
sdiconov@mail.ru

Иншакова Е. С.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
e.s.inshakova@gmail.com

Иомдин Л. Л.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
iomdin@gmail.com

Лазурский А. В.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
lazursky@mail.ru

Рыгаев И. П.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
irygaev@jent.org

Тимошенко С. П.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
timoshenko@iitp.ru

Фролова Т. И.
Институт проблем передачи информации им. А. А. Харкевича РАН,
Москва, Россия
tfrolova@gmail.com

Аннотация

Проект SemOntoCor ставит своей целью создание семантического корпуса русского языка на основе лингвистических и онтологических ресурсов. Этот проект развивается параллельно с разработкой семантического анализатора (SemETAP), нацеленного на построение семантических структур предложения и извлечение из них разного рода следствий. SemETAP используется для разметки SemOntoCor в полуавтоматическом режиме. С другой стороны, после того, как SemOntoCor достигнет достаточной зрелости, он сможет использоваться для разработки новых семантических анализаторов и для других семантических задач. SemOntoCor можно рассматривать как следующий шаг в развитии синтаксического корпуса SynTagRus, имеющего несколько уровней разметки. При разметке SemOntoCor на вход поступает морфо-синтаксическая разметка в формате SynTagRus, а на выходе строится базовая семантическая структура (BSemS). Эта структура представляет непосредственное значение предложения в терминах онтологических концептов, соединенных семантическими отношениями. Она абстрагируется от лексико-синтаксического многообразия естественного языка и во многих случаях осуществляет разложение лексического значения на более мелкие компоненты. Первая очередь SemOntoCor представляет собой разметку русского перевода повести-сказки Антуана де Сент-Экзюпери «Маленький принц» (1532 предложения, 13120 токенов).

Ключевые слова: семантический корпус; семантический парсер; онтология; Сент-Экзюпери

1 Introduction

Among various semantically annotated corpora, few combine multiple levels of annotation into one formalism. A well-known example is OntoNotes [Hovy et al. 2006] which is a resource comprising syntax, predicate-argument structure, word senses and co-reference. Another example is the Groningen Meaning Bank [Bos et al. 2017] that aims at integrating various linguistic phenomena, including predicate-argument structure, scope, tense, thematic roles, rhetorical relations and presuppositions within the formalism of the Discourse Representation Theory. On the other hand, an obvious fact is that most semantically annotated corpora that exist nowadays concentrate on English. This language is supported by many corpora built within different frameworks and annotated according to different annotation schemes. One of the rare exceptions that stands out in various respects is the Prague Dependency Corpus, which contains deeply annotated Czech texts (https://ufal.mff.cuni.cz/pdt3.0/data).

As far as Russian is concerned, it has a remarkable Russian National Corpus (RNC) which includes an extensive set of subcorpora, which cover over 2 billion words (ruscorpora.ru). Most of the subcorpora are annotated with morphological tags, and the main subcorpora (general subcorpus, newspaper subcorpus and a few others) have been recently supplied with syntactic features (including universal dependency features) and lexical semantic tags. Most of these annotation types have been produced automatically and have not been checked by experts on a mass scale. The SynTagRus subcorpus of RNC stands out as it contains several types of deep annotation, carried out in a uniform formalism according to the same theoretical framework. Namely, it is annotated with morphological features (including POS tags), dependency syntactic structures, word senses, anaphoric links, lexical functions (in terms of the Meaning-Text theory by Mel'čuk [1974]), micro-syntactic constructions, ellipsis, and temporal links. An important feature of these annotations is that, although many of them were carried out by means of software tools, all were thoroughly manually revised by the experts. Introducing a deeper level of annotation and aligning SynTagRus texts with semantic structures is a natural step further. In this paper, we describe an ongoing project aiming at performing this step. We began to compile a corpus annotated with what we call Basic Semantic Structures (BSemS). These structures are built on top of the existing morpho-syntactic annotation of SynTagRus and thus constitute the next higher level of sentence representation. It is important to note that our goal is not to annotate certain phenomena in a linguistically isolated way but to integrate all relevant semantic phenomena in a unified representation.

The structure of the paper is as follows. Section 2 outlines related work. Section 3 explains the framework of the project. Basic Semantic Structures we are constructing to annotate the corpus constitute an intermediate level of representation adopted in our semantic model. Section 4 presents BSemS in finer detail and explains some of its salient features. Section 5 describes the format of the corpus and shows how it is annotated by means of a special tool developed to facilitate the mark-up. Section 6 concludes the paper.

2 Related work

Existing semantic corpora may be classed into several groups by the types of information they provide.

1) The first group embraces semantically tagged corpora which annotate texts with word senses, ontology concepts or abstract semantic descriptors but do not provide information about any semantic relations. Some of them cover all non-grammatical words, others - only specific classes or words. Such corpora are used to test and train word-sense disambiguation tools. Some examples are:

- Semcor [Fellbaum et al, 1998] tags words with Wordnet synset references.
- Russian National Corpus [Raxilina et al, 2009, Kustova et al, 2005] contains automatically produced facet semantic tagging with semantic descriptors.
- Colorado Richly Annotated Full-Text (CRAFT) Corpus [Bada et al, 2012] is a collection of 97 full-length, open-access biomedical journal articles annotated with concepts from 9 different medical ontologies in parallel.

An overview of other resources of the same kind was presented in the paper «A Survey of WordNet Annotated Corpora» [Petrolito, Bond, 2014].

2) The second group consists of the corpora that explicate semantic relations between words or senses/concepts that replace the words. They provide some kind of semantic structures, which may be built in accordance with a certain linguistic theory or be theory-neutral. There are several theories

defining the representation for the sentence or whole text meaning that became the basis for corpus development projects. An overview of various semantic representations, including Abstract Meaning Representation (AMR), Discourse Representation Structures (DRS), Universal Networking Language (UNL), Tectogrammatical Representation (PDT) and more can be found in [Boguslavsky et al, 2021].

The most popular approach within this group is Frame Semantics [Fillmore, 1976]. There are multiple corpora that annotate propositional (predicate-argument "Who did what to whom?") structures within sentences using FrameNet [Baker, Fillmore et al., 1998, Ruppenhofer et al. 2007, 2016**]** or Verbnet **[**Kipper et al., 2008] dictionaries. These dictionaries serve as repositories of frames – prototypical situations that include a particular verbal sense (predicate), all necessary participants of the situation (arguments) and the roles they play. Such corpora are commonly called proposition banks or propbanks. They take different approaches towards the description of the roles. FrameNet has a very specific representation while the original PropBank corpus has the most general representation. The Verbnet approach takes the middle stand between the two. For example, given the ingestion sense of the verb *eat* in the sentence "Cynthia ate the peach with a fork", the respective representations for each would be:

- FrameNet: Cynthia(*ingestor*) ate(*predicate*) the peach(*ingestible*) with a fork(*instrument*).
- VerbNet: Cynthia(*agent*) ate(*predicate*) the peach(*patient*) with a fork(*instrument*).
- PropBank: Cynthia(*arg0*) ate(*predicate*) the peach(*arg1*) with a fork(*argm-manner*).

Some examples of proposition banks are:

- The original English Proposition Bank (PropBank) [Palmer et al, 2005, Palmer 2002] and a whole family of Propbanks in other languages than English.
- A similar resource called Nombank [Meyers et al, 2004] annotates predicative nouns.
- FrameNet Corpus [Bauer et al, 2012] contains parser-generated dependency structures (with POS tags and lemmas) for all FrameNet 1.5 sentences, with nodes automatically associated with FrameNet annotations.
- OntoNotes [Hovy et al, 2006, Weischedel et al, 2009] is a large multilingual corpus. The annotation includes parse trees, predicate argument structures (PropBank/NomBank style), word senses linked to an ontology, coreference, and named entities. The languages covered are English, Chinese, and Arabic with a significant amount of parallel data.
- FrameNet-Annotated Textual Entailment (FATE) [Burchardt, Pennacchiotti, 2008] is a manually crafted corpus for Recognizing Textual Entailment (RTE) tasks with FrameNet annotation. It features a new annotation schema based on full-text annotation of so-called relevant frame evoking elements. FATE annotates frames unknown in the FrameNet, anaphoric expressions in frames and constructions important for RTE, including support and copula verbs, existential constructions, modal expressions, metaphors.

3 SemETAP semantic model

The SemOntoCor corpus is a collection of Russian texts annotated with BSemSs built in accordance with the SemETAP semantic model [Boguslavsky 2017, Boguslavsky et al. 2018, Boguslavsky et al. 2019]. In its turn, this model is a component of a general-purpose rule-based linguistic processor ETAP-4, which implements basic linguistic competences of humans – text understanding and text production [Apresian et al. 2003]. ETAP-4 is built within the framework of the Meaning – Text Theory by I. Mel'čuk [1974, 2012, 2013, 2015]. SemETAP reuses the non-semantic modules of ETAP-4 – the morphological analyzer, the syntactic dependency parser, and the normalization submodule. SemETAP is used for annotating SemOntoCor in a semi-automatic mode (see Section 4 for details).

Our approach to semantics is in many respects similar to that of [McShane, Nirenburg 2021], although many linguistic and methodological solutions are different. We proceed from the assumption that the depth of understanding is growing with the number and sophistication of inferences we can draw from the text. In order to obtain inferences, we make intensive use of both linguistic and background knowledge. The former is incorporated in the dictionary and the grammar, and the latter is stored in the ontology. In many cases, explicit decomposition of words and ontology concepts is used to produce additional inferences and thus achieve a deeper understanding.

We distinguish two levels of our semantic structures. Basic semantic structure (BSemS) presents the direct meaning of the sentence, while Enhanced semantic structure (EnSemS) extends BSemS by means

of a series of inferences construed on the basis of linguistic and extralinguistic knowledge accessible to the model. The model produces both reliable inferences (*John forgot to take the pill ⇒ John did not take the pill*) and plausible expectations (*John went to Paris at moment t1 ⇒ John is in Paris at moment t2>t1*).

Both structure types (BSemS and EnSemS) are built from the elements of a language-independent ontology, OntoETAP, which thereby can be seen as a metalanguage of the semantic description. The ontological elements (concepts and individuals) have different kinds of properties in OntoETAP, such as class/subclass, class/individual, semantic slots a concept can take, etc.

From the formal point of view, semantic structures of both BSemS and EnSemS are Directed Acyclic Graphs (DAG) with individuals at the nodes and arrows labeled with semantic relations. They are represented in OWL and written in the RDF format, i.e. as sets of triples of the type *relation (Ontoelement-1, Ontoelement-2)*, where *relation* is an object or data property of the ontology, and *Ontoelement-i* is a variable or a constant denoting an individual. The RDF formalism was chosen because, on the one hand, it is very flexible and expressive, and on the other hand, it is supported by a wide range of tools and is easily integrated with many Semantic Web applications.

It should be noted that SemOntoCor, like SynTagRus and some other corpus projects, has double identity. On the one hand, it can be perceived as an autonomous semantic resource, and on the other hand it forms a unified complex with the SemETAP semantic parser, ontology and the inference engine. Below, we will concentrate on BSemS, because it is this type of structure that constitutes our corpus.

4 Basic Semantic Structures

When constructing a semantic representation of a natural language text, one of the most essential requirements is its ability to abstract away from formal and syntactic variation, namely to assign similar structures to different constructions that have a similar meaning, and to assign different structures to constructions that have different meanings, despite their surface similarity [Abend, Rappaport 2017].

In particular, this is manifested in the fact that grammar words (auxiliary and support verbs, strongly governed prepositions and conjunctions, or articles) are removed from the sentence, passive constructions are replaced by active ones, nouns derived from verbs are reduced to the base verbs, etc. The most important type of information that semantic structures seek to convey is the predicate-argument skeleton of the sentence, that is, the information about "who is doing what to whom". Here is the BSemS of one of the SemOntoCor sentences as an example:

(1) *Narisuj mne barashka...* ‘Draw me a lamb’

(2) Drawing
    hasAgent UtteranceAddressee
    hasObject Sheep
        hasGender Male
        isObjectOf HavingSize
            hasDegree LowDegree
    hasBeneficiary UtteranceSpeaker
    isTopicOf Urging
        hasAgent UtteranceSpeaker
        hasRecipient UtteranceAddressee
        hasTime SpeechTimePosition

BSemS (2) representing sentence (1) can be read as follows: «The speaker verbally encourages the addressee to draw a small male sheep for the speaker; the time of encouraging is the time of speech».

For all the simplicity of sentence (1), the structure (2) allows us to see how SemOntoCor addresses some of the main problems facing semantic corpora. These are the representation of word meanings (Subsection 4.1), the representation of relations between words (Subsection 4.2) and the representation of grammatical meanings (Subsection 4.3).

4.1 Word meaning

The semantic units (nodes) that form BSemS are not natural language words (in our case, Russian), but elements of the OntoETAP ontology. This makes BSemS largely language-neutral. In semantic corpora, there is often a one-to-one correspondence between full-fledged words in a sentence and semantic elements (with the precision up to synonymy) (see [Boguslavsky et al. 2021, section 3.2] for more details). In our example (1), such a correspondence exists for two words in the sentence – *narisuj* (Draw) and *mne* (UtteranceSpeaker). The third word in the sentence – *barashek* 'lamb' - is represented by multiple nodes of the structure simultaneously (Sheep hasGender Male isObjectOf (HavingSize hasDegree LowDegree) – 'a small male sheep'). Partial decomposition of the lexical meaning with a group of several semantic elements is widely used in SemOntoCor.

This approach has both advantages and disadvantages. One advantages is that it allows producing similar representations for different synonymous expressions: *zapel* ('began-to-sing') – *nachal pet'* ('began to sing') = Begin hasObject Singing, *ispugal ee* ('frightened her') – *zastavil ee bojat'sja* ('made her fear') = Cause hasObject Fear. Even if the expressions are not synonymous but have significant semantics in common, the decomposition makes it possible to make explicit both the common components and the differences. For instance, in the AMR corpora (https://github.com/amrisi/amr-guidelines/blob/master/amr.md), the name of a feature scale and a specific range of that scale (such as *age – old/young*, *weight – heavy/light*, *price – expensive/cheap*) are expressed as separate concepts that are not related to each other. In SemOntoCor, the links between such meanings are presented in a clear and graphic way: *age* – HavingAge, *old* – (HavingAge hasDegree HighDegree), *young* – (HavingAge hasDegree LowDegree).

Decomposition allows avoiding uncontrolled introduction of a large number of conceptually similar concepts. For example, the Russian word *tigr* 'tiger' denotes any animal of the given species, while the word *tigritsa* 'tigress' denotes only a female of such an animal. The principle of mutual one-to-one correspondence between words and concepts requires the introduction of two separate concepts – Tiger and FemaleTiger – in the ontology. Decomposition of the lexical meaning allows us to use only one concept, defining a female tiger as (Tiger hasGender Female). Similarly, *frantsuz* 'French person' is interpreted as the construction (Human hasNationality France) (= a person who is a French citizen), *frantsuzhenka* 'French woman'- as (Human hasNationality France hasGender Female), *parizhanin* 'Parisian' – as (Human livesIn Paris) (= a person living in Paris). Decomposition can be used to distinguish and simultaneously capture the similarity of social roles (for example, *monarx1* = MonarchRole) and people who perform social roles (*monarx2* = (Human hasSocialRole MonarchRole)).

The downside of the lexical meaning decomposition approach is that many words refer to complex phenomena, and decomposing them into smaller components would result in very unwieldy structures. To avoid this, we adopted a compromise approach in BSemS, where only words that allow for a small number of components, like the examples above, are decomposed.

The general principle is that the full-valued words are matched with concepts that are synonymous or quasi-synonymous with them. At the same time, we try not to clutter the ontology with different concepts that are similar in meaning. Therefore, when the meaning of a word can be reduced to a basic concept, supplemented by a small number of refinements, we resort to decomposition. As for more complicated decompositions, we postpone them to the level of EnSemS, where different kinds of inferences are performed, including those based on common sense. For more on EnSemS, the reader is referred to [Boguslavsky 2017, Boguslavsky et al. 2018, Boguslavsky et al. 2019].

To give the reader a better understanding of the non-isomorphism between the words of the sentence and their representation in BSemS, we provide a couple more examples.

One systematic example is the reduction of different words to the same concept through lexico-syntactic derivation. Many words refer to the same situation from different angles: *John is married to Ann – Ann is John's wife – John is Ann's husband.* Obviously, the words used in these sentences are not synonyms, but their semantic representations should make it clear that they denote the same situation. In SemOntoCor these words are assigned the same concept (in this case – Spouse), and the difference between them is accounted for by different relations between the concepts. In this example, we were dealing with nominal lexico-syntactic derivatives of the verb – *husband* and *wife*. There exist also adverbial and adjectival derivatives, that are also reduced to the main predicate. Cf. (3) and (4).

(3) *Ja dumaju, chto pojdet dozhd'* 'I think it is going to rain'

Believe
    hasExperiencer UtteranceSpeaker
    hasObject Raining

(4) *Po-moemu, pojdet dozhd'* 'in my opinion, it is going to rain'

Raining
    isObjectOf Believe
        hasExperiencer UtteranceSpeaker

Another example of the same type involves adjectives that refer to the same concept but characterize its different arguments. For example, *ispugannyj* 'frightened' is a property of the one who if afraid of something, while *strashnyj* 'scary' characterizes something that causes fear. In SemOntoCor, such adjectives are represented by the same concept but are connected to the concept they modify by different semantic relations:

ispugannyj malchik 'frightened boy' – Boy isExperiencerOf Fear
strashnaja situacija 'scary situation' – Situation isObjectOf Fear

Yet another example of non-isomorphism between the text and its BSemS are titles. A phrase like *roman Remarka "Tri tovarishcha"* meaning 'the novel by Remarque "Three Comrades"' is rendered by a structure that contains both the original title and its meaning:

Novel
    hasName "Tri tovarishcha"
    hasNameMeaning Friend
        hasQuantity 3

A frequent source of nodes that have no direct correspondence in the text is omission of arguments. For example, in the phrase *Petya xochet poprosit' Kolyu ujti* meaning 'Petya wants to ask Kolya to leave', the BSemS structure makes explicit the omitted subjects of the infinitives: *poprosit' – Petya* meaning 'ask – Petya', and *ujti – Kolya* meaning 'leave – Kolya'.

4.2 Relations between the nodes in BSemS

As mentioned above, BSemS is a graph consisting of OntoETAP ontology elements in its nodes and semantic relations as its edges. Several dozen relations are used. The most frequent ones are: hasAgent, hasObject, hasObject2, hasExperiencer, hasTime, hasLocation, hasDegree, hasQuantity, hasAttribute, etc[1].

Each relation can have an inverted variant, such as hasAgent – isAgentOf, hasObject – isObjectOf. Inverted relations allow expressing the difference in communicative dominance. More information on this can be found in [Mel'čuk 2015: 311-324]. Cf. *Malchik bezhit* 'the boy runs' – (Running hasAgent Boy). *Malchik, kotoryj bezhit (begushchij malchik)* 'the boy who runs (running boy)'– (Boy isAgentOf Running). Cf. also examples (3) and (4) above.

As far as the propositional content is concerned, the structures (Running hasAgent Boy) and (Boy isAgentOf Running) are completely equivalent. Therefore, for tasks that involve only propositional content, communicative dominance is irrelevant and inverted relations can be safely replaced with non-inverted ones. However, for other tasks, information on communicative dominance may be valuable. Structures that differ in communicative dominance are used in different contexts and are built into discourse in different ways. Cf. *Malchik bystro bezhal* meaning 'the boy was running fast' and *Skorost' bega malchika byla vysokoj* meaning 'the boy's running speed was high'. Obviously, the communicative

[1] A complete commented list of relations used for annotation can be found at https://docs.google.com/document/d/1W469sCt-ne7DB1yS3QM_hzpCJM_yuhJp

dominance is crucial for any task dealing with text generation and discourse cohesion. It is also important that the use of inverted relations makes BSemS easier to understand as it aligns more directly with the syntactic structure (cf. structures (3) and (4) above).

Besides inverting, each relation can be reified, i.e. it can come in the form of a concept. If additional information, such as time or modality, needs to be conveyed along with the relation, one cannot use the relation and should use a reified concept. For example, the meaning of localization in the noun phrase *dom v lesu* 'a house in the forest' can be conveyed by the relation: (House hasLocation Forest). However, if we need to characterize this situation as having taken place in the past, the relation cannot be used and we should use a concept (Location) instead:

Dom naxodilsja v lesu
Location
    hasObject House
    hasObject2 Forest
    hasTime TimeInterval
        before SpeechTimePosition

4.3 Uniform representation of lexical and grammatical meaning

For representing grammatical meanings, BSemS does not use grammatical labels such as present, past, imperative, interrogative, etc. Instead, we apply the same concepts that represent lexical meanings. In structure (2) above, the imperative form of the verb *narisovat'* 'to draw' is conveyed by the semantic structure meaning 'the speaker encourages the addressee to draw' by means of the concept Urging. This concept is also used to represent lexical (and not grammatical) markers of encouragement, for example *Peter encouraged (urged) Bill to stay.*

Similarly, to represent grammatical interrogation (*When will you come?*) we apply the concept Questioning, which is also used to represent such words as *ask, question, enquire, interrogate* etc. Thanks to the uniform representation of grammatical and lexical meanings, sentences of different structures obtain similar (or identical) BSemS. For example, sentences *Kogda ty pridesh'?* 'When will you come?' and *Ja sprashivaju tebja, kogda ty pridesh'* 'I am asking you when you will come' correspond to the same BSemS. Besides, there is also a natural opportunity to establish co-referentiality between grammatical and lexical expressions, such as: *Ty pridesh' zavtra? Etot vopros byl neumestnym.* 'Will you come tomorrow? The question was inappropriate'. If grammatical interrogation were conveyed by a grammatical label, it would be hard to infer that it is co-referential to the word *vopros* 'question'.

5 Corpus annotation

The first text annotated for SemOntoCor is the story "The Little Prince" by Antoine de Saint-Exupery, translated into Russian by Nora Gal. It was first published in French in 1943 and translated into more than 180 languages. It is one of the most popular books in the world literature, with over 80 million copies sold. The (Russian translation of the) text contains 1532 sentences (13120 tokens). It is included in at least two other known semantic corpora – AMR (https://amr.isi.edu/download.html) and UNL [Martins 2012]. It offers a rare opportunity to compare and assess different variants of semantic mark-up, as well as to develop a procedure of automatic (or semi-automatic) translation from one mark-up to another. The latter is interesting in that it opens up an enticing prospect of supplementing SemOntoCor with the data from other corpora.

As of today (May 2023), more than 1100 sentences of the story have been marked up. When the story is fully annotated, it will be made available to the public.

5.1 The BSemS format

The format in which BSemSs are stored is an extension of the XML format used for representing SynTagRus – the treebank of Russian which is an integral part of the National Corpus of Russian (https://ruscorpora.ru). The format used for SynTagRus is described in [Iomdin, Sizov 2009] and its extension for SemOntoCor is proposed in [Frolova, Rygaev 2022].

Each sentence and all linguistic information about it are stored inside the <S> tag, which contains <W> tags for the words of the sentence. They are followed by the <SEM> tag for the semantic structure.

Each <W> contains the information about the ID number of the word, its morphological features and incoming syntactic links.

In <SEM> tag the information is organized into <N> tags, each referring to a certain semantic node.

Each <N> tag has several attributes, ID, TYPE, and VALUE. ID attribute gives the number of the node, TYPE attribute shows the type of the node with respect to ontology, see more in [Frolova, Rygaev 2022]. VALUE attribute contains the name of the node.

If a node has outgoing links, they are listed in the <R> tag, which has attributes LINK (for the name of the link) and TO (for the ID of the target node).

Below is the BSemS of the sentence *Narisuj barashka* 'Draw a lamb'. It has two <W> tags according to the number of words in the sentence and nine <N> tags inside <SEM> according to the number of nodes in BSemS. For example, the node with ID="2" has value "Sheep", is connected to nodes 3 and 5 by links hasGender and isObjectOf respectively.

```
<S COMMENT="traced" DATE="09 02 2023 14:38:09" ID="1"
        <W DOM="_root" EXTRAFEAT="СЛР ЛИЧ" FEAT="V СОВ ПОВ ЕД 2-Л" ID="1"
KSNAME="РИСОВАТЬ" LEMMA="РИСОВАТЬ">Нарисуй</W>
        <W DOM="1" FEAT="S ЕД МУЖ ВИН ОД" HYPOT="1-компл.11" ID="2"
KSNAME="БАРАШЕК1" LEMMA="БАРАШЕК" LINK="1-компл">барашка</W>.
        <SEM>
                <N ID="1" TYPE="anonymous" VALUE="Drawing">
                        <R LINK="hasAgent" TO="4"/>
                        <R LINK="hasObject" TO="2"/>
                        <R LINK="isTopicOf" TO="7"/>
                </N>
                <N ID="2" TYPE="anonymous" VALUE="Sheep">
                        <R LINK="hasGender" TO="3"/>
                        <R LINK="isObjectOf" TO="5"/>
                </N>
                <N ID="3" TYPE="named" VALUE="Male"/>
                <N ID="4" TYPE="anonymous" VALUE="UtteranceAddressee"/>
                <N ID="5" TYPE="anonymous" VALUE="HavingSize">
                        <R LINK="hasDegree" TO="6"/>
                </N>
                <N ID="6" TYPE="anonymous" VALUE="LowDegree"/>
                <N ID="7" TYPE="anonymous" VALUE="Urging">
                        <R LINK="hasAgent" TO="8"/>
                        <R LINK="hasRecipient" TO="4"/>
                        <R LINK="hasTime" TO="9"/>
                </N>
                <N ID="8" TYPE="anonymous" VALUE="UtteranceSpeaker"/>
                <N ID="9" TYPE="named" VALUE="SpeechTimePosition"/>
        </SEM>
</S>
```

5.2 Annotation tool

To annotate SemOntoCor we use a custom Structure Editor (StrEd) software, originally developed for annotation of the SynTagRus treebank [Iomdin,Sizov, 2009] and later extended in order to annotate SemOntoCor. StrEd supports two annotation procedures – manual and semi-automatic. Fig. 1 shows sentence *Narisuj barashka* loaded into StrEd and its BSemS obtained automatically.

In the manual mode, the annotator enters each node individually and connects them by semantic relations using the context menu, shown in Fig. 1. The menu contains options of creating and deleting

nodes and links between them, as well as changing the correspondence between nodes and words. In the semi-automatic procedure, the annotator runs the SemETAP semantic parser for each sentence (cf. section 3 above). The BSemS obtained is loaded into StrEd for subsequent revision.

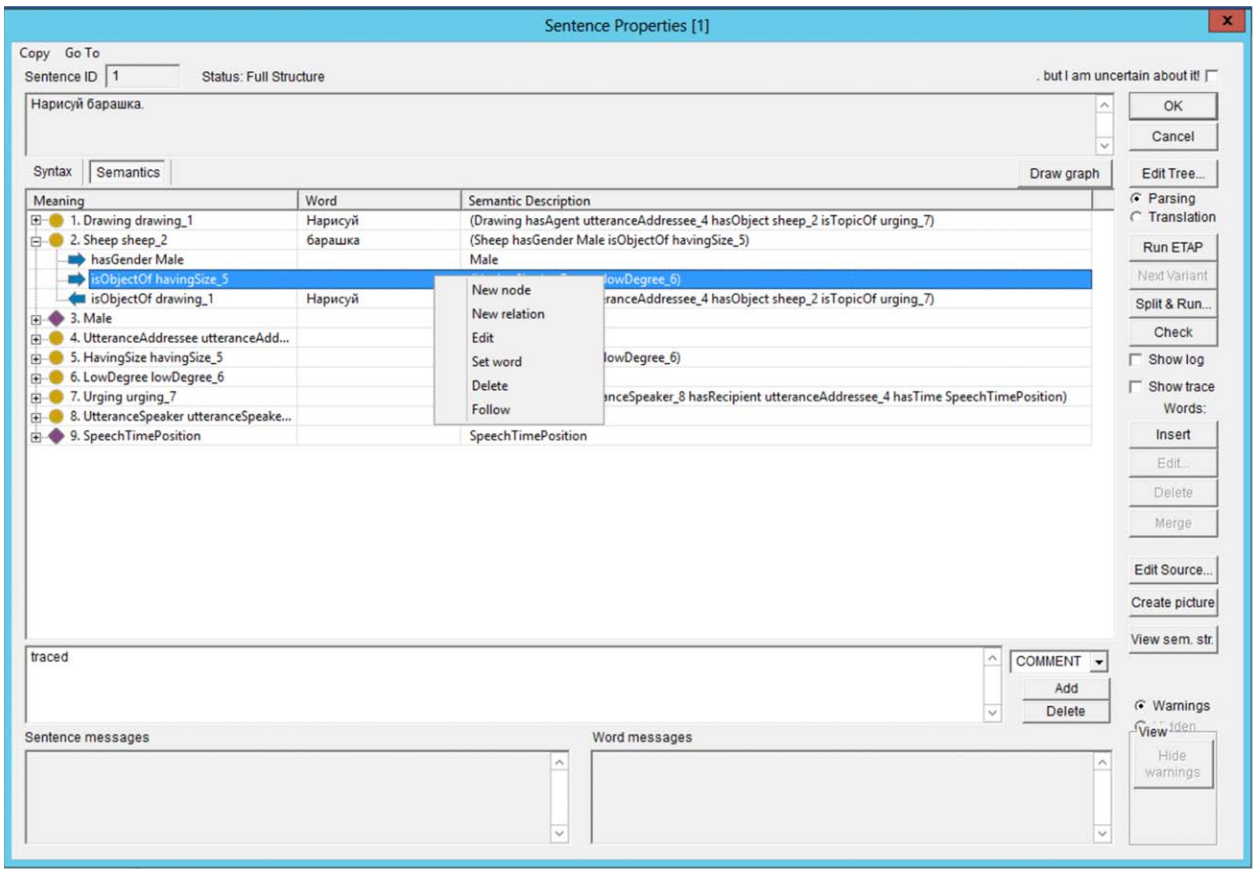


Figure 1: StrEd annotation tool

In the sentence properties window we can toggle between “Syntax” and “Semantics” tabs (upper left), which allows the user to choose between syntactic and semantic annotation.

The Semantics tab opens the table, where each line corresponds to a BSemS node. For each node the following information is available: the node number and its name, the word of the sentence which generated this node (if any) and a semantic description of the node, which is a fragment of BSemS containing outgoing links and nearest dependents.

A node can be expanded if it is connected to other nodes either with an incoming or an outgoing link. Node 2 (Sheep) in Fig.1 is expanded to reveal two outgoing links (blue arrows pointing to the right) and an incoming link (arrow pointing to the left) from Drawing.

For the annotator’s convenience, the “Draw graph” button in the upper right corner visualizes the BSemS. Fig. 2 shows the visualization of BSemS in Fig. 1.

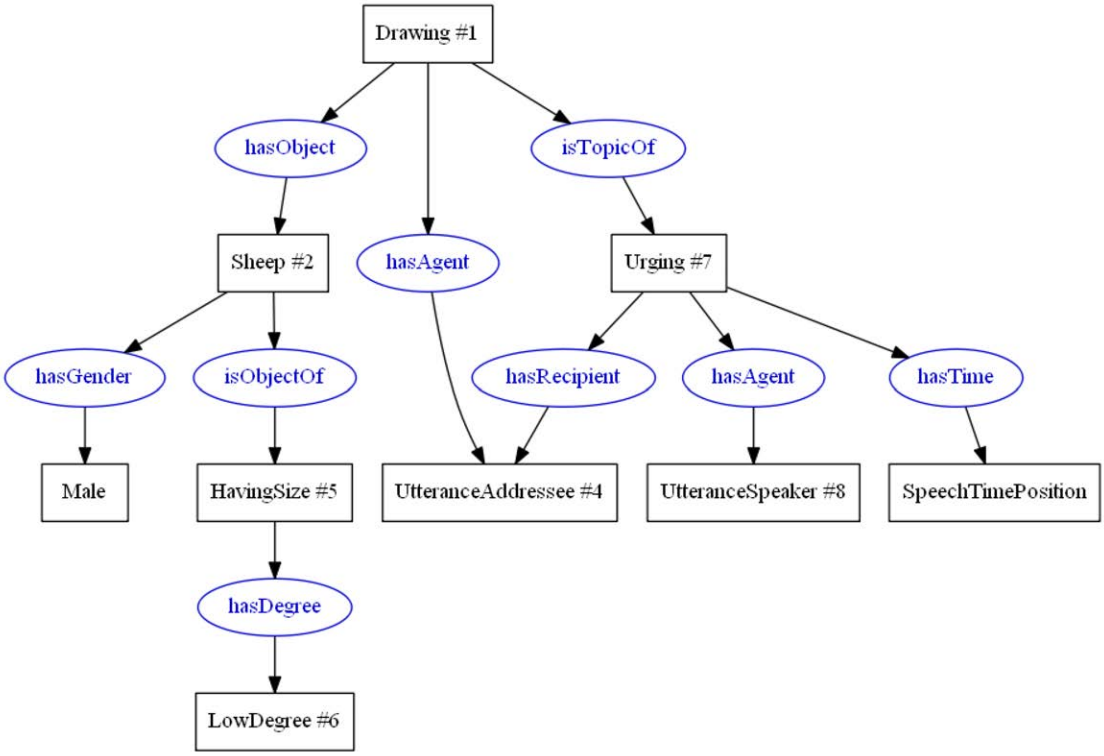


Figure 2: Visualization of the BSemS for *Narisuj barashka*

We performed a rough estimation of the current speed of annotation by selecting arbitrarily 10 sentences of average length (9-17 tokens) and complexity. The same annotator annotated 5 of them manually and the other 5 semi-automatically. The annotation proper (i.e. without the time spent on thinking) took 11,6 minutes/sentence for manual annotation and 8,4 minutes/sentence for semi-automatic annotation. New annotators, that will be recruited for the job, will undergo special training. At first, their annotation speed will probably be lower. However, given that annotators will acquire experience over time and the SemETAP parser will gain accuracy, we can expect the speed of annotation to increase both in manual and semi-automatic mode.

6 Conclusion

We present SemOntoCor - a new semantic corpus for Russian under construction at the Institute for Information Transmission Problems (RAS). It is the next step in the development of SynTagRus with its several types of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. SemOntoCor is unified with the SemETAP semantic parser, which produces two levels of semantic structures – Basic SemS and Enhanced SemS. The latter enriches BSemS with different types of inferences, based both on the linguistic and the common-sense knowledge. The annotation is done in a semi-automatic mode: SemETAP produces a draft BSemS, which is then revised by an expert. In the first version of SemOntoCor a novel by Antoine de Saint-Exupery “The Little Prince” is annotated.

Acknowledgements

This work was done with the financial support of the grant No. 22-28-01941 “Development of the infrastructure and the first phase of the semantic corpus for Russian” from the Russian Science Foundation.

References

[1] Abend O., Rappoport A. The state of the art in semantic representation. // Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). — Vancouver, Canada. Association for Computational Linguistics. — 2017. — p. 77–89.

[2] Abzianidze, L., J. Bjerva, K. Evang, H. Haagsma, R. van Noord, P. Ludmann, D.-D. Nguyen, and J. Bos (2017, April). The parallel meaning bank: Towards a multilingual corpus of translations annotated with compositional meaning representations. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, pp. 242–247. Association for Computational Linguistics.

[3] Abzianidze L. and Johan Bos. (2017). Towards universal semantic tagging. In: Proceedings of the 12th International Conference on Computational Semantics (IWCS 2017) – Short Papers, pages 1–6, Montpellier, France.

[4] Apresian Ju., Boguslavsky I., Iomdin L., Lazursky A., Sannikov V., Sizov V., Tsinman L. (2003). ETAP-3 Linguistic Processor: a Full-Fledged NLP Implementation of the MTT // First International Conference on Meaning-Text Theory (MTT’2003). June 16-18, 2003. Paris: Ecole Normale Supérieure. P. 279-288.

[5] Bada M., Eckert M., Evans D., Garcia K., Shipley K., Sitnikov D., Baumgartner W.A. Jr, Cohen K.B., Verspoor K., Blake J.A., Hunter L.E. (2012) Concept annotation in the CRAFT corpus. // BMC Bioinformatics. 2012 Jul 9 13:161. DOI: 10.1186/1471-2105-13-161. PMID: 22776079; PMCID: PMC3476437.

[6] Banarescu L., Bonial C., Shu Cai, Georgescu M., Griffitt K., Hermjakob U., Knight K., Koehn Ph., Palmer M., Schneider N. (2013) Abstract meaning representation for sembanking. // Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse. Sofia, Bulgaria. pp. 178–186.

[7] Baker C.F., Fillmore C., Lowe J.B. (1998) The Berkeley FrameNet project. // Proceedings of COLING-ACL, Montreal, Canada.

[8] Basile, V., J. Bos, K. Evang, and N. Venhuizen (2012). Developing a large semantically annotated corpus. In Proceedings of the 8th International Conference on Language Resources and Evaluation, Istanbul, Turkey, pp. 3196 – 3200.

[9] Bauer D., Fürstenau H., Rambow O. (2012) The Dependency-Parsed FrameNet Corpus. // In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, pp 3861–3867. European Language Resources Association (ELRA).

[10] Bin Li, Yuan Wen, Weiguang Qu, Lijun Bu, and Nianwen Xue. (2016) Annotating the Little Prince with Chinese AMRs. // In Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016), pages 7–15, Berlin, Germany. Association for Computational Linguistics.

[11] Bjerva J., Barbara Plank, and Johan Bos. (2016). Semantic tagging with deep residual networks. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 3531–3541, Osaka, Japan.

[12] Boguslavsky I. (2017). Semantic Descriptions for a Text Understanding System. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2017” Moscow, May 31—June 3, 201

[13] 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

[14] Boguslavsky I., Frolova T., Iomdin L., Lazursky A., Rygaev I., Timoshenko S. (2019). Knowledge-based approach to Winograd Schema Challenge. Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2019”. Moscow, May 29—June 1, 2019.

[15] Boguslavsky I., Dikonov V., Inshakova E., Iomdin L., Lazursky A., Rygaev I., Timoshenko S., Frolova T. (2021) Semantic Representations in Computational and Theoretical Linguistics: the Potential for Mutual Enrichment. pp. 127-141. DOI 10.28995/2075-7182-2021-20-127-141.

[16] Bos J., Abzianidze L. (2019). Thirty Musts for Meaning Banking. In Proceedings of the First International Workshop on Designing Meaning Representations. — Florence, Italy, August 1st. — 2019. — p. 15–27.

[17] Bos J., Basile V., Evang K., Venhuizen N.J., Bjerva J. (2017) The Groningen Meaning Bank. // In book: Handbook of Linguistic Annotation, pp.463-496, Springer Netherlands DOI:10.1007/978-94-024-0881-2_18

[18] Burchardt A., Pennacchiotti M. (2008) FATE: a FrameNet-Annotated Corpus for Textual Entailment. // In Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, Morocco. European Language Resources Association (ELRA).
[19] Carlson L., Marcu D., Okurowski M. E. (2001). Building a Discourse-Tagged Corpus in the Framework of Rhetorical Structure Theory. In Proceedings of 2nd SIGDIAL Workshop on Discourse and Dialogue, Eurospeech 2001. Aalborg, Denmark, 1–10.
[20] Fellbaum C., Landes S., Leacock C. (1998) Building semantic concordances. // In Fellbaum (1998), chapter 8, pp 199–216.
[21] Fillmore C. J., (1976) Frame semantics and the nature of language. // Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech, 280:20–32.
[22] Flickinger D., Yi Zhang, and Valia Kordoni. 2012. Deepbank: A dynamically annotated treebank of the Wall Street journal. In Proceedings of the Eleventh International Workshop on Treebanks and Linguistic Theories, pages 85–96. Edicoes Colibri.
[23] Frolova T.I., Rygaev I.P. (2022), Razrabotka formata razmetki i printsipov annotirovanija dlja semanticheskogo korpusa russkogo jazyka na materiale “Malen’kogo printsa” [Developing a mark-up format and annotation principles for a semantic corpus of Russian on the material of The Little Prince]. In: Sbornik trudov 46 mezhdistsiplinarnoj shkoly-konferentsii IPPI RAN “Informatsionnye texnologii i sistemy 2022 (ITiS'2022)”. M. IPPI, p. 122-136.
[24] Hajič J. (2002) Tectogrammatical Representation: Towards a Minimal Transfer In Machine Translation. // Proceedings of the Sixth International Workshop on Tree Adjoining Grammar and Related Frameworks. Association for Computational Linguistics. pp. 216–226.
[25] Hajič J., Hladká B., Pajas P. (2001) The Prague Dependency Treebank: Annotation Structure and Support. // IRCS Workshop on Linguistic Databases. pp. 105–114.
[26] Hovy E., Mitchell M., Palmer M., Ramshaw L., Weischedel R. (2006). *OntoNotes*: The 90% Solution. Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 57–60, New York, June 2006.
[27] Inshakova E.S., Iomdin L.L., Mitjushin L.G., Sizov V.G., Frolova T.I., Cinman L.L. (2019), SinTagRus segodnja [The SynTagRus Today]. In: Trudy Instituta russkogo jazyka im. V.V. Vinogradova [Proceedings of Vinogradov Russian Language Institute]. Vol. 21, Moscow, pp. 14–41. DOI: 10.31912/pvrli-2019.21.1.
[28] Iomdin L., Sizov V. Structure Editor: a Powerful Environment for Tagged Corpora // MONDILEX Fifth Open Workshop. Ljubljana, Slovenia, October 14–15, 2009. Ljubljana, 2009. P. 1-12. ISBN 978-961-264-012-5.
[29] Kamp H., Reyle U. (1993) From Discourse to Logic; An Introduction to Modeltheoretic Semantics of Natural Language, Formal Logic and DRT. // Kluwer, Dordrecht.
[30] Kipper K., Korhonen A., Ryant N., Palmer M. (2008) A large-scale classification of English verbs. // Language Resources and Evaluation Journal, 42:21–40
[31] Kustova G., Ljashevskaja O., Paducheva E., Raxilina E. (2005). Semanticheskaja razmetka leksiki v Natsional’nom korpuse russkogo jazayka: printsipy, problemy, perspektivy. [Semantic mark-up of words in Russian National Corpus: principles, problems, perspectives]. In: Natsional’nyj korpus russkogo jazyka: 2003-2005. Rezultaty i perspektivy. M., pp. 155-174.
[32] Martins R.T. (2012) Le Petit Prince in UNL. // International Conference on Language Resources and Evaluation (2012).
[33] Mann W.C., Thompson S.A. (1988) Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. // Text 8(3), pp 243–281.
[34] Martins, R. (2012). Le Petit Prince in UNL. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12), pages 3201–3204, Istanbul, Turkey. European Language Resources Association (ELRA).
[35] McShane M., S. Nirenburg. (2021). Linguistics for the Age of AI. The MIT Press. Cambridge, Massachusetts.
[36] Mel’čuk I.A. An essay of the theory of linguistic “Meaning ⇔ Text”models. Semantics. Syntax. [Opyt teorii lingvističeskix modelej “Smysl ⇔ Tekst”. Semantika, Sintaksis.]. — Science [Nauka], Moscow. — 1974.
[37] Mel’čuk I. Semantics: From Meaning to Text. Vol. 1. — Amsterdam/Philadelphia: John Benjamins. — 2012.
[38] Mel’čuk I. Semantics: From Meaning to Text. Vol. 2. — Amsterdam/Philadelphia: John Benjamins. — 2013.
[39] Mel’čuk I. Semantics: From Meaning to Text. Vol. 3. — Amsterdam/Philadelphia: John Benjamins. — 2015.
[40] Meyers A., Reeves R., Macleod C., Szekely R., Zielinska V., Young B., Grishman R., (2004) The NomBank Project: An Interim Report. // In proceedings of FCP@NAACL-HLT (2004).
[41] Oepen, S., D. Flickinger, K. Toutanova, and C. D. Manning (2004). LinGO Redwoods. A rich and dynamic treebank for HPSG. Research on Language and Computation 2(4), 575 – 596.
[42] O'Gorman, T. J., Regan M., Griffitt K., Hermjakob U., Knight K., Palmer M. (2018) AMR Beyond the Sentence: the Multi-sentence AMR corpus. // International Conference on Computational Linguistics (2018).
[43] Palmer M. (2002) From Treebank to PropBank. // In Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002), Las Palmas, Spain.

[44] Palmer M., Kingsbury P., Gildea D. (2005) The Proposition Bank: An Annotated Corpus of Semantic Roles. // Computational Linguistics. 31 (1): 71–106. CiteSeerX 10.1.1.136.8985. doi:10.1162/0891201053630264. S2CID 2486369.
[45] Ruppenhofer J., Ellsworth M., Petruck M.R.L, Johnson C.R, Baker C.F., Scheffczyk J. (2007) FrameNet II: Extended Theory and Practice // Book (Revised November 1, 2016). Available at https://framenet.icsi.berkeley.edu/fndrupal/the_book
[46] Petrolito T., Bond F. (2014) A Survey of WordNet Annotated Corpora. // Global WordNet Conference 2014.
[47] Raxilina E., Kustova G., Ljashevskaja O., Reznikova T., Shemanaeva O. (2009). Zadachi i printsipy semanticheskoj razmetki leksiki v NRRJa [Tasks and principles of the semantic mark-up of words in RNC]. In: Natsional'nyj korpus russkogo jazyka: 2006-2008. Rezultaty i perspektivy. Spb.: Nestor-Istorija, pp. 215-239.
[48] Taboada M., Mann W.C. (2006) Rhetorical Structure Theory: Looking Back and Moving Ahead. // Discourse Studies 8, pp. 423–459.
[49] Timoshenko S.P., Iomdin L.L., Gladilin S.A., Inshakova E.S. (2021), SinTagRus v sostave NKRJa: novye vozmozhnosti [The SynTagRus as part of RNC: new opportunities]. In: Trudy mezhdunarodnoj konferentsii "Korpusnaja lingvistika-2021" [Proceedings of the International conference "Corpus linguistics-2021"], p.31-43.
[50] Weischedel R., Hovy E., Mitchell M., Palmer M., Belvin R., Pradhan S., Ramshaw L., Xue N. (2009). OntoNotes: A Large Training Corpus for Enhanced Processing // In J. Olive, C. Chrisiansen, and J. McCrary (eds), Handbook of Natural Language Processing and Machine Translation.
[51] White A.S., Stengel-Eskin E., Vashishtha S., Govindarajan V.S., Reisinger D.A., Vieira T., Sakaguchi K., Zhang S., Ferraro F., Rudinger R., Rawlins K., Van Durme B.. (2020) The Universal Decompositional Semantics Dataset and Decomp Toolkit. // In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 5698–5707, Marseille, France. European Language Resources Association.
[52] Zeldes A. (2017) The GUM Corpus: Creating Multilayer Resources in the Classroom. // Language Resources and Evaluation 51(3), 581–612.
[53] Zeman D., Hajic J. (2020) FGD at MRP 2020: Prague Tectogrammatical Graphs. // In Proceedings of the CoNLL 2020 Shared Task: Cross-Framework Meaning Representation Parsing, pp. 33–39, Online. Association for Computational Linguistics.
[54] Zhao M., Wang Y., Lepage Y., (2022) Large-scale AMR Corpus with Re-generated Sentences: Domain Adaptive Pre-training on ACL Anthology Corpus // International Conference on Advanced Computer Science and Information Systems (ICACSIS), Depok, Indonesia, 2022, pp. 19-24, DOI: 10.1109/ICACSIS56558.2022.9923502.

Pseudo-Labelling for Autoregressive Structured Prediction in Coreference Resolution

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

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

Abstract

Coreference resolution is an important task in natural language processing, since it can be applied to such vital tasks as information retrieval, text summarization, question answering, sentiment analysis and machine translation. In this paper, we present a study on the effectiveness of several approaches to coreference resolution, focusing on the RuCoCo dataset as well as results of participation in the Dialogue Evaluation 2023. We explore ways to increase the dataset size by using pseudo-labelling and data translated from another language. Using such technics we managed to triple the size of dataset, make it more diverse and improve performance of autoregressive structured prediction (ASP) on coreference resolution task. This approach allowed us to achieve the best results on RuCoCo private test with increase of F1-score by 1.8, Precision by 0.5 and Recall by 3.0 points compared to the second-best leaderboard score. Our results demonstrate the potential of the ASP model and the importance of utilizing diverse training data for coreference resolution.

Keywords: Pseudo-Labelling, Autoregressive Structured Prediction, Coreference Resolution
DOI: 10.28995/2075-7182-2023-22-26-33

Псевдоразметка для разрешения кореферентности при использовании авторегрессионного структурированного предсказания

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

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

Аннотация

Разрешение кореферентности является важной задачей в области обработки естественного языка, поскольку она используется как элемент решения таких задач, как поиск информации, суммаризация текста, ответы на вопросы по тексту, анализ тональности текста и машинный перевод. В данной статье исследована эффективность различных подходов к разрешению кореферентности на русском языке, с фокусом на набор данных RuCoCo. Также представлены результаты участия в Dialogue Evaluation 2023. Исследованы способы увеличения размера набора данных с помощью псевдоразметки и перевода данных с другого языка. Используя такой подход, удалось утроить размер набора данных, сделать его более разнообразным и улучшить результаты авторегрессионного структурированного предсказания в задаче разрешения кореферентности. Такой подход позволил добиться наилучших результатов на частном тестовом наборе RuCoCo с повышением F1-меры, точности и полноты на 1.8, 0.5 и 3.0 процентных пункта соответственно по сравнению со вторым лучшим результатом. Наши результаты демонстрируют потенциал модели ASP и важность использования разнообразных обучающих данных для разрешения кореферентности на русском языке.

Ключевые слова: псевдоразметка, авторегрессионное структурированное предсказание, разрешение кореферентности

1 Introduction

1.1 Coreference Resolution

Coreference resolution is a natural language processing (NLP) task that involves identifying all the expressions in a text that refer to the same entity or concept, and then linking them together. It is typically modeled by identifying entity mentions (contiguous spans of text), and predicting an antecedent mention for each span that refers to a previously-mentioned entity, or a null-span otherwise. The goal is to determine which pronouns, nouns, and other expressions in a sentence or document refer to the same entity, and to group them into clusters accordingly. Coreference resolution is a challenging task because it requires good understanding of the context and the ability to recognize complex relationships between words and phrases. However, this task is crucial in many applications of NLP, such as information retrieval [1], text summarization [2], question answering [3], sentiment analysis [4] and machine translation [5]. In addition, coreference resolution can be used to improve the readability of a text, by replacing repeated mentions of the same entity with a pronoun or other reference.

1.2 Related Work

This section contains a brief overview of previous most recent coreference resolution models. Lee et al. [6] proposed an end-to-end model for coreference resolution that predicts an antecedent probability distribution over candidate spans. The model incorporates mention scores, coarse and fine coreference scores, and vector representations of the spans to learn a probability distribution over all possible antecedent spans for each span in the text. To improve computational efficiency while being competitive with other models Kirstain et al. [7] introduced a lightweight end-to-end coreference model that removes the dependency on span representations. Instead, they utilize the endpoints of a span (rather than all span tokens) to compute the mention and antecedent scores. But this approach still presents a computational challenge of $O(n^4)$ complexity over document length so the authors need to prune the resulting mentions. Dobrovolskii [8] considers coreference links between words instead of spans which reduces the complexity of the coreference model to $O(n^2)$ and allows it to consider all potential mentions without pruning any of them out. Instead of using mention or coreference scorer within search algorithms over possible spans of text, Bohnet et al. [9] proposed fundamentally different approach that uses a text-to-text (seq2seq) paradigm to predict mentions and links jointly. The T5-based model takes a single sentence as input, and outputs an action corresponding to a set of coreference links involving that sentence as its output. Liu T. et al. [10] proposed another seq2seq T5-based model for Autoregressive Structured Prediction, which is described in more detail in the next section.

1.3 Autoregressive Structured Prediction

Autoregressive Structured Prediction (ASP) represents structures as sequences of actions, which build pieces of the target structure step by step. For instance, in the task of coreference resolution, the actions build spans (contiguous sequences of tokens) as well as the relations between the spans.

The goal of ASP is to predict an action sequence $y = y_1, \ldots, y_N$, where each action y_n is chosen from an action space $\mathcal{Y}_n$ represented as $\mathcal{Y}_n \stackrel{\text{def}}{=} \mathcal{A} \times \mathcal{B}_n \times \mathcal{Z}_n$, where $\mathcal{A}$ is a set of structure-building actions, $\mathcal{B}_n$ is the set of bracket-pairing actions, and $\mathcal{Z}_n$ is a set of span-labeling actions.

The set of structure-building actions $\mathcal{A} = \{$`]`, `[*`, `copy`$\}$ allows to encode the span structure of a text, e.g., `[*Delaware]` encodes that Delaware is a span of interest. Specifically, the action `]` refers to a right bracket that marks the right-most part of a span. The action `[*` refers to a left bracket that marks the left-most part of a span. The superscript * on `[*` indicates that it is a placeholder for 0 or more consecutive left brackets. Finally, `copy` refers to copying a word from the input document. To see how these actions come together to form a span, consider the string `[*Delaware]`, which is generated from a sequence of structure-building actions `[*` , copy , and `]` and the input string `Delaware`.

The set of bracket-pairing actions consists of all previously constructed left brackets, i.e.:

$$\mathcal{B}_n = \{m | m < n \wedge a_m = \text{[*}\}$$

Thus, in general, $|\mathcal{B}_n|$ is $O(n)$. However, it is often the case that domain-specific knowledge can be used to prune $\mathcal{B}_n$. For instance, coreference mentions and named entities rarely cross sentence boundaries, which yields a linguistically motivated pruning strategy [11].

For the task of coreference resolution, the set of span-labelling actions is

$$\mathcal{Z}_{\mathrm{n}} = \{m | m < n \wedge a_m = \blacksquare\} \cup \{\epsilon\}$$

where ϵ by the convention set in [12] is the antecedent of the first mention in each coreference chain and $\{m | m < n \wedge a_m = \blacksquare\}$ is the set of all the previous spans, which allows the model to capture intra-span relationships.

The coreference structure built on top of a document D is first converted into an action sequence and then is modelled as a conditional language model

$$p_\theta(y|D) = \prod_{n=1}^{N} p_\theta(y_n | y_{<n}, D)$$

The model is built on the base of a pre-trained language model such as T5.

2 Preliminary Experiments and Baselines

During the competition we tested several approaches:

- SpaCy implementation[1] of coarse-to-fine model [8] with different backbone models;
- Different transformers pretrained with Longformer [13] architecture;
- Original implementation[2] of start-2-end model [7] with different backbone models;
- Original implementation[3] of ASP [10] with different backbone models.

2.1 SpaCy

We trained a word-level spacy-coref model on RuCoCo dataset with different transformer encoders. It is trained in two stages: coreference clustering model that use coarse and fine scores to form clusters of entities, than span resolution model that recover original span after word-level coreference resolution. The best backbone transformer model was cointegrated/LaBSE-en-ru[4]. Although this model is a great sentence encoder, it did a good job on the word-level task too. This approach managed to beat baseline of the competition (a 2.4 point higher F1-score).

2.2 Longformers

Since the documents in RuCoCo dataset are relatively long we considered Longformer models that are able to grasp a larger area of text and its context. We pretrained two Longformer models that were based on cointegrated/LaBSE-en-ru and sberbank-ai/ruRoberta-large[5]. Pretraining was done according to [13] using long documents from Russian part of Wikipedia. Using these models together with spacy-coref and increased input sequence length barely gave us a performance gain, while making models even more memory intensive.

3 Improving Autoregressive Structured Prediction Performance

To beat the results of our previous best model we used ASP without changes in the implementation. With that said, we can divide further improvements of the model in two parts:

1. Choosing the backbone model along with hyperparameters tuning;
2. Working on the dataset improvement.

While experimenting with ASP we used different ruT5 models, different training sequence lengths and hidden sizes of the structure-building action head. ASP based on the large ruT5 model became the best model so far (4.0 points higher F1-score than baseline of the competition).

As some studies report [14], different coreference resolution models often do not transfer well to unseen domains. Moreover, for datasets containing news, such as RuCoCo, situations often arise when

[1] https://github.com/explosion/projects/tree/v3/experimental/coref
[2] https://github.com/yuvalkirstain/s2e-coref
[3] https://github.com/lyutyuh/ASP
[4] https://huggingface.co/cointegrated/LaBSE-en-ru
[5] https://huggingface.co/ai-forever/ruRoberta-large

new words, concepts and entities are encountered in test split. Probably the only way to overcome this is to increase or augment the dataset. This will likely lead to improved generalization, better handling of rare words and phrases, reduction of overfitting, improved robustness, because with more training data, the model is exposed to a greater variety of language patterns, more instances of rare words and phrases, more diverse examples, such as different writing styles, genres, or domains. However, it is important to consider the quality of the data. Therefore, while training models we use loss, which is weighted according to the data quality. We distinguish three classes of datasets: “gold”, “silver” and “bronze”. The lower the data quality, the lower the weight. The following three sections provide information about the ways that we used to increase the dataset.

3.1 Adding More Russian Coreference Resolution Datasets

To increase the size of the dataset we used two previously known good quality coreference resolution datasets in Russian:

- RuCor [15] – 163 documents;
- AnCor [16] – 521 document.

These datasets are considered “gold” along with RuCoCo.

3.2 Translating OntoNotes from English

OntoNotes 5.0 is one of the most popular datasets for coreference resolution in English with high quality. In some of our experiments with multilanguage models (more specifically – models with Russian and English tokens) we used this dataset directly as “silver” training data. But our final model was Russian only, thus we used the translation as “bronze” dataset.

To accurately translate the dataset into Russian, we did the following:

1. Use Helsinki-NLP/opus-mt-en-ru[6] for machine translation;
2. Translate the sentence and its clusters with entity spans to Russian language;
3. For every translated entity span find the most similar part of translated sentence using sentence encoder for text similarity (we used cointegrated/LaBSE-en-ru);
4. Use that span in the sentence as proper translation of original entity.

This may not be the most efficient approach, but it helped to achieve a translation with about 3% of all entities lost (1.68 entities lost per entire document in average). After analysing the results, we were satisfied with such a translation. We got 3017 new documents.

3.3 Using Pseudo-Labelling

The last part of final dataset was gathered with pseudo-labelling. We considered texts from same and different domains, collected from Taiga or Web:

- Arzamas[7] (Fiction) – 140 documents;
- collection5[8] (News articles with manual PER, LOC, ORG markup) – 355 documents;
- Interfax[9] (News) – 638 documents;
- KP[10] (News) – 355 documents;
- Lenta[11] (News) – 602 documents;
- N+1[12] (News) – 538 documents;
- Plaintext Wikipedia dump 2018 (ru.txt.gz)[13] – 1550 documents.

[6] https://huggingface.co/Helsinki-NLP/opus-mt-en-ru
[7] https://linghub.ru/static/Taiga/Arzamas.zip
[8] http://www.labinform.ru/pub/named_entities/collection5.zip
[9] https://linghub.ru/static/Taiga/Interfax.rar
[10] https://linghub.ru/static/Taiga/KP.rar
[11] https://linghub.ru/static/Taiga/Lenta.rar
[12] https://linghub.ru/static/Taiga/nplus1.rar
[13] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-2735

By this point, the ASP model based on ai-forever/ruT5-large (former sberbank-ai/ruT5-large)[14] trained on RuCoCo was the best model that we had. It was used to produce labels, i.e. clusters of spans of entities for texts that we collected.

Initially, more documents were collected for each dataset, however, they were randomly selected in such a way that the distribution of text lengths was similar to that of the RuCoCo dataset. When pseudo-labelling procedure was done, all datasets were filtered in such a way, that entity count, cluster count, entities per text length and clusters per text length distributions were roughly similar to those of the RuCoCo dataset (Fig. 1, Fig. 2).

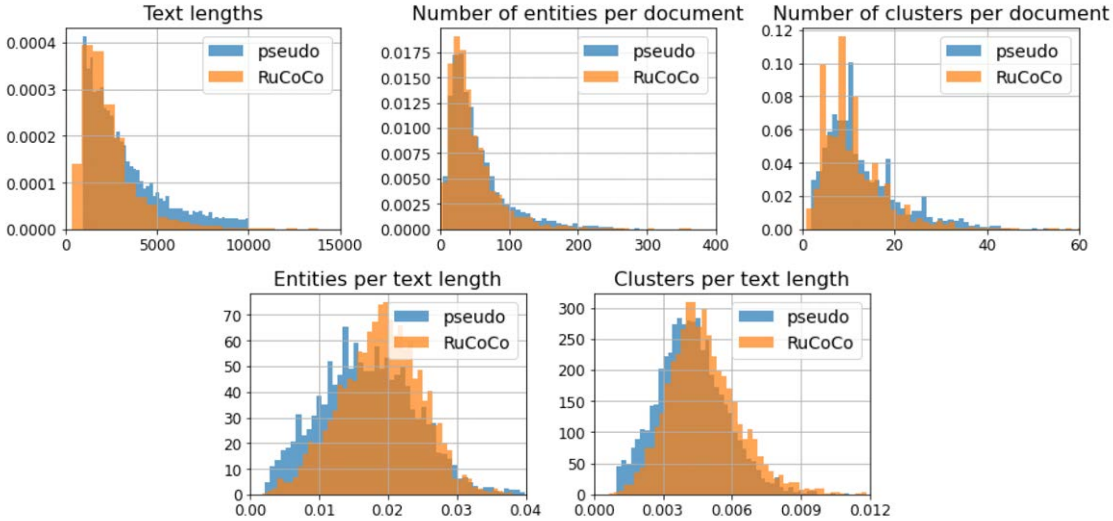


Figure 1: Comparative normalized histograms of two datasets: pseudo-labelled one and preprocessed RuCoCo

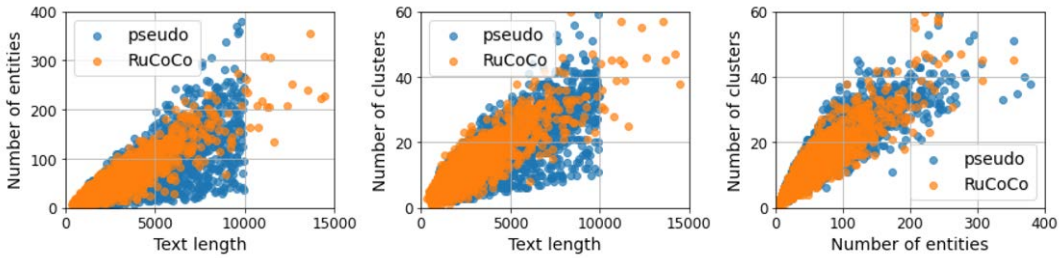


Figure 2: Scatterplots of different features of two datasets: pseudo-labelled one and preprocessed RuCoCo

A total of 3971 selected and labelled documents passed filtering and post processing. Final dataset is described in Table 1. OntoNotes Eng dataset was not used for final model. RuCoCo dataset is splitted into three sets – train, development and test (RuCoCo train split, RuCoCo dev split, RuCoCo test split, respectively) for local evaluation. All datasets together except OntoNotes Eng, RuCoCo dev split and RuCoCo test split are called “extra data” later in the paper.

[14] https://huggingface.co/ai-forever/ruT5-large

Dataset part	Number of documents	Dataset class	Loss weight
RuCoCo train split	2775	gold	1.0
RuCoCo dev split	150	gold	1.0
RuCoCo test split	150	gold	1.0
RuCor	163	gold	1.0
AnCor	521	gold	1.0
OntoNotes Rus	3017	bronze	0.1
Pseudo-labelled	3971	bronze	0.1
OntoNotes Eng[15]	3493	silver	0.5

Table 1: Final dataset parts and their sizes

4 Results and Analysis

For our final setup we used ASP based on ai-forever/ruT5-large utilizing transformers library [17]. We trained this model with input sequence length equal to 1550 tokens, hidden size of ASP action head equal to 4096, batch size equal to 1 for 18 epochs, which took 16 hours on a single nVidia RTX 3090Ti. Final dataset contained 10543 documents, including RuCoCo test split.

In Table 2 we present some results of different setups that we used during the competition. For evaluation we used LEA [18] as main metric for this competition. The last entry in bold is a result of the best model on private test of the competition.

Model	dev F1	test F1	leaderboard F1
baseline + cointegrated/LaBSE-en-ru	0.688	-	0.650
baseline + ai-forever/ruRoberta-large[16]	0.711	-	0.684
spacy-coref + cointegrated/LaBSE-en-ru	0.758	-	0.708
asp + cointegrated/rut5-base[17]	0.741	0.628	0.684
asp + cointegrated/rut5-base-multitask[18]	0.750	0.643	0.698
asp + ai-forever/ruT5-base[19]	0.765	0.650	0.699
asp + ai-forever/ruT5-large	0.791	0.664	0.727
asp + ai-forever/ruT5-large + extra data	0.786	0.667	0.733
asp + ai-forever/ruT5-large + extra data, test split, finetuned	0.799	-	0.738
asp, ai-forever/ruT5-large, extra data, test split, finetuned	**0.799**	**-**	**0.751**

Table 2: Evaluation results of different tested models

None of our models took into account split antecedents. That had an effect on recall metric. We tried to apply some simple models to handle this problem, and these models successfully increased recall, but all at cost of precision. Ultimately we could not achieve F1-score increase by handling split antecedents.

5 Ablation Study

Table 3 describes other experiments that included different base models and training techniques for ASP:

- Model 1 – final best model, added for comparison;
- Model 2 – one of the latest checkpoints of Model 1, but further trained a couple of epochs with only “gold” dataset, pseudo-labelled and translated data is excluded;
- Model 3 – google/mt5-large[20] model, but only with Russian and English tokens in dictionary, which is trained using entire available dataset, i.e. Model 1 dataset and OntoNotes Eng combined.

[15] https://huggingface.co/datasets/conll2012_ontonotesv5
[16] https://huggingface.co/ai-forever/ruRoberta-large
[17] https://huggingface.co/cointegrated/rut5-base
[18] https://huggingface.co/cointegrated/rut5-base-multitask
[19] https://huggingface.co/ai-forever/ruT5-base
[20] https://huggingface.co/google/mt5-large

Model	Dataset size	Training Epochs	Precision	Recall	F1	Private F1
Model 3	15584	4	0.7876	0.7972	0.7924	0.741
Model 2	11265 / 3797	13 / 5	0.7955	0.8083	0.8018	0.750
Model 1	11265	18	0.7925	0.8045	0.7985	0.751

Table 3: Evaluation results of top 3 final models on local development and global private split of RuCoCo. Model 2 was trained in two stages hence the separation in some columns. Dataset size is the number of documents after data preprocessing and it might be different with the initial number of documents in dataset because of long texts split

Model 3 is clearly undertrained and further experiments might bring some positive results. In addition, one can finetune Model 3 on some known tasks with sufficient multilanguage data before training it for coreference resolution.

Another experiments (Table 4) concerned the contribution of various dataset parts to the final result. Here we used ASP with cointegrated/rut5-base-multitask. RuCoCo train split was always a part of training data. Pseudo-labelled data was the same, i.e. acquired with ASP based on ai-forever/ruT5-large trained on RuCoCo.

Added data	dev split			test split			public test		
	P	R	F1	P	R	F1	P	R	F1
No added data	0.7468	0.7528	0.7498	0.753	0.561	0.643	0.739	0.652	0.693
RuCor, AnCor	0.7356	0.7594	0.7473	0.747	0.563	0.642	0.746	0.664	0.702
ONR	0.7417	0.7423	0.7420	0.753	0.544	0.632	-	-	-
PL	0.7589	**0.7895**	**0.7739**	0.748	0.577	0.652	-	-	-
ONR + PL	0.7431	0.7755	0.7590	0.735	**0.581**	0.649	-	-	-
ONE + ONR	0.7469	0.7595	0.7532	0.748	0.562	0.642	0.733	0.665	0.698
All data	**0.7658**	0.7657	0.7658	**0.765**	0.569	**0.653**	0.764	0.686	0.723

Table 4: Evaluation results with same model but different data. ONR – OntoNotes Rus, ONE – OntoNotes Eng, PL – Pseudo-Labelled. “All data” contains all unique datasets above in the table. Public test is what we managed to get while the development phase of the competition was active

6 Conclusions and Future Work

In this paper we present the research of approaches for coreference resolution in Russian language, results and details of our solution for Dialogue Evaluation 2023 RuCoCo competition. Our experiments reveal that the ASP model based on ai-forever/ruT5-large outperforms other coreference resolution models for Russian language, with the use of diverse and expanded training data, including translated OntoNotes and pseudo-labelled data, which significantly contributes to the model's performance. Our solution managed to take the first place in the competition. However, its performance still has room for improvement.

Future work should focus on exploring methods to handle split antecedents effectively. This is the most promising way to improve F1-score for such a task. Another aspect that our study highlights is the importance of diverse training data for model performance improvement. Training on pseudo-labelled data can be effective with small datasets within complex tasks. This technique also needs to be studied more precisely, since there are more ways to apply loss weighting and more data within different domains can be used. And last but not least, other backbone language models are applicable to this problem. One can use a multilanguage model with needed languages only [19] as a base transformer in ASP to more efficiently use datasets in another language, thus increasing the amount and diversity of training data even further.

Acknowledgements

The authors are grateful to colleagues at NTR Labs Machine Learning Research group for the discussions and support and to Prof. Sergey Orlov and Prof. Oleg Zmeev for the computing facilities provided.

References

[1] Brack A. et al. Coreference resolution in research papers from multiple domains //Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43. – Springer International Publishing, 2021. – pp. 79–97.

[2] Liu Z., Shi K., Chen N. F. Coreference-aware dialogue summarization //arXiv preprint arXiv:2106.08556. – 2021.

[3] Morton T. S. Using coreference for question answering //Coreference and Its Applications. – 1999

[4] Kobayashi H., Malon C. Analyzing Coreference and Bridging in Product Reviews //Proceedings of the Fourth Workshop on Computational Models of Reference, Anaphora and Coreference. – 2022. – pp. 22–30.

[5] Stojanovski D., Fraser A. Coreference and coherence in neural machine translation: A study using oracle experiments //Proceedings of the Third Conference on Machine Translation: Research Papers. – 2018. – pp. 49–60.

[6] Lee K., He L., Zettlemoyer L. Higher-order coreference resolution with coarse-to-fine inference //arXiv preprint arXiv:1804.05392. – 2018.

[7] Kirstain Y., Ram O., Levy O. Coreference resolution without span representations //arXiv preprint arXiv:2101.00434. – 2021.

[8] Dobrovolskii V. Word-level coreference resolution //arXiv preprint arXiv:2109.04127. – 2021.

[9] Bohnet B., Alberti C., Collins M. Coreference Resolution through a seq2seq Transition-Based System //Transactions of the Association for Computational Linguistics. – 2023. – T. 11. – pp. 212–226.

[10] Liu T. et al. Autoregressive Structured Prediction with Language Models //arXiv preprint arXiv:2210.14698. – 2022.

[11] Liu T. et al. A structured span selector. // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 2629–2641, Seattle, United States. Association for Computational Linguistics.

[12] Lee K., et al. End-to-end neural coreference resolution. //Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

[13] Beltagy I., Peters M. E., Cohan A. Longformer: The long-document transformer //arXiv preprint arXiv:2004.05150. – 2020.

[14] Toshniwal S. et al. On generalization in coreference resolution //arXiv preprint arXiv:2109.09667. – 2021.

[15] Ju T. S. et al. RU-EVAL-2014: Evaluating anaphora and coreference resolution for Russian //Komp'juternaja Lingvistika i Intellektual'nye Tehnologii. – 2014. – pp. 681–694.

[16] Budnikov A. E. et al. Ru-eval-2019: Evaluating anaphora and coreference resolution for russian //Computational Linguistics and Intellectual Technologies-Supplementary Volume. – 2019.

[17] Wolf T. et al. Transformers: State-of-the-art natural language processing //Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations. – 2020. – pp. 38–45.

[18] Moosavi N. S., Strube M. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric //Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). – 2016. – pp. 632-642.

[19] David Dale. How to adapt a multilingual T5 model for a single language. URL: https://towardsdatascience.com/how-to-adapt-a-multilingual-t5-model-for-a-single-language-b9f94f3d9c90

Light Coreference Resolution for Russian with Hierarchical Discourse Features

Elena Chistova and **Ivan Smirnov**
FRC CSC RAS
Moscow, Russia
{chistova, ivs}@isa.ru

Abstract

Coreference resolution is the task of identifying and grouping mentions referring to the same real-world entity. Previous neural models have mainly focused on learning span representations and pairwise scores for coreference decisions. However, current methods do not explicitly capture the referential choice in the hierarchical discourse, an important factor in coreference resolution. In this study, we propose a new approach that incorporates rhetorical information into neural coreference resolution models. We collect rhetorical features from automated discourse parses and examine their impact. As a base model, we implement an end-to-end span-based coreference resolver using a partially fine-tuned multilingual entity-aware language model LUKE. We evaluate our method on the RuCoCo-23 Shared Task for coreference resolution in Russian. Our best model employing rhetorical distance between mentions has ranked 1st on the development set (74.6% F1) and 2nd on the test set (73.3% F1) of the Shared Task[1]. We hope that our work will inspire further research on incorporating discourse information in neural coreference resolution models.

Keywords: coreference resolution, Rhetorical Structure Theory, referential choice, rhetorical distance, Russian
DOI: 10.28995/2075-7182-2023-22-34-41

Разрешение кореференции для русского языка с использованием признаков иерархического дискурса

Чистова Е. В., Смирнов И. В.
ФИЦ ИУ РАН
Москва, Россия
{chistova, ivs}@isa.ru

Аннотация

Разрешение кореференции – это задача выявления и группировки упоминаний, относящихся к одному и тому же объекту реального мира. При решении задачи методами глубокого обучения в первую очередь обращают внимание на проблемы обучения векторных представлений сущностей и оценки вероятности наличия кореферентной связи между ними. Однако существующие методы не позволяют в явном виде учитывать референциальный выбор в иерархическом дискурсе. В данной работе оценивается важность признаков, полученных на основе автоматического риторического анализа, применительно к нейросетевым моделям. В качестве базового метода реализована end-to-end архитектура с использованием мультиязычной языковой модели LUKE, учитывающей при кодировании текста границы сущностей. Лучшая модель, в которой используется признак риторического расстояния между сущностями, занимает первое место на валидационной (74.6% F1) и второе место на тестовой (73.3% F1) выборке соревнования RuCoCo-2023.

Ключевые слова: разрешение кореференции, теория риторических структур, референциальный выбор, риторическое расстояние, русский язык

[1]The code and models are available at https://github.com/tchewik/corefhd

1 Introduction

Coreference resolution is the task of identifying and grouping mentions referring to the same real-world entity. It is a challenging task in natural language processing, as it often requires both linguistic and common knowledge. In recent years, neural models have achieved remarkable success in coreference resolution. These models aim to identify mention spans and assign pairwise scores. However, they mostly rely on surface explicit features, such as the distance between entities in tokens, and overlook the hierarchical discourse structure. Contextual word embeddings, despite their morphosyntactic and semantic richness, also have limitations in capturing document discourse beyond local cues.

Our system for RuCoCo-2023, called CorefHD (**Coref**erence in **H**ierarchical **D**iscourse), enhances the classical neural architecture with automatically retrieved features that capture aspects of hierarchical discourse. It uses pretrained transformer-based contextualized word embeddings, along with dense embeddings of hierarchical discourse features: linear distance, rhetorical distance, and anaphor-to-LCA distance. To retrieve the discourse hierarchy of the text, we use an RST parser predicting constituency trees in accordance with the Rhetorical Structure Theory (Mann and Thompson, 1988).

The main contributions of this paper are:

- We propose a new method that incorporates discourse information into neural coreference resolution models.
- We test various discourse features that capture the distances between mentions on a large coreference resolution dataset in Russian.
- We apply a number of memory reduction techniques and demonstrate that high-quality coreference resolution can be done with standard neural architecture even with limited computational resources.
- We use a multilingual entity-aware LUKE (Yamada et al., 2020) language model and show that it performs competitively with the monolingual language models for Russian in coreference resolution.
- We join the RuCoCo-2023 Shared Task, and achieve 1st place on the development set and 2nd place on the test set of the contest with the model using the rhetorical distance feature.

The rest of this work is organized as follows: Section 2 reviews a concept of referential distance and current work on coreference resolution in hierarchical discourse. Section 3 describes our method in detail. Section 4 presents our experimental setup. Section 5 analyzes our results. Section 6 concludes the paper and discusses future work.

2 Related Work

Linear referential distance measures how many clauses separate an anaphor from its antecedent (Givón, 1983). However, not all phrases in discourse require the same level of attention. It is observed (Grosz and Sidner, 1986) that the discourse structure of a text contains discourse units inside and outside the intention and attention. Using a corpus of 30 manually annotated texts, it is shown (Cristea et al., 1999) that a hierarchical model of discourse has greater potential for improving the coreference resolution performance than a linear model of discourse. The most popular hierarchical discourse framework as of today is Rhetorical Structure Theory (Mann and Thompson, 1988). Within RST, one can consider in the referential distance the rhetorical structures, where attention focus is part of the definition (Moser and Moore, 1996) of subordinating (mononuclear) RST relations. An approach to computing referential distance with respect to the rhetorical tree is suggested by Kibrik (Kibrik, 1999): the rhetorical distance can be measured by counting the nodes in an RST tree that are visited while walking from the mention to its possible antecedent. A study on the RST Discourse Treebank[2] shows that while rhetorical distance does not imply the one and only referential choice, it is still one of the principal factors for referential choice prediction (Kibrik and Krasavina, 2005). Another study (Fedorova et al., 2010) uses six RST-annotated text fragments in Russian to demonstrate that rhetorical distance has a significant impact on the referent activation in working memory.

Closest to our work are (Khosla et al., 2021) implementing various features over an RST tree produced with a parser for English. However, their main concern is how general is the lowest common ancestor of

[2] https://catalog.ldc.upenn.edu/LDC2002T07

two mentions in the rhetorical constituency tree. While this is somewhat related to the working memory load of keeping two mentions active, they do not directly consider a concept of referential distance and, most importantly, ignore nuclearity (i.e. attention), which is a crucial feature in rhetorical structures.

In this paper, we apply the RST parser for Russian to build hierarchical discourse trees. The distance features obtained from these trees we use in a neural coreference resolution model. As far as we know, we are the first to model referential distances in hierarchical discourse with neural models. We also examine the impact of the RST features in coreference resolution for Russian on a large annotated corpus.

3 Approach

End-to-end coreference resolution involves finding entities in plain text and collecting them into clusters so that each cluster corresponds to a single real-world object. As a core method, we apply the classical (Lee et al., 2018)'s approach to end-to-end coreference parsing with a span-ranking architecture, except for the higher-order inference which has been proven to be ineffective (Xu and Choi, 2020). This approach to coreference resolution involves five main steps:

1. Collect the initial set of spans.
2. Rank the collected spans with a linear transformation of span embeddings and keep the top-k resembling entities.
3. Collect the coarse referent-to-antecedent probabilities for each possible pair of entities. This is calculated as a sum of corresponding span probabilities obtained in the previous step and a score obtained with a bilinear transformation of two *mention encodings*. Keep the top-n pairs with the highest prediction.
4. Compute the final coreference scores for each possible mention-antecedent pair that made it to this step. This is done with a feedforward layer processing *mention pair encodings*. Assign to each mention the antecedent with the highest predicted probability.
5. The predictions form connected chains of mentions that can be viewed as clusters.

The following gives the details of how our system encodes entities and their pairs.

Mention Encoding Each fine-grained token is encoded as an average of its subtoken representations obtained using a language model. The initial entity candidates are collected greedily, with the only parameter being the maximum length of the span. To adjust this parameter effectively, we use token representations instead of LM subtoken representations. Since language models work with a limited context, we collect each paragraph representation separately.

Mention Pair Encoding To calculate the final predictions for each pair of found mentions, we use a feedforward layer that takes a mention pair embedding as input. This embedding consists of the concatenation of two individual mention encodings and the embedding of the token count between them. For the models employing discourse hierarchy features, we represent them similarly to token distances and concatenate them to the pair embeddings.

3.1 Discourse Hierarchy Features

Given two spans i (a mention) and j (its possible antecendent), we first find the elementary discourse units u_i and u_j covering the corresponding spans in a predicted RST tree. Then we compute the discourse-related features and concatenate them with mention pair encoding.

Two metrics are used to measure referential distance in discourse, as outlined in (Kibrik, 1999):

- **Linear Distance (D_{Lin})** in our model is a number of predicted elementary discourse units (EDUs) occurring between two spans.
- **Rhetorical Distance (D_{Rh})** is a number of nuclear EDUs occurring between two spans in a hierarchical rhetorical tree.

We also adopt a feature estimating the amount of generality required to have two mentions in the same discourse subtree (Khosla et al., 2021):

- **Referent's distance to the LCA (D_{LCA})** Assuming mention i always appearing to the right of any possible antecedent j, and $LCA(u_j, u_i)$ being the lowest discourse unit covering both u_i and u_j in the constituency RST tree, $D_{LCA} = \text{dist}(u_i,\ LCA(u_j, u_i))$.

4 Experimental Setup

4.1 Pretrained Language Model

We employ the multilingual LUKE[3] (Ri et al., 2022). It is a language model that has been trained with both masked language modeling (MLM) and masked entity prediction (MEP) tasks. The entity annotations in the training corpus are collected from hyperlinks in Wikipedia dumps. This multilingual model has previously demonstrated significant improvement in question answering and cloze prompt tasks for Russian compared to mBERT (Devlin et al., 2019) and XLM-RoBERTa (Conneau and Lample, 2019). We hypothesize that explicit coreference resolution can also benefit from LM-ingrained entity encoding.

4.2 Factors Reducing Memory Consumption

Neural coreference resolution is a memory-intensive task. The common approach to end-to-end coreference resolution (Lee et al., 2017; Lee et al., 2018) requires computation over each and every span in a document. A number of recent works suggest more optimal alternative methods, in which the object of processing is not a span but a token (Kirstain et al., 2021; Thirukovalluru et al., 2021; Dobrovolskii, 2021). Despite this, the relevant research adopting language model fine-tuning still requires 40 to 80 GB of video memory (Dobrovolskii, 2021; Mæhlum et al., 2022). In our study, we investigate the extent to which the most classical span-based approach to coreference resolution can be scaled down.

Each our model is trained on a single 32GiB Tesla V100 GPU, with peak memory allocation of 98%. To achieve this, we modified the standard model architecture and implementation:

- The main factor that allows a coreference model to be trained on a large dataset with limited memory is *excluding full LM fine-tuning*. In our experiments, a language model is frozen except for the last k layers. The value of k is determined empirically by the amount of video memory available. In our setting, $k = 8$ of 23 layers.
- After initial token encoding, the bidirectional LSTM is used to obtain *lower-dimensional token embeddings*. The span embedding is computed from the sequence of compressed token embeddings using self-attention. In our experiments, $\mathbf{e}_{LM} \in \mathbb{R}^{1024}$ and $\mathbf{e}_{LSTM} \in \mathbb{R}^{100}$.
- *Each paragraph of the text is encoded with a language model separately.* This allows long news articles to be encoded without trimming and high-dimensional partially-trainable LM embeddings to be compressed in place, thereby saving memory.

We also use standard techniques reducing memory requirements:

- *Batch size* = 1. Gradient accumulation did not improve training results.
- All the calculations are performed with *mixed precision*.

4.3 Instruments for Linguistic Analysis

Tokenization and sentence splitting are performed with the Razdel[4] library. Named entities are recognized with the SpaCy[5] `ru_core_news_lg` model predicting BIO-tags from token embeddings. Discourse structures are produced with the IsaNLP RST[6] parser for Russian (Chistova et al., 2021). The parser generates trees for each paragraph; we merged these trees with a right-branching multinuclear Joint relation to construct the full-text RST trees.

4.4 Data

We perform the experiments on the RuCoCo-2023 Shared Task dataset described in (Dobrovolskii et al., 2022). It is a large corpus for coreference resolution collected from news articles in Russian. It contains annotated news in multiple categories, including finance, world news, sports, and more. The corpus

[3] `studio-ousia/mluke-large-lite`
[4] `https://github.com/natasha/razdel`
[5] `https://spacy.io/`
[6] `https://github.com/tchewik/isanlp_rst`

Original	Translation
Обитатели небоскребов Нью-Йорка спешат обзавестись [парашютами]$_{SA}$ Обитатели небоскребов Нью-Йорка спешат обзавестись [парашютами]$_{SA}$. Это связано с недавними терактами в этом городе. Одна из американских фирм по [их]$_{SA}$ производству сообщила, что в офисе не прекращают звонить телефоны. Владельцы квартир в высотных зданиях интересуются возможностью приобретения [новой модели парашюта]$_1$, [которая]$_1$ была разработана после трагических событий 11 сентября. [Он]$_1$ стоит около 800 долларов и раскрывается автоматически. [...]	Residents of New York skyscrapers rush to get [parachutes]$_{SA}$ Residents of skyscrapers in New York rush to get [parachutes]$_{SA}$. This is due to recent terrorist attacks in the city. One of [their]$_{SA}$ manufacturer reports that the phones in its office never stop ringing. Apartment owners in high-rise buildings are interested in buying [a new parachute]$_1$ [which]$_1$ is developed after the tragic events of September 11. [It]$_1$ costs about $800 and opens automatically. [...]

Table 1: Split-antecedent annotation example in the RuCoCo dataset, from `2001_world_new_003`.

includes both single one-to-one coreference annotation and split antecedents one-to-many coreference annotation. However, the distinguishing feature of the latter is that it is annotated among clusters (entities), not mentions (an example is shown in Table 1). It poses a challenge in identifying pairs of mentions from different groups that are connected by split-antecedent relations. To address this additional challenge, our model's architecture would require additional modifications. Although both tasks are evaluated jointly in the competition, this study's emphasis is on the standard coreference resolution. Here, we conduct some additional analyses of the data relevant to our methods.

Firstly, it is critical for our model to determine the maximum entity length in the corpus. The results on the train set are illustrated in Fig. 1. The mean entity length is 2, and the maximum is 42. The maximum mention length in our system is set to 13 tokens, which covers 99.7% of entities in the corpus.

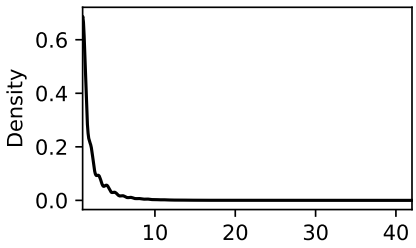


Figure 1: Entity lengths in tokens (train set).

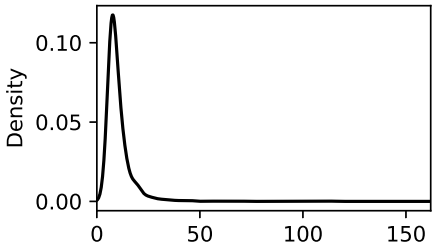


Figure 2: Number of paragraphs per annotation.

Secondly, we examine the number of paragraphs in the data. It will be identical to the number of trees in RST parser output. Thus, if we construct the text-level tree by merging paragraph trees, it could be critical for long discourse dependencies. The results are shown in Fig. 2. Every line split is considered a paragraph. The median paragraph count is 9, with the maximum number of separated lines being 162. Some news articles are exceptionally long, and some of them include enumerated lists. Combining multiple trees into one can affect the referential distance estimation in a few particularly long texts.

4.5 Evaluation

In the Shared Task, the coreference resolution F1 score is calculated using the Link-based Entity Aware (LEA) metric (Moosavi and Strube, 2016). In this metric, the weight of each entity is determined by its size, with larger entities being considered more important. It also evaluates resolved coreference relations instead of resolved mentions.

The models are validated during training using 5% of the official train set. We run random splitting 4 times and report the average result. The listed results on the official development and test sets of the competition are obtained with the exact same models.

5 Results and Discussion

In Table 2, we present the results of our experiments on the development set of the RuCoCo-2023 Shared Task. We also report the performance of our system on the test set (also called the final set) of the RuCoCo-2023 Shared Task in Table 3. Our baseline model noticeably outperforms the `RuRoBERTa-large`-based baseline provided by the organizers, which achieved 68.4% and 67.4% F1 on the development and test sets, respectively.

	Precision	**Recall**	**F1**	**Top-1 F1 (leaderboard)**
Baseline	78.7 ± 0.7	69.1 ± 0.7	73.5 ± 0.5	74.3
+D_{Lin}	78.6 ± 1.8	68.3 ± 2.2	73.0 ± 0.5	74.0
+D_{Rh}	78.5 ± 1.5	69.3 ± 1.0	**73.6 ± 0.9**	**74.6**
+D_{LCA}	75.0 ± 0.8	70.9 ± 1.0	72.9 ± 0.4	73.5

Table 2: Models evaluation on the official development set.

Due to the strict limit on the number of submissions in the final phase of the competition, we could only evaluate the two best performing models, Baseline and Baseline+D_{Rh}, on a private leaderboard.

	Precision	**Recall**	**F1**	**Top-1 F1 (leaderboard)**
Baseline	79.1 ± 0.8	66.9 ± 0.6	**72.5 ± 0.3**	72.8
+D_{Rh}	79.3 ± 1.6	66.6 ± 1.9	72.4 ± 0.5	**73.3**

Table 3: Models evaluation on the official test set ("Final").

Features D_{Lin} and D_{LCA} are not found to be effective for the task of neural coreference resolution on the development set (Table 2). Our hypothesis is that D_{Lin}, the linear distance in elementary DUs, may not offer much more information than the linear distance in tokens that the neural model already employs. D_{LCA}, the distance from the right-hand mention to the LCA, on the other hand, may not be accurate when we artificially merge the RST trees for each paragraph into a single right-branched tree. In this case, the depth of the right-hand branch depends more on the order of paragraphs than the actual discourse structure of the text.

The mean results of the model enhanced with the rhetorical distances D_{Rh} are not much different from the baseline results on both sets. However, its results vary more, hence the model with the best F1 score reached both leaderboards. This suggests to us that the rhetorical distance is more robust than the other features, even though it shares all the mentioned drawbacks of the other features.

6 Conclusion

In this paper, we propose a new method for neural coreference resolution that incorporates discourse information. We test our method on the RuCoCo-2023 Shared Task and demonstrate that it outperforms the competition baseline by a significant margin, while also ranking 1st on the development set and 2nd on the test set of the competition. The key findings of this work are:

1. We implemented various features related to distances in the text-level RST tree to study how the hierarchical discourse information obtained with discourse parser can help coreference resolution for Russian.
2. We observed a marginal improvement using the rhetorical distance feature. The model that uses this feature got the best result on the Shared Task development and test sets.
3. We used the multilingual entity-aware LUKE model and showed that it performs competitively with the monolingual language models for Russian in coreference resolution, even with limited computational resources.

These findings suggest that the multilingual entity-aware LUKE model is a viable option for coreference resolution in Russian, and despite the constraints of the current rhetorical analyzer for Russian that prevent full-text analysis, the features of hierarchical discourse can still be found useful. We hope that our work will inspire further research on incorporating referential distance information into neural coreference resolution models.

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). This study was conducted within the framework of the scientific program of the National Center for Physics and Mathematics, section №9 "Artificial intelligence and big data in technical, industrial, natural and social systems".

References

Elena Chistova, Artem Shelmanov, Dina Pisarevskaya, Maria Kobozeva, Vadim Isakov, Alexander Panchenko, Svetlana Toldova, and Ivan Smirnov. 2021. RST discourse parser for Russian: an experimental study of deep learning models. // *Analysis of Images, Social Networks and Texts: 9th International Conference, AIST 2020, Skolkovo, Moscow, Russia, October 15–16, 2020, Revised Selected Papers 9*, P 105–119. Springer.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. // *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, Red Hook, NY, USA. Curran Associates Inc.

Dan Cristea, Nancy Ide, Daniel Marcu, and Valentin Tablan. 1999. Discourse structure and co-reference: An empirical study. // *The Relation of Discourse/Dialogue Structure and Reference*.

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.

Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. RuCoCo: a new Russian corpus with coreference annotation. // *COMPUTATIONAL LINGUISTICS AND INTELLECTUAL TECHNOLOGIES*. RSUH, June.

Vladimir Dobrovolskii. 2021. Word-level coreference resolution. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 7670–7675, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Olga Fedorova, Ekaterina Delikishkina, and Anna Uspenskaya. 2010. Experimental approach to reference in discourse: Working memory capacity and language comprehension in Russian. // *Proceedings of the 24th Pacific Asia Conference on Language, Information and Computation*, P 125–132, Tohoku University, Sendai, Japan, November. Institute of Digital Enhancement of Cognitive Processing, Waseda University.

Talmy Givón. 1983. Topic continuity in discourse. *Topic continuity in discourse*, P 1–498.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Sopan Khosla, James Fiacco, and Carolyn Rosé. 2021. Evaluating the impact of a hierarchical discourse representation on entity coreference resolution performance. // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 1645–1651, Online, June. Association for Computational Linguistics.

Andrej A Kibrik and Olga N Krasavina. 2005. A corpus study of referential choice: The role of rhetorical structure. *Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference "Dialogue" (2005)*, P 561–569.

Andrej A Kibrik. 1999. Cognitive inferences from discourse observations: reference and working memory. // *Discourse studies in cognitive linguistics. Proceedings of the 5th International cognitive linguistics conference*, P 29–52.

Yuval Kirstain, Ori Ram, and Omer Levy. 2021. Coreference resolution without span representations. // *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 14–19, Online, August. Association for Computational Linguistics.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. // *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, P 188–197, Copenhagen, Denmark, September. Association for Computational Linguistics.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, P 687–692, New Orleans, Louisiana, June. Association for Computational Linguistics.

Petter Mæhlum, Dag Haug, Tollef Jørgensen, Andre Kåsen, Anders Nøklestad, Egil Rønningstad, Per Erik Solberg, Erik Velldal, and Lilja Øvrelid. 2022. NARC – Norwegian anaphora resolution corpus. // *Proceedings of the Fifth Workshop on Computational Models of Reference, Anaphora and Coreference*, P 48–60, Gyeongju, Republic of Korea, October. Association for Computational Linguistics.

William C Mann and Sandra A Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, 8(3):243–281.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 632–642, Berlin, Germany, August. Association for Computational Linguistics.

Megan Moser and Johanna D. Moore. 1996. Toward a synthesis of two accounts of discourse structure. *Computational Linguistics*, 22(3):409–419.

Ryokan Ri, Ikuya Yamada, and Yoshimasa Tsuruoka. 2022. mLUKE: The power of entity representations in multilingual pretrained language models. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 7316–7330, Dublin, Ireland, May. Association for Computational Linguistics.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 3921–3931, Online, August. Association for Computational Linguistics.

Liyan Xu and Jinho D. Choi. 2020. Revealing the myth of higher-order inference in coreference resolution. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 8527–8533, Online, November. Association for Computational Linguistics.

Ikuya Yamada, Akari Asai, Hiroyuki Shindo, Hideaki Takeda, and Yuji Matsumoto. 2020. LUKE: Deep contextualized entity representations with entity-aware self-attention. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 6442–6454, Online, November. Association for Computational Linguistics.

Partitive genitive in Russian: dictionary and corpus data

Oksana Iu. Chuikova
Herzen State Pedagogical University of Russia
oxana.chuykova@gmail.com

Abstract

The paper aims at comprehensive analysis of the verbs compatible with the partitive genitive object. Based on the Dictionary of Russian Language, the list of perfective verbal lexemes that are able to take the genitive object is compiled and semantic features that unite these verbs are revealed. The features are divided into two groups: aspectually relevant features and aspectually irrelevant features. The corpus-based analysis of the use of the verbs that take both genitive and accusative objects makes it possible to identify features that increase the likelihood of certain object case-marking.

Keywords: Russian language; partitive genitive; incremental relation, Aktionsart; corpus-based study
DOI: 10.28995/2075-7182-2023-22-42-50

Родительный партитивный в русском языке: словарные и корпусные данные

Оксана Юрьевна Чуйкова
Российский государственный педагогический
университет им. А. И. Герцена
oxana.chuykova@gmail.com

Аннотация

В статье предпринята попытка последовательного анализа глаголов, способных к употреблению с так называемым родительным партитивным в позиции прямого дополнения. По данным Малого академического словаря составлен перечень перфективных глагольных лексем, для которых отмечена возможность генитивного управления, и выявлены семантические признаки, объединяющие эти глаголы. Выявленные признаки подразделяются на аспектуально релевантные и аспектуально нерелевантные. Корпусный анализ примеров употребления глаголов, для которых в МАС зафиксирована возможность вариативного управления, позволяет определить семантические особенности, влияющие на предпочтительный выбор аккузативного или генитивного оформления объекта.

Ключевые слова: русский язык, родительный партитивный; накопительное отношение, способ действия; корпусное исследование

1 Теоретические предпосылки исследования

Исследование, результаты которого изложены в настоящей статье, посвящено рассмотрению особенностей употребления так называемого родительного (далее – род.) партитивного в позиции прямого дополнения в русском языке.

Следует отметить, что данная проблематика не нова и при этом редко вызывает дискуссии. Наиболее известным в литературе сюжетом является рассмотрение употребления род. партитивного в связи с грамматическим значением вида глагола, см. [7], [19] [10: 182–190], [11]. В частности, отмечается несочетаемость род. партитивного с глаголами несовершенного вида, при этом указывается, что это ограничение, как и запрет на употребление с количественными группами (*пить *молока*/**стакан молока*), касается только случаев использования несовершенного вида в актуально-длительном значении.

В ряде более современных работ также затрагиваются вопросы об особенностях употребления род. партитивного, в том числе в сопоставительном аспекте, напр., [16]. Представления об употреблении род. партитивного можно обобщить следующим образом.

1) Род. партитивный является одним из трех основных типов приглагольного употребления род. падежа (наряду с употреблением при интенсиональных[1] глаголах и в контексте отрицания) в русском и других славянских языках [15: 35-36, 415], [8], [18: 316–317].

2) Как правило, род. партитивный характеризуется рядом селективных ограничений, касающихся семантики глагола и свойств объекта. Отмечается, что род. партитивный употребляется с ограниченным и очень небольшим числом глаголов, обозначающих поглощение и перемещение [16: 356], [8: 28], а также накопление [18: 319–320], а в позиции дополнения употребляются именные группы, обозначающие вещественный или множественный объект.

3) В рамках двухкомпонентной теории вида условием для употребления род. партитивного, если объект является «накопителем эффекта» (градуальным пациенсом, инкрементальной темой), т.е. вовлекается в ситуацию последовательно, является характеристика глагола как перфективного (первый компонент), но «некульминативного» (non-culminating), т.е. обозначающего ситуации, не достигающие предела (второй компонент). В качестве примера, в частности, приводится русский делимитатив [16: 386].

4) Употребление род. партитивного может быть обусловлено наличием глагольных квантификаторов, роль которых выполняют глагольные префиксы. Наиболее часто в связи с род. партитивным упоминается префикс *на-* (напр., [5], [13], [14]); см. также о комбинации *по-на-* [16: 370].

Как представляется, по крайней мере некоторые из перечисленных выше утверждений нуждаются в уточнении. Например, какие именно глагольные лексемы входят в число способных к употреблению с род. партитивным? Если их число очень невелико, то они должны поддаваться исчислению. То же самое касается и других пунктов: ответы на вопросы, является ли делимитатив единственным случаем некульминативного перфектива, и каков набор глагольных приставок-квантификаторов, обусловливающий возможность употребления род. партитивного, можно получить путем установления множества глаголов, способных к употреблению с род. падежом, и анализа их реального употребления. Исследование, результаты которого представлены ниже, преследует следующие задачи: 1) установление круга глагольных лексем, способных к употреблению с род. партитивным, и составление перечня характеризующих их признаков; 2) выявление признаков/комбинаций признаков, определяющих предпочтительный выбор аккузативного или генитивного объекта при возможности вариативного управления.

2 Материал и процедура исследования

Для решения поставленных задач поэтапно была реализована следующая исследовательская процедура.

1) Установление круга глаголов, способных к управлению род. партитивным, по данным МАС (Малый академический словарь – [4]).

По МАС методом сплошной выборки был получен список неинтенсиональных переходных[2] глагольных лексем совершенного[3] вида (далее – СВ), для которых в рамках указанного словаря зафиксирована возможность управления род. падежом (в грамматическом комментарии дается указание на падежные вопросы: «что и чего» или «что и кого-чего», либо на переходность и падежный вопрос род. падежа: «перех. и чего»). Способность употребляться с род. падежом считается характеристикой лексемы (согласно определению Ю.Д. Апресяна, «слово, рассматриваемое

[1] Согласно определению Н.Д. Арутюновой, «[о]бъект интенсиональных глаголов обладает ментальным существованием» и противопоставлен реальному объекту [2: 57].

[2] Переходными считаются глаголы, способные к управлению винительным падежом без предлога по крайней мере в некоторых употреблениях.

[3] Отбор для анализа исключительно перфективных лексем обусловлен в первую очередь тем, что в МАС указание на возможность управления род. партитивным обнаруживается в грамматическом комментарии для глаголов СВ и отсутствует у соответствующих имперфективных глаголов. Можно предположить, что составители словаря последовательно придерживались распространенной точки зрения о наличии связи между грамматическим значением вида глагола и допустимостью род. партитивного, что отличает данный тип употребления род. падежа в позиции прямого дополнения от двух других (при интенсиональных глаголах и в отрицательных контекстах), где такой связи не наблюдается, см. [3: 28].

в одном из имеющихся у него значений» [1: 55]), при вхождении в словарную статью более одной лексемы сочетаемость / несочетаемость с род. падежом определяется отдельно для каждого лексического значения, а в случае если информация о сочетаемости с род. падежом дается в общем грамматическом комментарии, генитивное управление считается возможным для всех лексем, при этом в списке они учитываются отдельно.

2) Рассмотрение каждой глагольной лексемы с точки зрения характеризующих ее семантических признаков.

Перечень признаков составлен автором с опорой на существующие исследования, посвященные род. партитивному, а также на собственные наблюдения. Для оценки регулярности и продуктивности род. партитивного в условиях действия каждого из признаков рассматривались также примеры употребления в основном подкорпусе НКРЯ (Национальный корпус русского языка: https://ruscorpora.ru) и русскоязычном Интернете (рунете) глаголов, для которых в МАС не отражена возможность употребления с род. падежом[4]. Признаки условно подразделяются на аспектуально релевантные (т.е. связанные с характеристикой структуры ситуации и особенностями ее существования во времени) и аспектуально нерелевантные. К числу аспектуально релевантных признаков, например, относится последовательность вовлечения объекта в ситуацию и принадлежность к морфемно характеризованным способам действия (далее – СД)[5].

3) Установление по данным НКРЯ количественного соотношения род. и винительного (далее – вин.) падежей и наличия корреляции между предпочтительным выбором падежного управления и семантическими характеристиками глагола.

Отбор материала осуществлялся следующим образом: для глаголов из ранее определенного списка по данным НКРЯ методом ручного отбора был составлен перечень и подсчитано количество всех зафиксированных в корпусе употреблений с дополнением в форме род. и вин. падежей, выраженным кумулятивной именной группой без количественных модификаторов (существительным с вещественной, реже отвлеченной, семантикой либо существительным в множественном числе, напр., *попить воды/воду*, *прибавить скорости/скорость*, *нарубить дров/дрова*). За пределами рассмотрения остаются случаи возможной вариативности падежного оформления прямого дополнения в отрицательных контекстах, а также примеры, где падежная форма не определяется однозначно (например, *попить кофе*). Случаи употребления в составе устойчивых сочетаний (напр., *подлить масла в огонь* ‘усугубить’, *набрать в рот воды* ‘молчать’) также не учитываются либо учитываются отдельно.

3 Перечень перфективных глаголов, способных к употреблению с родительным партитивным, и их признаки

Путем сплошного поиска по МАС обнаружены 596 глагольных лексем СВ, в грамматическом комментарии к которым содержится указание на сочетаемость с род. падежом и которые не являются интенсиональными глаголами. При этом можно заметить, что глаголы, которые употребляются с род. партитивным, с точки зрения семантики объединены небольшим количеством неоднородных и невзаимоисключающих семантических признаков, таких как **инкрементальность** (последовательность вовлечения объекта в ситуацию: *потереть* <морковь>, *собрать* <грибы>) – 554 лексемы, принадлежность к **СД** (*покапать*, *понабрать*, *навалить*, *подсыпать*) – 535 лексем, отнесенность к одной из лексико-семантических групп: **поглощение** (**«ингестивные»** глаголы: *выпить*, *съесть*, *попить*, *испить*, *отглотнуть*) – 56 лексем, **приготовление/обработка** (как правило, продуктов питания: *почистить*, *наварить*, *поджарить*, *спечь*) – 53 лексемы, **передача/перемещение** (*добыть*, *призанять*, *задать* <корм лошади>) – 64 лексемы, **экспериенциальность** (немногочисленная группа лексем, часто метафорически связанных с ингестивными глаголами: *понюхать* <войны>, *хлебнуть* <тяжелой и непростой жизни>), а также такие признаки как

[4] В дополнение к основному перечню рассматриваемых глагольных лексем был составлен список глаголов, в отношении которых было сделано предположение о способности к употреблению с род. партитивным на основании наличия одного или нескольких признаков, характеризующих глаголы, для которых в МАС такая способность отмечена. Далее предположение проверялось на материале НКРЯ и рунета.

[5] В настоящей работе принимается такое определение СД: [с]пособами глагольного действия принято называть различные типы семантических модификаций глагола, выраженные определенными формальными средствами [20: 110]. При составлении перечня СД использовались классификации, представленные в работах Анны А. Зализняк, И. Л. Микаэлян и А. Д. Шмелева [20: 110–135] и М. А. Шелякина [17: 141–167], а также принципы, приведенные в [3], [6].

большое количество объекта (*понабрать, нагромоздить*) – 341 лексема, **экспрессивное/переносное** значение (*навернуть* <каши>, *напороть* <чепухи>) – 55 лексем.

Инкрементальность, или «накопительное отношение» (“incremental relation”, термин введен Д. Даути) – важное понятие аспектуальной композиции, которое предполагает, что объект ситуации, представляющий собой «накопитель эффекта» (в русскоязычной литературе понятие впервые использовано Е. В. Падучевой [12], см. также [9]), вовлекается в ситуацию последовательно, т.е. степень вовлечения объекта в ситуацию меняется пропорционально временно́й протяженности ситуации. Большинство глаголов, для которых в МАС зафиксирована возможность генитивного управления (554 из 596, 92,9%) выражают ситуации с накопительным отношением. При этом следует отметить, что в число таких глаголов включались лексемы, выражающие не только ситуации, в которых по мере ситуации меняется количество объекта, но и ситуации с абстрактным объектом, в которых протяженность ситуации связана с силой, интенсивностью, степенью проявления признака (напр., *набрать сил*). Можно заметить, что семантика поглощения, а также такие признаки как принадлежность к СД, экспериенциальность, большое количество объекта и экспрессивная окраска в нашей выборке лексем, для которых в словаре отмечена сочетаемость с род. падежом, обнаруживаются только у глаголов с накопительным отношением. Эту группу признаков отличает наличие аспектуальной релевантности, т.е. связи со структурой ситуации и характеристиками ее существования во времени.

Анализ материала НКРЯ и рунета позволяет обнаружить случаи реализации способности к генитивному управлению для каждого из 596 глаголов. Кроме того, в случае наличия перечисленных выше признаков род. падеж может употребляться и при глаголах, для которых в МАС отсутствует указание на сочетаемость с род. падежом. Приведем несколько примеров.

(1) экспериенциальность: *В некоторых из них впоследствии и мне пришлось* ***изведать*** *счастья.* [Анна Ларина (Бухарина). Незабываемое (1986-1990)]

(2) ингестивный: *В Абхазии очень много зелени и очень много мусора вдоль этих дорог, наверное это местный обычай — попил воды,* ***искурил*** *сигарет, швырнул мусор за борт* (http://idiot.fm/2011/08/16/a-vlasti-skryvayut-2)

(3) большое количество+экспериенциальность: *[Техник-смотритель, жен, техник-смотритель] Это я тебе говорю. Я этих контор* ***видела-перевидела****. Если наши кого примут — нигде не пропадешь.* [А. Н. Попов. Потом... потом... потом... (1975-1979)]

(4) приготовление/обработка: *Один предложил: штей бы ему.* ***Остудили*** *щей. Обмакнул Афанасий Петрович палец в щи и в рот ребенку.* [Вс. В. Иванов. Дитё (1922)]

(5) передача/перемещение: *На другой день я списался с Синусом,* ***отослал*** *ему денег для забронировать номерок в гостинице, ну и пожалуй, на этом подготовка была закончена*. (https://sinusmoto.ru/forum/showthread.php?t=35161#.YhW2yb1Bz_Q)

(6) инкрементальность + СД (делимитативный): *А может ещё "загадочных картинок"* ***попубликовать****?* (https://www.liveinternet.ru/users/1045060/post323873111)

(7) ингестивный + СД (семельфактивный) + переносное значение: *ещё советуют двинуть водки с пертцем и пропотеть под кучей одеял*. (https://pikabu.ru/story/pikabu_proshu_pomoshchi_2215011)

4 Количественное соотношение родительного партитивного и винительного падежей (по данным НКРЯ)

Ниже приведены результаты анализа данных об употреблении в НКРЯ глаголов с зафиксированной в словаре способностью к генитивному управлению с точки зрения количественного соотношения род. и вин. падежей. Рассмотрение осуществляется на выборке глаголов, для которых в

НКРЯ обнаруживаются случаи сочетания хотя бы с одной падежной формой (460 лексем), соотношение род. и вин. падежей определяется путем применения ряда статистических методов.

В первую очередь, оценим возможную корреляцию между предпочтительным выбором род. падежа при глаголе и набором его характеристик.

Порядок анализа: 1. для каждого глагола распределение род. и вин падежей сравнивалось с суммарными данными о распределении падежных форм при всех остальных глаголах выборки при помощи критерия согласия хи-квадрат; 2. из 460 рассматриваемых глаголов отобраны 176 лексем, которые демонстрируют статистически значимое различие в распределении падежных форм по отношению к суммарным данным (при $p<0.05$); 3. полученные глагольные лексемы разделены на два подмножества: глаголы со статистически значимым преобладанием род. падежа (gen_pref: род. > вин.) и глаголы со статистически значимым преобладанием вин. падежа (acc_pref: вин. > род.). 4. глаголы полученных подмножеств рассмотрены с точки зрения корреляции с перечисленными выше признаками на тепловой карте.

Тепловая карта (Рис.1) показывает коэффициенты корреляции Пирсона (r) [6] между предпочтительным выбором одной из моделей управления (gen_pref и acc_pref) и каждым из признаков, а также между самими признаками (попарно).

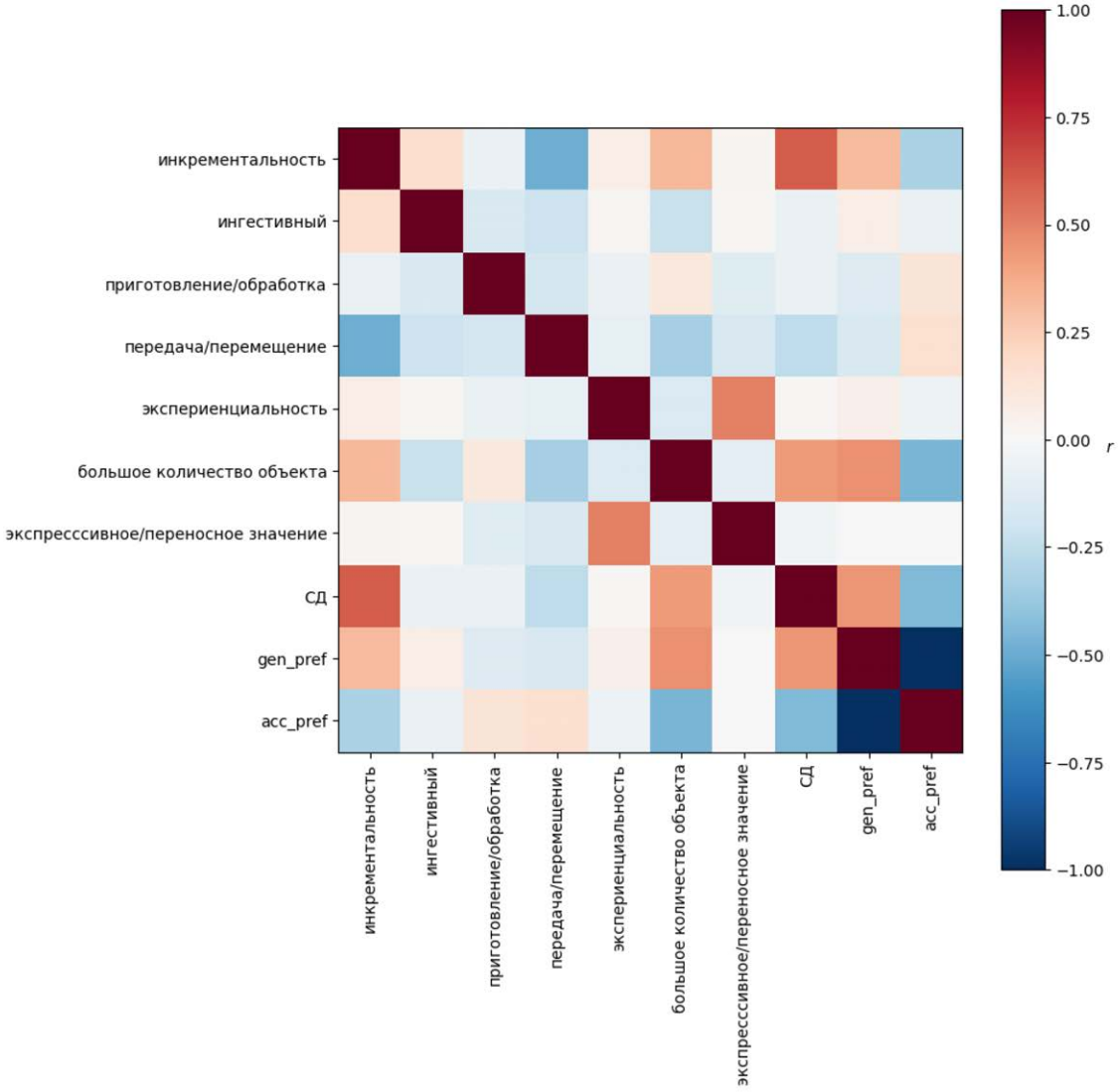


Рисунок 1: Матрица корреляции признаков и моделей управления в русском языке

[6] Значения коэффициентов корреляции r лежат в диапазоне от -1 (максимальная отрицательная корреляция) до 1 (максимальная положительная корреляция.

Рис.1 показывает наличие наибольшей положительной корреляции между предпочтительным выбором род. падежа прямого дополнения (gen_pref) и такими признаками, как инкрементальность, принадлежность глагола к СД и большое количество объекта, и наличие корреляции (меньшей степени) между выбором род. падежа и ингестивным и экспериенциальным значениями. Также можно констатировать наличие отрицательной корреляции между выбором генитивного управления и выражением глаголом значений передачи/перемещения и приготовления/обработки. Противоположная картина взаимодействия с признаками наблюдается для глаголов с предпочтительным выбором аккузативного управления (acc_pref). Кроме того, можно говорить о наличии связи между инкрементальностью с одной стороны и принадлежностью к СД, ингестивным и экспериенциальным глаголам и значением большого количества объекта – с другой. Значения передачи/перемещения и приготовления/обработки показывают отрицательную корреляцию практически со всеми признаками (исключение составляет значение большого количества объекта, что может объясняться наличием среди глаголов кумулятивного СД некоторого количества лексем с семантикой приготовления пищи/обработки: *нажарить*, *наварить*, *напечь*, *насушить* и др.).

Таким образом, Рис. 1 подтверждает высказанное выше предположение о наличии связи между значением инкрементальности и принадлежностью к СД (в наибольшей степени), ингестивным, экспериенциальным значением и значением большого количества, а также показывает, что этот набор признаков, которые можно условно охарактеризовать как аспектуально релевантные, коррелирует с предпочтительным выбором род. партитивного в позиции прямого дополнения.

Далее, распределение род. и вин. падежей при глаголах, характеризующихся перечисленными выше признаками, анализируется при помощи критерия согласия хи-квадрат. Анализ осуществляется на полной выборке глаголов, для которых в НКРЯ обнаруживаются случаи сочетания хотя бы с одной падежной формой (460 лексем). Результаты представлены в Табл. 1.

Данные Табл. 1 подтверждают сделанное на основании приведенной выше тепловой карты предположение, что такие признаки, как инкрементальность, принадлежность к СД, ингестивным и экспериенциальным глаголам и большое количества объекта повышают вероятность употребления род. падежа, а приготовление/обработка и перемещение/передача – вин. падежа.

значение признака	**РП**	**% РП**	**ВП**	**% ВП**	**p-value**
+ инкрементальность	21697	69.03	9732	30.97	$\chi2$= 4829.392, p<<0.01
- инкрементальность	1669	24.05	5271	75.95	
+ СД	18143	74.97	6057	25.03	$\chi2$= 5448.765, p<<0.01
- СД	5223	36.86	8946	63.14	
+ ингестивный	7232	71.47	2887	28.53	$\chi2$= 644.401, p<<0.01
- ингестивный	16134	57.11	12116	42.89	
+ большое количество	5731	84.73	1033	15.27	$\chi2$= 1957.041, p<<0.01
- большое количество	17635	55.80	13970	44.20	
+ экспериенциальность	622	67.90	294	32.10	$\chi2$= 19.042, p<<0.01
- экспериенциальность	22744	60.73	14709	39.27	
+ перемещение	4698	47.80	5131	52.20	$\chi2$= 951.682, p<<0.01
- перемещение	18668	65.41	9872	34.59	
+ приготовление	1071	52.45	971	47.55	$\chi2$= 64.291, p<<0.01
- приготовление	22295	61.37	14032	38.63	
Всего	**23366**	**60.90**	**15003**	**39.10**	

Таблица 1: Распределение род. и вин. падежей для глаголов с положительными (+) и отрицательными (-) значениями признаков

Наконец, проверим утверждение о преобладании род. партитивного при положительном значении таких признаков, как инкрементальность, принадлежность к СД и ингестивным глаголам, большое количество объекта и экспериенциальность. По крайней мере применительно к глаголам с положительным значением первых четырех признаков данная гипотеза подтверждается статистически, см. Табл. 2. Следует, однако, заметить, что в группах инкрементальных, ингестивных

глаголов и предикатов, предполагающих большое количество объекта, значительную часть составляют глаголы СД. При изъятии СД в каждой группе остаются глаголы, не демонстрирующие единой тенденции к предпочтительному выбору род. падежа. Также преобладания род. падежа не наблюдается в группе глаголов, не относящихся к СД в целом.

признак	**Кол-во лексем (всего)**	**Кол-во лексем (из 460)**	**W**	**p-value**
СД	535	400	61654.5	2.94e-32
не СД	61	60	526.5	0.99
инкрементальный:				
всего	554	421	65216.5	3.20e-27
СД	524	392	59249	5.31e-32
не СД	30	29	174.5	0.74
ингестивный:				
всего	56	51	944	0.00013
СД	42	37	553.5	4.66e-05
не СД	14	14	57	0.21
большое количество:				
всего	341	252	26714.5	2.64e-32
СД	340	251	26480.5	3.96e-32
не СД	1	1	–	–

Таблица 2: Результаты применения W-критерия для проверки гипотезы М(вин. п.) < М(род. п) для глаголов СД / не СД

Таким образом, можно сделать вывод, что принадлежность к СД является важнейшей характеристикой глагола, влияющей на выбор род. падежа как предпочтительного средства оформления прямого дополнения. В качестве косвенного подтверждения данного вывода можно привести и тот факт, что на долю СД приходится 89,8% всех глаголов, для которых в МАС зафиксирована возможность генитивного управления (535 из 596), среди глаголов с ненулевым числом употреблений с дополнением в форме род. и/или вин. падежа в НКРЯ доля СД составляет 86,96 (400 из 460), а также то, что при глаголах отдельных СД, как и по данной группе с целом, наблюдается преобладание случаев употребления с род. партитивным, см. [3]. Двумя указанными факторами – преобладанием среди глаголов, для которых в МАС отмечена возможность управления род. партитивным, лексем, относящихся к СД, и предпочтительным выбором род. падежа прямого дополнения при глаголах данной группы – можно объяснить наблюдаемое выше в Табл. 1 суммарное распределение случаев употребления род. и вин. падежей при рассматриваемых глаголах по данным НКРЯ.

5 Заключение

Подведем итоги исследования.

В первую очередь следует отметить, что существующие представления о принципах употребления род. партитивного могут быть уточнены и дополнены путем последовательного анализа словарных и корпусных данных.

В результате сплошного поиска по МАС составлен список всех перфективных лексем, для которых в рамках данного словаря отмечена возможность генитивного управления. Отмечено, что глаголы, способные к управлению род. партитивным, объединены небольшим количеством неоднородных семантических признаков, таких как инкрементальность (последовательность вовлечения объекта в ситуацию), принадлежность к морфемно характеризованным СД, отнесенность к одной из лексико-семантических групп: поглощение, приготовление/обработка, передача/перемещение, экспериенциальная семантика, большое количество объекта, экспрессивное/переносное значение. Для каждого глагола из полученного списка в НКРЯ и/или рунете обнаруживаются примеры употребления с род. партитивным. При наличии одного или нескольких из

перечисленных признаков случаи генитивного оформления объекта обнаруживаются также при глаголах, для которых в МАС не отмечена возможность управления род. падежом.

Признаки, характеризующие глаголы, способные к употреблению с род. партитивным, подразделяются на две группы, которые можно определить как аспектуально релевантные (т.е. связанные со структурой ситуации и особенностями ее существования во времени, к ним относится инкрементальность как объединяющий признак, а также принадлежность к способам действия, семантика поглощения, экспериенциальная семантика и др.) и аспектуально нерелевантные (значения приготовления/обработки, передачи/перемещения). Анализ показывает, что признаки, относящиеся к первой группе, влияют на предпочтительный выбор род. партитивного как средства оформления прямого дополнения, при этом наиболее значимой в этом отношении характеристикой является принадлежность к СД.

Благодарности

Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 19-312-60006 «Прямое дополнение и аспектуальные характеристики славянского глагола».

Acknowledgements

The reported study was funded by the Russian Foundation for Basic Research (RFBR), project number 19-312-60006.

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] Arutyunova N. D. Sentence and its meaning. Logical and semantic problems [Predlozhenie i ego smysl. Logiko-semanticheskie problemy]. — Moscow: Nauka, 1976.

[3] Chuikova O. (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, 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, Moscow. Access mode: http://feb-web.ru/feb/mas/mas-abc/14/ma239217.htm

[5] Filip Hana. On accumulating and having it all, H. Verkuyl, H. de Swart, & A. van Hout (eds.). Perspectives on aspect. — Dordrecht: Springer, 2005. — P. 125–148.

[6] 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.

[7] Jakobson R. (1936), Beitrag zur allgemeinen Kasuslehre. Gesamtbedeutung der russischen Kasus, Travaux du cercle linguistique de Prague, 6, pp. 240–288.

[8] Kagan O. Semantics of genitive objects in Russian: A study of genitive of negation and intensional genitive case (Studies in Natural Language and Linguistic Theory: Vol. 89). — Dordrecht: Springer, 2013.

[9] Mehlig H. R. The interaction between the verbal aspect and "incremental themes" in Russian [Vzaimodejstvie mezhdu vidom i "nakopitelyami" v russkom yazyke]. // A. V. Bondarko, G. I. Kustova, R. I. Rozina (eds.). Dynamic Models: Word, sentence, text. In honour of E. V. Paducheva [Dinamicheskie modeli: Slovo. Predlozhenie. Tekst: Sbornik statej v chest' E. V. Paduchevoj]. — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul'tury], 2008. — P. 562–593.

[10] Paducheva Elena V. Semantic studies: Semantics of tense and aspect in Russian; Semantics of the narrative [Semanticheskie issledovaniya: Semantika vremeni i vida v russkom yazyke. Semantika narrativa]. — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul'tury], 1996.

[11] Paducheva Elena V. (1998), On non-compatibility of partitive and imperfective in Russian, Theoretical linguistics, vol. 24 (1), pp. 73–82.

[12] Paducheva Elena V. (2004), The "incremental theme" and Russian aspectology [«Nakopitel' effekta» i russkaya aspektologiya], Topics in the study of language [Voprosy yazykoznaniya], 5, pp. 46-57.

[13] Pereltsvaig Asya (2006), Small nominlas, Natural Language and Linguistic Theory, vol. 24, pp. 433–500.

[14] Romanova Eugenia. Constructing Perfectivity in Russian: Ph.D. Dissertation. — University of Tromsø, Tromsø, 2006
[15] Russian grammar [Russkaya grammatika]. Ed. by N. Yu. Shvedova). — Moscow: Nauka, 1980. Vol. 2.
[16] Seržant Ilja A. (2015), Independent partitive as a Circum-Baltic isogloss, Journal Language Contact, 8, pp. 341–418.
[17] Shelyakin M. A. Category of aspectuality of the Russian verb [Kategoriya aspektual'nosti russkogo glagola]. — Moscow: URSS, 2008.
[18] Timberlake A. A reference grammar of Russian. — Cambridge: Cambridge University Press, 2004.
[19] Wierzbicka Anna. On the semantics of the verbal aspect in Polish // To honor Roman Jakobson. — The Hague–Paris: Mouton, 1967. — P. 2231–2249.
[20] 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.

Bimodal sentiment and emotion classification with multi-head attention fusion of acoustic and linguistic information

Anastasia Dvoynikova
St. Petersburg Federal Research Center of the Russian Academy of Sciences / Saint-Petersburg, Russia
dvoynikova.a@iias.spb.su

Alexey Karpov
St. Petersburg Federal Research Center of the Russian Academy of Sciences / Saint-Petersburg, Russia
karpov@iias.spb.su

Abstract

This article describes solutions to couple of problems: CMU-MOSEI database preprocessing to improve data quality and bimodal multitask classification of emotions and sentiments. With the help of experimental studies, representative features for acoustic and linguistic information are identified among pretrained neural networks with Transformer architecture. The most representative features for the analysis of emotions and sentiments are EmotionHuBERT and RoBERTa for audio and text modalities respectively. The article establishes a baseline for bimodal multitask recognition of sentiments and emotions – 63.2% and 61.3%, respectively, measured with macro F-score. Experiments were conducted with different approaches to combining modalities – concatenation and multi-head attention. The most effective architecture of neural network with early concatenation of audio and text modality and late multi-head attention for emotions and sentiments recognition is proposed. The proposed neural network is combined with logistic regression, which achieves 63.5% and 61.4% macro F-score by bimodal (audio and text) multitasking recognition of 3 sentiment classes and 6 emotion binary classes.

Keywords: sentiments; emotions; CMU-MOSEI; attention mechanism; bimodal; multitask
DOI: 10.28995/2075-7182-2023-22-51-61

Бимодальная классификация сентимента и эмоций на основе объединения акустической и лингвистической информации с помощью механизма внимания

Двойникова А. А.
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук / Санкт-Петербург, Россия
dvoynikova.a@iias.spb.su

Карпов А. А.
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук / Санкт-Петербург, Россия
karpov@iias.spb.su

Аннотация

Данная статья посвящена описанию решений нескольких задач: предобработка базы данных CMU-MOSEI для улучшения качества данных и bimodal multitask классификация эмоций и сентимента. С помощью экспериментальных исследований выявляются репрезентативные признаки для акустической и лингвистической информации среди предобученных нейронных сетей с архитектурой Transformer. Наиболее репрезентативными признаками для анализа эмоций и сентимента являются EmotionHuBERT и RoBERTa для аудио и текстовой модальности, соответственно. В статье устанавливается baseline для бимодального многозадачного распознавания сентимента и эмоций – 63,2 % и 61,3 % macro F-score, соответственно. Также проводятся эксперименты с различным подходами к объединению модальностей – конкатенация multi-head attention. Предлагается наиболее эффективная архитектура нейронной сети с ранней конкатенацией аудио и текстовой модальности и позднем multi-head attention для распознавания эмоций и сентимента. Предложенная нейронная сеть объединяется с логистической регрессией, с помощью чего достигается 63,5 % и 61,4 % macro F-score при бимодальном (аудио- и текстовый) многозадачном распознавании 3 классов сентимента и 6 бинарных классов эмоций.

Ключевые слова: сентимент; эмоции; CMU-MOSEI; механизм внимания; бимодальность; многозадачность

1 Introduction

Many existing studies are devoted to recognition of emotions and sentiments, because this area is in demand and there are still many unsolved problems [1-4]. People express emotions through visual, verbal and non-verbal manifestations. Based on this, developing a system for recognizing emotions and sentiments, it is necessary to analyse as many different sources of emotion manifestation as possible (video, audio, text modality) [5]. Based on the specifics of the data and the task, each modality can make a different contribution to the reliability of the system [6]. Therefore, it is important to conduct experimental studies to identify representative modalities for each task.

There are several approaches to emotions and sentiments recognition: emotions and sentiments analyses separately [2, 3, 5] and together (multitask) [4]. Multitasking systems have advantages in summarizing information better and finding correlations between different tasks.

In this article, we solve 2 important tasks: preprocessing the CMU-MOSEI [7] database, and bimodal multitask recognition of emotions and sentiments with different approaches to fusion modalities. A multimodal CMU-MOSEI database was used for experimental studies. This corpus has some problems, such as incorrect timings of speech utterances, extracted subtitles from videos instead of transcriptions of speakers' speech. Therefore, the CMU-MOSEI data corpus has been significantly modified by semi-automatic methods. This was done to improve the quality of the data. However, the experimental studies carried out on the modified data corpus become incomparable with existing studies with this database. Therefore, in this article we are setting a baseline for multitask recognition of multiclass sentiments (3 classes) and multilabel emotions (6 classes) by acoustic and linguistic information of speech utterances.

The article contains the following structure: Section 2 presents an analysis of existing solutions in the field of multimodal and multitask classification of emotions and sentiments on the CMU-MOSEI database. Section 3 describes the CMU-MOSEI data and the data processing algorithm. Section 4 contains experimental studies aimed at identifying relevant acoustic and linguistic features, establishing a baseline and bimodal multitask approach to emotions recognition and sentiments using various methods of merging modalities. Conclusions are given in Section 5.

2 Related Work

All the articles described in this section relate to research with the CMU-MOSEI database. Some researchers [1, 2] suggest later or early fusion of modalities using concatenation. This approach supposes the equal importance of information in each modality. In the real world, the relevance of information from each modality is unbalanced. This is due to presence of noise in the data, equipment malfunctions during recording, etc. More complex approaches to fusion modalities, such as hierarchical [8], attention mechanisms [1-4, 6, 9], allow to get a more reliable automatic system for recognizing emotions and sentiments. The authors of most recent studies [1-3, 10] use multi-head attention (MHA) to combine modalities in the tasks of emotion and sentiments classification. The advantages of MHA are that the algorithm uses several parallel streams of self-attention (head attention). This allows to find more relationships in the information. There are various areas of research: the application of MHA to each modality separately [1], to several modalities at an early stage [1-3], later combining several modalities [1, 6, 10]. [1] compares several ways of combining modalities (video, audio, text): with early concatenation and with early and late MHA. Studies show that various methods with early concatenation show emotion recognition accuracy on average 1-2% lower than methods with MHA. In [2], on the contrary, a simple concatenation for combining modalities (audio text) shows the accuracy of emotion recognition on the CMU-MOSEI corpus by 1% better than with concatenation after MHA for each modality. Experiments in the article [10] prove that later combining of modalities using MTA can improve the accuracy of emotion recognition by 4% than using late concatenation. Despite a large number of experiments in this field, researchers have not been able to come to a conclusion on the most effective ways of combining modalities for recognizing emotions or sentiments. This is due to the different nature of information, modalities and their combinations, neural network architecture, features, etc.

A large number of experimental studies in the field of emotions and sentiments recognition are conducted on the data of the CMU-MOSEI corpus [7]. The authors of many studies [4, 6, 9] used the features provided by the authors of the database – Emotive FACET, OpenFace, COVAREP, Glove for visual, acoustic and linguistic modalities. However, there are works by [1-3] in which the authors used other

features, for example BERT, spectrograms, pre-trained VGG16, ResNet50, etc., and achieved the highest accuracy. Some researchers used video, audio and text modalities simultaneously [1, 4, 6, 9], some bimodal recognition (audio and text) [2, 3]. Using the database, the tasks of emotions recognition are solved [1, 3, 6] and sentiments [9], both separately and simultaneously (multitask approach) [4]. In [4], using video, audio and text modalities and multitask approach, recognition 78.6% and 78.8% F-scores, 62.8% and 80.5% weighted accuracy is achieved for 6 classes of emotions and 2 classes of sentiments, respectively. Research by [11] shows that with an increase in the number of modalities, the accuracy of sentiments recognition on the CMU-MOSEI corpus increases. The article [4] uses a multimodal (video, audio, text) and multitask (using cross-modal attention) approach for recognizing sentiment and emotions on the CMU-MOSEI. It achieves recognition F-score of 78.6% and 78.8% for 6 classes of emotions (multilabel) and 2 classes of sentiments (multiclass), respectively. The presented results can be considered as the baseline of existing studies, since the work is as close as possible to the present research in this article. However, this study uses a modified CMU-MOSEI corpus, so the research results are not comparable to the baseline of existing studies.

3 Multimodal Database

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) database was used to conduct experimental studies [7]. The CMU-MOSEI includes video monologues from YouTube. These videos contain only one person in the frame who discusses one topic of interest. The videos were selected by 250 tags such as reviews, debates, business, products, speech, politics, etc. The volume of the corpus is 3,228 videos, the number of unique speakers is 1000. The authors of the corpus manually extracted transcriptions of speech from speakers' monologues (in fact, subtitles from videos). Then 23453 sentences were selected, timestamps of the beginning and end of the phrase were set. After that, each sentence was annotated by sentiment [-3; 3] and 6 basic emotions according to Ekman [12] (joy, sadness, anger, fear, disgust, surprise) on a scale of [0; 3] each. One utterance can contain several emotions. Each phrase was annotated by 3 annotators from a crowdsourcing platform. The annotators were not provided with instructions on how to annotate emotions so that they could interpret "how they feel".

3.1 Data preprocessing

In the original database, some phrases have incorrect timings: the first phrase in the monologue has a negative start time or the timings of phrases do not correspond to the pronunciation time. Such timings were adjusted manually. There were also situations when the timings of two adjacent phrases intersected. Such timings were corrected automatically by calculating the mean between the end of one phrase and the start of the next and subtracting (or adding) the resulting mean from these two timings.

The authors of the CMU-MOSEI provided their own data separation into training, validation and test sets. However, not all files from this distribution have an annotation. Therefore, 2 files from the test set were deleted.

The authors of the database extracted transcriptions of speech from videos. These transcriptions are subtitles from the video and sometimes these subtitles do not match the speaker's speech. In addition, the database contains videos of monologues of a speech-impaired person using gestures, so there is no acoustic and linguistic information. The analysis of speech transcriptions, rather than subtitles, is more correct for the analysis of emotional speech utterances. Therefore, speech transcriptions were extracted from the corrected audio timings using the automatic speech recognition System (ASR) - Vosk[1]. The modified data can be found in [2].

This study solves the problem of multilabel classification. Therefore, the labels of each emotion for each phrase were converted to binary categorization (0 – no emotion, >0 – there is an emotion). Continuous sentiment labels were transformed into 3 category classes: negative (<0), neutral (=0), positive (>0).

[1] https://alphacephei.com/vosk/

[2] https://github.com/Dvoynikova/CMU-MOSEI-modified.git

3.2 Data analysis

After the data preprocessing stage, the volume of the database has changed. Information about the total duration of the audio and the number of sentences in each of the training, validation and test sets is shown in Figure 1. Figures 2 (top) and 2 (bottom) show the distribution of sentiments and emotions in the modified CMU-MOSEI database.

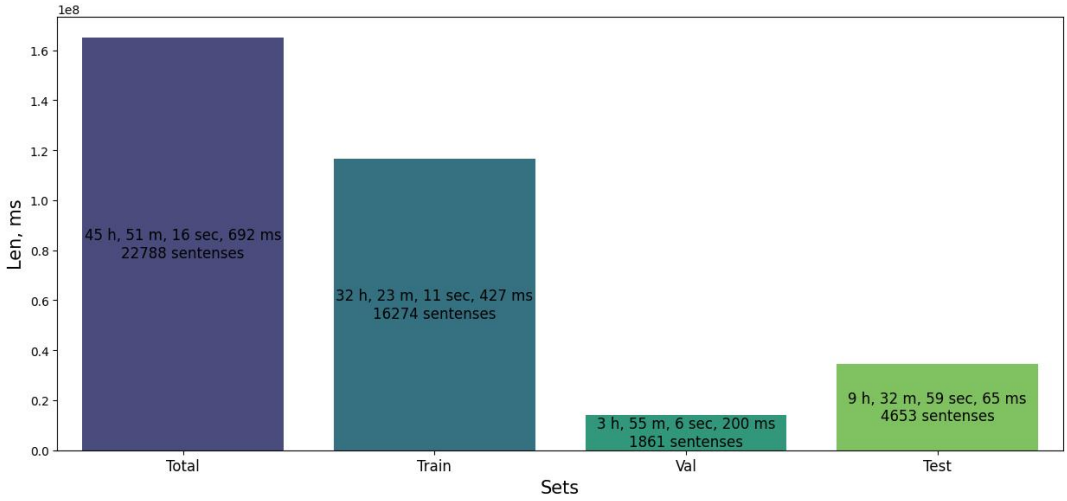


Figure 1: Modified CMU-MOSEI data distribution

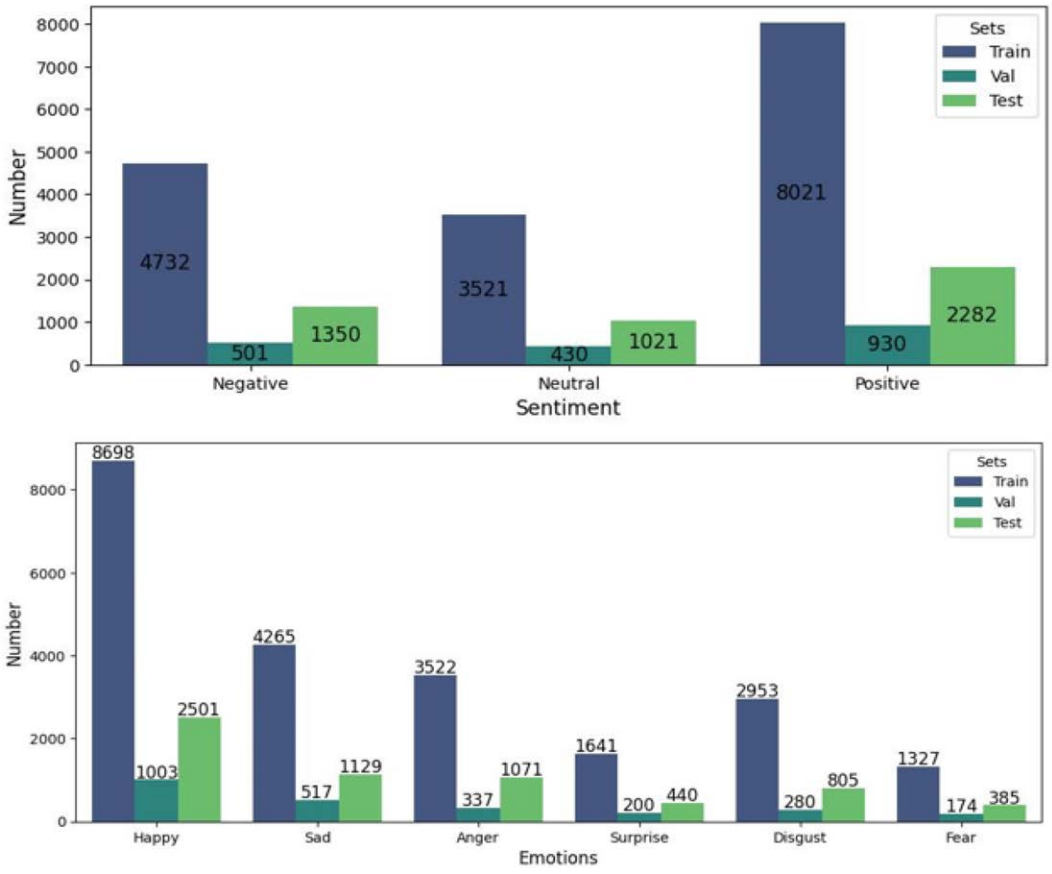


Figure 2: Distribution of sentiments in the modified CMU-MOSEI (top); Distribution of emotions in the modified CMU-MOSEI (bottom)

As mentioned above, each sentence might have several emotion labels. Figure 3 shows the number of pairwise intersections of emotions (on a binary scale) in the statements of the entire database. As can be seen from the heatmap matrix, the emotions of anger, sadness and disgust are most often found to overlap on the statements of the CMU-MOSEI corpus. Also, the labels of some statements may contain both joy and sadness. This raises doubts about the quality of the annotation, because the nature of these emotions has different valences, which makes it difficult to manifest these emotions simultaneously or with a short time interval.

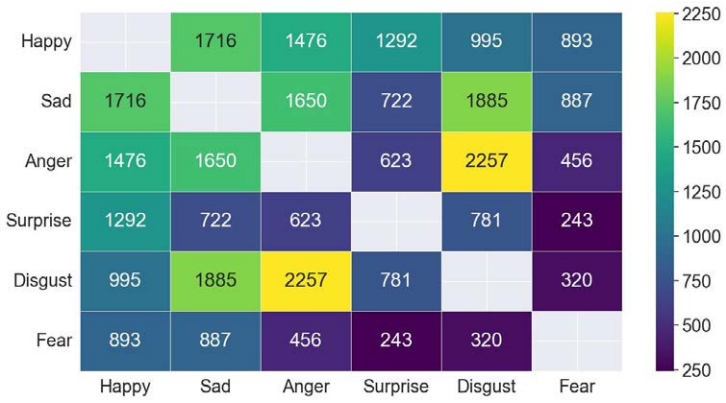


Figure 3: Numbers of intersections of emotions in sentences

Transcriptions of speech were extracted from speech utterances using ASR. Figure 4 shows the word cloud of the resulting linguistic information. It can be seen from the picture that the speakers' speech was free, and one of the most popular topics was the discussion of films. At the same time, it can be noted that words with emotional content are rarely found in statements.

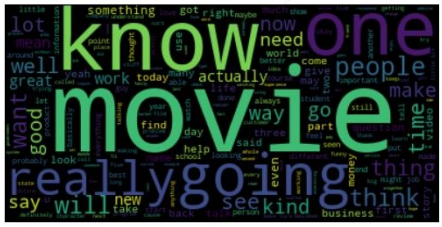


Figure 4: WordCloud ASR transcriptions of the modified CMU-MOSEI

4 Experimental studies

In this study, the bimodal multitask classification problem is solved. The acoustic and linguistic information of speech utterances from the CMU-MOSEI database is analyzed. The developed system simultaneously recognizes 3 classes of sentiments (negative, neutral, positive) – multiclass task, and 6 classes of emotions (joy, sadness, anger, surprise, disgust, fear) – multilabel task.

4.1 Baseline

The first stage of experimental research is the identification of representative features from acoustic and linguistic information. To date, the use of pretrained neural networks based on Transformer is the most effective solution for many tasks [13]. Therefore, we are experimenting with various Transformer architectures, namely HuBERT [14], Wav2Vec [15], EmotionHuBERT [16] for audio and BERT [17], RoBERTa [18], ALBERT [19] – for ASR transcriptions. For each speech utterance, we extract mean and std from a vector of dimension 1 x sequence_lenght x size_vector. As a result, we get vectors of dimension 2048 (for EmotionHuBERT) and 1536 (for the rest of Transformers) for each utterance. Then these vectors are fed to the MultiOutputClassifier with Logistic Regression. The training models with the selection of hyperparameters to identify the best macro F-score on the validation set. The choice of the F-score metric is determined by the unbalanced classes in the database, as well as the importance of the

absence of predictions of false negative results. In addition, the F-score metric is used in many existing research with the CMU-MOSEI database. The use of macro F-score is necessary in order to abstract from unbalanced classes and get a more objective estimation. The results of recognition of 3 classes of sentiments and 6 binary emotions on the text set are shown in Table 1. The Average Emotion in the Table means the average between the macro F-score for each individual emotion.

Features	**Sentiment**	**Average Emotion**
Acoustic features		
HuBERT	50.6	57.9
Wav2Vec	49.8	48.8
EmotionHuBERT	*58.9*	*59.9*
Linguistic features		
BERT	57.7	53.3
ALBERT	57.6	56.7
RoBERTa	*61.9*	*59.2*
Acoustic + Linguistic features		
EmotionHuBERT + RoBERTa	*63.2*	*61.3*

Table 1: Baseline results of sentiments and emotions recognition on MultiOutputClassifier with Logistic Regression, macro F-score, %

As can be seen from Table 1, the most representative features are EmotionHuBERT and RoBERTa for acoustic and linguistic information, respectively. It can also be noted that the text modality is more informative for the analysis of sentiments, and the audio modality is for the analysis of emotions. Combining modalities makes it possible to increase the accuracy of recognition of sentiments and emotions comparing to unimodal classifiers by macro F-score = 1.4% and 2%, respectively. With the help of the conducted experiments, we establish a baseline for recognizing 3 classes of sentiments – 63.2%, and 6 binary emotions – 61.3% macro F-score on the modified CMU-MOSEI corpus. The dummy classifier recognizes sentiments and emotions macro F-score = 21.9% and 43.7%, respectively. Thus, the proposed baseline based on the MultiOutputClassifier with Logistic Regression exceeds the dummy classifier by 41.3% and 17.6% of sentiments and emotions recognition.

4.2 Approaches to modality fusion

At the second stage, we conduct experimental studies with different approaches to combining modalities. We explore different stages of fusion (early and later) and approaches of fusion – concatenation and multi-head attention (MHA). Figure 5 (Neural Network block) shows the best neural network architecture for bimodal recognition of sentiments and emotions.

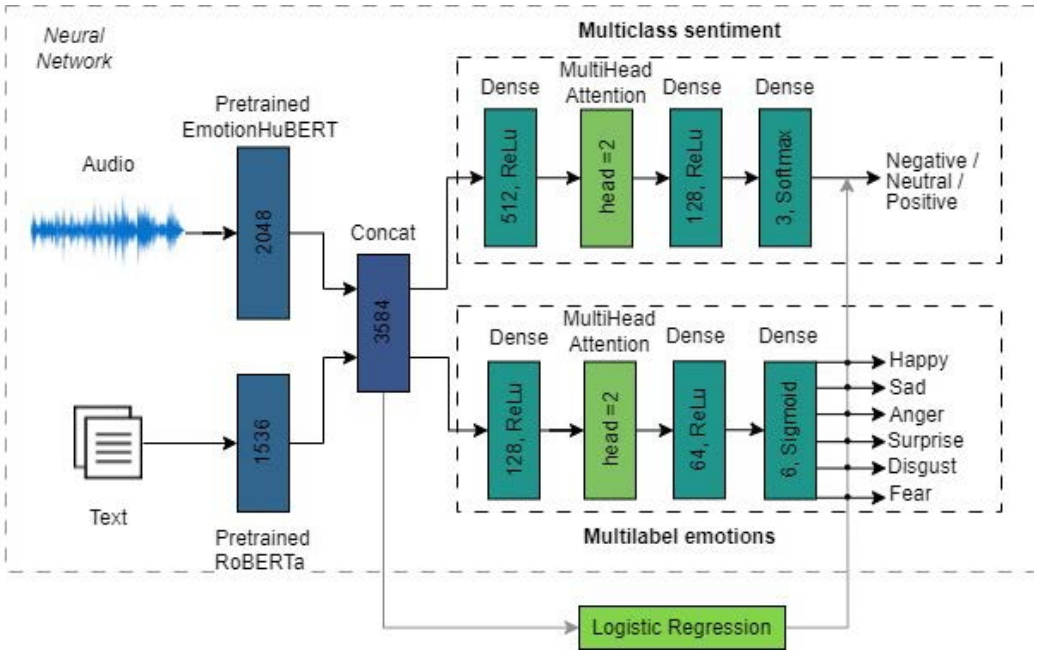
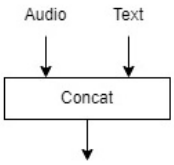


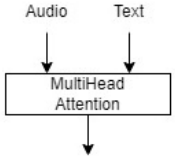
Figure 5: Bimodal multitask system architecture

We use pretrained EmotionHuBERT and RoBERTa to extract acoustic and linguistic features. The proposed MHA blocks contain 1 attention block and 2 attention heads. We use ReLU activation after each fully connected layer, except the last ones. Neural network training takes place with Adam optimizer with a learning rate of 0.01, batch size = 128. For sentiments recognition, categorical cross-entropy loss and softmax activation on the last layer are used, since a speech utterance can have only one sentiments label. Binary cross-entropy loss and sigmoid activation on the last layer are used for emotion recognition, because a speech utterance can have several binary emotion labels. Training takes place at 150 epochs, but it can stop when the loss on validation ceases to decrease during training.

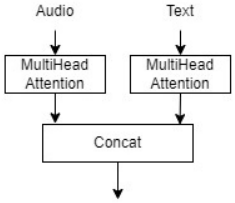
Figure 6 shows various approaches to modality fusion that were used to conduct the experiments. The early fusion of features (Fig. 6 a), b), d), e) f)) allow to immediately analyze information from two modalities [2]. Combining the modalities using MHA (Fig. 6 b), d), f)) allow to highlight the most relevant information among the 2 information flows and focus on the more important one [1]. Later fusion (Fig. 6 c), d)) allows to first highlight the relevant information for each modality, and then at later steps to combine it [10]. Using MHA (Fig. 6 e), f)) for each branch of sentiments and emotions recognition, it helps to highlight only the information that is necessary for a specific task [4].



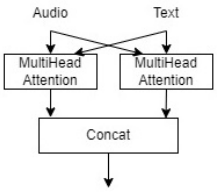
a) Early concatenation



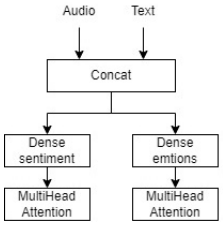
b) Early single MHA



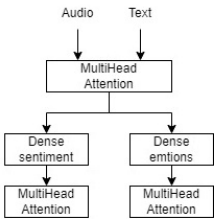
c) Early MHA for each modality



d) Early MHA for each cross-modality



e) Early concatenation with late MHA



f) Early MHA with late MHA

Figure 6: Different approaches to audio and textual modality fusion

Each approach from Figure 6 is applied and trained on the parameters described above. The results of experimental studies and conclusions on them are presented in the discussion section.

5 Experimental Results and Discussion

In Section 4, various approaches to fusion modalities using concatenation and the MHA mechanism at early and late stages are proposed. To determine the effectiveness of combining several modalities, experiments were also conducted with only one modality. The architecture of the system for unimodal multitask sentiments and emotions recognition is similar to the architecture in Figure 5, with the exception of only one modality. Table 2 presents results of experimental studies with the proposed approaches (Figure 5).

Modality fusion	Sentiment	Happy	Sad	Anger	Surprise	Disgust	Fear	Average Emotion
Unimodal Audio	58.8	69.0	62.8	57.3	47.5	66.4	47.8	58.5
Unimodal Text	61.0	64.9	55.1	63.9	47.5	70.7	47.8	58.3
Early concatenation	61.3	67.1	59.6	58.9	52.5	69.7	50.6	59.7
Early single MHA	59.1	65.3	58.4	62.4	47.5	69.3	47.8	58.5
Early MHA for each modality	56.6	68.8	61.9	62.8	47.5	67.0	47.8	59.8
Early MHA for each cross-modality	**63.0**	68.5	59.8	54.4	47.5	69.9	47.8	58.0
Early concatenation with late MHA	61.1	67.1	62.5	66.0	47.5	71.5	50.0	**60.8**
Early MHA with late MHA	61.8	68.5	58.7	55.5	47.5	70.5	47.8	58.1
Early concatenation with late MHA + LR	**63.5**	68.4	61.7	62.0	53.8	68.8	53.5	**61.4**

Table 2: Results of various fusion of modalities, macro F-score, %

Based on the results in Table 2, it can be concluded that combining the modalities makes it possible to obtain a more robust system, as well as increase in accuracy of recognizing sentiments and emotions by an average of 1-2% compared to unimodal systems. Also from the results obtained we can say, that MHA cannot be unambiguously called an effective approach to fusion modalities. For example, in our experiments, early concatenation showed a higher recognition result of emotions and sentiments, than a simple early fusion using MHA. But in general, the mechanism of attention in most cases allows to achieve higher accuracy. Also, it can be noted that applying a cross MHA to each modality is the most effective approach with respect to modality fusion using the Attention block. Using the late MHA for each task (sentiments and emotions) separately allows to achieve the highest recognition results.

The results in the Table 2 show that there is no unambiguously better approach among neural networks for recognizing emotions and sentiments at the same time. Early MHA for each cross-modality approach allowed to achieve maximum accuracy of sentiments recognition (grade 3) 63.0% macro F-score. At the same time, emotion recognition (6 binary classes) with a maximum average macro F-score of 60.8% was obtained using the early concatenation with late MHA approach.

It is also worth noting that the recognition of emotions such as surprise and fear occurs quite poorly with all the approaches considered. This may be due to the complex nature of the origin of emotions, as well as the fact that these emotions have the most unbalanced classes relative to other emotions. The emotion of disgust is always recognized better, than other emotions with all the approaches considered. Because this emotion has vivid manifestations in acoustic characteristics, as well as specific antropophones, which can manifest themselves in a linguistic modality.

To choose the most effective approach among neural networks to the recognition of emotions and sentiments, it is necessary to analyze the accuracy of recognition of each emotion separately. The disadvantage of the early MHA for each cross-modality approach is that it does not recognize emotions such as anger, surprise and fear well. However, the emotion of anger is important in the analysis of emotions, because an angry person can pose a danger to others. Therefore, it is important that automatic

systems recognize the emotion of anger as best as possible. Among neural networks, the most effective approach is Early connect with late Attention (it is shown in Figure 5 of the Neural Network block). It's disadvantage is the lower accuracy of sentiments recognition relative to other approaches. The advantage of this approach is that it predicts all emotions more reliably. However, this approach does not exceed the baseline from section 4.1.

The most effective approach for the task of bimodal multitasking of emotions and sentiments is an approach based on the neural network Early concat with late Attention and Logistic regression (Figure 5). Concatenated acoustic and linguistic features are fed to the input of the neural network and logistic regression. The two models are combined at the decision-making level. The proposed approach allows achieving 63.5% and 61.4% macro F-score for recognizing 3 sentiments classes and 6 binary emotions classes, respectively. The obtained results are 0.3% and 0.1% higher than the established baseline with Logistic regression. More detailed results of emotion and sentiments recognition using the proposed approach are presented in Table 3.

Metrics	Senti-ment	Happy	Sad	Anger	Surprise	Disgust	Fear	Average Emotion
Macro precision	63.5	68.4	61.7	61.6	54.9	67.1	55.2	61.5
Macro recall	63.8	68.8	65.2	64.9	61.3	74.9	64.0	66.5
Macro F-score	63.5	68.4	61.7	62.0	53.8	68.8	53.5	61.4

Table 3: Results of emotions and sentiments recognition by early concatenation with the late Attention approach

When recognizing emotions, the multilabel task was solved, in other words, there could be several emotion labels in one phrase. Thus, it is impossible to analyze classifier errors. Sentiments recognition occurred in the multiclass task, i.e. there could be only one sentiments label in one phrase. The matrix of errors in the recognition of the sensor is shown in Figure 7.

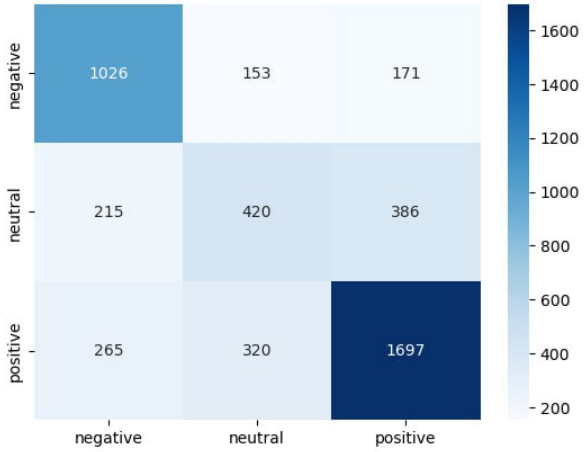


Figure 7: Matrix of sentiments recognition errors

As can be seen from Figure 7, the negative and positive classes of sentiments are recognized quite well. The neutral class is most often confused with the positive one.

As mentioned in the Related work section, all existing research in the field of emotion and sentiment recognition using the CMU-MOSEI database was conducted on the original data set. We conducted experiments on a modified data set, which makes our results completely incomparable with other works. The best accuracy of multitask recognition of emotions (6 binary classes) and sentiments (2 classes) using modal analysis (video, audio, text) F-score 78.6% and 78.8% Fare achieved in [4]. Our results (63.5% of emotions and 61.4% of sentiments) are lower than the existing ones. It is worth noting that we recognize 3 classes of sentiments when other studies carry out the recognition of 2 classes of sentiments. In addition, we only analyze audio and textual modality, while another work analyzes 3 modalities. It is also worth saying that only the baseline is set in this article, the improvement of this baseline is planned in our next studies.

6 Conclusions

The article is devoted to a bimodal multitask approach to the recognition of emotions and sentiments. The CMU-MOSEI database was used as data for experimental studies. One of the main tasks that we solved in this study is the semi-automatic preprocessing of CMU-MOSEI in order to improve data quality. We also identified representative features for acoustic and linguistic information – EmotionHuBERT and RoBERTa, respectively. Using these features, we have established a baseline for bimodal multitasking recognition of emotions and sentiments – 63.2% and 63.3% macro F-score, respectively.

In our study, we conducted experiments to identify the most effective approach to fusion (concatenation and multi-head attention) audio and text modality for the multitask of recognizing sentiments and emotions. Based on the results of the experiment, we can conclude that the use of early concatenation of acoustic and linguistic information and MHA for each task (sentiments and emotions) separately allows achieving the highest recognition results. Using EmotionHuBERT and RoBERTa as features and the MHA mechanism for each task, we achieve 61.1% and 60.8% macro F-score for a bimodal (audio and text) multitask approach to recognize 3 sentiments classes and 6 binary emotion classes.

The proposed bimodal (audio and text) approach using the Early concat with late Attention neural network and Logistic regression makes it possible to achieve 63.5% and 61.4% macro F-score for recognizing sentiments and emotions.

The prospect of further research may manifest itself in the addition of a video modality. Facial expressions, gestures, postures are also representative information in the manifestation of emotions. Therefore, it is assumed that the analysis of the video modality will help to increase the accuracy of the recognition of sentiments and emotions.

Acknowledgements

This research was supported by the Russian Science Foundation (project No. 22-11-00321, https://rscf.ru/project/22-11-00321/).

References

[1] Lee S., Han D.K., Ko H. (2021), Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification, IEEE Access, Vol. 9, pp. 94557-94572.

[2] Siriwardhana S., Reis A., Weerasekera R, Nanayakkara S. (2020), Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition, arXiv preprint arXiv:2008.06682.

[3] Ho N.H., Yang H.J., Kim S.H. (2020), Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network, IEEE Access. Vol. 8, pp. 61672-61686.

[4] Akhtar M.S., Chauhan D.S., Ghosal D., Poria S., Ekbal A., Bhattacharyya P. (2019), Multi-task learning for multi-modal emotion recognition and sentiment analysis, arXiv preprint arXiv:1905.05812.

[5] Verkholyak O., Dvoynikova A., Karpov A. (2021), A Bimodal Approach for Speech Emotion Recognition using Audio and Text, Journal of Internet Services and Information Security. Vol. 11, no. 1, pp. 80-96.

[6] Mittal T., hattacharya U., Chandra R., Bera A., Manocha D. (2020), M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, Proceedings of the AAAI conference on artificial intelligence. Vol. 34, no. 02, pp. 1359-1367.

[7] Zadeh A.B., Liang P.P., Poria S., Cambria E., Morency L.P. (2018), Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 1, pp. 2236-2246.

[8] Velichko A., Markitantov M., Kaya H., Karpov A. (2022), Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 4735-4739.

[9] Kumar A., Vepa J. (2020), Gated mechanism for attention based multi modal sentiment analysis, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4477-4481.

[10] Choi W. Y., Song K. Y., Lee C. W. (2018), Convolutional attention networks for multimodal emotion recognition from speech and text data, Proceedings of grand challenge and workshop on human multimodal language, pp. 28-34.

[11] Ghosal D., Akhtar M. S., Chauhan D., Poria S., Ekbal A., Bhattacharyya B. (2018), Contextual inter-modal attention for multi-modal sentiment analysis, Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 3454-3466.

[12] Ekman P., Freisen W.V., Ancoli S. (1980), Facial signs of emotional experience, Journal of personality and social psychology. Vol. 39, no. 6, pp. 1125.
[13] Alibaeva K., Loukachevitch N. (2022), Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006”. Pp 8-17.
[14] Hsu W.N., Bolte B., Tsai Y.-H.H., et al. (2021), Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 29, pp. 3451–3460.
[15] Baevski A., Zhou Y., Mohamed A., Auli M. (2020), Wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems. Vol. 33, pp. 12449-12460.
[16] Wagner J., Triantafyllopoulos A., Wierstorf H., Schmitt M., Burkhardt F., Eyben F., Schuller B.W. (2022), Dawn of the transformer era in speech emotion recognition: closing the valence gap, arXiv preprint arXiv:2203.07378.
[17] Devlin J., Chang M., Lee K., Toutanova K. (2018), Bert: Pre-training of deep bidirectional transformers for language understanding,arXiv preprint arXiv:1810.04805.
[18] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019), Roberta: A robustly optimized bert pretraining approach,arXiv preprint arXiv:1907.11692.
[19] Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. (2019), Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942.

Introduction model in Russian «Pear reportages»: The role of common ground

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

Abstract

In this study, the peculiarities of the character introduction in the genre of live reportage were studied. The participants were 25 students oh the Lomonosov Moscow State University. Speech production was elicited by means of the "Pears Film" by W. Chafe. Different types of the collective common ground were considered. It turned out that, unlike narratives of other genres, the chronological scale is more important for the introduction than the status scale. It was also shown that the collected reportages from the point of view of the introduction peculiarities are more similar to classical retellings than to the sports reportages.

Keywords: introduction; common ground; reportage; working memory; "Pears Film"; speech production
DOI: 10.28995/2075-7182-2023-22-62-68

Модель интродукции в русских «Репортажах о грушах»: роль общей позиции

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

Аннотация

В этом исследовании изучались особенности интродукции персонажей в нарративах в жанре репортажей. В эксперименте участовали 25 студентов МГУ имени М.В. Ломоносова. В качестве стимульного материала был использован «Фильм о грушах» У. Чейфа. Были рассмотрены разные типы интродуктивной коллективной общей позиции. Оказалось, что в отличие от нарративов других жанров, для интродукции более важна хронологическая шкала, чем статустная. Было показано также, что собранные репортажи с точки зрения особенностей интродукции больше похожи на классические пересказы, чем на собственно (спортивные) репортажи.

Ключевые слова: интродукция; общая позиция; репортаж; «Фильм о грушах»; порождение речи

1 Введение. Интродукция референта в текстах разных дискурсивных жанров

В серии работ, опубликованных почти тридцать лет назад на материале анализа русских, немецких и шанских[1] сказок, был исследован вопрос о типологии средств интродукции референта в письменных текстах [1], см. раздел 1.1. В 2015 г. разработанная модель интродукции была впервые применена к устному дискурсу [2], см. раздел 1.2. Настоящее исследование продолжает данную серию работ и вводит в рассмотрение интродуктивную модель применительно к жанру устного **репортажа**, с особым вниманием к понятию общей позиции, см. разделы 2 и 3.

[1] Шанский язык – один из тайских языков, распространен в основном в Мьянме и Китае.

В качестве стимульного материала был выбран известный **«Фильм о грушах»** У. Чейфа [3]; для настоящей работы важно, что на основании разработанных в [1] критериев в нем однозначно выделяются следующие пять персонажей: (1) главный герой, который действует на протяжении всего фильма (МАЛЬЧИК); (2) герой, который участвует во многих ключевых эпизодах фильма (САДОВНИК); (3) второстепенный персонаж, который появляется в нескольких эпизодах (ТРИ МАЛЬЧИКА); а также (4) и (5) эпизодические персонажи, которые принимают участие только в одном эпизоде фильма (МУЖЧИНА С КОЗОЙ и ДЕВОЧКА). Краткий сюжет фильма таков:

(1) Садовник на дереве собирает груши. Мимо проходит мужчина с козой. Приезжает маль чик на велосипеде, берет одну корзину с грушами и уезжает. По дороге навстречу маль чику едет девочка на велосипеде; у мальчика слетает шляпа; велосипед наезжает на ка мень; мальчик падает; груши рассыпаются. Три подошедших мальчика помогают собрать груши и отдают потерянную шляпу. Мальчик дает им три груши; они уходят в разные стороны. Садовник на дереве продолжает собирать груши; он обнаруживает пропажу од ной корзины с грушами. Мимо проходят три мальчика, жуя подаренные груши. Садовник смотрит им вслед.

Ключевым понятием данной работы является понятие **общей позиции** (ОП, common ground, CG). Вслед за Г. Кларком мы будем называть ОП «сумму общих (common), совместных (joint) и взаимных (mutual) знаний (knowledge), мнений (beliefs) и допущений (suppositions)» [4]. Г. Кларк выделяет два типа ОП: коллективную ОП (communal CG), которая связана с культурной средой, к которой принадлежат собеседники, и личную общую позицию (personal CG), которая проистекает из личного опыта взаимодействия данных конкретных собеседников [4]. В дальнейшей работе мы будем иметь в виду только первую из них.

Учет фактора общей позиции является важным компонентом успешности речевой коммуникации. В частности, в [5] было показано, что использование определенного артикля является успешной стратегией номинации в том случае, когда собеседники обладают общей позицией относительно данного референта; в противном же случае, то есть если говорящий использует определенную референцию при отсутствии общей позиции, общение заканчивается коммуникативной неудачей.

Важно отметить также, что понятие общей позиции может быть использовано при описании не только диалогической речи, но и монологической, о потенциальных адресатах которой см. раздел 2.2.

1.1 «Сказочная» интродуктивная модель

Интродукцией мы будем называть введение нового референта в долговременную память адресата. В прототипическом случае интродуктивное предложение состоит из трех основных элементов: ОП (в работе [1] был использован термин «привязка») – бытийный оператор – номинация референта. В работе [1] на материале сказок был обоснован постулат о принципиальной небходимости ОП; были выделены эксплицитные (выраженные вербально) и имплицитные (подразумеваемые, но не выраженные вербально) ОП, а также ступенчатая ОП и псевдоОП – введение персонажа через другие детали (ступенчатая ОП) или других персонажей (псевдоОП), особенно частотно это явление в абсолютном начале сказочных текстов, например: *Жили-были старик* да *старуха, у них была дочка Аленушка да сынок Иванушка.* Бытийный оператор указывает на интродуктивную неопределенную референцию имени. Номинация состоит из интродуктива (который во многих европейских языках обычно выражается неопределенным артиклем) и собственно номинации. На материале сказок было показано также, что

- чем выше место персонажа в иерархии «главный герой – герой – второстепенный персонаж – эпизодический персонаж», тем больше разнообразных средств используется для его введения, то есть тем больше сила интродукции;
- ОП по времени / месту оказывается характерна для важных персонажей, а ступенчатая и псевдоОП – для неважных.

Если следовать канонической сказочной модели интродукции, введение персонажей «Фильма о грушах» могло бы выглядеть так:

(2) [САДОВНИК] В одной мексиканской деревне жила-была одна семья, у которой был боль шой грушевый сад. Однажды солнечным летним днем глава семьи – крупный усатый мужчина средних лет в шляпе и белом фартуке – отправился в сад собирать груши. <...> [МУЖЧИНА С КОЗОЙ] В этот момент мимо дерева проходит мужчина, который вдет козу. <...> [МАЛЬЧИК] Тут к грушевому дереву на большом красном велосипеде подъезжает мальчик, он тоже в шляпе. <...> [ДЕВОЧКА] В это время навстречу мальчику по дороге едет девочка на велосипеде. <...> [ТРИ МАЛЬЧИКА] Тут к мальчику подходят трое ребят и помогают ему подняться, один из мальчиков играет пинг-понговым шариком.

1.2 Интродуктивная модель в «Рассказах о грушах»

Данное исследование [2] было выполнено на материале 25 пересказов «Фильма о грушах»; корпус интродуктивных предложений содержал 125 единиц, по 25 для каждого из пяти персонажей. В целом, типичная усредненная интродукция персонажей при пересказе «Фильма о грушах» выглядела таким образом (символами *ээ* и *мм* обозначены заполненные паузы хезитации, в скобках обозначена длительность пауз):

(3) [САДОВНИК] (.5) ээ(.5) фильм начинается с того что мм(.6) усатый мужчина в шляпе со бирает груши. <...> [МУЖЧИНА С КОЗОЙ] (.6) в это время ээ(.3) мимо проходит какой-то мужчина с козой, коза упирается. <...> [МАЛЬЧИК] (.8) ээ(.4) затем ээ(.2) приезжает ма ленький мальчик на велосипеде, в большой шляпе. <...> [ДЕВОЧКА] (.4) ээ(.5) через поле, (.3) едет девочка, (.4) тоже на велосипеде. <...> [ТРИ МАЛЬЧИКА] (1.0) это видят (.5) трое других мальчиков, (.2) мм(.3) они подходят и помогают ему подняться, (.4) и собирают груши в корзину.

Рассмотрим более подробно типы ОП в «Рассказах о грушах». ПсевдоОП и ступенчатая ОП оказались не характерны для пересказов, такие конструкции встретились только по одному разу. Наиболее частотной оказалась так называемая кинематографическая ОП, то есть использование в качестве общего знания того факта, что рассказчик пересказывает сюжет фильма. Например, интродукция персонажа [МУЖЧИНА С КОЗОЙ] в одном из пересказов выглядела так:

(4) ээ(.2) потом мы слышим (.3) звук блеяния козы,
(.5) видим человека,
(.4) который ведет козу на веревке.

Вопрос использования кинематографического взгляда в «Рассказах о грушах» был впервые описан еще в [6]. Оказалось, что американские испытуемые чаще пересказывали видеоролик как фильм, в то время как греческие испытуемые просто рассказывали историю, не упоминая о том, что действие происходит в фильме. Согласно работе [7] в большинстве «Рассказов о грушах» кинематографическая лексика встречается хотя бы один раз; в частности, для нидерландских пересказов эта цифра составляет 79%, для греческих 80%, для польских 85%.

В русских «Рассказах о грушах» кинематографический взгляд также использовался хотя бы один раз в каждом пересказе [2]. Более того, была выявлена важная закономерность: количество его использования для нужд интродукции последовательно сокращалось от начала к концу пересказа независимо от статуса персонажа: [САДОВНИК] был введен при помощи кинематографической ОП в 23 случаях из 25, [МУЖЧИНА С КОЗОЙ] в 10 случаях, [МАЛЬЧИК] в 9 случаях, [ДЕВОЧКА] в 3 случаях, [ТРИ МАЛЬЧИКА] в 1 случае (всего 46 случаев).

Частотность других типов ОП следовала законам сказочной интродукции: ОП по времени / месту была характерна для важных персонажей, а другие ОП – для неважных. Однако в отличие от сказочной интродукции доминирующим средством в «Рассказах о грушах» оказывается кинематографическая ОП.

2 Интродуктивная модель в «Репортажах о грушах»

Исследование жанра репортажа в отечественной лингвистике еще только начинается, однако см. книгу Е. Г. Малышевой (2011) [8] о спортивном комментировании, в которой автор, в частности, выделяет семь коммуникативных типов языковых личностей спортивных комментаторов, подробнее см. ниже. В работе [9] авторы предположили, что жанр репортажа требует от человека довольно больших когнитивных ресурсов, поэтому начали исследование этого жанра с изучения особенностей вербальной рабочей памяти (вРП) испытуемых. Для первого исследования были выбраны три критерия «успешности» репортажа: «успешный» репортаж – это такой репортаж, в котором испытуемый говорит (1) непрерывно, (2) быстро и (3) лексически разнообразно. Статистический анализ 16 репортажей показал, что, действительно, мы наблюдаем положительную корреляцию между объемом вРП и темпом речи и лексическим разнообразием. Однако корреляция между объемом вРП и непрерывностью репортажа вопреки ожиданиям оказалась отрицательной; авторы предположили, что навык избегать незаполненных пауз хезитации относится к коммуникативной сфере и поэтому не требует больших затрат когнитивных ресурсов.

2.1 Испытуемые, гипотезы

Настоящее исследование было проведено весной 2022 года с 25 студентами ОТиПЛа филфака МГУ имени М.В. Ломоносова. Каждый из испытуемых прошел два теста: тест по комментированию «Фильма о грушах» (см. раздел 2.2) и тест по определению объема вРП при порождении речи (результаты второго теста в данной работе рассматриваться не будут). Как и в [2], корпус интродуктивных предложений содержал 125 единиц, по 25 для каждого из пяти персонажей.

В данной работе мы априори предполагаем, что особенности жанра репортажа будут предопределять две следующие особенности интродуктивного поведения испытуемых:

- большое количество имплицитных ОП (так как испытуемый описывает то, что непосредственно видит на экране);
- небольшое количество кинематографических ОП (так как испытуемый ведет репортаж, а не пересказывает сюжет только что увиденного фильма).

2.2 Сбор корпуса репортажей

Каждый испытуемый смотрел фильм длительностью 5 мин 55 с на экране компьютера. В инструкции было сказано: «Вам надо будет комментировать фильм по ходу развития действия как можно более подробно. Представьте себе, что рядом с вами сидит незрячий человек и вам нужно детально описать ему все, что происходит на экране. Старайтесь описывать не только происходящие события, но и окружающую обстановку». Помимо аудиозаписи при помощи айтрекера Tobii Spectrum с частотой 600 Гц велась регистрация движений глаз.

Все записи были аннотированы в программе Praat; тексты были разбиты на ЭДЕ по [10]; при разбиении на ЭДЕ аннотаторы прежде всего руководствовались длительностью пауз. Интродуктивные высказывания, которые анализировались в настоящей работе, состояли из одной или нескольких ЭДЕ; вопрос о соотношении важности персонажа и количества ЭДЕ, использованных при его описании, заслуживает дополнительного исследования.

При помощи программных продуктов фирмы Tobii были получены окуломоторные данные о распределении зрительного внимания испытуемых во время просмотра фильма. Важной особенностью интродуктивных описаний оказалось отсутствие больших задержек между фиксацией взгляда испытуемого на том или ином персонаже и началом его описания; в среднем эта задержка составляла около 1с. В 4 записях задержки при описании некоторых второстепенных персонажей оказались больше 2с; эти 4 записи были заменены на новые. Таким образом, мы можем быть уверены, что в каждый момент времени испытуемые описывали то, что непосредственно видели на экране. Длительность фиксаций в каждом случае была не меньше 200мс, а суммарная длительность на каждом из персонажей при интродуктивном описании составила не меньше 1с, что вполне достаточно для того, чтобы не только посмотреть на референт, но и увидеть его.

2.3 Результаты

В табл. 1 представлены сводные результаты данного исследования.

	САДОВНИК	МУЖЧИНА С КОЗОЙ	МАЛЬЧИК	ДЕВОЧКА	ТРИ МАЛЬЧИКА	всего
кинематогр.	25	19	11	8	3	66
псевдо	0	1	0	13	19	33
ступенч.	16	0	0	0	0	16
по времени	0	4	6	4	2	16
имплицитн.	0	2	6	4	1	13
по месту	0	7	4	0	1	12

Таблица 1: Распределение типов ОП по персонажам

2.4 Обсуждение результатов

Рассмотрим более подробно ключевые моменты интродуктивного репортажа в сопоставлении с другими жанрами с точки зрения описания ОП.

Абсолютное начало (6). В сказочных интродукциях в абсолютном начале обычно используются псевдоОП и ступенчатая ОП. В «Рассказах о грушах» доминирует кинематографическая ОП. В «Репортажах о грушах» чаще других встречается совмещенный вариант, см. (5):

(5) [САДОВНИК] (.6) мы видим лестницу, (.4) на которой стоит мужчина и (.2) собирает что-то на дереве.

(6) Абсолютное начало

сказка	псевдо, ступенч.
пересказ	кинематограф.
репортаж	кинематограф. & ступенч.

Роль хронологической шкалы (7). Для сказочной интродукции хронологическая шкала оказывается неважной, за исключением рассмотренного абсолютного начала повествования. Для интродуктивных высказываний в «Рассказах о грушах» распределение кинематографической ОП определяется именно хронологической шкалой. В «Репортажах о грушах» мы видим ту же самую строгую закономерность, см. табл. 1.

(7) Хронологическая шкала
[САДОВНИК] > [МУЖЧИНА С КОЗОЙ] > [МАЛЬЧИК] > [ДЕВОЧКА] > [ТРИ МАЛЬЧИКА]

Роль статусной шкалы (9). В сказках при интродукции важных персонажей обычно используется ОП по времени / месту, при интродукции неважных – другие ОП. В американских «Рассказах о грушах», согласно транскриптам из [3], во всех случаях доминирует использование ОП по времени / по месту. Интродукция в русских «Рассказах о грушах» повторяет сказочную тенденцию. В «Репортажах о грушах» ОП по времени / месту встречаются в интродуктивных высказываниях независимо от статуса персонажа. Отметим также, что в собранных репортажах кинематографическая ОП нередко сочетается с ОП по времени / по месту или даже с обоими, например:

(8) [МУЖЧИНА С КОЗОЙ] (.4) сейчас вдалеке виден (.2) еще один человек, мм(.3) с осликом.

(9) Статусная шкала
[МАЛЬЧИК] > [САДОВНИК] > [ТРИ МАЛЬЧИКА] > [ДЕВОЧКА] > [МУЖЧИНА С КОЗОЙ]

Роль псевдоОП. В сказках псевдоОП иногда используется при интродукции неважных персонажей; в «Рассказах о грушах» псевдоОП практически не используется. В «Репортажах о грушах» количество псевдоОП значительно возрастает к концу репортажа при интродукции двух последних (неважных) персонажей – второстепенного и эпизодического; таким образом, мы не можем однозначно определить, что играет главную роль – хронологическая шкала или статусная.

Роль имплицитной ОП. В сказках имплицитная ОП используется редко, исключительно при интродукции неважных персонажей; в «Рассказах о грушах» наблюдается примерно та же тенденция. В «Репортажах о грушах» вопреки нашим ожиданиям имплицитная ОП тоже используется нечасто, причем несколько чаще при интродукции персонажа [МАЛЬЧИК] – главного героя повествования.

3 Интродуктивная модель в «Репортажах о грушах»: повторные репортажи

В отличие от пересказов, которые испытуемые порождают после окончания просмотра фильма, при комментировании они описывают то, что видят в данный момент, не имея возможности оценить значимость происходящего в масштабах всего сюжета. Однако даже самый словоохотливый комментатор не может успевать описывать во всех деталях все происходящее в данный момент, поэтому поневоле испытуемым приходится фильтровать информацию с точки зрения ее важности. Таким образом, при таком задании мы получаем репортаж, напоминающий спортивный, когда комментаторы сами не знают, чем закончится спортивное событие. Именно поэтому, на наш взгляд, статусная шкала может играть в репортажах такую незначительную роль.

Однако ситуация может измениться, когда при повторном комментировании фильма с интервалом в полгода испытуемые забудут некоторые детали, однако еще будут помнить сюжет и будут строить свои репортажи, исходя из знаний этого сюжета. Таким образом в данном дополнительном исследовании мы проверяем гипотезу, что известность сюжета фильма повышает роль статусной шкалы и понижает роль хронологической шкалы.

Данное дополнительное исследование было проведено осенью 2022 года с 9 испытуемыми, которые проходили весной то же тестирование; таким образом, сопоставительный корпус интродуктивных предложений содержит 80 единиц, по 18 для каждого из персонажей, см. табл. 2.

	САДОВНИК	МУЖЧИНА С КОЗОЙ	МАЛЬЧИК	ДЕВОЧКА	ТРИ МАЛЬЧИКА	всего
кинематогр.	9 / 9	8 / 7	4 / 4	2 / 2	1 / 1	24 / 23
псевдо	0 / 0	1 / 0	0 / 0	5 / 6	7 / 7	13 / 13
ступенч.	6 / 7	0 / 0	0 / 0	0 / 0	0 / 0	6 / 7
по времени	0 / 0	2 / 1	2 / 3	2 / 1	0 / 0	6 / 5
имплицитн.	0 / 0	0 / 1	2 / 2	2 / 1	0 / 0	4 / 4
по месту	0 / 0	2 / 3	2 / 1	0 / 0	0 / 0	4 / 4

Таблица 2: Распределение типов ОП по персонажам в первичных / повторных репортажах

Оказалось, однако, что распределение типов ОП по персонажам в первичных и повторных репортажах практически не различается, то есть наша гипотеза о важности статуса персонажа в повторных репортажах не подтверждается.

4 Заключение

Как мы показали, интродуктивное поведение испытуемых в «Репортажах о грушах» заметно отличается от классического спортивного репортажа типа (10) по работе [8], в котором основной акцент делается на ключевых событиях спортивного соревнования, определенных заранее:

(10) Итак ↑ внимание! ↓ / Стрельба Гараничева в эстафетной гонке! ↑ // Гараничев вторым выстрелом мажет ↑ // У словенца много промахов ↓ // Уходит немец ↑ / это Лессер! ↑ // Гараничев один промах! ↑

Типичная усредненная репортажная интродукция наших персонажей скорее выглядит таким образом:

(11) [САДОВНИК] (.3) ээ(.2) вижу дерево, к нему приставлена лестница, мм(.6) на которой стоит мужчина. <...> [МУЖЧИНА С КОЗОЙ] (.4) мы видим человека, ээ(.3) в футболке и шляпе, (.4) он ведет козу. <...> [МАЛЬЧИК] ээ(.4) на заднем плане едет мальчик на красном велосипеде, (.5) он тоже в шляпе. <...> [ДЕВОЧКА] (.3) ему навстречу едет (.3) девочка на велосипеде, (.4) с двумя косичками. <...> [ТРИ МАЛЬЧИКА] (.6) мальчик слышит стук, (.5) поднимает голову, (.3) и видит трех других мальчиков.

Подведем некоторые итоги: в абсолютном начале репортажей испытуемые часто используют кинематографическую ступенчатую ОП, при этом распределение персонажей регулируется хронологической шкалой. Статусная шкала, по-видимому, не имеет особого значения, как и имплицитная ОП, а псевдоОП служит для введения персонажей ближе к концу фильма. Таким образом, репортажи о грушах, имея некоторые свои особенности, все же остаются намного больше похожи на классические «Рассказы о грушах», чем на истинные (спортивные) репортажи. За рамками рассмотрения в настоящей работе, однако, осталось много других особенностей «Репортажей о грушах», которые, возможно, в будущем прольют свет на такие немного загадочные результаты.

Благодарности

Автор выражает благодарность студентам ОТиПЛ филфака МГУ имени М.В. Ломоносова за участие в экспериментах и помощь в обработке результатов.

References

[1] Fedorova O.V. (1994), A typology of referent introduction means [Tipologiya sredstv introduktsii referenta], unpublished. — Moscow.
[2] Fedorova O.V. (2015), Referent introduction in Russian spoken narrratives // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2015). — Issue 14. — P. 131–140.
[3] Chafe W. (ed.). (1980), The pear stories: Cognitive, cultural, and linguistic aspects of narrative production. — Norwood: Ablex.
[4] Clark H. (1996), Using language. — Cambridge: Cambridge University Press.
[5] Clark H.H., Marshall C.R. (1981), Definite reference and mutual knowledge // A.K. Joshi, B. Webber, I. Sag (eds.) Elements of discourse understanding. Cambridge: Cambridge University Press. P. 10–63.
[6] Tannen D. (1980), A comparative analysis of oral narrative strategies: Athenian Greek and American English // W. Chafe (ed.), The Pear Stories: Cognitive, cultural and linguistic aspects of narrative production. — Norwood: Ablex. — P. 51–87.
[7] Mazur I., Chmiel A. (2012), Towards common European audio description guidelines: Results of the Pear Tree Project // Perspectives: Studies in Translatology. — Vol. 20(1). — P. 5–23.
[8] Malysheva E.G. (2011), Russian sports discourse [Russkiy sportivnyy diskurs]. — Moscow: Flinta, 2011.
[9] Fedorova O.V. (2022), “Pears Film” live: Cognitive peculiarities of the reportage // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue” (2022). — Issue 21. — P. 203–210.
[10] Kibrik A. A., Podlesskaya V. I. (eds.) (2009), Night Dream Stories: A corpus study of spoken Russian discourse [Rasskazy o snovideniyah: korpusnoe issledovanie usntogo russkogo diskursa]. — Moscow: Jazyki Slavyanskih Kul’tur.

Foreground and background in Russian Sign Language narratives: the role of aspect and actionality

Elizaveta Filimonova
Russian State University for the Humanities;
Institute of linguistics, Russian Academy of Sciences,
Moscow, Russia
ev.filimonova@list.ru

Abstract

The paper explores the role of aspect and actionality in foregrounding and backgrounding of clauses in Russian Sign Language narratives. Corpus study shows similarities to functions of aspectual markers and actionality in spoken languages. Besides grammatical markers and predicate types, non-manual marking and prosodic features of verbal sign can contribute to clause foregrounding and backgrounding.

Ключевые слова: russian sign language, narrative, aspect, actionality
DOI: 10.28995/2075-7182-2023-22-69-78

Основная линия и фон в нарративах в русском жестовом языке: роль аспектуальности и акциональности

Филимонова Е. В.
Российский государственный гуманитарный университет;
Институт языкознания РАН,
Москва, Россия
ev.filimonova@list.ru

Аннотация

В работе исследуется роль аспектуальных и темпоральных показателей и акциональных типов предиката в выдвижении клауз на первый план нарратива в русском жестовом языке. Корпусное исследование обнаруживает явления, сходные со теми, что характерны для звуковых языков. Помимо грамматических показателей и акциональных типов предиката выделенности клаузы в русском жестовом языке могут также способствовать немануальные маркеры и просодические характеристики глагольного жеста.

Ключевые слова: русский жестовый язык, нарратив, аспектуальность, акциональность

1 Введение

Русский жестовый язык (РЖЯ) представляет собой естественный язык визуальной модальности, который используется глухими и слабослышащими на территории России и некоторых стран СНГ. Языковые средства, которые способствуют выделенности клаузы в нарративе, не были предметом специального исследования в РЖЯ. Использование данных корпуса текстов РЖЯ [2] позволило проанализировать нарративы от носителей разных диалектов, различных социальных и возрастных характеристик.

Под нарративом понимают текст, который характеризуется условной временной последовательностью изложения событий и ориентирован на агенса [14: 9]. В нарративах выделяют основную линию (foreground, storyline, mainline) и фон (background, supportive material). События, принадлежащие основной линии, следуют в хронологическом порядке и продвигают историю вперед, образуя «нарративную цепочку», в то время как ситуации,

относящиеся к фону, не находятся в хронологической последовательности ни по отношению к событиям основной линии, ни по отношению друг к другу; они поддерживают, развивают, комментируют основное повествование [8; 14: 21]. Различие между основной линией и фоном отражается в использовании языковых средств: аспектуальных и временных показателей, порядка слов, залога [8], акциональности, агентивности, модальности и т. п. [9].

Важную роль в выдвижении клауз в нарративе во многих языках играют темпоральные и аспектуальные показатели, а также акциональная семантика предиката. Так, П. Хоппер демонстрирует, что в языках, где имеется оппозиция перфектива и имперфектива, перфектив используется для изображения событий основной линии, тогда как имперфективные формы описывают фон нарратива. Он указывает также на использование форм прошедшего времени в основной линии нарратива [8]. С. Уоллас отмечает, что в случае если настоящее время используется вместо прошедшего (настоящее историческое), то это способствует выдвижению на первый план; напротив, если события настоящего описаны формами прошедшего времени, ситуация становится фоновой [22: 210]. В работе П. Хоппера и С. Томпсон [9] говорится о характеристиках глаголов, которые более характерны для клауз основной линии: это переходные, агентивные, предельные глаголы, пунктивы. В фоновых клаузах, напротив, преобладают стативные, неагентивные, непереходные глаголы.

2 Исследования нарратива и средств выдвижения в жестовых языках

Существуют исследования, посвященные различным аспектам и особенностям нарратива в жестовых языках. Так, в работе Б. Баркуэй-Браун [1] анализируется структура нарративов в новозеландском жестовом языке и делается вывод о их соответствии структуре У. Лабова, предложенной для звуковых языков. Важной особенностью текстов на жестовых языках также является «сконструированное действие» — рассказчик принимает на себя роль участника ситуации, передавая его выражение лица, жестикуляцию и манеру поведения [16]. Также рассматривается такое явление, как «разделение тела» (body partitioning): говорящий использует пространство и части своего тела для отображения нарратора и персонажа. Разделение тела и направление взгляда могут использоваться и для переключения между временными планами в нарративе [10]. Отмечается также важность использования в нарративе жестикуляции: в статье [15] авторы показывают, как нарративный дискурс складывается из использования лексем, мануальной жестикуляции, направления взгляда, пантомимы и выражения лица.

Средства выдвижения рассматриваются в работе Р. Уилбур на материале американского жестового языка: она выделяет топикализацию, клефт, псевдоклефт, дислокацию влево [24], а также в диссертации В. Киммельмана на материале РЖЯ: он обнаруживает функцию выдвижения у модели X Y X, где удвоенный элемент X выдвигается на первый план [11]. Во многих исследованиях также отмечается специфическое для жестовых языков средство выдвижения на первый план — использование одновременных конструкций: две руки выполняют разные мануальные жесты, обозначающие две ситуации, одна из которых принадлежит к основным событиям нарратива, другая является фоном [5; 13; 17; 19].

Функции аспектуальных показателей и акциональных типов предиката в нарративе рассматриваются в работе К. Раттмана, посвященной аспектуальности в американском жестовом языке. Согласно его выводам, продвижение истории вперед связано с предельностью глагола, поэтому для обозначения ситуаций основной линии используются предельные глаголы и глаголы с показателем перфектива FINISH ‘закончить’. Использование этого показателя указывает на хронологический порядок действий, описываемых глаголами [19].

3 Корпус текстов русского жестового языка

Исследование выполнено на материале корпуса текстов русского жестового языка. Корпус текстов РЖЯ был создан в рамках проекта «Корпусное исследование морфосинтаксиса и лексики русского жестового языка» (2012–2014 гг.), поддержанного Российским фондом фундаментальных исследований, под руководством С.И. Бурковой и при участии автора данной статьи.

Корпус РЖЯ содержит более 230 аннотированных видеотекстов разных видов: спонтанные нарративы, диалоги, пересказы мультфильмов и рассказы по комиксам Бидструпа. Тексты были записаны от 43 информантов — носителей языка. Возраст информантов варьируется от 18 до 63 лет. В корпусе также представлены все типы носителей русского жестового языка: глухие, слабослышащие, CODA (слышащие дети глухих родителей). Запись видеотекстов осуществлялась в Новосибирске и в Москве, поэтому, предположительно, они принадлежат к «московскому» и «сибирскому» диалектам РЖЯ с возможным включением других региональных вариантов, так как некоторые информанты длительно проживали в различных регионах России (Алтай, Казань, Красноярск, Томск и т. п.), а также в Казахстане.

Разметка корпуса включает три слоя: «Правая рука», «Левая рука» и «Перевод». Слои «Правая рука» и «Левая рука» содержат условный перевод каждого жеста на русский язык. В глоссах также обозначены особенности словообразования (компаунд, дактильное заимствование и др.) и грамматические показатели (императив, множественное число, субъектно-объектное согласование жеста и др.). Слой «Перевод» содержит литературный перевод текста на русский язык в виде предложений. Корпус текстов РЖЯ позволяет осуществлять поиск с учетом следующих критериев: слой, место записи текста, вид текста, пол и возраст информанта.

Для данного исследования из корпуса было выбрано 50 спонтанных нарративов. В выборку не вошли нарративы по техническим причинам (отсутствует начало или конец истории), а также монологи, не содержащие истории с сюжетом.

4 Функции аспектуальных показателей и акциональных типов предикатов в организации нарратива с точки зрения основной линии и фона в РЖЯ

Данное исследование опирается на уже существующие работы по русскому жестовому языку, посвященные исследованию темпоральности, аспектуальности и акциональности. Так, средства выражения прошедшего времени изучались в статье Е. Шамаро [20], аспектуальная система русского жестового языка и акциональность были описаны в диссертации автора данной статьи [6]. Русский жестовый язык не имеет грамматической категории времени и четкого противопоставления перфектива и имперфектива, хотя и имеет средства выражения этих значений. Прошедшее и будущее время выражается с помощью аналитических показателей — частично грамматикализованных жестов БЫЛО и БУДЕТ, обычно занимающих позицию после глагола (но не являющихся обязательными). Также используются различные показатели для перфективных (жест УЖЕ/ВСЕ для инхоатива, комплетива, жест ГОТОВО для результатива) и имперфективных (редупликация для итератива, жест ВСЕГДА для хабитуалиса) значений.

4.1 Жест УЖЕ/ВСЕ как показатель перфективных значений и перфекта

Жест УЖЕ/ВСЕ в знаменательной функции имеет значения ‘все’, ‘уже’, ‘довольно’, ‘кончено’ [20: 184; 7]. Как грамматический показатель он выражает значение перфекта и ряд перфективных значений: комплетив, инхоатив, перфектив. В служебной функции он занимает позицию строго после глагола. Форма жеста и некоторые пути его грамматикализации обнаруживают сходство с американским жестом FINISH ‘закончить’, упомянутым выше.

Глаголы, маркированные УЖЕ/ВСЕ, описывает перфективную ситуацию, относящуюся к основной линии нарратива (1, 2).

(1) CLF:ЧЕЛОВЕК.ИДТИ ЧИСТИТЬ ГРЯЗНЫЙ ПОДТЕРЕТЬ **ЧИСТИТЬ ВСЕ**
‘Пошел подмыть запачкавшегося [ребенка], подтер, подмыл’.

(2) INDX ЧЕЛКА PRTCL PRTCL **СТРИЧЬ ВСЕ** ЧИСТЫЙ ВСЕ
‘Там челку, ладно, постриг, все, чисто’.

Часто этот жест употребляется именно там, где нужно указать на последовательность действий: предыдущее действие закончено, началось следующее (3).

(3) **КУТАТЬ ВСЕ** Я ГОТОВИТЬСЯ СТИРАТЬ ВЕШАТЬ НУЖНО
‘Ребенка укутал, готовлюсь — надо стирать и вешать [белье]’.

Однако данные показывают, что в качестве показателя перфектива этот показатель употребляется в нарративах реже, чем в других функциях. Большинство его употреблений в нарративах связано с дискурсивными функциями: 1) маркирование конца эпизода, 2) указание на строгий порядок действий, процедуру, 3) маркирование пика нарратива. Об этом свидетельствует его употребление не после глагола, а в конце клаузы, появление в отрицательных клаузах, отсутствие перфективизации глагола. Таким образом, этот жест может структурировать не только последовательность отдельных ситуаций, но и последовательность эпизодов.

Как и в звуковых языках, в РЖЯ показатель перфективных значений проявляет тенденцию к использованию в клаузах основной линии. Функционирование данного показателя в РЖЯ сходно с показателем перфектива *le* в китайском языке: согласно работам [3; 4], он также развивает функции маркирования конца эпизода и пика нарратива. Исследования схожих показателей в жестовых языках показывают, что они также употребляются для смены темы и маркирования конца эпизода/нарратива [12; 18], указания на порядок действий или процедуру [18; 19].

4.2 Жест БЫЛО как показатель прошедшего времени

Большинство лингвистов, изучающих различные жестовые языки, сходятся во мнении, что маркирование времени в них не является обязательным, соответственно, категория времени не является грамматической. Темпоральная интерпретация ситуации может зависеть от различных факторов: лексических средств (например, наречий времени), аспектуальных показателей и контекста [23, 25]. В русском жестовом языке, однако, есть специализированный показатель прошедшего времени — частично грамматикализованный жест БЫЛО, который в качестве полнозначного жеста переводится как ‘был/была/было/были’. В функции показателя прошедшего времени он обычно находится в постпозиции по отношению к глагольному жесту, но в спонтанной речи может встречаться и перед глаголом.

Его употребление, однако, также не является обязательным; он может не встречаться в тексте ни разу, а может маркировать несколько глаголов. Е. Шамаро отмечает, что жест БЫЛО может быть опущен, если в предложении временная отнесенность ситуации уже выражена лексически [20]. Это позволяет предположить, что употребление жеста БЫЛО как показателя прошедшего времени может быть связано именно с дискурсивной организацией текста. Анализ нарративов в корпусе РЖЯ показывает, что он выполняет две функции: задает временную интерпретацию в интродуктивном фрагменте нарратива (4) и маркирует выпадение ситуации из нарративной цепочки как относящейся к прошлому (5).

(4) ТАКОЙ.ЖЕ БЫЛО Я ДЕРЕВНЯ **ЕХАТЬ БЫЛО**
‘Еще было — **я поехал** в деревню’.

(5) ПОНЯТЬ ФАКТ Я **ДАТЬ БЫЛО** 500 РУБЛЬ
‘[Девушка дала таксисту деньги]. Поняла, что на самом деле я **дала** 500 рублей’.

Часто маркирование прошедшего времени с помощью жеста БЫЛО встречается в диалогах (6), поскольку персонажи обсуждают события, которые были изложены до этого.

(6) Я БОЛЬНИЦА **ПРОВЕРИТЬ БЫЛО** / Я РОЖАТЬ НИКОГДА
‘[Мужчина прошел обследование, вернулся домой и говорит жене]: Меня **проверили** в больнице, у меня не может быть детей’.

4.3 Простая редупликация как показатель итератива

Итеративное значение в РЖЯ выражается с помощью простой редупликации глагольного жеста и/или редуплицированного жеста БЫЛО в значении ‘бывает’. Использование простой

редупликации для выражения итератива ограничено предикатами со значением событий, тогда как редуплицированный жест БЫЛО сочетается с предикатами всех типов.

Редупликация как показатель итератива чаще маркирует те глаголы, которые описывают ситуации, относящиеся к фону нарратива. Это соответствует выводам П. Хоппера о том, что итеративные ситуации, будучи в основном имперфективными, относятся к фону нарратива [8: 215]. В примерах (7, 8) редуплицированный жест БЫЛО появляется, когда информанты переходят от основной линии к комментарию.

(7) **БЫЛО+** **БОЛЕТЬ** СПИНА INDX
‘**Бывает, болит** спина’.

(8) А ГОВОРИТЬ **БЫЛО+** **СМОТРЕТЬ++** PRTCL
‘А слышащие [букв. говорящие], **бывает, смотрят** [на глухих, когда те общаются]’.

В (9) клауза с редуплицированным глаголом ПРИЙТИ становится фоновой по сравнению с остальными, содержащими нередуплицированные предикаты событий СДАТЬ, ДАТЬ, ПРИЙТИ, АРЕСТОВАТЬ, ПНУТЬ. В (10) информант отклоняется от истории своего выздоровления, чтобы рассказать, как в принципе работает прибор по очистке воды.

(9) МАГАЗИН СДАТЬ / ДАТЬ ПОЛОЖИТЬ В КАРМАН /
ВИДЕО ПРИЙТИ ГЛАЗЕТЬ // ТАКЖЕ КАЖДЫЙ П-О-Ч-Т-И
ЧАСТО **ПРИЙТИ+** / ПОКА ДО ПОКА Я АРЕСТОВАТЬ ПНУТЬ
‘В магазин сдали [бутылки], мне дали [денег], положил в карман, мы пошли смотреть видео [в видеосалон]. Мы ходили [на завод] часто, почти каждый [день], пока нас не арестовали и не дали пинка’.

(10) ВОДА ЛЕЧИТЬ ЕСТЬ // СПЕЦИАЛЬНЫЙ УСТРОЙСТВО СТОИТЬ
ТЫСЯЧА:ПЯТЬ ШЕСТЬСОТ РУБЛЬ // **ПИТЬ+++** ТУАЛЕТ **БЕЗ+++**
‘Вода меня вылечила. Это специальное устройство, стоит 5600 рублей. Пьешь, ходишь в туалет’.

В отношении использования показателя итератива РЖЯ демонстрирует сходство со звуковыми языками, в которых употребление имперфективных форм более характерно для фоновых клауз.

4.4 Жест ВСЕГДА как показатель хабитуалиса

Жест ВСЕГДА грамматикализован в роли показателя хабитуалиса, по крайней мере в «сибирском» варианте РЖЯ. Анализ нарративов в корпусе показывает, что этот показатель маркирует ситуации, относящиеся к фону. В (12) рассказчик описывает быт хозяев, у которых он остановился. В (13) рассказывает об устройстве общежития, описывая решетку, которая играет важную роль в дальнейшем сюжете. В (14) рассказчик описывает, чем он занимается в течение нескольких дней с друзьями, пока самолет задерживается.

(12) Л-А-Й-К-А РЯДОМ СОБАКА INDX **ВСЕГДА** **СПАТЬ**
ВМЕСТЕ СЕМЬЯ ДЕТИ ВМЕСТЕ
‘Лайка обычно спит вместе с семьей и детьми’.

(13) РЕШЕТКА **ВСЕГДА** МАЛЬЧИК ДЕВОЧКА **БОЛТАТЬ** ЛЮБИТЬ ЛЮБОВЬ
‘Через решетку мальчики с девочками обычно болтают, влюбленные’.

(14) К-А-Ф-Е **ПРИЙТИ** **ВСЕГДА** СТОЛОВАЯ ПРИЙТИ ЕСТЬ:DISTR
‘Ходим в кафе, в столовую, едим’.

4.5 Акциональные типы предикатов

В диссертационном исследовании [6] нами было выделено 8 типов предикатов на основании их сочетаемости с различными аспектуальными показателями: событие, сильный предельный процесс, слабый предельный процесс, дискретный непредельный процесс, дискретный процесс, процесс-событие, эпизодическое состояние, вневременное состояние. Также было показано, что акциональная семантика отражается в фонологической форме предиката и связана с наличием/отсутствием движения, повтора, контакта с корпусом. Акциональная семантика предикатов играет важную роль в перфективной/имперфективной интерпретации предложения, предикаты со значением событий обозначают перфективные ситуации, предикаты со значением состояний — имперфективные ситуации, а предикаты со значений процессов могут получать перфективную/имперфективную интерпретацию в зависимости от контекста.

Для описания ситуаций основной линии в РЖЯ используются предикаты со значением событий (15) и предельных процессов (16).

(15) INDX СОБАКА СПАТЬ **ВСКОЧИТЬ** СТРАХ // **ГАВКНУТЬ** Г-А-В // ЖЕНЩИНА КРИЧАТЬ **ИСПУГАТЬСЯ** КРИЧАТЬ // CLF:МНОЖЕСТВО:БЕГАТЬ ПОЛНОСТЬЮ **РОДИТЬ**
‘Там собака спала, она вскочила, испугавшись. Гавкнула. Женщина кричит, испугалась. Все быстро забегали, засуетились, [женщина] родила’.

(16) CLF:ПРИНЕСТИ.ПОДНОС // Я **ЕСТЬ** PRTCL
‘Принесли поднос. Я поел, ладно’.

Ситуации, относящиеся к фону, могут описываться разными типами предикатов, но преобладают состояния (17) и непредельные процессы (18).

(17) Я РЫБА **НЕ.ЛЮБИТЬ**
‘[Принесли второе. Еще хуже. Вареная рыба]. Я рыбу не люблю’.

(18) Я ОДНА ГЛУХОЙ / ГОВОРЯЩИЙ **ЛЕЖАТЬ:DISTR** МАЛЬЧИК ДЕВОЧКА МАЛЕНЬКИЙ // **ИГРАТЬ** ПОПАСТЬ **РИСОВАТЬ+**
‘Я одна глухая. Там лежали слышащие мальчики и девочки. Мы играли, рисовали’.

Предикаты со значением событий также могут использоваться в клаузах, относящихся к фону. Так, например, в (19) среди предикатов со значением процесса, описывающих быт монахов, встречается два предиката со значением событий, которые явно имеют итеративную интерпретацию, при этом они не маркированы редупликацией. Интерпретация предиката в таком случае зависит от контекста.

(19) ОКАЗЫВАТЬСЯ INDX **ВСТАТЬ** УТРО РАНО ПЯТЬ // МОЛИТЬСЯ // ПОТОМ CLF:МНОЖЕСТВО:ИДТИ РАБОТАТЬ ПОЛЕ // INDX СВОЙ ПОЛЕ // САМ ПОСАДИТЬ+ САМ СОБИРАТЬ // ДЕНЬГИ ПРОДАВАТЬ МЕНЯТЬ СТРОИТЬ // НИКТО **ПОМОЧЬ**
‘Оказывается, они встают рано, в пять. Молятся. Потом идут работать в поле. Там свое поле. Сами сажают, сами собирают, на деньги от того, что продали, строят. Никто не помогает’.

В классификации акциональных предикатов в [6] не рассматриваются специально классификаторы — жесты, обозначающие класс объекта, с которыми можно образовать практически неограниченное количество конструкций, изменяя параметры жеста (локализация, движение, конфигурация, ориентация). Классификаторные конструкции активно используются в нарративах. Опираясь на статью Т. Суппалы, где он делит классификаторные конструкции на пропозиции существования, нахождения и движения [21], можно предположить, что первые две

могут быть рассмотрены как предикаты состояния, а классификаторы, передающие движение объекта, могут обозначать процессы или события.

Действительно, материал РЖЯ показывает, что для описания фоновых ситуаций используются классификаторные конструкции без движения, обозначающие состояние или местонахождение (20), или с повторяющимся движением, выражающим семантику итератива (21).

(20) **CLF:МНОЖЕСТВО.СИДЕТЬ** СМЕЯТЬСЯ МОЛЧАТЬ
‘Люди сидят, смеются, молчат’.

(21) ЭТАЖ:ПЯТЬ=ОТРЕЗОК РЯДОМ ЭТАЖ:ПЯТЬ= ЭТАЖ:ПЯТЬ ЭТАЖ:ПЯТЬ= **CLF:ЧЕЛОВЕК.ДВИГАТЬСЯ++** ЭТАЖ:ПЯТЬ=INDX СТОЛОВАЯ ЕСТЬ ЭТАЖ:ПЯТЬ=ПРИЙТИ ЭТАЖ:ПЯТЬ= **CLF:ЧЕЛОВЕК.ДВИГАТЬСЯ++** ЭТАЖ:ПЯТЬ=INDX ОБЩЕЖИТИЕ
‘Там рядом стоят два дома, в одном столовая, туда ходили есть, в другом общежитие, вот и ходили между ними туда-сюда’.

Соответственно, классификаторные конструкции с одиночным движением выступают в роли предикатов, обозначающих события, и продвигают историю вперед (например, в (22) жест CLF:САМОЛЕТ:ВЗЛЕТЕТЬ). Классификаторы с другими типами движения (в (22) в жесте CLF:ИДТИ перебирание пальцами изображает движение ног) могут, как и процессы, быть интерпретированы перфективно или имперфективно в зависимости от контекста, соответственно, появляться и в ситуациях основной линии, и в ситуациях фона.

(22) НОРМАЛЬНО **CLF:ИДТИ** // **CLF:САМОЛЕТ:ВЗЛЕТЕТЬ**=ПОВЕРХНОСТЬ УДАЧНО ВСЕ
‘[Таможню] нормально прошла. Самолет взлетел, [все вышло] удачно’.

Так, в отношении использования акциональных типов предикатов в клаузах основной линии и фона РЖЯ демонстрирует закономерности, характерные для звуковых языков: в клаузах основной линии чаще встречаются предикаты со значением событий и предельных процессов, в фоновых клаузах — предикаты со значением процессов и состояний.

4.6 Особенности исполнения глагольного жеста

Так как темпоральные и аспектуальные показатели не всегда маркируют глагольный жест в спонтанных нарративах (что, скорее всего, связано с незавершенным процессом грамматикализации и ненормированностью РЖЯ), а большинство типов предикатов могут использоваться, хоть и с разной частотностью, для обозначения и ситуаций основной линии, и фона, также обращают на себя внимание мануальные и немануальные особенности исполнения глагольного жеста, которые тоже могут способствовать выделенности клаузы.

Часто жест, описывающий фоновую ситуацию, так или иначе редуцируется: уменьшается амплитуда движения, уменьшается время исполнения, утрачивается повтор. Наиболее ярко это проявляется при сравнении одного и того же предиката в одном нарративе в ситуациях, относящихся к основной линии и к фону. Так, в (23) жест КОЛЕСО.СПУСТИТЬ представляет собой предикат события и описывает одно из ключевых событий истории, принадлежащее к основной линии: у машины рассказчика спустило колесо. В (24) этот жест употребляется в клаузе, относящейся к фону, — «там, где у меня спустило колесо». Рисунок 1 демонстрирует различие в амплитуде жеста. В ситуации основной линии длительность жеста составляет 1,06 секунды, в ситуации фона — 0,7 секунды.

(23) INDX ВЕСТИ.МАШИНА СЛОМАТЬСЯ **КОЛЕСО.СПУСТИТЬ**
‘Туда еду-еду, и тут сломалось, колесо спустило’.

(24) ПРЫГНУТЬ КАНАЛ **КОЛЕСО.СПУСТИТЬ** CLF:ЧЕЛОВЕК.ИДТИ ВОДА CLF:ЧЕЛОВЕК.ИДТИ ПОВЕРХНОСТЬ / ПЛЫТЬ

‘Прыгнул там, где колесо спустило, над каналом, плыву’.

Рисунок 1: Жест КОЛЕСО.СПУСТИТЬ в ситуации основной линии и фона

Также один и тот же предикат, описывающий ситуации основной линии и фона, может различаться немануальными маркерами (движения лица и тела, мимика). В нарративах жестовых языков активно используется «сконструированное действие» ролевой сдвиг: когда рассказчик описывает какую-то ситуацию, он принимает на себя роль субъекта этой ситуации и передает его манеру говорить, поведение и выражение лица [16]. Так, например, в (25) одно из ключевых событий нарратива — рассказчица открутила лампочку, и вся гирлянда погасла — передается с выражением, изображающим себя — ребенка в прошлом, а жест ПОГАСНУТЬ исполняется резко и сопровождается выражением шока и испуга на лице. В (26) рассказчица делает отступление и поясняет принцип работы гирлянды, у нее нейтральное выражение лица, взгляд направлен на собеседника, жест ПОГАСНУТЬ исполняется более плавно.

(25) **КРУТИТЬ** **ОТКРУТИТЬ** ЦОКОЛЬ **ПОГАСНУТЬ**
‘Кручу, открутила [лампочку], [весь свет] погас’.

(26) НУЖНО ОТКУДА ФИЗИКА POSS // ОДИН **ОТКРУТИТЬ** INDX
ВСЕ ВСЕ **ПОГАСНУТЬ**
‘Нужно физику знать. Одну выкрутишь — все погаснет’.

Рисунок 2: Немануальные маркеры жестов ОТКРУТИТЬ и ПОГАСНУТЬ
в ситуации основной линии

Рисунок 3: Немануальные маркеры жестов ОТКРУТИТЬ и ПОГАСНУТЬ в ситуации фона

5 Выводы

Корпусный анализ спонтанных нарративов русского жестового языка показывает, что аспектуальные и темпоральные показатели, не будучи строго обязательными, не могут последовательно использоваться для выдвижения ситуаций на передний план или, наоборот, уменьшения степени их выделенности. При этом они могут указывать на выпадение ситуации из нарративной цепочки или отнесенность ее к пику нарратива. Более последовательно для разграничения основной линии и фона используются акциональные типы предиката: для ситуации основной линии характерны предикаты со значением событий и предельных процессов, для фона — предикаты со значением процессов и состояний. В данном аспекте РЖЯ обнаруживает закономерности, свойственные звуковым языкам. Однако в РЖЯ также есть специфические для языка визуальной модальности средства, которые могут выделять ситуацию из других, — большая амплитуда, длительность или резкость исполнения жеста, а также немануальное маркирование — направление взгляда, передача эмоций и манеры поведения субъекта ситуации. Можно предположить, что данные особенности будут характерны для жестовых языков в целом, так как обусловлены визуально-кинетическим каналом передачи информации, в то время как в функционировании показателей времени и аспектуальности в нарративном дискурсе в жестовых языках обнаружатся различия, так как эти показатели развиваются независимо в каждом жестовом языке. Во многих аспектах жестовые языки проявляют большее типологическое сходство, чем звуковые, поэтому полученные данные будут полезны для установления того, насколько отличается устройство нарративов в различных жестовых языках, а также причин данных сходств или различий.

Благодарности

Автор благодарит рецензентов за полезные замечания и конструктивную критику. Работа выполнена при поддержке гранта РНФ №22-18-00120.

References

[1] Barkway-Brown B. What is it like to “hear” a hand? Deaf Narratives from the New Zealand Deaf Community. Master’s thesis. — Dunedin: University of Otago, 2011. — 234 p.

[2] Burkova S.I. Russian Sign Language Corpus [electronic source]. (Korpus russkogo zhestovogo yazyka [Elektronnyy resurs]. Project manager Svetlana Burkova (Rukovoditel′ proyekta Burkova S. I.). Novosibirsk, 2012–2015. http://rsl.nstu.ru/

[3] Chang V. The particle *le* in Chinese narrative discourse: an integrative description. Doctoral dissertation. — University of Florida, 1986.

[4] Chappell, Hilary. 1986. Restrictions on the use of ‘double le’ in Chinese. — Cahiers de Linguistique Asie Orientale 15. — 1986. — P. 223–252.

[5] Engberg-Pedersen E. Some simultaneous constructions in Danish Sign Language // Brennan M. and Turner G. (eds) Word-order Issues in Sign Language. Working Papers. — Durham: International Sign Linguistics Association, 1994. — P. 73–88.
[6] Filimonova E.V. Functional-semantic category of aspect in Russian Sign Language. Doctoral dissertation (Funktsional′no-semanticheskaya kategoriya aspektual′nosti v russkom zhestovom yazyke: dissertatsiya kandidata filologicheskikh nauk). — Moscow, 2016. — 314 p.
[7] Filimonova E.V. Functions of the sign ALREADY in Russian Sign Language (Funktsii zhesta UZHE v russkom zhestovom yazyke). Cross-cultural communication: linguistic and linguodidactic aspects: collection of materials of 4th international scientific and methodological conference (Mezhkul′turnaya kommunikatsiya: lingvisticheskiye i lingvodidakticheskiye aspekty: sbornik materialov 4-y mezhdunarodnoy nauchno-metodicheskoy konferentsii). Novosibirsk: publishing company of Novosibirsk State Technical University, 2013. P. 161–169.
[8] Hopper P. Aspect and foregrounding in discourse // Talmy Givon (ed.) Discourse and syntax (Syntax and semantics, 12). — New York: Academic Press, 1979. — P. 213–241.
[9] Hopper P., Thompson S. Transitivity in grammar and discourse. Language. — 1980. — Vol. 56 (2). — P. 251–299.
[10] Janzen T. Shared Spaces, Shared Mind: Connecting Past and Present Viewpoints in American Sign Language Narratives // Cognitive Linguistics, 30. — 2019. — P. 253–79.
[11] Kimmelman V. Information structure in Russian Sign Language and Sign Language of the Netherlands. Amsterdam: University of Amsterdam PhD dissertation. — 2014.
[12] Lee W.F. Aspect in Hong Kong Sign Language. — Chinese University of Hong Kong. — 2002.
[13] Leeson L., Saeed J. I. Windowing of attention in simultaneous constructions in Irish Sign Language (ISL) // Proceedings of the Fifth Meeting of the High Desert Linguistics Society, 2004. — P. 1–18.
[14] Longacre R. The Grammar of Discourse. — New York: Springer Science & Business Media, 1996. — 362 p.
[15] McCleary L., Viotti E. Sign-gesture symbiosis in Brazilian Sign Language narrative // Parrill F., Tobin V., Turner M. (eds) Meaning, Form and Body. — Chicago: University of Chicago Press, 2010. — P. 181–201.
[16] Metzger M. Constructed dialogue and constructed action in American Sign Language // Lucas C. (ed.) Sociolinguistics in Deaf Communities. — Washington, DC: Gallaudet University Press, 1995. — P. 255–271.
[17] Miller C. Simultaneous constructions in Quebec Sign Language // Ahlgren I., Bergman B., and Brennan M. (eds) Proceedings of the Perspectives on Sign Language Structure: Papers from the 5th International Symposium on Sign Language Research. — Durham: International Sign Linguistics Association, 1994. — P. 89–112.
[18] Palfreyman N. Form, function, and the grammaticalisation of completive markers in the sign language varieties of Solo and Makassar. — NUSA: Linguistic studies of languages in and around Indonesia, 55. — 2013. — P. 153–172.
[19] Rathmann C. Event Structure in American Sign Language. PhD thesis. — Austin: The University of Texas, 2005. — 279 p.
[20] Shamaro E. Some facts of temporal-aspectual system of Russian Sign Language (Nekotoryye fakty vidovremennoy sistemy RZHYA). Komarova A. (ed.), Current aspects of sign language (Sovremennyye aspekty zhestovogo yazyka (sost. A. A. Komarova). Moscow, 2006. P. 180–191.
[21] Supalla T. The classifier system n American Sign Language // Craig C. (ed.) Noun classes and categorization: Typological Studies in Language. — Philadelphia: John Benjamins Publishing, 1986. — P. 182–214.
[22] Wallace S. Figure and Ground: The Interrelationships of Linguistic Categories // Hopper P. (ed.) Tense-Aspect: Between semantics & pragmatics (Typological Studies in Language, 1). — 1982. — P. 201–223.
[23] Wilbur R. American Sign Language and Sign Systems. — Baltimore: University Park Press. — 1979.
[24] Wilbur R. Foregrounding structures in ASL. — Journal of Pragmatics, 22(6). — 1994. — P. 647– 672.
[25] Zucchi S. Tense, Time, and Adverbs in Italian Sign Language // Universit'a degli Studi di Milano. — 2005.

Приложение А. Условные обозначения

DISTR — дистрибутив; CLF — классификаторный жест; INDX — указательный жест; POSS — притяжательное местоимение; PRTCL — частица; REC — реципрок; / — пауза; // — длинная пауза; + – повторение жеста при редупликации; : — несегментное выражение нескольких значений в рамках одной формы; = — одновременное исполнение жестов разными мануальными артикуляторами; П-О-Ч-Т-И — передача слов с помощью дактильной азбуки.

Multimodal Discourse Trees in Forensic Linguistics

Boris Galitsky
Knowledge Trail Inc.
San Jose, CA, USA
bgalitsky@hotmail.com

Dmitry Ilvovsky
NRU HSE
Moscow, Russia
dilvovsky@hse.ru

Elizaveta Goncharova
NRU HSE, AIRI
Moscow, Russia
egoncharova@hse.ru

Abstract

We extend the concept of a discourse tree (DT) in the discourse representation of text towards data of various forms and natures. The communicative DT to include speech act theory, extended DT to ascend to the level of multiple documents, entity DT to track how discourse covers various entities were defined previously in computational linguistics, we now proceed to the next level of abstraction and formalize discourse of not only text and textual documents but also various kinds of accompanying data. We call such discourse representation Multimodal Discourse Trees (MMDTs). The rational for that is that the same rhetorical relations that hold between text fragments also hold between data values, sets and records, such as Reason, Cause, Enablement, Contrast, Temporal sequence. MMDTs are evaluated with respect to the accuracy of recognition of criminal cases when both text and data records are available. MMDTs are shown to contribute significantly to the recognition accuracy in cases where just keywords and syntactic signals are insufficient for classification and discourse-level information needs to be involved.

Keywords: natural language processing; discourse structure; multimodality
DOI: 10.28995/2075-7182-2023-22-79-87

Построение мультимодальных дискурсивных деревьев для анализа судебных документов

Б.А. Галицкий
Knowledge Trail Inc.
Сан-Хосе, Калифорния, США
bgalitsky@hotmail.com

Д.А. Ильвовский
НИУ ВШЭ
Москва, Россия
dilvovsky@hse.ru

Е.Ф. Гончарова
НИУ ВШЭ, AIRI
Москва, Россия
egoncharova@hse.ru

Аннотация

В работе исследуется концепция построения мультимодального дискурсивного дерева для структурированного представления текста, обогащенного дополнительной информацией из источников различной природы. В более ранних работах были введены понятия коммуникативных дискурсивных деревьев, расширенных с помощью теории речевых актов, а также расширенных дискурсивных деревьев, которые отражают структуру не одного текста, а набора связанных документов; в данной работе мы исследуем возможность расширения дискурсивной структуры за счет включения данных из дополнительных (нетекстовых) модальностей. Мы называем подобное дерево мультимодальным дискурсивным деревом и показываем, что отношения, которые можно установить между частями текста (дискурсивными единицами), также переносятся на данные, дополняющие текст, к которым можно отнести записи из баз данных (например, истории вебпоиска или финансовых операций и т.д.). Мы показываем, что построение мультимодального дискурсивного дерева помогает улучшить качество решения задач поиска на примере анализа судебных документов, которые в большинстве случаев сопровождаются информацией из различных дополнительных источников, по сравнению с поиском по ключевым словам или поиском по стандартному (текстовому) дискурсивному дереву.

Ключевые слова: обработка естественного языка; дискурсивные деревья; мультимодальность

1 Introduction

Discourse analysis plays important role in constructing a logical structure of thoughts expressed in text. Discourse trees are means to formalize textual discourse in a hierarchical manner, specifying rhetorical relations between phrases and sentences. Discourse trees (DTs) are a high-level representation compromise between complete logical representations like logical forms and informal, unstructured representations in the form of original text. Learning DTs has found a number of applications in content generation, summarization, machine translation and question answering (Amgoud et al., 2015; Joty et al., 2015; Joty et al., 2019). The limitation of DT's employment in a general data analysis task is that they are designed to represent the discourse of a text rather than a causal relationship between components of an abstract data item. In this paper, we will address this limitation and propose a solution to generalize DTs towards arbitrary data types with applications to health management and security. In previous works, the authors took DTs to the higher level of abstraction with the goal to form a unified structure for interactive knowledge discovery (Galitsky, 2020). The authors believed that a knowledge exploration should be driven by navigating a discourse tree built for the whole corpus of relevant content. They called such a tree as an extended discourse tree (EDT, (Galitsky, 2019)). It is a combination of discourse trees of individual paragraphs first across paragraphs in a document and then across documents. In this paper, we demonstrate application areas of a discourse representation with a higher level of abstraction and generality. We extend the concept of a discourse tree in the discourse representation of text towards data of various forms and natures. Having defined communicative DT (CDT) to include speech act theory, extended DT to ascend to the level of multiple documents (Ilvovsky et al., 2020) and entity DT to track how discourse covers various entities, we now proceed to the next level and discourse abstraction and formalize discourse of not only text and textual documents but also various kinds of accompanying data. The motivations here are that the same rhetorical relations that hold between text fragments also hold between data values, sets and records, such as *Reason*, *Cause*, *Enablement*, *Contrast*. We call DTs for text and other data forms Multimodal DTs (MMDTs) and apply them in the domains of forensic linguistics (Svartvik and Evans, 1968). Forensic linguistics examines language as it is used in cross-examination, evidence presentation, judge's direction, police cautions, police testimonies in court, summing up to a jury, interview techniques, the questioning process in court, and in other areas such as police interviews (Solan and Tiersma, 2005; Coulthard, 2014).

2 Multimodal Discourse Representation

In this work, we present the notion of Multimodal Discourse Tree (MMDT) that operates on the text level supported with the additional information derived from various sources, where the data is kept in more structural way rather than simple raw texts. Our objective is to recover chains of events from logs of transactions of various sorts including textual descriptions. We show a simple idea of merging various data sources in Figure 1. The trick is how to retain an original structure inherent to each source and merge it with the logical structure of text (an original story). We are motivated by the fact that any coherent text such as patients' complaints or description of the crime scene from the police report is structured so that we can derive and interpret the information.

Discourse analysis aims to reveal the logical structure of some coherent text. This structure shows how discourse units (text spans such as sentences or clauses) are connected and related to each other. In this work, we utilize the Rhetorical Structure Theory (Mann and Thompson, 1988) (RST) as a framework to derive this structure. RST divides a text into elementary discourse units (EDUs). It then forms a tree representation of a discourse called a discourse tree using rhetorical relations such as *Elaboration* and *Explanation* as edges, and EDUs as leaves. EDUs are linked by a rhetorical relation and are also distinguished based on their relative importance in conveying the author's message; nucleus is the central part, whereas satellite is the peripheral part.

In the multimodal setup, we propose to extend the original DT derived from plain text with additional information retrieved from the external sources, such as various logs (financial, call, driving, etc.). The discourse tree extended with this additional information is called MMDT.

Let us consider the motivation that lies behind MMDT construction. A user of some system is not

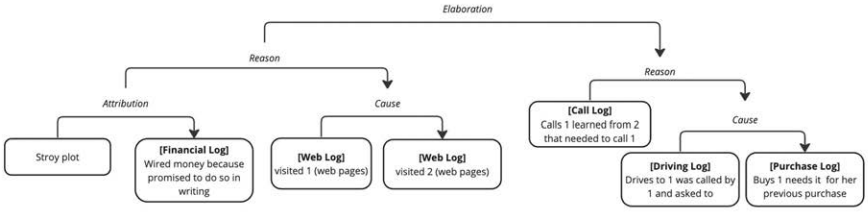


Figure 1: A scheme for a MMDT extended with additional data sources.

always aware of sharing the whole information about his needs, thus, his or her real agenda can be hidden from the system. By the analysis of the external sources of information and merging it with the initial user's request we can reconstruct the whole story and provide more relevant responses for user's request.

We consider various logs such an example of these external sources. It can be bank transactions logs, driving or call logs, web logs, and others. These sources of data can be parsed and represented as the EDUs in the MMDT as shown in Figure 1.

For example, let us imagine a situation, where user wants to make a money transfer, and bank manager is not sure whether this operation is fraud or not. By analysis of the log information kept for this user, the manager can know that this particular user promised to wire money in written form (known from the financial log), then the user visited some web pages from bank system (web logs) in order to make money transaction that he promised to. Thus, this transaction is not fraud and can be performed by the bank. This part of history of the user's log can be hierarchically linked and presented in the form of MMDT presented in Figure 1 (left branch). There we can see that an inner relation within a given data source are combined with interrelations between sources. The same relations hold within a source and between them. The overall logical structure of data is now independent of its nature. A numerical record for banking can be rhetorically connected with a numerical record for calling, which is in turn connected with that of for driving.

Let us now proceed with another example of crime scene description using the MMDT representation. We have a formal description of the crime scene described in the police report. We build the MMDT that can describe this situation within the supported facts represented as the additional modalities. The sources of the extra modalities are shown as the pictograms on the figures.

Let us split the police report about the crime scene into small chunks and build the MMDT supported with extra data for each of them. The MMDT for the first part of the original story is shown in Figure 2. There, the pictograms show the sources for the multimodal data, such as (driving logs, financial logs, etc.).

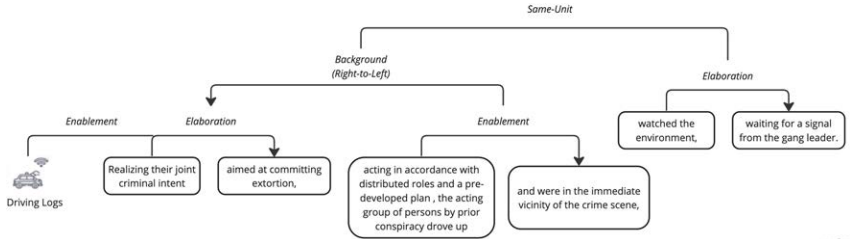


Figure 2: A MMDT for a preparation for a crime. The data record for driving is shown by a pictogram connected with the textual DT by *Enablement*.

Original story. Part 1.

Realizing their joint criminal intent aimed at committing extortion, acting in accordance with distributed roles and a predeveloped plan, the acting group of persons by prior conspiracy drove up and were in the immediate vicinity of the crime scene, watched the environment, waiting for a signal from the gang leader.

We proceed to the start of the extortion crime (Figure 3).

Original story. Part 2.

The victim, unaware of the impending crime against him, at about 5pm, arrived at the house number 162. He was awaited by Jones, acting by a group of persons in a prior conspiracy with Smith and Clark, who, under the pretext of taking out the garbage, left the house and went out to call Smith and Clark that the victim was now located indoors. Thus, Jones gave the signal to start committing the crime.

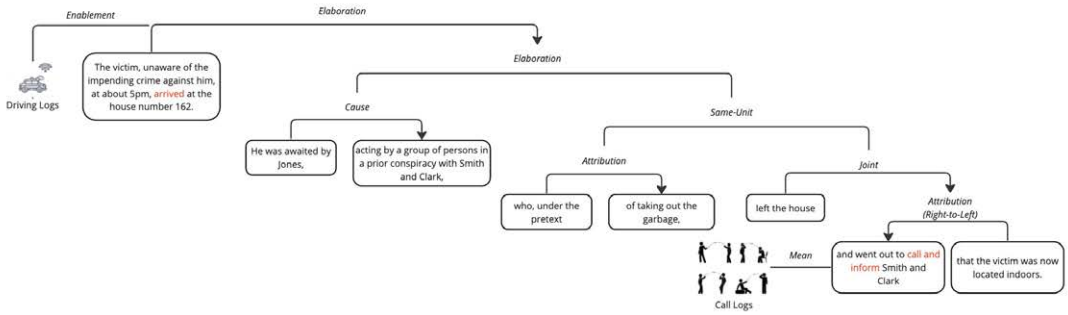


Figure 3: A MMDT for initiation of the crime extended with the multimodal data from the driving logs and call logs.

Rhetorical relations link text EDUs as well as discourse units with information chunks of other modalities: call logs and driving logs, connected with rhetorical relations of Means and Enablement. We now proceed with the crime description (Figures 4-5).

Original story. Part 3.

In the continuation of her joint criminal intent aimed at committing extortion, Jones returned to the house. She did not lock the front door with a key, in order for the gang to enter the house. Smith and Clark acted with her jointly and in agreement. Then they entered the house through the unlocked front door, yelling at the victim.

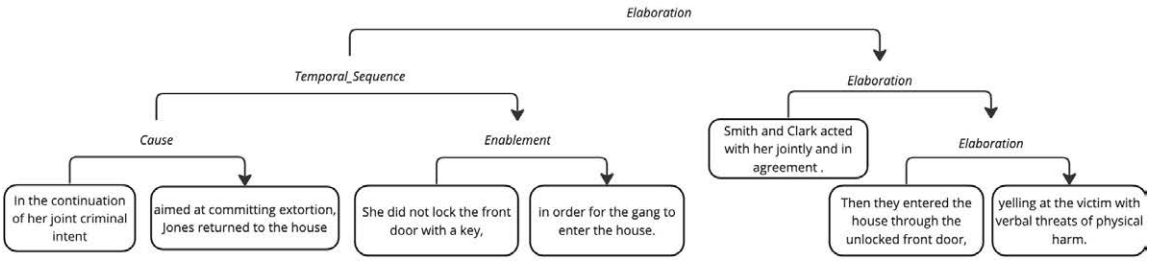


Figure 4: A MMDT for the start of the crime.

Original story. Part 4.

Smith pointed the knife to the victim's elbow and Gereyhanov pointed the handgun to the victim's back, threatening the victim with the murder, unless he does a money wire from his Chase account to Jones's Sberbank account. As the attackers needed more time to have the wire completed, they decided to move the victim to another house to continue money transfer. Jones made a call, making sure certain arrangement were made. Then the attackers pulled the victim out of the house and lead him to the car to drive 35 miles north.

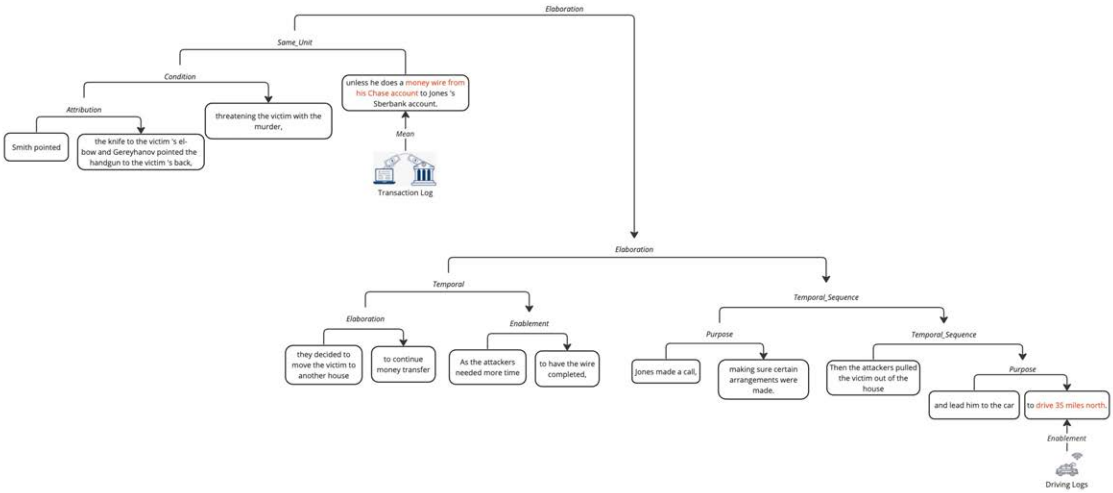


Figure 5: A MMDT for the main stage of the crime.

We now show a tree-like visualization of an arbitrary MMDT which can represent a crime scenario as well as a legal behavior one (Figure 6). This is an example of MMDT where discourse units are data elements such as phone calls, automated number plate recognition (ANPR) records, financial transactions, and texts.

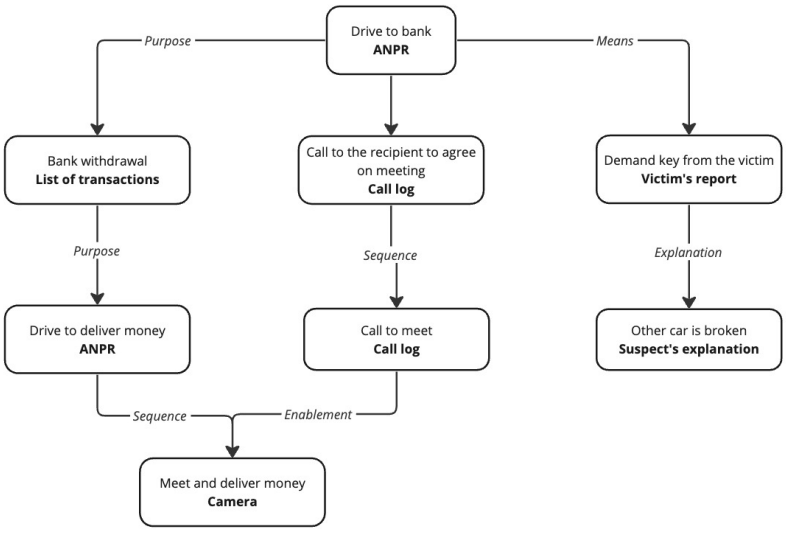


Figure 6: A Multimodal Discourse Tree. The source of additional information are in bold, while the corresponding discourse relations are in italics.

2.1 Multimodal Data Sources and References between them

In this section we analyze the use-case example of crime scene description introduced in the previous section. Via analysis of the MMDTs presented in Figures 2-6, we can observe the whole story in more structural way that can enable us to retrieve answers to the questions by navigating this discourse tree and automatically finding the correct part of the text where this answer can be found. By extending the discourse tree with the additional sources of data we provide extra logical links among the text parts.

Various data sources are not only connected via a discourse tree but are also linked with each other directly. It is hard to form a meaning from a single source, but once we can correspond event parameters from multiple sources and build a whole picture, the constructed event becomes meaningful. For example, if two cars follow each other with a short interval (as determined from the ANPR system), it means that their movement is coordinated.

Once the detective establishes that the gang member drove through a certain point one after another, she would look for confirmation from other sources. If people in one car can see another car, they do not need to call with the purpose of coordination or orientation; if they do make calls then they can have another communication purpose.

It is hard to prove that extortion occurs as the victim could possibly meet the demands of the attackers voluntarily, or the demands did not occur. Sources like calling logs, banking transfers and web logs can indicate whether extortion indeed occurred or not.

Once the extortion process starts, one would expect the victim to be deprived of communication means including phones and the Internet to avoid her calling for help. The call log can easily confirm or reject this expectation. If the frequency of calls of the victim is zero or much lower than that of the gang members, this is a confirmation of an extortion process. The web log can confirm the activity of the victim directly by showing how the victim logged into different accounts and made transfers. Corresponding weblog activities of the attackers who check the receipt of money would be informative as well. Moreover, victims' calls to the banker to perform a transaction that cannot be completed online can also be tracked and matched against the transactions themselves. IP addresses of bank requests can be matched against IP addresses of weblogs. Bank branch locations can be matched with ANPR locations (not used in this particular case).

When a financial transaction happens, a sender and a recipient need to call each other. Also, they likely drove together, or met at some location, as determined by ANRP and call log. Hence for two sources and events in each of them, there are frequently causal links between these events (shown as arrows in Figure 6).

The multimodal DT can be used as the additional source of information that allows us to answer the question based on the extended discourse tree and also ask questions w.r.t. the constructed DT. We now can enumerate multimodal DT-based questions that can be formalized and asked against a MMDT:

For a given individual, find people who visited at one point any location visited by a given person and transferred money to him or back

- Find all pairs of people who drove on different cars following each other within a kilometer of each other
- Find people who call each other and then meet
- Find people who call each other and then transfer money
- Find all people who were once in a location where a given person stayed/visited
- If A calls B who is in a branch in location L to check on account B?

All these questions are relevant for practical applications, where one can easily navigate through the connected textual corpora represented in the form of MMDT.

3 System architecture

We build a conventional CDT from text, convert into MMDT using available structured sources, and then put it into the index for classification and search. The steps of converting a DT into the MMDT are as follows:

1. Once we build an individual CDT for each portion of text, we build a single DT for the whole corpus.
2. As the DT is available, we start preparing accompanying data to incorporate it into DT to form the MMDT. Each data source is converted into a unified, canonical form with normalized named entities: time, date, location, person name, phone number, account number (if available). A scheme for multimodal data transformation is shown in Figure 7.
3. Iterate through each EDU of DT, identifying candidate phrases that can potentially be associated with accompanying data. Extract name entities with their types. Form a list of candidate EDUs for linking with data record.
4. For each candidate EDU, attempt to match entity values against those in data records.
5. In data records taken separately from DT, match records with each other and establish causal links, employing R-C reasoning framework.
6. Iterating through all causal links (and other link types), including internal in data records and external (DT - data records) links, confirm or reject each.
7. For confirmed causal links, insert respective edges in DT to obtain MMDT without relation labels.
8. Recognize types of rhetorical relations between DT and data records. Also, recognize rhetorical relations between data records.
9. Determine if the data record as EDU is connected with DT as nucleus or satellite.
10. Convert obtained labelled MMDT into a normalized MMDT.

As the additional multimodal data sources we use specific logs that provide us with the structured textual descriptions of the described event. For example, if a data record is linked to a pair of text EDUs connected with Elaboration, then Cause is inserted to strengthen the nucleus. We show the scheme for the normalization procedure that turns the data record into a regular EDU below in Figure 7.

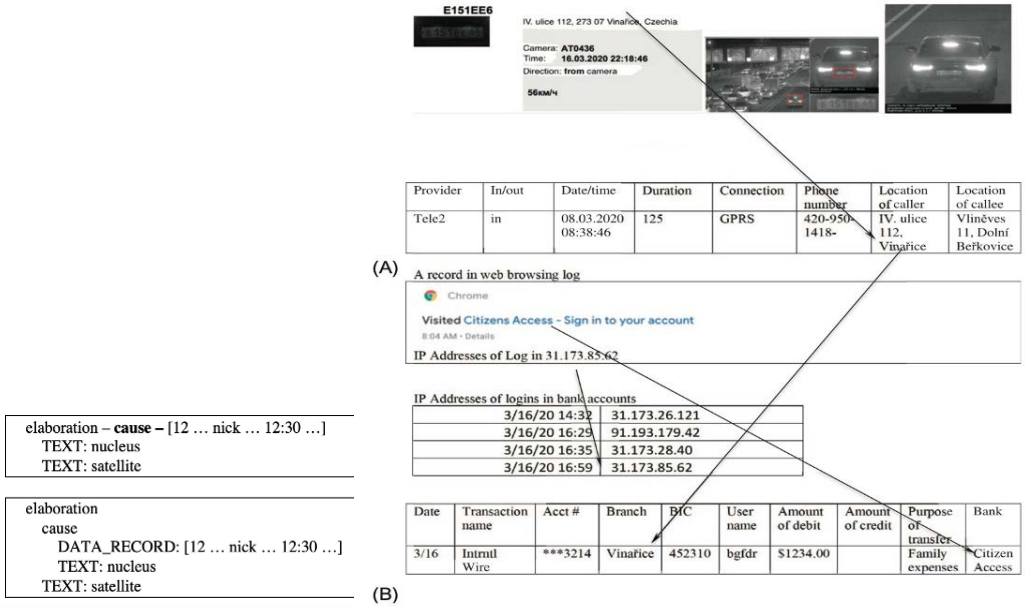


Figure 7: Data records in a call log, web browsing, IP Addresses of financial logins and banking transactions.

4 Evaluation

For the evaluation we considered one practical domain, where the MMDTs can be used for the analysis of texts with the supported structural information (such as log data). To evaluate the contribution of MMDT relative to DT for recognizing scenarios such as criminal cases, we classify them with respect to the felony category such as robbery, theft, abduction and extortion.

Recognition method	Keyword-based	DT	CDT	MMDT–brief description	MMDT - structured
Extortion vs Robbery	58.2	65.0	66.8	70.3	74.3
Robbery vs Theft	61.9	66.8	68.2	71.2	73.7
Extortion vs Theft	66.3	70.4	71.6	73.4	75.4
Abduction vs Extortion	69.1	74.2	76.0	78.5	80.3
Abduction vs Robbery	66.7	72.2	73.6	75.1	81.3
average	64.4	69.7	71.2	73.7	77.0
improvement		5.3	1.5	2.5	3.3

Table 1: Recognition accuracy for felony classes.

It should be noted that in such practical use-cases simple yet effective key-words based search is often applied that allows interpretable and fast search of relevant information. However, using keywords is usually insufficient, as the crime descriptions in court decisions are written in the same or similar keywords for all these crimes: property, car, guns, threats, violence. Discourse level considerations are required, and the more accurate and richer the representation is, the higher the expected recognition accuracy.

We form a dataset of criminal court decisions and attempt to automatically classify the entries in this dataset with respect to felony class. We mine the case site for criminal cases based on statute number and retain the description of corpus delicti to automatically relate it to the statute number. Case descriptions are mined from http://www.sudrf.ru and translated from Russian into English by Bing Translation API.

Due to the lack of complete data on criminal cases other than anonymized textual decision documents, evaluation of the contribution of MMDT is difficult. We build a hybrid dataset of genuine anonymized textual descriptions and attach the same randomized multi-source set of data records. We autogenerate a generic dataset of data records (GDDR) of phone calls, ANPR, weblog, and bank transactions. Having the names, dates, locations and other entities anonymized in both GDDR by the authors and in the public criminal dataset by the court authorities, we insert random entity value to associate actual criminal cases with randomized, hypothetical data records to obtain the complete criminal case data. We recognize one felony category against another, where there is a high similarity in how a crime in a given category is described. Each class there contains 500 documents with 3000 words on average.

Our baseline is keyword-based recognition and regular DTs (columns two and three). In the fourth column we include the phone, drive and money transfer data as a brief description rather than a complete data record and there are no inter–data record rhetorical relations. Finally, in the fifth column, more complete, structured multimodal information is included with built internal data record — data record rhetorical relations.

One can observe that DTs yield more than 5% recognition accuracy compared to keywords, and as we proceed to CDT we gain just 1.5% (Table 1). The next step of enhancement towards the ‘light’ MMDT delivers 2.5% while the ‘complete’ MMDT gives further 3.3%. The recognition rate does not vary significantly across GDDR with the felony class. The contribution of MMDT to an accurate representation of criminal case turns out to be significant and we expect this representation to not depend significantly on the machine learning method.

5 Discussions and Conclusion

In this paper, we took the discourse representation via trees to the next level of abstraction, going beyond textual data and enforcing rhetorical relations between arbitrary components of data items. This allowed us to treat computationally complex scenarios of inter-human interactions described in text and also as numerical and string vectors, once a causal relationship between the latter elements is established.

Complex scenarios of interactions such as GDDR also appear in such domains as security and health management, beyond criminalistics, where textual descriptions need to be merged with numerical values and the logical structure of these data sources must be analyzed together. In this work in progress, we consider the practical application of the MMDTs that can be used to organize complex texts derived from the various source in structural logically-organized format. In future research, we plan to extend the applicability of the introduced framework for more domains, showing where this approach to build the MMDT instead of simple DT or the analysis of plain texts can be preferable.

We computationally evaluated that complex scenarios of inter-human interactions described in plain words and also in data records can be adequately represented via MMDTs in the forensic analysis domain. MMDTs can be employed in other domains involving complex interactions between people or complex correlation between parameters such as customer and patient complaints, prediction of patients' behavior at pandemic times, control of a military unit and prediction of market behavior. These domains are hybrid in the sense that textual information is combined with numerical data and needs to be organized in a uniform way that is invariant with respect to the nature of features used in problem-solving. Statistical learning including deep learning families of approaches encodes all information numerically and certain meanings expressed in text are always lost. Even with a high recognition accuracy of statistical methods, explainability cannot be achieved because numerical representation cannot always be converted back into an interpretable form.

Conversely, MMDTs attempt to encode all available information via a graph with the focus on a high-level logical flow irrespectively of the learning machine which would be applied. Therefore, the MMDT – based approach fully supports explainability and avoids information loss under knowledge representation. MMDT can be naturally combined with additional characteristics of numerical data as well as syntactic and semantic representations.

Acknowledgements

The publication was supported by the grant for research centers in the field of AI provided by the Analytical Center for the Government of the Russian Federation (ACRF) in accordance with the agreement on the provision of subsidies (identifier of the agreement 000000D730321P5Q0002) and the agreement with HSE University No. 70-2021-00139.

References

Leila Amgoud, Philippe Besnard, and Anthony Hunter. 2015. Representing and reasoning about arguments mined from texts and dialogues. *// 13th European Conference on Symbolic and Quantitative Approaches with Uncertainty (ECSQARU 2015)*, volume 9161 of *Lecture Notes in Computer Science book series (LNCS)*, P 60–71, Compiègne, France, July.

Malcolm G. Coulthard. 2014. Whose text is it? on the linguistic investigation of authorship.

Boris Galitsky. 2019. *Discourse-Level Dialogue Management*. Springer International Publishing, Cham.

Boris Galitsky, 2020. *Recognizing Abstract Classes of Text Based on Discourse*, P 379–414. Springer International Publishing, Cham.

Dmitry Ilvovsky, Alexander Kirillovich, and Boris Galitsky. 2020. Controlling chat bot multi-document navigation with the extended discourse trees. *// Proceedings of the 4th International Conference on Computational Linguistics in Bulgaria (CLIB 2020)*, P 63–71, Sofia, Bulgaria, September. Department of Computational Linguistics, IBL – BAS.

Shafiq Joty, Giuseppe Carenini, and Raymond T. Ng. 2015. CODRA: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, 41(3):385–435, September.

Shafiq Joty, Giuseppe Carenini, Raymond Ng, and Gabriel Murray. 2019. Discourse analysis and its applications. *// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, P 12–17, Florence, Italy, July. Association for Computational Linguistics.

William Mann and Sandra Thompson. 1988. Rethorical structure theory: Toward a functional theory of text organization. *Text*, 8:243–281, 01.

Lawrence M. Solan and Peter M. Tiersma. 2005. *Speaking of Crime. The Language of Criminal Justice*. Chicago Series in Law and Society. University of Chicago Press.

J. Svartvik and T.J. Evans. 1968. *The Evans Statements: A Case for Forensic Linguistics*. Acta Universitatis Gothoburgensis. University of Göteborg.

Incremental Topic Modeling for Scientific Trend Topics Extraction

Nikolai Gerasimenko
Sberbank, MSU Institute for Artificial Intelligence
nikgerasimenko@gmail.com

Alexander Chernyavskiy
National Research University Higher School of Economics
alschernyavskiy@gmail.com

Maria Nikiforova
Sberbank
labenzom@gmail.com

Anastasia Ianina
Moscow Institute of Physics and Techonology (MIPT)
yanina@phystech.edu

Konstantin Vorontsov
MSU Institute for Artificial Intelligence, MIPT
vokov@forecsys.ru

Abstract

Rapid growth of scientific publications and intensive emergence of new directions and approaches poses a challenge to the scientific community to identify trends in a timely and automatic manner. We denote trend as a semantically homogeneous theme that is characterized by a lexical kernel steadily evolving in time and a sharp, often exponential, increase in the number of publications. In this paper, we investigate recent topic modeling approaches to accurately extract trending topics at an early stage. In particular, we customize the standard ARTM-based approach and propose a novel incremental training technique which helps the model to operate on data in real-time. We further create the Artificial Intelligence Trends Dataset (AITD) that contains a collection of early-stage articles and a set of key collocations for each trend. The conducted experiments demonstrate that the suggested ARTM-based approach outperforms the classic PLSA, LDA models and a neural approach based on BERT representations. Our models and dataset are open for research purposes.

Keywords: Topic modeling, Trend Extraction, Additive Regularization of Topic Models, Incremental Topic Modeling.

DOI: 10.28995/2075-7182-2023-22-88-103

Инкрементальная тематическая модель для выделения научных тематических трендов

Аннотация

Быстрый рост количества научных публикаций и интенсивное внедрение новых направлений и подходов исследований значительно усложняет проблему автоматического выделения научных трендов. Мы определяем тренд как семантически однородную тему, которая характеризуется постепенно эволюционирующим лексическим ядром, а также резким, часто экспоненциальным, скачком количества публикаций в начале развития тренда. В этой статье мы применяем тематическое моделирование для выделения трендовых тем на раннем этапе их развития. Визоизменив стандартный подход APTM, мы создали новую технику инкрементального обучения тематических моделей, которая может дообучаться с использованием актуальных статей в режиме реального времени. Также мы представляем датасет трендов по искусственному интеллекту (Artificial Intelligence Trends Dataset, AITD), который содержит коллекцию статей и набор ключевых слов для каждого тренда. Проведенные эксперименты показывают, что предложенный подход на основе APTM превосходит классические алгоритмы (PLSA, LDA) и нейронные подходы на основе BERT. Наши модели и датасет доступны для исследовательских целей.

Ключевые слова: тематическое моделирование, выделение трендов, аддитивная регуляризация тематических моделей, инкрементальное тематическое моделирование.

1 Introduction

The rapid growth of scientific publications, journals, and conferences makes it effortful to reconstruct a complete purview of specific subject areas. Nowadays, people have to keep track of numerous emerging areas and domains, for which the global scientific importance is not always explicit at the first sight. In

this regard, more attention is paid to methods that solve the research trend identification task (Ho et al., 2014; Rotolo et al., 2015; Prabhakaran et al., 2016; Färber and Jatowt, 2019; Uban et al., 2021).

In this study, we consider the task of trend-like topic detection in real-time. The resulting topics should comply with the following conditions:

1. They should contain as many trending topics as possible. Here, we apply the definition of a trend proposed by (Kontostathis et al., 2004), where the emerging trend is defined as a topic, interest to which was strongly increasing in a particular time interval.
2. Trend-like topics should be identified as early as possible by the time they appear.
3. Each topic should be semantically homogeneous and impartible. This formulation imposes specific restrictions.

In our experiments, we extract trending publications in the field of Artificial Intelligence (AI), but the proposed approach can be applied to other scientific fields a well.

Let us consider an example with the ELECTRA model (Clark et al., 2020). It immediately aroused great interest among the scientific community and began to be actively used in various applications. Hence, more than 200 articles referring to it had already been published in 2020 alone. This certainly conforms to our definition of a trend, and our goal is to build a system that will also be able to highlight such trends as early as possible.

It should be emphasized that the trend can be not only a *model* but also a *task* (like the fact-checking task) or a *method* (like Dropout or AdamW). Moreover, we do not aim to utilize models only for retrospective analysis and highlight the main trends in the past. Thus, we set the problem statement in such a way that the system is allowed to make predictions into the future, that is, to distinguish research areas that are currently developing most actively.

In order for the final model to operate in real-time, we suggest incremental training. At each timestamp, we aim to generate new topics as distant as possible from existing ones, which is not implied a priori in some topic models. Further, many current topic modeling approaches have issues associated with the dilution of topics and terms, and the decorrelation of terms. To overcome these and other similar problems, we apply a topic model with additive regularization, namely ARTM (Vorontsov and Potapenko, 2015). Moreover, we offer several ways to customize it that contributes to achieve the best quality.

Despite active research in the field, there is no single quality metric for comparing trend detection models. Thus, we propose our intuitive metric in accordance with the assigned task.

Apart from that, we create a special expertly assembled dataset for comparison, which we issue in the public domain. We called it Artificial Intelligence Trends Dataset (AITD).

Our contributions can be summarized as follows:

- We propose the incremental mechanism of ARTM training to detect trend topics in real-time.
- We propose the novel ARTM-based approach that outperforms popular neural network and topic modeling approaches in the task of early trend detection.
- We create a specialized dataset to validate trend topic detection approaches, which we release for the research community.
- We make our approach and the created dataset open releasing code and the data there: `https://drive.google.com/file/d/1ueb9OgTdeITk0Cly7doKO4KwL7G_YjT_/view?usp=share_link`.

2 Related Work

Trend detection systems generally can be divided into two groups: semi-auto and auto approaches. We investigate only approaches that do not require human interaction.

Generally, automatic detection of trends involves two stages: topic detection (or identification) and topic evolution (with emerging trend classification). The first stage is needed to construct the set of topics from which the trends will be selected. The following types of approaches can be distinguished: statistical, knowledge-based, and hybrid. Statistical approaches use only the given textual context without any additional meta-information. Various models have been already investigated in this direction: topic modeling (Prabhakaran et al., 2016; Uban et al., 2021; Krivenko and Vasilyev, 2009), clustering approaches

(Mei and Zhai, 2005; Behpour et al., 2021), and so forth. Among the aforementioned models a sequential variant of LSI (Krivenko and Vasilyev, 2009) is the approach most similar to ours in terms of problem formulation. Apart from that, other models utilize information from knowledge bases like the web (Roy et al., 2002) or citation graphs (Erten et al., 2004; Chang and Blei, 2010). Hybrid approaches (Jo et al., 2007; He et al., 2009; Ma et al., 2010) combine term-based topic detection and co-citation/co-authorship graph analysis.

There also has been some research on neural approaches for topic modeling (e.g. Transformer-based) (Grootendorst, 2020; Angelov, 2020). However, these approaches are not directly applicable due to the specifics of our collection, namely the length of the full texts of articles. For instance, the BERT model (Devlin et al., 2018a) has a limit on the length of input sequences of 512 tokens. Thus, it is needed to use either aggregation or additional models for text summarization to construct embeddings of entire texts. Nonetheless, we use the BERTopic model (Grootendorst, 2020) for comparison and show its inefficiency compared to our approach.

Topic evolution is utilized to consider topic emergence in time. Here, some approaches use custom metrics based on the topic characteristics (Ho et al., 2014; Prabhakaran et al., 2016; Grosso et al., 2017; Färber and Jatowt, 2019; Behpour et al., 2021). Another category of approaches considers citations-based metrics. In this way, (Le et al., 2006) proposed to use various temporal citation-based features to evaluate the growth in interest and utility of topics over time. In this work, we do not investigate classification of topics into trends and non-trends and mainly focus on the first part of the trend extraction pipeline. However, experiments with the trend evolution analysis are a subject for the further research.

To track topic emergence in real time, we investigate incremental topic models. Some researchers suggested online techniques for LDA (Canini et al., 2009; Hoffman et al., 2010). Nevertheless, due to the qualitative limitations of LDA-based approaches which are confirmed by our experiments, we use the ARTM model (Vorontsov and Potapenko, 2015) and propose a method of its incremental training. Our incremental mechanism is based on trend keywords detection. Similar to our approach, (Färber and Jatowt, 2019) proposed a method to estimate the impact index of keywords but did not integrate it into the trend detection pipeline.

Later (Sivanandham et al., 2021) proposed a model for analyzing research trends using topic modeling (LDA) and vector auto regression. On the contrary, (Lee et al., 2021) applied language modeling (BERT) and t-SNE algorithm for future prediction of growth potential of technologies.

The most recent studies mostly focus on neural topic models bridging the gap between probabilistic dynamic topic models based on matrix factorization techniques and the power of large language models. For example, Aligned Neural Topic Model (ANTM) (Rahimi et al., 2023) uses document embeddings to compute clusters of semantically similar documents at different periods of time and then aligns document clusters to represent their evolution. ANTM outperforms models Dynamic Embedded Topic Models (Dieng et al., 2019; Dieng, 2020) and significantly improves topic coherence and diversity over other existing dynamic neural topic models (e.g. BERTopic (Grootendorst, 2020)).

Another interesting research direction that can be easily applied for trend extraction is Graph Neural Networks. Such approaches preserve document dynamics and network adjacency by saving document relatedness via graph edges. For example, (Liang et al., 2023) fuses the graph topology structure and the document embeddings, while (Zhang and Lauw, 2022) proposes two neural topic models aimed at learning unified topic distributions that incorporate both document dynamics and network structure.

3 Trend Topic Detection

3.1 Task Definition

We consider the task of trending topic detection in real-time. In order to experiment not only with models based on matrix factorization but also with other popular approaches (e.g. clustering-based), we suggest to reduce the topic detection task to a search problem. Broadly speaking, we have a query for each topic (a topic name), and the goal is to get relevant lists of terms and documents associated with it. In our case, the queries are hidden, but we can still solve the recommendation task for them. Thus, the system should return ranked lists of per-topic documents and words for each predefined timestamp.

3.2 Approach

To obtain real-time predictions and reduce training time, we suggest incremental training of the topic model. The model leverages an incremental approach to create new topics based on words and collocations appearing in the last time interval, which contributes to more accurate trend extraction. The incremental model solves two subtasks: choosing the number of new topics, initializing new topics and adjusting them later.

We chose ARTM (Vorontsov and Potapenko, 2015) as the base model since it allows to build multi-objective models adding multiple criteria in a form of regularizers.

Base Model The ARTM model, in contrast to the LDA model that considers only a Dirichlet regularizer, allows to regard nonstandard important regularizers: smoothing and thinning of distributions of terms and topics, decorrelating distributions of terms in topics. Thus, we chose it as the base model for our topic modeling approach.

Initialization Let D be a collection of documents and W be a dictionary of words. After a new collection of documents D' appears, the model considers a set of emerging words W' and updates current topics T by adding new topics T' to it.

Generally, topic modeling approaches operate with matrices Φ and Θ representing word-topic and topic-document distributions respectively.

We suggest an incremental update to each of them. So, we initialize the matrices Φ_{n+1} and Θ_{n+1} in the current step using the matrices Φ_n and Θ_n from the previous step. More specifically, we copy Φ_n to the $\{W \leftrightarrow T\}$ submatrix of the matrix Φ_{n+1}, and Θ_n — to the $\{T \leftrightarrow D\}$ submatrix of the matrix Θ_{n+1}. All other values are filled according to the uniform distribution.

Number of New Topics The number of new topics for updating can be chosen in various ways (based on new documents collection, new terms or some combination of them). Here, we consider two of them: (*i*) a base straightforward approach that adds a fixed number of topics, (*ii*) a customized approach based on the emerging trend vocabulary V that is constructed based on impact scores similar to scores from (Färber and Jatowt, 2019).

In the base approach, we firstly count the mean value of terms related to each topic. This can be done by training one topic model for the first timestamp. Next, a new topic is created when the corresponding number of new terms appears in the vocabulary of key terms. This is because we are changing the current vocabulary to maintain a fixed size. Thus, some of collocations removed or added over time.

In the custom approach, the emerging trend vocabulary V consists of terms that have become much more commonly used compared to the moment of the last update of the topic model.

Let $w \in W \cup W'$ be a word from the current corpus. At the current timestamp, this word is added to V if it appears in at least $mindf$ documents and it satisfies the trend condition:

$$\frac{\text{tf}_{\text{new}} - \text{tf}_{\text{old}}}{\text{tf}_{\text{old}}} > \alpha \tag{1}$$

Here, tf_{old} is the count of the occurrence of w in documents D, and tf_{new} is the count of the occurrence of w in $D \cup D'$. $\alpha \in (0, 1)$ is a regulation hyper-parameter that sets the degree of increase in the occurrence of words to classify them as trending.

$$|T'| = |T_{\text{start}}| + \left\lfloor \frac{|V|}{\beta} \right\rfloor \tag{2}$$

In (2), T_{start} determines the number of topics at the initial timestamp, $\beta \in \mathbb{N}$ limits the number of added topics, and $\lfloor \cdot \rfloor$ denotes an integer part.

In other works the strategy of choosing number of topics in topic models include approaches based on simple heuristics and grid search (Ianina and Vorontsov, 2019; Ianina and Vorontsov, 2020), minimax optimal guarantees (Bing et al., 2020) or Bayesian approach and GNNs (Loureiro et al., 2023). Detailed comparison between methods of choosing the right number of topics is beyond the scope of this work.

Training Document Collections The result is also affected by the set of documents used for the model retraining at each timestamp for update: there are options to take either all documents in the history, or only new ones, or some intermediate option (with overlapping).

Formally, we have several options for the training document collection $\hat{D}$ at each step t:

$$\hat{D} = \begin{cases} D_t \\ D_t \setminus D_{t-1} \\ D_t \setminus D_{t-k} \quad \text{for } 1 < k < t \end{cases} \tag{3}$$

All these options affect the training time, and the second allows us to learn in real-time. In our experiments, we analyze these options in terms of quality and efficiency for our task.

Topic Models as Recommendation Systems To solve the recommendation task, we leverage probability scores from Φ and Θ to rank documents and words in the most appropriate way for each topic. A higher probability indicates that the model considers the document or word to be more important.

3.3 Evaluation

We propose a matching stage to map the labeled trends to the detected topics. At each iteration of the additional training of the incremental model, the search for the best topic for each trend is performed as follows.

Let D_{trend} and W_{trend} be the labeled sets of documents and words associated with the given trend respectively. Apart from that, we consider "golden" set of topic names S_{trend} . Here, S_{trend} contains from one to three synonymous collocations, each of which can be used as the trend name. At the output stage of the model each topic is represented by two ranked lists denoted as D_{topic} and W_{topic}. Also, we define $S_{\text{topic}} := W_{\text{topic}}$.

To perform matching, we firstly calculate three Recall@k based metrics:

$$\text{XRecall@k} = \frac{|X_{\text{topic}}[:\text{k}] \cap X_{\text{trend}}|}{k} \tag{4}$$

Here, $X[:\text{m}]$ denotes first m elements of the list X, where X is W, D or S respectively. We use three different values of the parameter k for documents, words and topic names, which are denoted as k_D, k_W and $k_S \leq k_W$ respectively.

We combine DRecall@k, WRecall@k and SRecall@k scores to estimate the relevance of the selected topic to the selected trend. We consider the trend to be detected once it has been matched with one of the extracted topics.

Since our goal is to minimize time delay for the trend detection, the final quality metric is the average number of days (or timestamps) that elapsed from the inception of a trend to its detection by the model. In our case, the inception date is the date of the earliest publication from the dataset.

4 Dataset

4.1 Background

To validate topic models, we collected a dataset of scientific trends. The closest work to us is the TRENDNERT benchmark proposed by (Moiseeva and Schütze, 2020), where the first public baseline for detecting (down)trends was presented. The dataset was constructed from a subset of papers published from 2000 to 2015.

Despite the large volume, the TRENDNERT benchmark has several drawbacks. Firstly, due to the fact that stratification was used for documents selection, some key papers that had a high impact on the trends at the beginning of their evolution could be lost. Secondly, the trends presented in this benchmark can be obtained by mapping internal identifiers proposed by the authors of the paper to an identifier from the Semantic Scholar database. However, we found this mapping outdated and results cannot be 100% replicated.

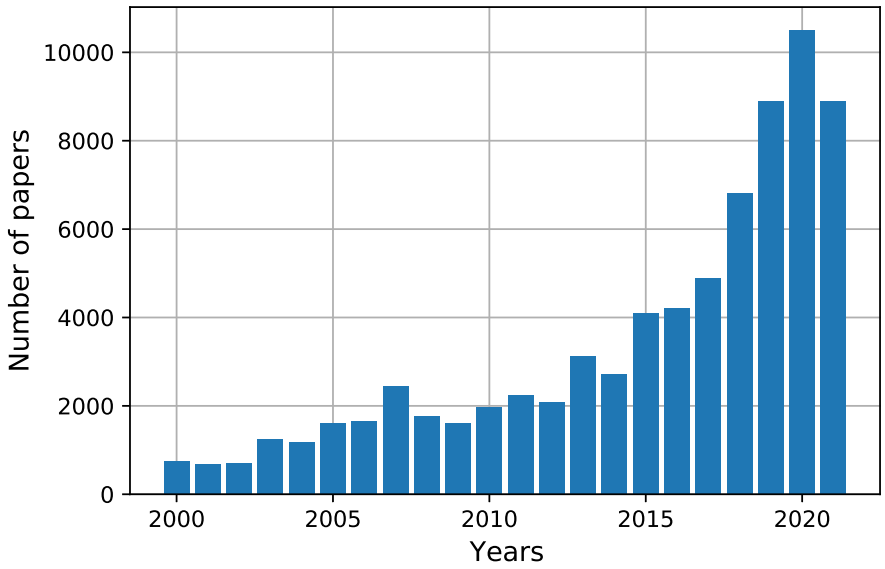


Figure 1: Papers distribution from selected conferences.

To overcome the drawbacks listed above, we present a new dataset, namely `AITD`. It focuses on trends in the Artificial Intelligence field across 2009-2021 years.

4.2 Data Sources

We used the part of Semantic Scholar Open Research Corpus as the main source of scientific publications, namely the Computer Science section (∼18M articles) with publications from 2000 to 2021. To filter the dataset, we considered only publications from 11 conferences that were selected based on data of top venues of Google Scholar[1] (Artificial Intelligence, Computational Linguistic, Computer Vision & Pattern Recognition sections were chosen) and h-index exceeding 100. Further, we filtered publications that did not contain any information about the corresponding conference name or the year of publication.

Figure 1 demonstrates the number of papers published by years from 2000 to 2021. It can be seen that almost every year the number of papers increases, and most of them were published relatively recently.

Final list of conferences and number of papers from each of it presented in Table 1. It demonstrates that the most of papers were presented on the computer vision CVPR conference. Apart from that, most of the dataset (more than 40%) is made up of articles from general conferences: NeurIPS, AAAI, IJCAI and ICML.

We enriched our dataset by adding information from the arXiv dataset[2], and updated years for some publications. Thus, we solve the problem of data leakage for trend detection. That is, we exclude the situation when the article was first published on the arXiv site and became available to the scientific community and only after some time appeared in the proceedings of some conference.

Eventually, our dataset contains the following attributes: the paper id on Semantic Scholar, the title, authors' ids, venue, ids of publications it refers to, ids of papers that refer to it, the date of publication on arXiv, and the date of the conference. For the dataset construction, we utilized SciPDF Parser4 and PyMuPDF5 to extract text layer from the downloaded PDF files. We extracted collocations using the TopMine (El-Kishky et al., 2014) algorithm. To avoid "looking into the future" data leaks, the dataset was divided into subsets by two-week time intervals from March 2000 to December 2021.

[1] `http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=eng`
[2] `https://www.kaggle.com/Cornell-University/arxiv`

Conference	Number of papers	% in the dataset
CVPR	11547	15.61
NeurIPS	11423	15.45
AAAI	9408	12.72
IJCAI	7523	10.17
ICML	7143	9.66
ACL	6752	9.13
ICCV	4976	6.73
EMNLP	4855	6.56
ECCV	4030	5.45
NAACL	3192	4.32
ICLR	3110	4.21

Table 1: Distribution of papers in dataset per conference.

4.3 Labeling

Trends Generation To prepare the validation dataset, we used the reference graph from the Semantic Scholar dataset. Initially, we generated manually 91 trends (for “model”, “method”, and “task” types) in the field of Machine Learning and Artificial Intelligence (e.g. CNN, RNN, BERT). For each of them, we found a paper with which this trend began or revived, as we call it “first story”. This concept can be illustrated with Fig.**??**: the evolution of each trend is described with a number of relevant papers being published at the selected period of time. There may be clear upword trends (e.g. "CNN" in Fig. 2) or trends with more complicated evolutional paths (e.g. "PCA" in Fig. 3)

Further, for each trend, we expertly selected at least 10 relevant publications based on the citation graph and used collocations. For the chosen papers, we analyzed the most frequent collocations and selected only those that are directly related to the topic of the trend (more than 5 keywords per trend).

To this end, we firstly collected a list of machine learning concepts frequently mentioned in scientific publications. We considered the exponential growth of mentions from some point in time as the condition for a topic to be a trend. The year starting from which the mentions rapidly grew was considered as trend start date. The dataset is not balanced and the maximum number of trends (more then 17) appeared in 2015. The distribution has light tails with relatively few trends (less than 5).

Papers and Keywords Selection After the first paper of the trend is found (the paper that created or re-invented the trend topic), we select related articles for each trend. The following conditions were used: (1) the selected articles should be directly related to the trend; (2) the articles should be published no later than two years after the first paper of the trend. For each trend, at least 10 articles that satisfy the conditions were selected. Further, collocations were selected from those papers to create keyword lists (at least 10 keywords for each trend).

Trend Names Labeling The last step was to choose alternative names for the trends based on the general knowledge or keywords. All the names were selected from the fixed collocations vocabulary.

Final Dataset Thus, we collected the dataset with the following structure: trend name, a subset of papers related to the trend, trend keywords, possible trend names. During the dataset construction, we utilized SciPDF Parser[3] and PyMuPDF[4] to extract text layer from downloaded PDF files. Unparsed articles are not further considered.

[3] `https://github.com/titipata/scipdf_parser`
[4] `https://github.com/pymupdf/PyMuPDF`

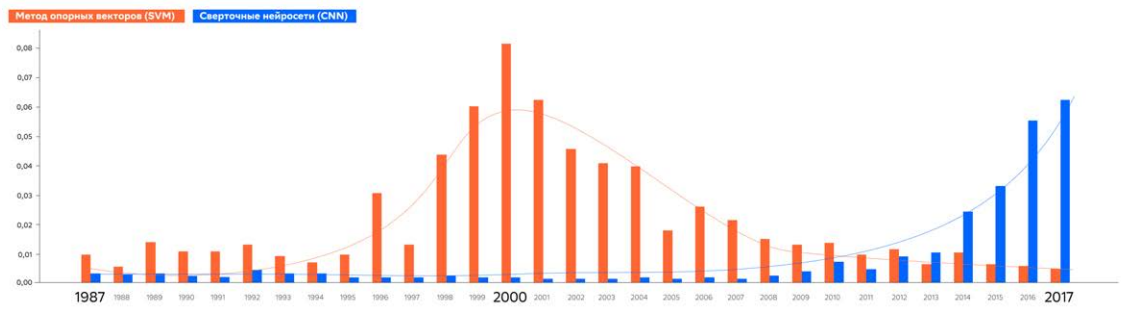


Figure 2: The example of two trends (CNN and SVM) evolving in time.

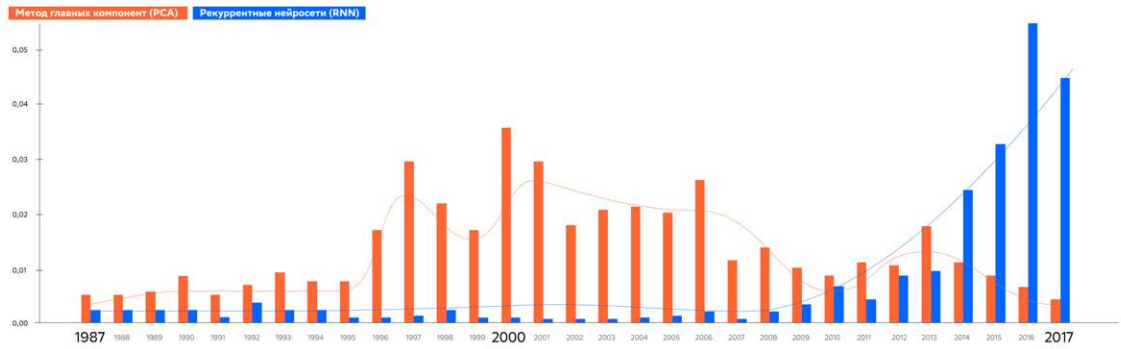


Figure 3: The example of two trends (PCA and RNN) evolving in time.

We extracted collocations using the TopMine (El-Kishky et al., 2014) algorithm. To avoid looking into the future, the dataset was divided into subsets by two-week time intervals from March 2000 to December 2021. We passed these batches to TopMine to extract collocations with maximal length of 5 words.

The dataset is publicly available here: `https://drive.google.com/file/d/1ueb9OgTdeITkOCly7doKO4KwL7G_YjT_/view`.

5 Experiments

5.1 Implementation Details

We trained our model on a sequence of time periods. We chose these periods in such a way that each period was at least two weeks long and contained at least 1000 published documents. As a result, we obtained 82 periods for training.

The open-source BigARTM library (Vorontsov et al., 2015) was used to train PLSA (Hoffman, 1999), LDA (Blei et al., 2003) and ARTM (Vorontsov and Potapenko, 2015) models. For ARTM, we used the regularizer named Decorrelator Φ that contributes to the decorrelation of columns in the Φ matrix. The regularization coefficient was set to 0.2. We also used the SmoothSparse Θ regularizer for which regularization coefficient was set to -1.

5.2 Models

Code for experiments was written on Python 3.

We conducted our experiments for sequence of timestamps, updating every 2 weeks, if at least 1000 new documents had been published in this period. For our dataset we got 82 timestamps and for each of them the batch of documents was created.

The open-source BigARTM library (Vorontsov et al., 2015) was used to train PLSA, LDA and ARTM models. In the case of the ARTM model, we used the regularizer named Decorrelator Φ that contributes to the decorrelation of columns in the Φ matrix. The regularization coefficient was set to 0.2. We also used the SmoothSparse Θ regularizer and regularization coefficient was set to -1.

In the process of the incremental learning when the sparsity of matrix Φ achieved 0.9 the Decorrelator Φ turned off. Similarly, when the sparsity of matrix Θ achieved 0.9 SmoothSparse Θ turned off. If sparsity drops below 0.9, then regularizers turn back on. We also used the same procedure with LDA model, because Dirichlet Regularizers had poor sparse effect on Φ and Θ matrices.

In the incremental learning process we used early stopping criteria. If within three passes the topic models perplection changes by less than 5% over subcollection of current incremental steps, then the learning process ends and the model proceeds to a new incremental step. Also model goes to the next incremental step upon reaching 24 collection passes on a incremental steps subset.

5.3 Baselines

We consider several baselines to compare our solution with.

Probabilistic Latent Semantic Allocation (PLSA) PLSA (Hoffman, 1999) is historically the first probabilistic topic model. Within PLSA one finds an approximate representation of counter matrix $F = (\frac{n_{dw}}{n_d})_{W \times D}$ (n_{dw} and n_d are counters of occurrences of term w in document d and overall number of terms in document d respectively) into a product of two unknown matrices — matrix Φ of term probabilities for the topics and matrix Θ of topic probabilities for the documents. In ARTM formulation PLSA corresponds to the model with no regularizers.

For consistency, in our experiments we used implementation of PLSA from BigARTM library (Vorontsov et al., 2015). The number of topics was chosen to be 200.

Latent Dirichlet Allocation (LDA) LDA (Blei et al., 2003) is a three-level hierarchical Bayesian model, in which documents are represented as random mixtures over latent topics, where each topic is characterized by distribution over words. Following the formulation of the problem from PLSA, in LDA parameters Φ and Θ are constrained by an assumption that vectors ϕ_t and θ_d are drawn from Dirichlet distributions with hyperparameters $\beta = (\beta_w)_{w \in W}$ and $\alpha = (\alpha_t)_{t \in T}$ respectively. In ARTM formulation LDA corresponds to the model with two regularizers that force an assumption that Φ and Θ columns are generated from Dirichlet distribution with hyperparameter β and α respectively. Following formulas represent corresponding regularizers within LDA model:

$$R(\Phi) = \sum_{t \in T} \sum_{w \in W} (\beta - 1) \ln \phi_{wt} \to \max$$

$$R(\Theta) = \sum_{d \in D} \sum_{t \in T} (\alpha - 1) \ln \theta_{td} \to \max$$

We used LDA implementation from BigARTM library (Vorontsov et al., 2015). Hyperparameters for LDA model were set to default values for symmetric Dirichlet distribution: $\alpha = \frac{1}{|T|}$, $\beta = \frac{1}{|T|}$, whete $|T|$ is the number of topics.

ARTM with decorrelation regularizer Another baseline is ARTM model with just one regulizer: decorrelation of matrix Φ. It is used to determine the lexical kernel of each topic which distinguishes it from the other topics. It minimizes covariations between columns of the Φ matrix:

$$R(\Phi) = \tau \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max$$

In our experiments coefficient of regularization τ is equal to 0.2.

Statistic	*Config1*				*Config2*				*Config3*			
	PLSA	**LDA**	**BERTopic**	**ARTM**	**PLSA**	**LDA**	**BERTopic**	**ARTM**	**PLSA**	**LDA**	**BERTopic**	**ARTM**
mean	295	268	**76**	181	526	519	685	586	731	761	608	**538**
min	1	1	**0**	**0**	**4**	**4**	10	**4**	**4**	11	10	11
25%	45	45	**14**	22	38	**23**	176	52	190	152	176	**110**
50%	126	114	**45**	56	443	**361**	484	476	556	504	**420**	479
75%	282	249	**95**	120	847	**827**	966	867	1074	1156	989	**761**
max	2907	2659	**871**	3433	**1921**	2273	2319	2711	2907	2659	2131	**1949**
# extracted	70	76	**90**	74	51	**53**	36	**53**	34	**39**	28	30

Table 2: Statistics of delays in days: max, mean and percentiles over all extracted trend topics for the considered approaches and matching configurations. *Config1* matches trends based on documents only (DRecall@k > 0.1); *Config2* matches trends based on keywords only (WRecall@k > 0.3 and SRecall@k > 0); *Config3* is a joint option (DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0).

BERTopic Apart from probabilistic topic models, we compared our solution to a neural-based model called BERTopic (Grootendorst, 2020) that leverages the token embeddings retrieved from BERT model (Devlin et al., 2018b). BERTopic is a topic modeling technique that uses transformers and c-TF-IDF to create dense topical clusters. First, BERTopic transforms document into embeddings. BERTopic supports many embedding models, including ones from Sentence-Transformers, Flair, Spacy, Gensim, USE. We used sentence-transformers package to get document-level embeddings. Second, BERTopic performs dimensionality reduction on the embeddings as a preparation step for clustering. Specifically, it uses UMAP (McInnes et al., 2018) as it keeps a significant portion of the high-dimensional local structure in lower dimensionality. Third, BERTopic clusters the documents with HDBSCAN (McInnes et al., 2017). Having the topical clusters, one may want to get the tokens of most importance from each cluster. Class-based TF-IDF (c-TF-IDF) is used to solve this. c-TF-IDF treats all documents in a topic as a single document and then applies TF-IDF, so that resulting TF-IDF scores demonstrate the important words in a topic.

Although BERTopic supports dynamic topic modeling, it did not fit to our purposes at all. First, BERTopic DTM creates a general topic model as if there were no temporal aspect in the documents. Then for each topic and timestep, it calculates the c-TF-IDF representation, resulting in different formulations of the same topics at different timesteps. To detect and track how new topics emerge, we trained 82 separate models, one for each timestamp respectively.

5.4 Comparison with the Baselines

We compare our solution to the aforementioned baselines using the base elements of the approach: a basic way of choosing the number of new topics and using the full history of documents for training at each step. We matched the extracted topics with the labeled trend topics using several metrics based on DRecall@k, WRecall@k and SRecall@k scores described in Section 3.3.

Three combinations of thresholds were used for matching at each timestamp:

- *Config1*: DRecall@k > 0.1, matches trends based on documents only;
- *Config2*: WRecall@k > 0.3 and SRecall@k > 0 matches trends based on keywords only;
- *Config3*: DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0 (joint option).

Table 2 shows the calculated statistics for the day delay metric. It can be seen that BERTopic model achieves the best scores for *Config1*, extracting almost all the trends in this configuration: 90 out of 91. This is due to the fact that this model has a larger number of topics and it is able to successfully distinguish documents among them. However, BERTopic is very bad at keywords extraction, since this is not its primary purpose. Therefore, for the other two configurations, its quality is much worse.

If we compare only topic models, then there is no single approach that stands out. From the table, we can conclude that PLSA is not the best choice for our task. The LDA model seems more apposite for

Config2, but ARTM is better in terms of *Config1* and *Config3*. The ARTM model generates the correct topics quickly enough even with the rigid configuration *Config3* compared to other topic models, although it can sometimes extract fewer trends in total. In the configuration *Config1*, when the main goal is to correctly divide documents by topics, ARTM extracts almost half of the trends in the first two months. Thus, it is well suited for qualitative identification of trends in the problem of early detection.

It is also worth noting that the BERTopic and ARTM models are able to extract a trend right at the moment of its inception for *Config1* (zero values for the "min" row). This is due to the fact that some of the trends are tasks in which there is no clear first paper.

To analyze the evolution of the quality metric depending on time, we explored the dependence of the proportion of detected trends from the time elapsed since their inception.

Figure 4 demonstrates the corresponding results for *Config1*. It can be used to rank models by quality in terms of document evaluation. In this case, the BERTopic model is superior to others at each timestamp, while PLSA is inferior to others. Further, ARTM is better than LDA because it extracts trends faster, although it compares later in total.

Figure 5 shows similar results for *Config2* and analyzes the quality in terms of the ranked keywords. In this case, as it was shown earlier, the BERTopic model performs much worse than the aforementioned topic models and extracts much fewer trends at any given timestamp. The quality for topic models increases approximately to the same extent. The LDA model has a slight advantage in the first months, but after a year and a half, the PLSA and ARTM models occasionally overtake it. These conclusions are consistent with Table 2.

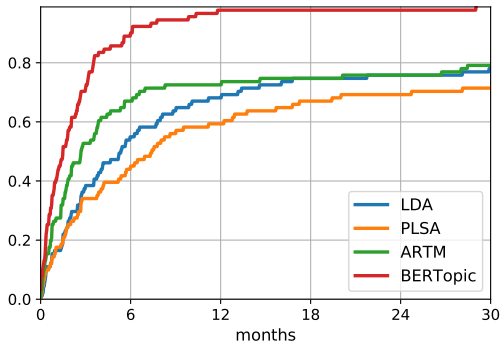


Figure 4: The dependence of the proportion of extracted trends on the months since their inception for *Config1* (DRecall@k > 0.1).

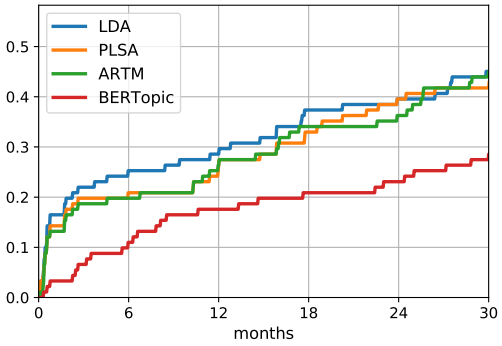


Figure 5: The dependence of the proportion of extracted trends on the months since their inception for *Config2* (WRecall@k > 0.3 and SRecall@k > 0).

In general, LDA was expected to perform better in some cases (e.g. *Config2*) because it considers the sparsity of the matrix Θ. Thus, we conducted experiments to integrate this into the model as one way of the base model modification.

Further, it should be emphasized that topic models require much less training time compared to BERTopic even though the latter is trained on GPU.

We tried to analyze why models extract some trends too late (after more than 2000 days) or not at all in some cases. Generally, the quality is limited by several factors: the sizes of topics and their presence in the validation dataset (for instance, "EM-algorithm" and "pattern recognition" present quite weakly); the occurrence of keywords in articles (the keyword "GPT" usually appears in a paper only a couple of times); the quality of the dataset and internal components of the approach (e.g. the matching procedure).

5.5 Approach Customization

Incremental Dataset In our approach, we have several options for choosing a dataset for retraining at each step. Experiments were conducted for two possible extremes (the first two options from 3): training

Statistic	*Config1*		*Config2*		*Config3*	
	B	***I***	***B***	***I***	***B***	***I***
mean	181	**67**	586	**576**	548	**498**
min	0	0	4	4	11	**4**
25%	22	**16**	**52**	214	110	**79**
50%	56	**41**	476	**452**	479	**443**
75%	120	**81**	867	**841**	**761**	793
max	3433	**514**	2711	**1921**	1949	1949
# extracted	74	**85**	53	**58**	30	**33**

Table 3: Statistics of delays in days for incremental and non-incremental dataset options. *B* denotes the base non-incremental approach (ARTM) and *I* denotes the incremental one (ARTMi). *Config1* matches trends based on documents only (DRecall@k > 0.1); *Config2* matches trends based on keywords only (WRecall@k > 0.3 and SRecall@k > 0); *Config3* is a joint option (DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0).

on the whole document history and training on the new ones only (incrementally).

Table 3 demonstrates results for ARTM. We denoted the incrementally trained approach as ARTMi. Firstly, ARTMi extracts more trends in total than ARTM in all matching configurations. For *Config1*, it is significantly superior to the LDA model and close to BERTopic. At the same time, statistics on the delay in days for it is also less than for ARTM almost in all cases. For instance, the number of days required for trend detection has decreased by almost 10 percent compared to the base model in the configuration *Config3*.

Algorithm Complexity ARTMi as well as ARTM is trained on CPU. We used Intel(R) Xeon(R) Gold 6348 CPU (24-cores) to train both ARTM and ARTMi. ARTMi can be trained in approximately 40 minutes using 16 cores in parallel. We also compared the training time for ARTM and ARTMi and found that the ARTMi model can be trained about 50 times faster. This result can be also confirmed analytically. The models take 99 batches of approximately the same size as an input. Each model runs an average of 5 times for each batch. Thus, we get $5 \cdot 99 = 495$ passes for ARTMi. The ARTM model overlaps over all previous batches at each new step. Thus, we get $5 \cdot \sum_{n=18}^{99} n = 5 \cdot 4797$ passes for ARTM, that is, 50 times more. Thus, training on an incremental dataset helps ARTMi to extract more trends in total. ARTMi is much faster than ARTM and can be effectively applied in real time. We use the incremental dataset in all the further modifications.

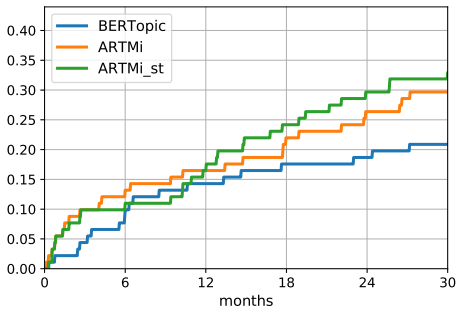


Figure 6: The dependence of the proportion of extracted trends on the months since their inception for *Config3* (DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0).

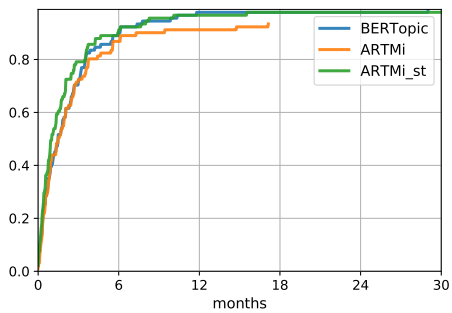


Figure 7: The dependence of the proportion of extracted trends on the months since their inception for *Config1* (DRecall@k > 0.1).

Statistic	*Config1*		*Config2*		*Config3*	
	B	***C***	***B***	***C***	***B***	***C***
mean	68	**65**	**584**	604	**554**	626
min	0	0	9	9	**9**	17
25%	12	**11**	**302**	303	256	**186**
50%	30	30	**510**	533	475	**457**
75%	74	**62**	906	**900**	**906**	912
max	949	**942**	1921	1921	**1921**	2187
# extracted	90	90	56	**57**	**40**	39

Table 4: Statistics of delays in days for two options of the number of new topics selection: *B* denotes the base approach and *C* – customized. *Config1* matches trends based on documents only (DRecall@k > 0.1); *Config2* matches trends based on keywords only (WRecall@k > 0.3 and SRecall@k > 0); *Config3* is a joint option (DRecall@k > 0.1 , WRecall@k > 0.3 and SRecall@k > 0).

Sparsity of Matrix Θ As described in Section 5.4, we have added the Θ matrix sparsity regularizer to the standard ARTM model. We denoted this model as ARTMi_st. BERTopic was also used for comparison since it: *(1)* is different from the topic models in substance and does not have any regularizations; *(2)* obtained the best results for *Config1*.

Figure 6 shows the dependence of the proportion of extracted trends on the months since their inception for the balanced configuration *Config3*. It can be seen that ARTMi_st outperforms both BERTopic and ARTMi almost for all the timestamps. It is able to extract more trends more quickly even for complex matching, evaluating both documents and keywords. In total, ARTMi_st extracted 40 trends, whereas ARTM and BERTopic – only 33 and 28 respectively.

Moreover, in the first months, the ARTMi_st model immediately overtakes BERTopic (Fig. 7, *Config1*), despite the fact that the latter outperformed the base models by a large margin. Therefore, even for distinguishing documents by topics, the topic model with decorrelation and regularization performs better than the neural BERT-based approach. Thus, adding Θ sparsity regularizer is one of the crucial components of the topic model to achieve the best quality.

Number of New Topics We experimented with two ways of the number of new topics selection (described in Section 3.2) for the ARTMi_st model. The results are demonstrated in Table 4. For *Config1*, the customized option is better than the base one. It extracts the same amount of trends in total, but it does so earlier in time. For the matching configurations associated with the presence of the correct keywords, the results are about the same as in the base case. The number of extracted topics differs by one, and the difference between delays in days is not statistically significant. Thus, the customized way of choosing the number of new topics improves the quality for some matching configurations, but it does not provide significant advantages for others.

6 Future Work

Firstly, we highlight a direction related to the trend identification subtask. We are going to leverage the document-topic distribution matrix to construct trend profiles in time. Such profiles will allow us to track the evolution of topics over time and, in case of exponential growth, serve as one of the trend indicators. Secondly, we are going to analyze and visualize the current results of the early trend detection. Besides, we plan to apply the proposed approach to other scientific areas except for machine learning and AI.

Possible applications of the proposed technology include news monitoring and extraction of the most relevant trends in different domains, assistance with scientific research (e.g. automatic tracking of emerging topics of interest) and help with literature review composing. Furthermore, such a technique may appear useful not only in scientific or news monitoring areas, but also in corporate segment for structuring and analysing large piles of legal documentation and technical requirements.

7 Conclusion

In this paper, we investigated the topic modeling approaches to the scientific trend topics detection task. The main goal was to make predictions in real-time. To this end, we customized the standard ARTM-based approach and proposed incremental training consisting of incremental initialization, incremental dataset and the number of topics updating based on the current vocabulary of trend collocations. Apart from that, we integrated sparsity regularization into our approach which increased the model quality. Our method is universal and is not model-specific.

We described the validation process and proposed a method for matching labeled trends and extracted topics. We collected the expertly labeled specialized dataset, namely AITD, to validate approaches solving early trend topic detection task. The dataset consists of 91 groups of machine learning and AI articles (each group corresponds to one trend topic) with corresponding keywords selected from publications from top conferences and alternative trend names.

The evaluation demonstrated that the basic ARTM model achieves one of the best results compared to the other baselines using different matching configurations. Moreover, incremental training techniques and additional regularization led to a significant improve in the base model quality regarding early trend detection. The final ARTM-based approach extracts the largest number of trends at the early stages of their evolution, and can operate in real-time since it requires the least training time.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Sahar Behpour, Mohammadmahdi Mohammadi, Mark V. Albert, Zinat S. Alam, Lingling Wang, and Ting Xiao. 2021. Automatic trend detection: Time-biased document clustering. *Knowledge-Based Systems*, 220:106907.

Xin Bing, Florentina Bunea, and Marten Wegkamp. 2020. A fast algorithm with minimax optimal guarantees for topic models with an unknown number of topics.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Kevin Canini, Lei Shi, and Thomas Griffiths. 2009. Online inference of topics with latent dirichlet allocation. *Journal of Machine Learning Research - Proceedings Track*, 5:65–72, 01.

Jonathan Chang and David M. Blei. 2010. Hierarchical relational models for document networks. *The Annals of Applied Statistics*, 4(1), Mar.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *ArXiv*, abs/2003.10555.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Adji Bousso Dieng. 2020. *Deep Probabilistic Graphical Modeling*. Columbia University.

Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R. Voss, and Jiawei Han. 2014. Scalable topical phrase mining from text corpora. *ArXiv*, abs/1406.6312.

C. Erten, P. J. Harding, S. G. Kobourov, K. Wampler, and G. Yee. 2004. Exploring the computing literature using temporal graph visualization. *Proceedings of SPIE - The International Society for Optical Engineering*, 5295:45–56. Visualization and Data Analysis 2004 ; Conference date: 19-01-2004 Through 20-01-2004.

Michael Färber and Adam Jatowt. 2019. Finding temporal trends of scientific concepts. // *BIR@ ECIR*, P 132–139.

Maarten Grootendorst. 2020. Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.

Minor Eduardo Quesada Grosso, Edgar Casasola, and Jorge Antonio Leoni de León. 2017. Trending topic extraction using topic models and biterm discrimination. *CLEI Electron. J.*, 20(1):3:1–3:13.

Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. 2009. Detecting topic evolution in scientific literature: How can citations help? // *ACM 18th International Conference on Information and Knowledge Management, CIKM 2009*, International Conference on Information and Knowledge Management, Proceedings, P 957–966. ACM 18th International Conference on Information and Knowledge Management, CIKM 2009 ; Conference date: 02-11-2009 Through 06-11-2009.

Jonathan C. Ho, Ewe-Chai Saw, Louis Y.Y. Lu, and John S. Liu. 2014. Technological barriers and research trends in fuel cell technologies: A citation network analysis. *Technological Forecasting and Social Change*, 82(C):66–79.

Matthew Hoffman, David Blei, and Francis Bach. 2010. Online learning for latent dirichlet allocation. volume 23, P 856–864, 11.

Thomas Hoffman. 1999. Probabilistic latent semantic indexing. // *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval, 1999*, P 50–57.

Anastasia Ianina and Konstantin Vorontsov. 2019. Regularized multimodal hierarchical topic model for document-by-document exploratory search. // *2019 25th Conference of Open Innovations Association (FRUCT)*, P 131–138. IEEE.

Anastasia Ianina and Konstantin Vorontsov. 2020. Hierarchical interpretable topical embeddings for exploratory search and real-time document tracking. *International Journal of Embedded and Real-Time Communication Systems (IJERTCS)*, 11(4):134–152.

Yookyung Jo, Carl Lagoze, and C. Lee Giles. 2007. Detecting research topics via the correlation between graphs and texts. // *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '07, P 370–379, New York, NY, USA. Association for Computing Machinery.

April Kontostathis, Leon M. Galitsky, William M. Pottenger, Soma Roy, and Daniel J. Phelps, 2004. *A Survey of Emerging Trend Detection in Textual Data Mining*, P 185–224. Springer New York, New York, NY.

Mikhail Krivenko and Vitaly Vasilyev. 2009. Sequential latent semantic indexing. // *Proceedings of the 2nd Workshop on Data Mining using Matrices and Tensors*, P 1–9.

Minh-Hoang Le, Tu Bao Ho, and Yoshiteru Nakamori. 2006. Detecting emerging trends from scientific corpora.

June Young Lee, Sejung Ahn, and Dohyun Kim. 2021. Deep learning-based prediction of future growth potential of technologies. *Plos one*, 16(6):e0252753.

Dingge Liang, Marco Corneli, Charles Bouveyron, and Pierre Latouche. 2023. The graph embedded topic model.

Manuel V Loureiro, Steven Derby, and Tri Kurniawan Wijaya. 2023. Topics as entity clusters: Entity-based topics from language models and graph neural networks. *arXiv preprint arXiv:2301.02458*.

Huifang Ma, Zhixin Li, and Zhongzhi Shi. 2010. Combining the missing link: An incremental topic model of document content and hyperlink. // Zhongzhi Shi, Sunil Vadera, Agnar Aamodt, and David B. Leake, *Intelligent Information Processing V - 6th IFIP TC 12 International Conference, IIP 2010, Manchester, UK, October 13-16, 2010. Proceedings*, volume 340 of *IFIP Advances in Information and Communication Technology*, P 259–270. Springer.

Leland McInnes, John Healy, and Steve Astels. 2017. hdbscan: Hierarchical density based clustering. *Journal of Open Source Software*, 2(11):205.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Qiaozhu Mei and ChengXiang Zhai. 2005. Discovering evolutionary theme patterns from text - an exploration of temporal text mining. P 198–207, 01.

Alena Moiseeva and Hinrich Schütze. 2020. Trendnert: A benchmark for trend and downtrend detection in a scientific domain. // *AAAI*.

Vinodkumar Prabhakaran, William L Hamilton, Dan McFarland, and Dan Jurafsky. 2016. Predicting the rise and fall of scientific topics from trends in their rhetorical framing. // *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1170–1180.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2023. Antm: An aligned neural topic model for exploring evolving topics. *arXiv preprint arXiv:2302.01501*.

Daniele Rotolo, Diana Hicks, and Ben R. Martin. 2015. What is an emerging technology? *Research Policy*, 44(10):1827–1843.

Soma Roy, David Gevry, and William Pottenger. 2002. Methodologies for trend detection in textual data mining. 2, 10.

S Sivanandham, A Sathish Kumar, R Pradeep, and Rajeswari Sridhar. 2021. Analysing research trends using topic modelling and trend prediction. // *Soft Computing and Signal Processing: Proceedings of 3rd ICSCSP 2020, Volume 1*, P 157–166. Springer.

Ana Sabina Uban, Cornelia Caragea, and Liviu P Dinu. 2021. Studying the evolution of scientific topics and their relationships. // *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, P 1908–1922.

Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Machine Learning*, 101(1):303–323.

Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. 2015. Bigartm: Open source library for regularized multimodal topic modeling of large collections. // *International Conference on Analysis of Images, Social Networks and Texts*, P 370–381. Springer.

Delvin Ce Zhang and Hady Lauw. 2022. Dynamic topic models for temporal document networks. // *International Conference on Machine Learning*, P 26281–26292. PMLR.

Fine-tuning Text Classification Models for Named Entity Oriented Sentiment Analysis of Russian Texts

Anna Glazkova
University of Tyumen
Tyumen, Russia
a.v.glazkova@utmn.ru

Abstract

The paper presents an approach to named entity oriented sentiment analysis of Russian news texts proposed during the RuSentNE evaluation. The approach is based on RuRoBERTa-large, a pre-trained RoBERTa model for Russian. We compared several types of entity representation in the input text, and evaluated strategies for handling class imbalance and resampling entity tags in the training set. We demonstrated that some strategies improve the results of pre-trained models obtained on the dataset presented by the organizers of the evaluation.

Keywords: targeted sentiment analysis, named entities, named entity oriented sentiment analysis, text classification, RuSentNE, RuRoBERTa.

DOI: 10.28995/2075-7182-2023-22-104-116

Дообучение моделей классификации текстов для анализа тональности к именованным сущностям в русскоязычных текстах

Анна Глазкова
Тюменский государственный университет
Тюмень, Россия
a.v.glazkova@utmn.ru

Аннотация

В статье описывается подход к анализу тональности к именованным сущностям в новостном тексте на русском языке, предложенный в рамках соревнования RuSentNE. Подход основан на использовании RuRoBERTa-large, предобученной модели RoBERTa для русского языка. Мы сравнили эффективность нескольких типов представления именованных сущностей в тексте и оценили ряд стратегий преодоления дисбаланса классов и типов сущностей в исходном датасете. Некоторые из рассмотренных стратегий улушили качество моделей классификации текстов на текстовом корпусе, предоставленном организаторами соревнования.

Ключевые слова: анализ тональности к сущностям и аспектам, именованные сущности, анализ тональности к именованным сущностям, классификация текстов, RuSentNE, RuRoBERTa.

1 Introduction

Designing effective methods for different levels of sentiment analysis is a crucial task of natural language processing. Recently, there is a growing interest in detecting sentiment for entities instead of the whole sentence or document (Li and Lu, 2017). The task of entity-level sentiment analysis is more challenging but is more useful in many applications such as content analysis and opinion mining systems.

The paper describes a system developed for the Dialogue 2023 shared task on Targeted Sentiment Analysis for the Russian Language — RuSentNE (Golubev et al., 2023). The task aims to predict sentiment labels towards named entities in Russian news texts. In this work, we compared several pre-trained language models, types of entity representation, and strategies for processing imbalanced datasets. We found that some strategies for handling class imbalance and resampling entity tags can improve the performance of pre-trained models. Our approach based on the use of RuRoBERTa-large achieved a high

result during the evaluation phase. For the final submission, we utilized a soft-voting ensemble of the models fine-tuned on the augmented dataset containing the official training set provided by the organizers, and the development set with silver labels.

The paper is organized as follows. Section 2 contains a brief review of related works. In Section 3 we describe the RuSentNE task. In Section 4 we present the methods we used. Section 5 provides and discusses the results. Some examples of the model's errors are demonstrated in Section 6. Section 7 concludes this paper.

2 Related Work

The problem of named entity oriented sentiment analysis relates to the field of targeted sentiment analysis. Target-based sentiment analysis involves opinion target extraction and actual target sentiment classification. Most of the existing studies usually explored one of these two sub-tasks alone (Wan et al., 2020). For example, the task of detecting the opinion target mentioned was solved using unsupervised (Yin et al., 2016; Giannakopoulos et al., 2017; Wu et al., 2018) and supervised (Xu et al., 2018; Yang et al., 2020) methods. The second sub-task, which is the target sentiment classification, aims to determine the entity-level sentiment for specific entities in each input text. In recent years numerous studies have extensively studied the target sentiment classification task. Most of the approaches were based on deep learning, including Recurrent Neural Networks (RNN) (Ye and Li, 2020), Long Short-Term Memory networks (LSTM) (Ma et al., 2018a; Ma et al., 2018b), Gated Recurrent Units (GRU) (Liu et al., 2018; Setiawan et al., 2020), and Bidirectional Encoder Representations from Transformers (BERT) (Sun et al., 2019; Wan et al., 2020; Mutlu and Özgür, 2022).

The concept of targeted sentiment analysis is relatively rarely found in works on the analysis of Russian texts. However, in recent years, a number of authors conducted research in related fields, such as aspect-based sentiment analysis and stance detection for the Russian language. In contrast to target sentiment analysis, which determines the opinion polarity towards the target entity in a given text, aspect-based sentiment analysis evaluates the polarity towards different aspects of a single entity (Saeidi et al., 2016). Stance detection aims to determine the position of a person from a piece of text towards a target (a concept, idea, event, etc.) either explicitly specified in the text or only implied (Küçük and Can, 2021). The general state of sentiment analysis research for the Russian language is reflected in (Smetanin, 2020; Loukachevitch, 2021).

SentiRuEval, the first sentiment analysis evaluation for Russian, was organized in 2015 (Loukachevitch et al., 2015). One of the tasks was the aspect-oriented analysis of the reviews about restaurants and automobiles. The participants utilized the methods based on LSTM (Tarasov, 2015), Support Vector Machines (SVM) (Ivanov et al., 2015; Mayorov et al., 2015), Conditional Random Fields (CRF) (Rubtsova and Koshelnikov, 2015), rule-based techniques (Vasilyev et al., 2015), and the use of Pointwise Mutual Information (PMI) and semantic similarity measures (Blinov and Kotelnikov, 2015). The SentiRuEval dataset was later used as a part of the official dataset during the international SemEval aspect-based sentiment evaluation (Pontiki et al., 2016) and utilized for evaluating deep-learning models. In (Kotelnikova et al., 2022), the authors compared several lexicon-based methods with RuBERT (Kuratov and Arkhipov, 2019). Within this comparison, the best result for the SentiRuEval dataset was obtained using the Russian adaptation of a Semantic Orientation CALculator (SO-CAL) (Taboada et al., 2011).

Studies in the field of aspect-based sentiment analysis on other text corpora were also carried out. In (Naumov et al., 2020), the authors presented an approach to aspect-based sentiment analysis where a named entity is considered as an aspect. The paper describes the dataset collected using a crowdsourcing platform and a deep neural model with Embeddings from Language Models (ELMo) (Peters et al., 2018) for word vector representation. The dataset for aspect-based sentiment analysis of Russian users' comments about COVID-19 was presented in (Nugamanov et al., 2021). The best result on this corpus was obtained using the RuBERT model in the Natural Language Inference (NLI) formulation. In (Makogon and Samokhin, 2022), a multilingual Ukrainian and Russian dataset for entity-oriented sentiment analysis was presented. The best result in terms of the F1-score for this dataset was obtained by RuBERT. The same model was applied for named entity oriented sentiment analysis in media texts in (Salnikova

Characteristic	Train	Development	Test
Number of sentences	6,637	2,845	1,947
Avg number of tokens	33.07±17.74	33.56±16.37	31.44±14.5
Distribution of tags			
Country	1,274	533	363
Nationality	276	116	110
Organization	1,487	653	484
Person	1,934	857	480
Profession	1,666	686	510

Table 1: The data statistics.

and Kyrychenko, 2021).

As targeted sentiment analysis involves determining the point of view of the text's author in relation to the given entity, it is related to the stance detection task. In (Vychegzhanin and Kotelnikov, 2017), several traditional machine-learning methods were evaluated on the dataset containing opinions of users about the topic of vaccinating children. Later, the dataset was complemented by the texts concerning other socially significant issues (Vychegzhanin and Kotelnikov, 2019). In (Lozhnikov et al., 2020), RuStance, a new dataset of Russian tweets and news comments from multiple sources, was presented. In 2022, the first evaluation on stance detection for Russian was organized (Kotelnikov et al., 2022). The participants analysed VKontakte users' comments discussing COVID-2019 news texts. The highest F1-score was obtained by the NLI-BERT system (Alibaeva and Loukachevitch, 2022) based on COVID-Twitter-BERT (Müller et al., 2020).

3 Task Description

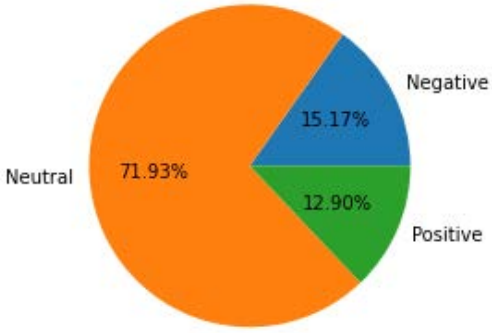
The purpose of the task is to identify sentiments for named entities. The task belongs to the class of targeted sentiment analysis tasks. Based on (Mutlu and Özgür, 2022), the problem of targeted sentiment analysis can be defined as follows. Let E denote all entities in a document D. Each e indicates an entity, $E = \{e_1, ..., e_l\}$, $l \in Z^+$. $D = \{w_1, ..., w_k\}$, $k \in Z^+$, where w denotes a word. The objective of targeted sentiment analysis is to find all sentiment pairs (s_i, t_i) in document D where t_i is a target from T, $T = \{t_1, ..., t_m\}$, $t_i \in E$, $m, i \in Z^+$, and s_i is the sentiment toward t_i.

The dataset provided for the task contains sentences from mass-media news texts in Russian. Each sentence is annotated by:

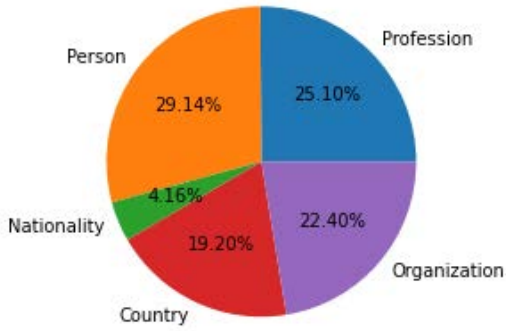
- *entity*, the object of sentiment analysis;
- *entity_tag*, the tag for the entity (Country, Nationality, Organization, Person, or Profession);
- *entity_pos_start_rel*, *entity_pos_end_rel*, the indices of the initial and next symbols for the entity occurrence;
- *label*, the sentiment label (negative, neutral, or positive)

Figure 1 shows the distribution of the labels and entity tags in the training set. As can be seen from the figure, most of the entries (71.93%) relate to the neutral class. Some tags are also dominant over others. The most common tag is Person (29.14%), while Nationality is the least abundant (4.16%). The distribution of labels within tags also varies. The texts with the tag Person include the largest proportion of sentiment labels (positive and negative, 40.54%). The smallest proportion of sentiment labels is contained in the tag Profession (12.6%). The breakdown between the training, development, and test sets is shown in Table 1. The number of tokens is obtained using the tokenizer of RuRoBERTa-large[1].

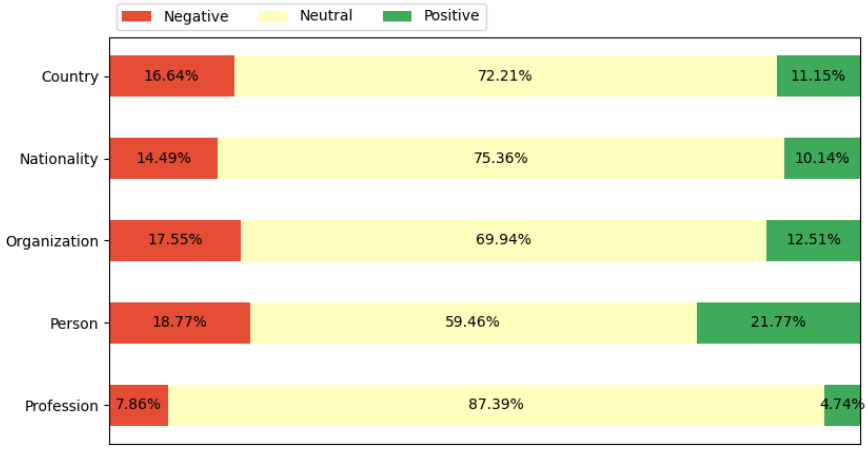
[1] `https://huggingface.co/sberbank-ai/ruRoberta-large`



a) labels



b) tags



c) labels per tags

Figure 1: The distribution of the labels and tags in the training set.

4 Methods

4.1 Entity Representation

Following previous research (Zhou and Chen, 2022; Alibaeva and Loukachevitch, 2022), we compared several groups of entity representation methods.

- **Entity mask**. This type of entity representation introduces new special tokens for masking the named entity in the source text. We compared two ways to implement this technique. In the first case, we replaced all target entities with a special token *[NE]*. In the second case, we used special tokens *[TYPE]*, where *TYPE* denotes one of the five entity tags.
- **Entity markers**. This representation type introduces new special tokens *[NE]* and *[/NE]* to enclose the named entity. We experimented with the use of one token to enclose the named entity (*[NE] entity [NE]*), as well as two tokens (*[NE] entity [/NE]*).
- **Entity markers (punct)**. This technique encloses the named entity using punctuation (* *entity* *). In this case, we did not introduce new special tokens into the model's vocabulary. The variant of this technique is adding entity types without introducing special tokens (* @ *TYPE* @ *entity* *).
- **Typed entity markers**. This technique is similar to the previous ones, but it uses control codes to highlight named entities. We consider the four types of typed entity markers: replacing target entities with the control code <|*NE*|>; enclosing entities with two similar codes (<|*NE*|>*entity*<|*NE*|>); enclosing entities with different codes (<|*NE*|>*entity*<|*/NE*|>); adding entity types to the control

code (<|*NE:TYPE*|>*entity*<|*NE:TYPE*|>).

All entity representation types are illustrated in Table 2 on the example of the text "Apple и Samsung нарушали патенты друг друга" (*Apple and Samsung infringed on each other's patents*) from the official training set of RuSentNE. In this entry, the target named entity is "Samsung", the entity type is Organization and the sentiment label is -1 (negative).

4.2 Models

We compared three pre-trained language models for the Russian language on the named entity oriented sentiment analysis task.

- **RuBERT-base**[2] (Kuratov and Arkhipov, 2019), a BERT-based model for the Russian language with 180M parameters trained on the Russian part of Wikipedia and news data. A multilingual version of BERT-base (Devlin et al., 2019) was used as an initialization.
- **RuBERT-large**[3], a large version of RuBERT containing 427M parameters trained on the Russian part of Wikipedia, news texts, books, and a fragment of the Taiga corpus (Shavrina and Shapovalova, 2017).
- **RuRoBERTa-large**[4], a modification of RuBERT that is pre-trained using dynamic masking (Liu et al., 2019), 355M parameters.

4.3 Handling Class Imbalance

Since news texts contain numerous named entities with a neutral sentiment, the neutral class largely dominates in the training set. We experimented with the following methods to reduce the impact of class imbalance on classification performance.

- **Weighted Inverse of Number of Samples (WINS)**, a class weighting technique that weights the samples as the inverse of the class frequency for the class they belong to and then normalizes them over different classes. The weight for the particular class (w_j) is calculated as follows:

$$w_j = \frac{n}{c \cdot n_j}, \tag{1}$$

  where n is the number of entries in the dataset, c is the number of classes, n_j is the number of samples of the particular class.
- **Effective Number of Samples (ENS)** (Cui et al., 2019), a class weighting scheme that calculated the weight for a particular class as follows:

$$w_j = \frac{n}{c \cdot E_{n_j}}, E_{n_j} = \frac{1 - \beta^{n_j}}{1 - \beta}, \tag{2}$$

  where E_{n_j} represents the Effective number of Samples, β is a hyperparameter ($\beta \in [0, 1)$). We experimented with the β values equal to 0.999 and 0.9999.
- **Random Oversampling for Classes (RO$_c$)**, the technique which consists if that randomly selecting entries from minority classes and adding them to the training set until the classes become the same size.
- **Data Augmentation (DA)**, we used back translation (Sennrich et al., 2016) as a data augmentation technique. For each entry from the minority classes we produced new training examples using the public translation engine, Google Translate[5], and the deep-translator Python tool[6].

4.4 Resampling Entity Tags

Since the number of entities of different types is not the same, we also investigated resampling methods to balance the number of entity tags. The following approaches were evaluated:

[2] https://huggingface.co/DeepPavlov/rubert-base-cased
[3] https://huggingface.co/sberbank-ai/ruBert-large
[4] See footnote 1
[5] https://translate.google.com/
[6] https://github.com/nidhaloff/deep-translator

№	Input representation	Example
1	Entity mask - Replacement	Apple и [NE] нарушали патенты друг друга (*Apple and [NE] infringed on each other's patents*)
2	Entity mask - Type	Apple и [ORGANIZATION] нарушали патенты друг друга (*Apple and [ORGANIZATION] infringed on each other's patents*)
3	Entity markers - 1	Apple и [NE] Samsung [NE] нарушали патенты друг друга (*Apple and [NE] Samsung [NE] infringed on each other's patents*)
4	Entity markers - 2	Apple и [NE] Samsung [/NE] нарушали патенты друг друга (*Apple and [NE] Samsung [/NE] infringed on each other's patents*)
5	Entity markers (punct)	Apple и * Samsung * нарушали патенты друг друга (*Apple and * Samsung * infringed on each other's patents*)
6	Entity markers (punct) - Type	Apple и * @ ORGANIZATION @ Samsung * нарушали патенты друг друга (*Apple and * @ ORGANIZATION @ Samsung * infringed on each other's patents*)
7	Typed entity markers - Replacement	Apple и <|NE|> нарушали патенты друг друга (*Apple and <|NE|> infringed on each other's patents*)
8	Typed entity markers - 1	Apple и <|NE|>Samsung<|NE|> нарушали патенты друг друга (*Apple and <|NE|>Samsung<|NE|> infringed on each other's patents*)
9	Typed entity markers - 2	Apple и <|NE|>Samsung<|/NE|> нарушали патенты друг друга (*Apple and <|NE|>Samsung<|/NE|> infringed on each other's patents*)
10	Typed entity markers - Type	Apple и <|NE:ORGANIZATION|>Samsung <|NE:ORGANIZATION|> нарушали патенты друг друга (*Apple and <|NE:ORGANIZATION|>Samsung<|NE:ORGANIZATION|> infringed on each other's patents*)

Table 2: Types of entity representation.

№	RuBERT-base	RuBERT-large	RuRoBERTa-large
1	65.29	69.64	73.17
2	67.34	70.65	71.28
3	65.25	**71.66**	71.88
4	67.25	69.95	73.18
5	66.1	70.72	**73.62**
6	66.52	70.33	72.71
7	67.1	70.42	72.56
8	**68.15**	70.12	73.3
9	65.99	70.37	73.16
10	65.99	70.32	73.27

Table 3: Comparison of entity representations and models (macro F1-score, %).

- **Random Oversampling for Tags (RO$_t$)**, the technique is similar to Random Oversampling for Classes, but the purpose is to balance the number of entries with different tags.
- **Sentence-Level Resampling (SLR)** (Wang and Wang, 2022), the technique was proposed for named entity recognition to increase the number of tokens of a particular entity type in the training set. The resampling function f_s can be adapted for our task in the following way. Let us denote the set of all target entity tags as T. Let $c(t, s)$ be the number of tokens of the target named entity in sentence s. The rareness r of the entity tag is measured as follows:

$$r_t = -log_2 \frac{\sum_{s \in S} c(t, s)}{N}, \tag{3}$$

where S is the set of all sentences in the training set, $\sum_{s \in S} c(t, s)$ is the total number of tokens included in the target named entities with the type t in the training set, N is the number of all tokens in the training set.

$$f_s = \frac{r_t \cdot \sqrt{c(t, s)}}{\sqrt{l_s}}, \tag{4}$$

where l_s is the number of tokens in the particular text. The resampling function f_s shows the number of times a sentence s should be resampled in a training set. The greater the number of tokens of the target entity, and the less often the entity tag is presented in the training set, the more frequently the sentence is resampled.

5 Results

5.1 Development Phase

During the development phase, we evaluated the techniques presented in Section 4. The training set was split into training and validation subsets in a ratio of 70:30. We fine-tuned each model for 6 epochs with a learning rate of 5e-6, a maximum sequence length of 200, and a batch size of 8. To evaluate the results on the validation subset, we used the macro F1-score.

Table 3 presents the results of the comparison of the models and entity representations. The highest scores for each model are shown in bold. The three best results across all models are highlighted. For better presentation, a correspondence between the types of entity representation utilized in this work and their sequential numbers is listed in Table 2. RuRoBERTa-large demonstrated the highest scores across all entity representation types. None of the entity representations showed a clear advantage over others. For instance, entity representation type 3 (Entity markers - 1) demonstrated the highest F1-score for RuBERT-large (71.66%) and the lowest for RuBERT-base (65.25%). For RuBERT-base, the best result was obtained using entity representation type 8 (Typed entity markers - 1). For RuRoBERTa-large,

Technique	RuRoBERTa-large (5)	RuRoBERTa-large (8)	RuRoBERTa-large (10)
WINS	**73.83** ↑	**73.44** ↑	72.8
$ENS_{\beta=0.999}$	73.18	**74.42** ↑	**73.62** ↑
$ENS_{\beta=0.9999}$	72.64	72.49	71.84
RO_c	73.32	71.75	72.37
DA	72.46	72.98	71.71
RO_t	73.51	72.8	72.81
SLR	**74.23** ↑	**73.34** ↑	71.94

Table 4: Comparison of strategies for handling class imbalance and resampling entity tags (macro F1-score, %).

the highest score was achieved with entity representation type 5 (Entity markers (punct)). Since many models showed very similar results, we selected three models with the highest values of the F1-score for further experiments. The selected models include Ru-RoBERTa-large with entity representation types 5 (Entity markers (punct), 73.62% of F1-score), 8 (Typed entity markers - 1, 73.3%), and 10 (Typed entity markers - Type, 73.27%).

In Table 4, the results for comparing strategies for class weighting and resampling entity tags are presented. The numbers of the corresponding entity representation types are given in brackets in the names of the columns. The results that exceeded the result of the corresponding model without the use of the strategy are shown in bold and marked with an arrow (↑). It can be seen from the values in the table that no strategy gave an advantage on all compared models. WINS showed a slight improvement with the entity representations 5 (+0.21%) and 8 (+0.14%). ENS with the value of β equal to 0.9999 ($ENS_{\beta=0.9999}$) increased the RuRoBERTa-large performance using the entity representation types 8 (+1.12%) and 10 (+0.35%). Other strategies for handling class imbalance ($ENS_{\beta=0.999}$, RO_c, and DA) led to a performance decrease in our experiments. Concerning the issue of resampling entity tags, RO_t worsened scores for all the considered models while SLR increased the F1-score with the entity representation types 5 (+0.61%) and 8 (+0.04%).

The development phase showed that the results may vary depending on the type of entity representation. Nevertheless, in our experiments, the best result for each entity representation type was achieved by RuRoBERTa-large. The choice of the entity representation type may not be obvious due to the close results obtained by models. Some strategies for handling class imbalance and resampling entity tags demonstrated an improvement in performance on the validation subset. However, not a single strategy showed an increase for all models and the growth value was often small. Therefore, during the phase, several strategies and entity representation types were selected for use in the evaluation phase.

5.2 Evaluation Phase

During this phase, we experimented with the models fine-tuned, using the techniques that showed an improvement in the development phase (WINS, $ENS_{\beta=0.9999}$, SLR) and the entity representations that demonstrated the best results during the development phase (5, 8, and 10). To increase the results on the test set, we also utilized ensemble learning and produced silver labels for the unlabelled development set provided by the organizers. Our best submission for the evaluation phase represents a system based on RuRoBERTa-large, fine-tuned using WINS with entity representation type 8. We utilized the augmented dataset consisting of the official training set and the development set with silver labels. The total size of the augmented dataset was 9,482. To combine the predictions of fine-tuned models, we used a soft-voting technique.

The official results are presented in Table 5. The models were evaluated in terms of the macro $F1_{p,n}$-score (the main performance metric), which is averaged over two sentiment classes, and the macro $F1_{p,n,0}$-score for three-class classification. Our system demonstrated the best $F1_{p,n,0}$-score out of nine

Score	Metric	
	$F1_{p,n}$-score	$F1_{p,n,0}$-score
F1-score, %	66.64	74.29
rank	2	1
baseline	40.92	56.71
avg F1-score	58.27	67.12

Table 5: Official results.

submitted teams and the second $F1_{p,n}$-score (0.03% from the first-place team).

6 Error Analysis

In this section, we provide some error examples produced by RuRoBERTa-large, fine-tuned on the official development set. The gold labels for the development set were released by the organizers of the evaluation after the end of RuSentNE. Since the gold labels for the test set had not been published by the time this paper was submitted, we cannot analyse the errors of our final model. However, an analysis of errors on the development set makes it possible to empirically trace the general trends.

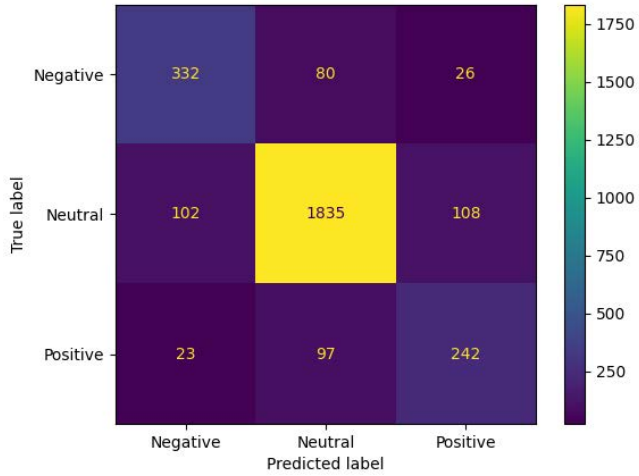


Figure 2: Confusion matrix (the development set).

The confusion matrix for the development set is presented in Figure 2. As can be seen from the figure, most of all errors are associated with the classifying entries from the neutral class as positive or negative. Examples of such errors are given in Table 6 (1 and 2). In the first sentence, the model predicts a positive class, probably, due to the availability of positive information ("universally recognized record"), however, the general meaning of the sentence is interpreted incorrectly. Perhaps this error is because the sentence is quite long and contains co-reference expressions ("Jeanne Calment", "who", "whose"). The second example is classified as negative in view of the presence of a negative fact ("was deprived of all victories"). Sentences 3-5 in Table 6 illustrate the opposite situation when sentences from the sentiment classes are classified as neutral. In the third sentence, as in the first, the presence of co-reference expressions ("Rusnok", "new prime minister") leads to an error. In addition, in examples 3 and 4, the lack of knowledge of the context complicates the classification. Examples 2 and 5 look thematically similar, but they contain entities with different tags (Person and Profession respectively). Finally, sentences 6 and 7 illustrate the situation when the model predicts the opposite sentiment class.

№	Sentence	Predicted label	Actual label
1	Общепризнанный рекорд долголетия принадлежит француженке Жанне Кальман, скончавшейся в 1997 году в возрасте 122 лет и 164 дней, возраст которой подвергается сомнению (*The universally recognized record of longevity belongs to a French woman Jeanne Calment, who died in 1997 at the age of 122 and 164 days, whose age is questioned*)	Positive	Neutral
2	Лэнса Армстронга лишили всех побед на "Тур де Франс" (Lance Armstrong was deprived of all victories in the "Tour de France")	Negative	Neutral
3	Земан назначил Руснока под предлогом, что новый премьер - хороший экономист, который займется подготовкой бюджета следующего года (*Zeman appointed Rusnok under the pretext that the new prime minister is a good economist who will engage in the preparation of the budget for the next year*)	Neutral	Positive
4	Через Германию пролегали маршруты нелегальных самолётов, которые перевозили заключённых (*The routes of illegal aircraft that transported prisoners ran through Germany*)	Neutral	Negative
5	Восемь бадминтонисток были дисквалифицированы на Олимпийских играх (*Eight badmintonists were disqualified at the Olympic Games*)	Neutral	Negative
6	Россия и Китай заблокировали резолюцию ООН, направленную против правительства Сирии (*Russia and China blocked the UN resolution directed against the Government of Syria*)	Negative	Positive
7	Лебедев признал свое участие в драке, но отверг обвинения в хулиганстве и политической ненависти. (*Lebedev admitted his participation in a fight, but rejected accusations of hooliganism and political hatred*)	Positive	Negative

Table 6: Error examples (the development set). The target entity is highlighted.

In general, in such cases, the model pays more attention to the nearest context of the entity, without analysing the general meaning of the sentence.

7 Conclusion

In this paper, we present our approach to performing named entity oriented sentiment analysis of Russian news texts. The proposed method is based on the use of RuRoBERTa-large using class weighting, data augmentation with silver data, and ensemble learning. We also studied the impact of the use of different entity representation types and strategies for handling class imbalance and resampling the dataset and provided the results of error analysis. We foresee two directions for future work. One potential direction is to investigate the impact of co-reference resolution as a pre-processing step for named entity sentiment analysis of Russian texts. Another future direction is exploring approaches for the inclusion of contextual-semantic information.

References

Kamila Alibaeva and Natalia Loukachevitch. 2022. Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference. // *Computational Linguistics and Intellectual Technologies*, P 8–16.

Pavel Blinov and Evgeniy V Kotelnikov. 2015. Semantic similarity for aspect-based sentiment analysis. *Russ. Digit. Libr. J.*, 18(3-4):120–137.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. 2019. Class-balanced loss based on effective number of samples. // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, P 9268–9277.

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.

Athanasios Giannakopoulos, Claudiu Musat, Andreea Hossmann, and Michael Baeriswyl. 2017. Unsupervised Aspect Term Extraction with B-LSTM & CRF using Automatically Labelled Datasets. *EMNLP 2017*, P 180.

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"*.

V.V. Ivanov, E.V. Tutubalina, N.R. Mingazov, and I.S. Alimova. 2015. Extracting aspects, sentiment and categories of aspects in user reviews about restaurants and cars. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 22–33.

Evgeny Kotelnikov, Natalia Loukachevitch, Irina Nikishina, and Alexander Panchenko. 2022. RuArg-2022: Argument Mining Evaluation. // *Computational Linguistics and Intellectual Technologies: papers from the Annual conference "Dialogue"*.

Anastasia Kotelnikova, Danil Paschenko, Klavdiya Bochenina, and Evgeny Kotelnikov. 2022. Lexicon-based methods vs. BERT for text sentiment analysis. // *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 71–83. Springer.

Dilek Küçük and Fazli Can. 2021. Stance detection: Concepts, approaches, resources, and outstanding issues. // *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, P 2673–2676.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for Russian language. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 333–339.

Hao Li and Wei Lu. 2017. Learning latent sentiment scopes for entity-level sentiment analysis. // *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Fei Liu, Trevor Cohn, and Timothy Baldwin. 2018. Recurrent Entity Networks with Delayed Memory Update for Targeted Aspect-based Sentiment Analysis. // *Proceedings of NAACL-HLT*, P 278–283.

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. *arXiv preprint arXiv:1907.11692*.

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.

Natalia Loukachevitch. 2021. Automatic sentiment analysis of texts: the case of Russian. *The Palgrave Handbook of Digital Russia Studies*, P 501–516.

Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. 2020. Stance prediction for Russian: data and analysis. // *Proceedings of 6th International Conference in Software Engineering for Defence Applications: SEDA 2018 6*, P 176–186. Springer.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018a. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive LSTM. // *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yukun Ma, Haiyun Peng, Tahir Khan, Erik Cambria, and Amir Hussain. 2018b. Sentic LSTM: a hybrid network for targeted aspect-based sentiment analysis. *Cognitive Computation*, 10:639–650.

Iuliia Makogon and Igor Samokhin. 2022. Targeted sentiment analysis for Ukrainian and Russian news articles. // *ICTERI 2021 Workshops: ITER, MROL, RMSEBT, TheRMIT, UNLP 2021, Kherson, Ukraine, September 28–October 2, 2021, Proceedings*, P 538–549. Springer.

V. Mayorov, I. Andrianov, N. Astrakhantsev, V. Avanesov, I. Kozlov, and D. Turdakov. 2015. A high precision method for aspect extraction in Russian. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 34–43.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. COVID-Twitter-BERT: A natural language processing model to analyse COVID-19 content on Twitter. *arXiv preprint arXiv:2005.07503*.

Mustafa Melih Mutlu and Arzucan Özgür. 2022. A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, P 467–472.

Aleksandr Naumov, R Rybka, A Sboev, A Selivanov, and A Gryaznov. 2020. Neural-network method for determining text author's sentiment to an aspect specified by the named entity. // *CEUR Workshop Proceedings*, P 134–143.

Eduard Nugamanov, Natalia Loukachevitch, and Boris Dobrov. 2021. Extracting sentiments towards COVID-19 aspects. // *Supplementary 23rd International Conference on Data Analytics and Management in Data Intensive Domains, DAMDID/RCDL 2021*, P 299–312.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 2227–2237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. // *ProWorkshop on Semantic Evaluation (SemEval-2016)*, P 19–30. Association for Computational Linguistics.

Y.V. Rubtsova and S.A. Koshelnikov. 2015. Aspect Extraction Using Conditional Random Fields. *SentiRuEval-2015*.

Marzieh Saeidi, Guillaume Bouchard, Maria Liakata, and Sebastian Riedel. 2016. SentiHood: Targeted Aspect Based Sentiment Analysis Dataset for Urban Neighbourhoods. // *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, P 1546–1556.

Svitlana Salnikova and Roman Kyrychenko. 2021. Sentiment analysis based on the BERT model: Attitudes towards politicians using media data. // *Proceedings of the International Conference on Social Science, Psychology and Legal Regulation (SPL 2021)*, P 39–44. Atlantis Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Edinburgh Neural Machine Translation Systems for WMT 16. // *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, P 371–376.

Esther Irawati Setiawan, Ferry Ferry, Joan Santoso, Surya Sumpeno, Kimiya Fujisawa, and Mauridhi Hery Purnomo. 2020. Bidirectional GRU for targeted aspect-based sentiment analysis based on character-enhanced token-embedding and multi-level attention. *Computing*, 1:2.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning: "Taiga" syntax tree corpus and parser. // *Proceedings of "CORPORA-2017" International Conference*, P 78–84.

Sergey Smetanin. 2020. The applications of sentiment analysis for Russian language texts: Current challenges and future perspectives. *IEEE Access*, 8:110693–110719.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing BERT for Aspect-Based Sentiment Analysis via Constructing Auxiliary Sentence. // *Proceedings of NAACL-HLT*, P 380–385.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Denis Tarasov. 2015. Deep recurrent neural networks for multiple language aspect-based sentiment analysis of user reviews. // *Proceedings of the 21st international conference on computational linguistics dialog*, volume 2, P 53–64.

V.G. Vasilyev, A.A. Denisenko, and D.A. Solovyev. 2015. Aspect extraction and twitter sentiment classification by fragment rules. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 76–86.

Sergey V. Vychegzhanin and Evgeny V. Kotelnikov. 2017. Stance Detection in Russian: a Feature Selection and Machine Learning Based Approach. // *AIST (Supplement)*, P 166–177.

Sergey V. Vychegzhanin and Evgeny V. Kotelnikov. 2019. Stance detection based on ensembles of classifiers. *Programming and Computer Software*, 45:228–240.

Hai Wan, Yufei Yang, Jianfeng Du, Yanan Liu, Kunxun Qi, and Jeff Z Pan. 2020. Target-aspect-sentiment joint detection for aspect-based sentiment analysis. // *Proceedings of the AAAI conference on artificial intelligence*, volume 34, P 9122–9129.

Xiaochen Wang and Yue Wang. 2022. Sentence-level resampling for named entity recognition. // *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 2151–2165.

Chuhan Wu, Fangzhao Wu, Sixing Wu, Zhigang Yuan, and Yongfeng Huang. 2018. A hybrid unsupervised method for aspect term and opinion target extraction. *Knowledge-Based Systems*, 148:66–73.

Hu Xu, Bing Liu, Lei Shu, and S Yu Philip. 2018. Double Embeddings and CNN-based Sequence Labeling for Aspect Extraction. // *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 592–598.

Yunyi Yang, Kun Li, Xiaojun Quan, Weizhou Shen, and Qinliang Su. 2020. Constituency lattice encoding for aspect term extraction. // *Proceedings of the 28th international conference on computational linguistics*, P 844–855.

Zhihao Ye and Zhiyong Li. 2020. A variant of recurrent entity networks for targeted aspect-based sentiment analysis. // *ECAI 2020*, P 2268–2274. IOS Press.

Yichun Yin, Furu Wei, Li Dong, Kaimeng Xu, Ming Zhang, and Ming Zhou. 2016. Unsupervised word and dependency path embeddings for aspect term extraction. // *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, P 2979–2985.

Wenxuan Zhou and Muhao Chen. 2022. An improved baseline for sentence-level relation extraction. // *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, P 161–168.

Aspect-based Argument Generation in Russian

Valeriya Goloviznina
Vyatka State University, Kirov, Russia
golovizninavs@gmail.com

Irina Fishcheva
Vyatka State University, Kirov, Russia
fishchevain@gmail.com

Tatiana Peskisheva
Vyatka State University, Kirov, Russia
peskisheva.ta@gmail.com

Evgeny Kotelnikov
Vyatka State University, Kirov, Russia
kotelnikov.ev@gmail.com

Abstract

The paper explores the argument generation in Russian based on given aspects. An aspect refers to one of the sides or property of the target object. Five aspects were considered: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". Various approaches were used for aspect-based generation: fine-tuning, prompt-tuning and few-shot learning. The ruGPT-3Large model was used for experiments. The results show that traditionally trained model (with fine-tuning) generates 51.6% of the arguments on given aspects, with the prompt-tuning approach – 33.9%, and with few-shot learning – 10.6%. The model also demonstrated the ability to generate arguments on new, previously unknown aspects.

Keywords: argumentation mining; controlled texts generation; GPT; fine-tuning; prompt-tuning; few-shot learning

DOI: 10.28995/2075-7182-2023-22-117-129

Аспектно-ориентированная генерация аргументов на русском языке

Головизнина В. С.
Вятский государственный университет, Киров, Россия
golovizninavs@gmail.com

Фищева И. Н.
Вятский государственный университет, Киров, Россия
fishchevain@gmail.com

Пескишева Т. А.
Вятский государственный университет, Киров, Россия
peskisheva.ta@gmail.com

Котельников Е. В.
Вятский государственный университет, Киров, Россия
kotelnikov.ev@gmail.com

Аннотация

В статье исследуется генерация аргументов на русском языке с учетом аспектов. Под аспектом понимается одна из сторон или свойство целевого объекта. Рассматривались пять аспектов: «Безопасность», «Влияние на здоровье», «Надежность», «Деньги», «Удобство и комфорт». Для аспектно-ориентированной генерации применялись различные подходы: fine-tuning, prompt-tuning и few-shot learning. Для экспериментов использовалась модель ruGPT-3Large. Результаты показывают, что модель, дообученная традиционным способом (fine-tuning), генерирует 51.6% доводов по требуемым аспектам, при подходе prompt-tuning – 33.9%, а при few-shot learning – 10.6%. Также модель продемонстрировала способность генерировать аргументы по новым, ранее неизвестным аспектам.

Ключевые слова: анализ аргументации; управляемая генерация текстов; GPT; fine-tuning; prompt-tuning; few-shot learning

1 Introduction

One of the important directions in the field of controlled text generation is the generation of argumentative texts [2], [10], [12]. An argument is a combination of a claim and at least one premise supporting or refuting that claim [13] (see Figure 1). The claim expresses the author's point of view on the controversial issue. The point of view includes the author's stance and the topic (or target). For example, in the claim "Electric cars are better than ordinary cars", the target is electric cars and the stance is "for".

To support or refute the claim, premises[1] "for" or "against" can be given, respectively.

Each premise describes one or more aspects of target. Aspect is a word or phrase that indicates one of the sides or property of the target. For example, the rebuttal premise "Battery costs have more than halved in the last four years alone" mentions the "Money" aspect.

Aspect-based argument generation allows to tune the meaning of the generated premises. However, at present there are very few studies in the field of aspect-based argument generation for the English language [12], and, to the best of our knowledge, there are no such studies for the Russian language.

We are trying to fill this gap. Applying various approaches, we train the ruGPT-3Large model on the Russian-language corpus of arguments with annotated aspects. Five aspects were considered: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". For aspect-based generation, the following methods were used: fine-tuning, prompt-tuning and few-shot learning.

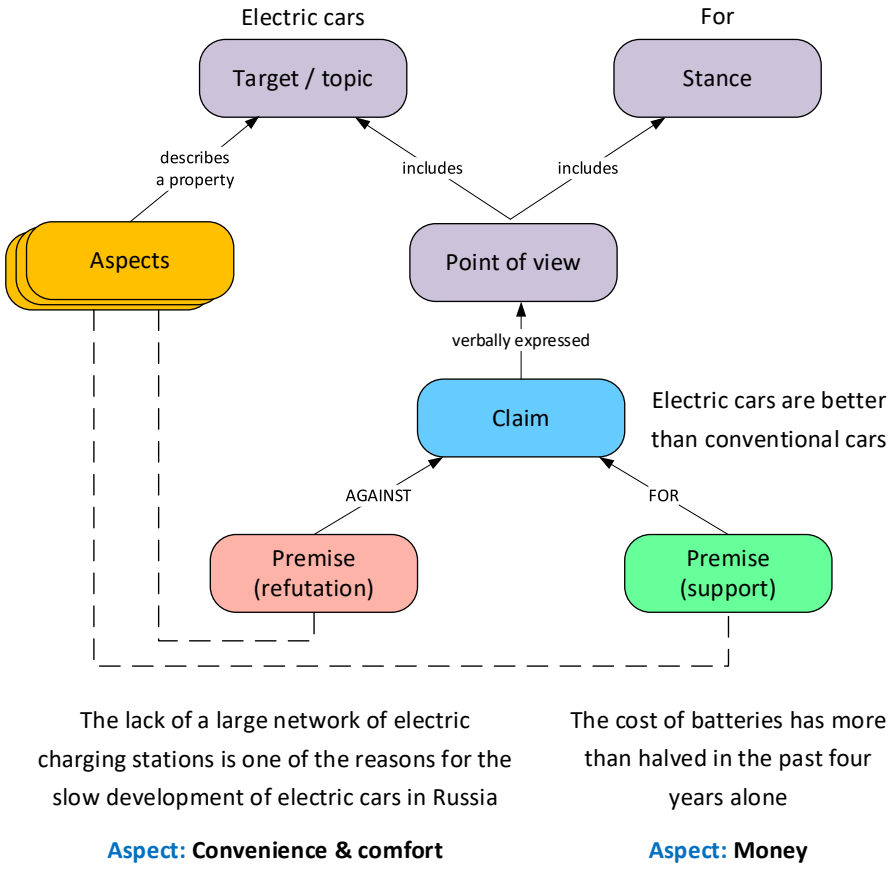


Figure 1: Argument structure. The claim "Electric cars are better than conventional cars", which expresses the stance "for" regarding the target (electric cars), is supported by the premise with the aspect "Money" and is refuted by the premise with the aspect "Convenience and comfort"

[1] Often a "premise" is called an "argument" when it is clear from the context which claim it is being referred to.

The contributions of our work are as follows:

- for the first time in the Russian language the methods of aspect-based argument generation are studied;
- the possibilities of models for arguments generation for new, unfamiliar aspects are analyzed;
- the best-scoring fine-tuned model is made publicly available [2].

2 Previous work

In this section, we first review general approaches to controlled text generation and then provide an overview of the work on aspect-based argument generation.

2.1 Controlled Text Generation

Controlled text generation refers to the task of generating text according to a given controlled element [14]. The main idea of controlled text generation based on pre-trained language models is to give the model a control signal in an explicit or implicit way to control the generation of text that satisfies given conditions. Zhang et al. [14] identify several approaches to controlled text generation.

Fine-tuning consists in tuning the parameters of the whole model or a part of it to generate text that meets specific conditions. In addition to traditional fine-tuning, there are other methods: adding an adapted module, using a prompt, and reinforcement learning.

Adding an adapted module is the construction of an additional module for solving a specific problem [14]. During the training process, the parameters of the language model are frozen, only a special module is trained.

Using a prompt is selecting an input sequence template and using it as a control hint for the language model to generate the required texts. Templates can be selected manually or automatically. The few-shot and zero-shot methods [1] involve manual selection of the prompt. The prefix tuning [5], p-tuning [6], or prompt tuning [4] methods allow to select the prompt automatically. In this case, the vectors corresponding to the prompt are tuned during the training process, while the parameters of the language model remain unchanged.

The main idea of methods based on reinforcement learning is to get feedback on whether the control conditions are achieved as a reward for fine-tuning of the language model [14].

Retrain or refactoring is a change in the original architecture of the language model or retraining of the model from scratch in accordance with the characteristics of a given task [14]. This approach can improve the quality and controllability of text generation, but is limited by the lack of tagged data and the high consumption of computing resources.

During **post-processing**, the parameters of the language model are fixed [14]. For the input sequence, the language model creates an initial distribution of tokens, the post-processing module re-ranks this distribution, ensuring that the model selects the desired token, thus controlling the generation of text.

2.2 Aspect-based Argument Generation

The problem of aspect-based generation of arguments has not yet been studied enough. Schiller et al. [12] apply fine-tuning of the CTRL model on sequences that include control codes [Topic][Stance][Aspect] (for example, *Nuclear Energy CON radioactive waste*) and a premise (for example, *Nuclear reactors produce radioactive waste…*) for the controlled generation of premises on a given topic, stance and aspect.

In paper [2], to generate premises on economic topics, the original ruGPT-3Large model and the same model fine-tuned on an argument corpus containing 3,500 sentences were used. As a result of manual evaluation, 63.2% of the sentences generated by the fine-tuned ruGPT-3Large model turned out to be premises, while the original model without fine-tuning was able to generate only 42.5% of premises.

In our work, in contrast to [2] and [12], in addition to the traditional fine-tuning of the whole model, the following methods for controlled argument generation are studied:

- traditional fine-tuning of the whole model and fine-tuning of the last layer only;
- using of prompt-tuning;
- using a few-shot manual prompt.

[2] https://tinyurl.com/452euk4w.

In [2], the aspect of the premise is not taken into account, and in [12], the aspect is part of the premise. In our work, the aspect reflects the semantic orientation of the premise and is not part of the sentence containing the premise. For example[3]:

- topic: school uniforms,
- premise: outsiders who do not belong to the campus are easy to identify and therefore do not pose much of a threat to students.

In [12], the following aspects are indicated: [outsiders, easy to identify, threat]; in our work, such an premise would have the "Security" aspect.

3 Materials and Methods

3.1 Corpora

We use the existing corpus of premises with aspects specified for them[4].

The corpus contains 548 premises that have from 1 to 3 aspects from the list containing 20 aspects, such as "Safety", "Living standard", "Quality", etc. A complete list of aspects with their frequency is given in Appendix A. We have combined the most similar aspects "Impact on health" and "Impact on the psyche", "Price" and "Profitability", replacing them with aspects "Impact on health" and "Money" respectively. From this corpus, we have identified the most frequently aspects (they met more than 80 times) and the sentences corresponding to them. Thus, we have formed a corpus for the generation, which includes 418 unique argumentative sentences. For each sentence, from 1 to 3 aspects are selected from the following list: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". Since one premise can have several aspects, we used 507 argumentative sentences (with repetitions) to train the models.

The corpus contains 14 topics. Each topic is reflected in the claim about which the argument is built. For example, the topic "Cryptocurrency" in the corpus is represented by the claim "We need to use and invest in cryptocurrencies", the topic "Children's video blogs" is represented by the claim "Children should be encouraged to create vlogs", and the topic "Esports" is represented by the claim "Esports should be made an Olympic sport". A complete list of topics, claims to them and the distribution of topics by aspects are presented in Appendix B.

3.2 Language Model

Models of the GPT (Generative Pre-trained Transformer) family consist of a Transformer decoder with a different number of layers [8]. The ruGPT-3 model is a Russian-language model from Sber, based on GPT-2 [9], available in five versions of different sizes (ruGPT-3Small, ruGPT-3Medium, ruGPT-2Large, ruGPT-3Large, ruGPT-3XL) [11]. The model was trained on 80 billion tokens. In our experiments, we used the ruGPT-3Large model (760M parameters).

3.3 Training Methods

For controlled argument generation, we explore several methods of model training: fine-tuning, prompt-tuning, and few-shot learning.

In traditional fine-tuning, the weights of the model change, adjusting to the required task – generating premises. The ruGPT-3Large model is fine-tuned on text sequences containing a claim, an aspect, and a premise. The input of the model is a sequence of the form:

Claim: {*claim*}; **Aspect:** {*aspect*}; **Premise:** {*premise*}

Parts of the sequence in bold are keywords; instead of parts in curly brackets, real claims, aspects, and premises are substituted. When testing, the input of the model is the following sequence:

Claim: {*claim*}; **Aspect:** {*aspect*}; **Premise:**

[3] Example is taken from the UKP corpus [12].

[4] https://github.com/kotelnikov-ev/RuArgumentMining/tree/main/AspectCorpus.

The model generates a premise by continuing the sentence. Two variants of this approach are considered:

- fine-tuning of the whole model,
- fine-tuning of only the last layer.

In the prompt-tuning method, the vectors that serve as the prompt for the generation control are trained, the model weights are frozen. The prompt vectors obtained during the training process are fed to the input of the model along with embeddings representing text tokens. The input sequence looks like:

$$\textbf{<P*}\boldsymbol{n}\textbf{>}\{claim\}\textbf{<P*}\boldsymbol{m}\textbf{>}\{aspect\}\textbf{<P*}\boldsymbol{k}\textbf{>},$$

where $m, n, k \geq 0$ are numbers that indicate the number of special <**P**> tokens in a specific prompt format, claims and aspects are substituted for curly braces.

In the few-short learning method, a prompt is supplied to the model input, including generation examples that describe the task. In our work, these are examples of arguments that include a claim, an aspect, and a premise. In this method, neither the model nor additional vectors are trained. The model input in the few-shot learning contains a prompt that includes 16 examples (limited by GPU memory) written as:

Claim: {*claim*}; **Aspect:** {*aspect*}; **Premise:** {*premise*}

and ending with the sequence:

Claim: {*claim*}; **Aspect:** {*aspect*}; **Premise:**

The model is asked to generate a premise on the last specified claim and aspect.

In this method, we use two types of prompts:

- the prompt contains examples of premises for all aspects in accordance with the distribution of aspects in the original corpus,
- the prompt contains examples of premises only on the aspect of the generated premise.

4 Experiments

4.1 Experimental Setup

We considered several options for formats of prompt in prompt-tuning:

- <P*100>{claim}<P*20>{aspect}<P*100>,
- <P*100>{claim}<P*4>{aspect}<P*20>,
- <P*60>{claim}<P*4>{aspect}<P*60>,
- <P*20>{claim}<P*20>{aspect}<P*20>,
- <P*60>{claim}<P*1>{aspect}.

To implement this approach, we used the ru-prompts library[5]. When choosing the number of special tokens, we were guided by training examples provided by the library developers, which used sequences of 100, 20, and 4 special tokens. We also added variants of the prompt formats, with the same or close total value of the number of special tokens, but arranged differently in the sequence. The second and third formats have the same number of special tokens, but they have a different arrangement in the sequence, similarly for the third and fourth options.

With the help of 5-fold cross-validation we selected the best two prompt formats:

- <P*100>{claim}<P*20>{aspect}<P*100>,
- <P*100>{claim}<P*4>{aspect}<P*20>.

[5] https://github.com/ai-forever/ru-prompts.

For experiments, the NVIDIA RTX A6000 video card and transformers library[6] were used. For each method, we selected a number of training epochs on a 5-fold cross-validation from the following ranges:

- fine-tuning the whole model = [1…5],
- fine-tuning the last layer = [1…20],
- prompt-tuning = [1…300].

The best were 2 epochs for training the whole model, 20 epochs for training the last layer, and 300 epochs for prompt-tuning. The learning rate $5 \cdot 10^{-5}$ and batch size 4 were the same for all the models. We used the following parameters to generate[7]: top_p=0.95, top_k=50, do_sample=True, max_new_tokens=150, no_repeat_ngram_size=3. The generated sequence was segmented into sentences using the natasha library[8]. The first sentence was used for annotation.

Thus, we test six models:

- a fine-tuned whole model,
- a model with fine-tuned last layer,
- an original model with 220 special tokens in prompt,
- an original model with 124 special tokens in prompt,
- an original model with a prompt containing various aspects,
- an original model with a prompt containing one aspect of interest.

4.2 Results and Discussion

Using each of the six models, 254 sentences were generated for 5 aspects of the corpus, that is, 1,524 sentences were obtained for annotation. We determined the number of sentences generated in accordance with the distribution of aspects by claims in the original corpus (Appendix B), excluding rare <claim-aspect> pairs, which accounted for only 1 or 2 premises.

The annotation was done by three annotators, for which the method of generating sentences was hidden. The annotator evaluated the generated sentence according to several parameters:

- whether the sentence is a premise for the specified claim (0/1);
- if the sentence is a premise, then what aspect is expressed in it;
- whether the sentence contains grammatical errors (0/1);
- whether the sentence contains a violation of logic (0/1).

For each generated premise, the annotator could indicate 3 aspects from the full list of aspects (Appendix A).

We considered as premises those sentences that were annotated as premises by at least two annotators, the same rule was applied to identify sentences with grammatical errors and violations of logic. Examples of premises with grammatical and logical errors (for the claim: "Need to save for retirement"):

- *быть осторожным. это означает: иметь накопления и ни в коем случае не делать пенсионные сбережения втом банке, в котором они, по крайней мере в ближайшие пять лет являются гарантированным вкладом и,если такая ситуация возникнет, по истечении такого же времени с целью последующего ихсчета в банке, где они будут гарантированы на все случаи жизни, но ни в каких других.*
  [*to be careful. this means: to have savings and in no case make pension savings in a bank in which they are, at least for the next five years, a guaranteed deposit and,if such a situation arises, after the same time with a view to their subsequent account in a bank where they will be guaranteed for all occasions, but in no other.*];
- *Следует делать пенсионную сбережения, которые вернут в виде процентов, полученных за активы, в том числе активы, которые были начислены в результате обесценения;*
  [*One should make retirement savings, which will return in the form of interest received on assets, including assets that have been accrued as a result of impairment;*].

[6] https://huggingface.co/docs/transformers.
[7] https://huggingface.co/docs/transformers/main_classes/text_generation.
[8] https://github.com/natasha/razdel.

If at least one annotator indicated among the three aspects the one on which the sentence was generated, then we considered this sentence as a premise generated by the required aspect.

The inter-annotator agreement, calculated on the basis of Krippendorff's alpha, was 0.5781 for premise annotation, and 0.4911 for aspect annotation (whether the required aspect is present in the premise). The most difficult topics for premise annotation were "Retirement savings" (0.3601) and "Online shopping" (0.4578); the easiest topics were "Distant work" (0.8059) and "Shooters" (0.7648). The most difficult topics for aspect annotation were also "Retirement savings" (0.2356) and "Esports" (0.2639); the easiest topics coincided with premises: "Distant work" (0.6290) and "Shooters" (0.7185).

Table 1 shows the number of trainable parameters for each model and the statistics of the generated premises: the number of premises generated by the model; the number of premises generated for the required aspect and the number of premises generated for the required aspect that do not contain defects (grammatical errors or violation of logic). The table also shows the proportion (in percent) of such premises among all sentences generated by the model (indicated in brackets).

Model	# trainable parameters	# generated sentences (%)	Premises		Premises on aspect		Premises on aspects without defects	
			#	%	#	%	#	%
Fine-tuned whole model	760,300,032	254 (100%)	**158**	**62.2**	**131**	**51.6**	33	13.0
Fine-tuned last layer	77,194,752		36	14.2	26	10.2	9	3.5
Prompt-tuned-220	337,920		125	49.2	86	33.9	**75**	**29.5**
Prompt-tuned-124	190,464		101	39.8	72	28.4	63	24.8
Few-shot learning all aspect	0		49	19.3	27	10.6	21	8.3
Few-shot learning selected aspect	0		41	16.1	26	10.2	21	8.3

Table 1: Number of trainable parameters and statistics of generated premises for each model

The largest number of premises (column 3) and premises on the required aspects (column 4) was generated by fine-tuned whole model. The quality of the generated premises (column 5) is higher for prompt-tuned and few-shot learning models than for fine-tuned models. Perhaps this is due to the modification of the model weights during fine-tuning, which to some extent violates the grammatical correctness of the generated texts.

Table 2 shows examples of generated premises. Each example was labeled by three annotators as a premise on the required aspect.

Claim: Video games have a positive effect on people	
Aspect: Impact on health	
Model	Generated premise
Fine-tuned whole model	*Однако исследования, в которых участвовали 15-летние школьники, показали, что многие из них играли в видеоигры в отсутствие реальной угрозы жизни или здоровью и таким образом получали возможность выплеснуть свою агрессию.* [*However, studies involving 15-year-old high school students showed that many of them played video games in the absence of a real threat to life or health and thus got an opportunity to express their aggression.*]
Fine-tuned last layer	*это связано с тем, что многие подростки проводят время за видеоиграми, в которых им дают шанс использовать способности мозга и сознания в определенных областях своего восприятия реальности, повышая степень своего развития и, соответственно, шанс успешного получения результата в том или ином из многих случаев, о котором в них говорится.* [*this is due to the fact that many teenagers spend time playing video games in which they are given a chance to use their brain and consciousness abilities in certain areas of their perception of reality, increasing their degree of development and thus their chance of success in one or another of the many cases they are talking about.*]
Prompt-tuned-220	*В исследовании было показано, что те, кто играет в видеоигры, в 2,5 раза чаще страдают заболеванием сосудов головного мозга по сравнению с теми, кто не играет в игры.* [*The study showed that those who play video games are 2.5 times more likely to suffer from cerebral vascular disease compared to those who do not play games.*]
Prompt-tuned-124	*В ходе исследования медики изучили влияние игр на людей и пришли к выводу, что люди, игравшие в видеоигры, страдают от различных болезней чаще.* [*During the study, doctors have studied the effect of games on people and concluded that people who played video games, suffer from a variety of diseases more often.*]
Few-shot learning all aspect	*ьы часто используют видеоигры для развития моторики и координации движений;* [*often use video games to develop motor skills and motor coordination;*]
Few-shot learning selected aspect	*ыысбые 3D-экранные видеоигры могут привести к психическим заболеваниям. ...* [*3D screen video games can lead to mental illness. …*]

Table 2: Examples of generated premises. Defects are highlighted in red, the symbol "..." means that the continuation of the sequence does not refer to the premise and contains errors

When evaluating generation methods, it is important to understand whether the model replicates the training data. To determine the similarity between the generated premises and the training data, we calculated the average and maximum ROUGE-L and cosine similarity between the generated premises of each model and the premises of the training corpus (Table 3). Cosine similarity was calculated using the RuBERT model [3] and the Sentence Transformers library[9].

[9] https://www.sbert.net/

Model	ROUGE-L		Cosine similarity	
	mean	max	mean	max
Fine-tuned whole model	0.0468	**0.4800**	0.6255	0.9400
Fine-tuned last layer	0.0445	0.3590	0.5859	0.9288
Prompt-tuned-220	0.0487	0.4000	0.6107	**0.9501**
Prompt-tuned-124	0.0448	0.3478	0.5913	0.9332
Few-shot learning all aspect	0.0459	0.2632	0.5998	0.9245
Few-shot learning selected aspect	0.0487	0.2500	0.6018	0.9279

Table 3: ROUGE-L and cosine similarity between the generated premises and the premises of the training corpus

With the maximum mean cosine similarity of 0.6255 across the models, the mean ROUGE-L of the fine-tuned whole model does not exceed 0.05. The maximum value of ROUGE-L among all models was shown by the fine-tuned whole model, while the maximum cosine similarity value was shown by the prompt-tuned model. However, even similar premises have clear differences. We give examples of premises that obtained maximum ROUGE-L and cosine similarity scores:

- **Corpus sentence**: *В 2014 году глава Банка Эстонии осторожно отмечал отсутствие доказательства того, что Биткойн не является финансовой пирамидой* [*In 2014, the Governor of the Bank of Estonia was careful to point out the lack of proof that Bitcoin is not a pyramid scheme*].
  Generated sentence: *заявил, что биткойн не является самостоятельной финансовой пирамидой* [*stated that Bitcoin is not a financial pyramid in its own right*].
- **Corpus sentence**: *Если ты будешь платить ренту в биткойнах, ты можешь обанкротиться в случае если он сильно пойдет вверх, а твой доход привязан к фиатным деньгам* [*If you pay your rent in bitcoins, you can go bankrupt if it goes up strongly and your income is tied to fiat money*].
  Generated sentence: *Это приводит к риску вывода активов и денежных сумм с кратной целью, т.е. при первоначальной сделке с биткойнами инвестор рискует получить убыток и подвергнуть риску свои накопления* [*This leads to the risk of withdrawing assets and sums of money in multiples, i.e. in an initial bitcoin transaction an investor risks making a loss and putting his savings at risk*].

Thus, we can conclude that the models generate different premises from the training examples.

We also tested the ability of the resulting models to generate premises for unfamiliar claims and aspects. This means that the training data for the fine-tuned and prompt-tuned models did not contain these claims and aspects, and the prompt of the few-shot model did not contain examples for these claims and aspects. To test this, we generated three five-aspect sentences for two claims using each model:

- claim: "Children's camp affects the child positively",
  aspects: "Intellectual development", "Moral aspect", "Communication with people", "Popularity", "Entertainment";
- claim: "Free money is better spent than saved",
  aspects: "Standard of living", "Legal aspect", "Efficiency", "Attitudes of the authorities", "Prospects".

Thus, we obtained 150 sentences for annotation. This annotation was conducted similarly to the previous one. The results of the analysis are presented in Table 4.

Inter-annotator agreement, calculated with the help of Krippendorff's alpha, is 0.4538 for the annotation of premises, and 0.3006 for the annotation of aspects (whether there is a required aspect in the premise).

Model	# generated sentences (%)	Premises		Premises on aspect		Premises on aspects without defects	
		#	%	#	%	#	%
Fine-tuned whole model	30 (100%)	**18**	**60.0**	**11**	**36.7**	1	3.3
Fine-tuned last layer		3	10.0	0	0.0	0	0.0
Prompt-tuning_220		7	23.3	6	20.0	**6**	**20.0**
Prompt-tuning_124		6	20.0	4	13.3	3	10.0
Few-shot learning all aspect		8	26.7	2	10.0	1	3.3

Table 4: Statistics of generated premises on unfamiliar aspects for each model

In this experiment, the fine-tuned whole model also showed the best result.

Table 5 shows examples of generated premises for unfamiliar aspects. For each example, the number of annotators who considered the sentence to be a premise and whether the premise expresses the required aspect are given.

Claim: Children's camp has a positive effect on the child			
Aspect: Communication with people			
Model	Generated premise	Voting	Aspect
Fine-tuned whole model	*Но в реальности дети зачастую лишены такой возможности, потому что находятся в стесненных условиях, их постоянно окружают чужие люди, от которых они не могут укрыться.вольте думать позитивно!...* [*But in reality, children are often deprived of this opportunity, because they are in cramped conditions, they are constantly surrounded by strangers, from whom they can not hide.*]	3 from 3	+
Fine-tuned last layer	*Утверждение: В лагере есть работа и льготы для детей;* [*The camp has jobs and benefits for children;*]	2 from 3	–
Prompt-tuned-220	*Коротко, в отличие от традиционных лагерей, для детей проводят так называемое коммуникативно ориентированное пространство или пространство деятельности.* [*Briefly, unlike traditional camps, a so-called communication-oriented space or activity space is held for children.*]	2 from 3	+
Prompt-tuned-124	*Влияние на детей в детском возрасте может быть позитивным, когда люди, с которыми он сталкивается, являются нормальными, но может быть и негативным, когда он сталкивается с ненормальными.* [*The impact on children at an early age can be positive when the people he encounters are normal, but it can also be negative when he encounters the abnormal.*]	3 from 3	+
Few-shot learning	*Детский лагерь является лучшим местом для раскрытия творческих способностей у детей, что позволяет им чувствовать себя более комфортно;* [*Children's camp is the best place to unleash creativity in children, making them feel more comfortable;*]	3 from 3	–

Table 5: Examples of generated premises on unfamiliar aspects. Defects are highlighted in red, the symbol "..." means that the continuation of the sequence does not refer to the premise and contains errors.

5 Conclusion

The paper explored aspect-based argument generation in Russian. Five aspects were considered: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". The models were trained using different approaches (fine-tuning, prompt-tuning and few-shot learning). The best model was wholly fine-tuned on the aspect-based corpus of premises. This model generated 51.6% of the premises on the given aspects, the model obtained using the prompt-tuning approach gives 33.9% of the premises on the given aspects, and with the few-shot learning approach – 10.6%.

The problem of fine-tuned models is the low level of grammatical correctness of the generated premises compared to the prompt-tuning and few-shot learning models. For example, for the best fine-tuned model out of 131 generated premises without grammatical errors and logic violations, there were 33 premises (25.2%), and for the prompt-tuned-220 model, 75 out of 86 premises (87.2%) were correct. When the post-processing of the generated sentences is complicated or impossible for some reason, prompt-tuned models become preferable.

It is important to note that fine-tuned models are able to generate premises on new, unfamiliar aspects. For example, the fine-tuned whole model was able to generate 36.7% of premises (11 premises for the required aspect out of 30 generated sentences). This allows us to hope for the potential application of such models for a wide range of topics.

In the future, we plan to expand the annotated corpus of premises and aspects and use reinforcement learning [7].

We made the best fine-tuned model for generating premises on the given aspects available to the public.

Acknowledgements

This work was supported by Russian Science Foundation, project № 22-21-00885, https://rscf.ru/en/project/22-21-00885/.

References

[1] Brown T., Mann B., Ryder N., Subbiah M., Kaplan J.D. et al. (2020), Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems, Vol. 33, pp. 1877–1901.

[2] Fishcheva I., Osadchiy D., Bochenina K., Kotelnikov E. (2022), Argumentative Text Generation in Economic Domain // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue-2022", Issue 21, pp. 211–222.

[3] Kuratov Y., Arkhipov M. (2019), Adaptation of deep bidirectional multilingual transformers for Russian language, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue -2019", Issue 18, pp. 333–340.

[4] Lester B., Al-Rfou R., Constant N. (2021), The Power of Scale for Parameter-Efficient Prompt Tuning // Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 3045–3059.

[5] Li X. L., Liang P. (2021), Prefix-Tuning: Optimizing Continuous Prompts for Generation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pp. 4582–4597.

[6] Liu X., Zheng Y., Du Z., Ding M., Qian Y., Yang Z., Tang J. (2021), GPT Understands, Too. arXiv:2103.10385.

[7] Ouyang L., Wu J., Jiang X., Almeida D., Wainwright C. et al. (2022), Training language models to follow instructions with human feedback // NeurIPS 2022.

[8] Radford, A., Narasimhan, K., Saliman, T., Sutskever I. (2018), Improving Language Understanding by Generative Pre-Training. OpenAI Technical report.

[9] Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. (2019), Language Models are Unsupervised Multitask Learners, OpenAI Technical report.

[10] Ruckdeschel M., Wiedemann G. (2022), Boundary Detection and Categorization of Argument Aspects via Supervised Learning // International Conference on Computational Linguistics. Proceedings of the 9th Workshop on Argument Mining, pp. 126–136.

[11] Sber: Russian GPT3 models. Retrieved March 15, 2023, from https://github.com/sberbank-ai/ru-gpts.

[12] Schiller B., Daxenberger J., Gurevych I. (2021), Aspect-Controlled Neural Argument Generation // Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 380–396.

[13] Stede M., Schneider J. (2018), Argumentation Mining, Synthesis Lectures on Human Language Technologies, San Rafael: Morgan and Claypool Publishers.
[14] Zhang H., Song H., Li S., Zhou M., Song D. (2022), A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models // arXiv:2201.05337.

Appendix A

Aspect	Frequency in the corpus, sentences
Safety	**133**
Impact on health (Impact on health + Influence on the psyche)	**107 (56 + 51)**
Reliability	**90**
Money (Price + Profitability)	**89 (55 + 34)**
Convenience and comfort	**88**
Attitude of the authorities	78
Prospects	72
Efficiency	39
Standard of living	26
Legal aspect	26
Environmental friendliness	23
Communication with people	22
Popularity	15
Quality	12
Career	9
Intellectual development	7
Entertainment	7
Moral aspect	4

Table 6: Frequency of aspects in the corpus[10]

[10] The arguments in the table are not unique, since one sentence can have several aspects.

Appendix B

Topic	Claim	Aspect					Total
		Safety	Impact on health	Reliability	Money	Convenience and comfort	
Paper and e-books	Paper books are better than e-books	1	0	0	0	7	8
Children's video blogs	Children should be encouraged to create vlogs	1	11	0	2	0	14
Children's gadgets	Gadgets have a positive effect on children	12	59	0	0	0	71
Blood donation	Donation is necessary for society and safe	6	14	0	0	0	20
Esports	Esports should be made an Olympic sport	0	1	1	4	0	6
Cryptocurrency	You need to use cryptocurrency and invest in it	87	1	59	30	24	201
Online education	Online education can compete with traditional education	0	0	0	5	2	7
Retirement savings	Need to save for retirement	0	0	15	4	1	20
Online shopping	Should shop online	11	0	1	0	10	22
Supermarkets and food markets	It is better to buy products in the supermarket, not in the market	6	0	2	7	11	26
Distant work	Remote work is preferable to office work	3	1	0	6	11	21
Freelance	Freelancing is better than being hired	0	1	1	4	2	8
Shooters	Video games have a positive effect on a person	0	19	0	0	0	19
Electric cars	Electric cars are better than regular cars	6	0	11	27	20	64
Total		133	107	90	89	88	507

Table 7: Distribution of corpus topics by aspects[11]

[11] The arguments in the table are not unique, since one sentence can have several aspects.

RuSentNE-2023: Evaluating Entity-Oriented Sentiment Analysis on Russian News Texts

Anton Golubev
Lomonosov Moscow State University
antongolubev5@yandex.ru

Nicolay Rusnachenko
Bauman Moscow State Technical University
rusnicolay@gmail.com

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Abstract

The paper describes the RuSentNE-2023 evaluation devoted to targeted sentiment analysis in Russian news texts. The task is to predict sentiment towards a named entity in a single sentence. The dataset for RuSentNE-2023 evaluation is based on the Russian news corpus RuSentNE having rich sentiment-related annotation. The corpus is annotated with named entities and sentiments towards these entities, along with related effects and emotional states. The evaluation was organized using the CodaLab competition framework. The main evaluation measure was macro-averaged measure of positive and negative classes. The best results achieved were of 66% Macro F-measure (Positive+Negative classes). We also tested ChatGPT on the test set from our evaluation and found that the zero-shot answers provided by ChatGPT reached 60% of the F-measure, which corresponds to 4th place in the evaluation. ChatGPT also provided detailed explanations of its conclusion. This can be considered as quite high for zero-shot application.

Keywords: Targeted Sentiment Analysis, Named Entity, News Texts, ChatGPT
DOI: 10.28995/2075-7182-2023-22-130-141

RuSentNE-2023: анализ тональности по отношению к сущностям в русскоязычных новостных текстах

Голубев А. А.
МГУ им. Ломоносова
Москва, Россия
antongolubev5@yandex.ru

Русначенко Н. Л.
МГТУ им. Баумана
Москва, Россия
rusnicolay@gmail.com

Лукашевич Н. В.
МГУ им. Ломоносова
Москва, Россия
louk_nat@mail.ru

Аннотация

В статье описывается тестирование RuSentNE-2023, посвященное таргетированному анализу тональности в русскоязычных новостных текстах. Задача участников состояла в том, чтобы предсказать тональность по отношению к именованному объекту в предложении. Датасет тестирования RuSentNE-2023 основан на корпусе российских новостей RuSentNE, в котором размечено несколько типов явлений, связанных с тональность. Корпус аннотирован именованными сущностями, тональностью по отношением к этим сущностям, размечены последствия для сущностей в связи описываемыми ситуациями и эмоциональные состояниями сущностей.Тестирование было организовано на основе специализированного сайта для тестирований CodaLab. Основной мерой оценки было макроусреднение положительных и отрицательных классов. Наилучший результат, полученный участниками, был 66% макро-F-мере (положительные + отрицательные классы). Мы также протестировали ChatGPT на тестовом наборе нашего тестирования и обнаружили, что zero-shot (без обучения) ответы ChatGPT достигают 60% F-меры, что соответствует 4-му месту в тестировании RuSentNE. Модель ChatGPT также предоставила подробные пояснения к своему заключению. Этот результат можно считать достаточно высоким для применения в условиях zero-shot.

Ключевые слова: таргетированный анализ тональности, именованная сущность, новостные тексты, ChatGPT

1 Introduction

Sentiment analysis studies began with the general task setting, in which the general sentiment of a sentence or a text should be detected. Currently, so-called targeted sentiment analysis is intensively discussed, in which a model should determine the attitude towards specific entities, their aspects (properties), or topics.

Targeted sentiment analysis is especially important for news flow processing. It is assumed that news texts should be neutral, but in fact they contain a variety of information about positions on various issues of state bodies, companies, opinions of individuals, positive or negative attitudes of the mentioned subjects to each other. All these opinions are important for collecting and analysis.

Currently, news sentiment seems understudied. For example, the search on paper titles in Google Scholar shows that number of papers devoted to sentiment in social media is three times larger than the number of paper discussing news sentiment.

The specific features of news texts from the point of view of sentiment analysis are as follows (Loukachevitch and Rusnachenko, 2018):

- these texts contain various opinions conveyed by different subjects, including the author(s)' attitudes, positions of cited sources, and relations of mentioned entities to each other;
- some sentences are full of named entities with different sentiments, which makes it difficult to determine sentiment for a specific named entity;
- news texts contain numerous named entities with neutral sentiment, which means that the neutral class largely dominates;
- significant share of sentiment in news texts is implicit, for example can be inferred from some actions of entities.

The authors of (Hamborg et al., 2021) annotated a news corpus with sentiment and stated that sentiment in the news is less explicit, more dependent on context and the reader's position, and it requires a greater degree of interpretation. It was found that state-of-the-art approaches to targeted sentiment analysis perform worse on news articles than in other domains. The authors also point out that in 3% of cases, the sentiment is depends on the position of the reader.

In this paper we present the Russian News corpus RuSentNE annotated with named entities and sentiment towards these entities, related effects and emotional states. We used the sentiment annotation of the RuSentNE corpus to organize the shared task RuSentNE-2023 within the framework of Dialogue evaluation series.

2 Related Work

There were several international evaluations devoted to targeted sentiment analysis.

In 2012-2014 within the CLEF conference, RepLab events devoted to the evaluation of online reputation management systems were organized (Amigó et al., 2012; Amigó et al., 2014). The task was to determine if the tweet content has positive or negative implications for the company's reputation.

In the SemEval evaluation workshops 2015, 2016, studies were devoted to aspect-based sentiment analysis (ABSA) in several domains. The task was to determine sentiment towards specific characteristics discussed in users' reviews such as food or service in restaurants (Pontiki et al., 2015; Pontiki et al., 2016). In 2016-2017, topic-oriented sentiment analysis tasks were studied in the SemEval series of Twitter sentiment evaluations (Nakov et al., 2016; Rosenthal et al., 2017).

The latest trend in targeted sentiment analysis is the so-called Structured Sentiment Analysis, which involves extracting tuples from texts that describe opinions of the following form $\langle h, t, e, p \rangle$, where h is the opinion holder, p represents sentiment (positive or negative), in relation to the entity t, expressed by means of the word or phrase e. Structured sentiment analysis can be divided into five subtasks: i) sentiment expression extraction, ii) sentiment object extraction, iii) sentiment subject extraction, iv) determination of the relationship between these elements, and v) sentiment extraction (positive or negative). Modern approaches aim to address these problems in a unified manner, generating the required tuples (Lin et al., 2022). In 2022, the competition on structural sentiment analysis competition was organized as part of the SemEval-2022 international evaluation workshop (Barnes et al., 2022).

The relatively recent advent of *transformers* (Vaswani et al., 2017) cause a significant breakthrough in machine learning application across the variety of natural language processing tasks, including targeted sentiment analysis. Within the last five years, the significant amount of works were aimed on application of transformer components, namely the encoder (Devlin et al., 2019) and decoder (Alt et al., 2019a). In the case of the application of BERT, complemented by classification layer on top, has resulted in the appearance of a variety pretrained models employing different pretrain techniques (Zhuang et al., 2021; Alt et al., 2019b). Another group of authors studies conveying the structure of target in context, emphasizing the target in texts and forming a prompt-based input constructions (Sun et al., 2019a; Shin et al., 2020). To the best of our knowledge, the structured representation of the input contexts (Morio et al., 2022) as a part of sequence-to-sequence and graph-based models represents the latest advances in target oriented sentiment analysis.

For Russian, targeted sentiment evaluations were organized as a two-year series in 2015 and 2016. In 2015, aspect-based sentiment evaluation in restaurant and car reviews was explored (Loukachevitch et al., 2015). During two years, methods for reputation monitoring tasks towards banks and telecommunication companies in tweets were evaluated (Loukachevitch and Rubtsova, 2015). The best classification results in 2015-2016 competitions were mainly based on the SVM method and the GRU neural network. Later, the results were significantly improved with the application of the BERT model (Golubev and Loukachevitch, 2020; Golubev and Loukachevitch, 2021).

We see that most international and Russian evaluations are devoted to targeted sentiment analysis applied to social media and user reviews, not to news analysis. The general task of sentiment analysis devoted to classification of Russian news quotes was studied in the ROMIP-2013 evaluation (Chetviorkin and Loukachevitch, 2013).

The work closest to ours is the NewsTSC corpus (Hamborg et al., 2021). The authors selected articles of 14 US newspapers from the Common Crawl news crawl (CC-NEWS). Mentions of named entities such as PERSON or ORG were automatically identified, and corresponding sentences were extracted. The annotators should read the sentence and determine sentiment towards a highlighted named entity. In total, 3002 sentences were annotated, the neutral sentiment was most frequent (2087). Several BERT models were used in the experiments. The best results (58.8 macro F measure) were achieved LCF-BERT (Zeng et al., 2019), where BERT was tuned on the Common Crawl news corpus.

3 RuSentNE Sentiment-Annotated Corpus

The dataset RuSentNE-2023 constructed for the shared task, is based on the RuSentNE corpus, in which diverse sentiment-related phenomena are annotated.

The corpus RuSentNE includes news texts from the NEREL dataset, annotated with 29 named entity types (Loukachevitch et al., 2021). 400 NEREL texts with the largest relative share of sentiment words according to the RuSentiLex lexicon (Loukachevitch and Levchik, 2016) were selected for sentiment annotation in RuSentNE. The annotated named entities were used as targets of sentiment in the RuSentNE corpus.

Then sentiment-related tags and relations towards NEREL entities were labeled in the RuSentNE. The sentiment annotation includes 12 entity tags and 11 relation types, which describe sentiment, arguments, effects and emotions of entities mentioned in the text.

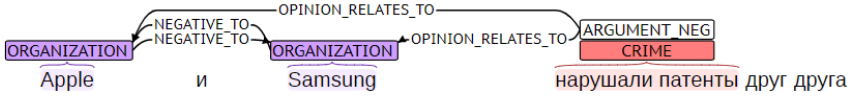


Figure 1: Example of sentiment annotation in sentence "Apple and Samsung infringed each other's patents" shows negative relations between companies and negative argument towards them

Figures 1, 2 and 3 show examples of annotation in the RuSentNE corpus. In Figure 1, we can see that two companies are negative to each other (`NEGATIVE_TO` relation); also negative argument ("enfringed patents") is annotated that explains the attitude of two companies towards each other

(OPINION_RELATES_TO relation).

In Figure 2, we see that the Matteo Renzi is positive to Fo (POSTITVE_TO relation); the explanation for this attitude is given (*major figure in cultural life*).The emotional state of Matteo Renzi is negative because of the Fo death. Also negative effect for Dario Fo is annotated, which originated from his death. In Figure 3 we see the author position towards Berlusconi depicted as tag AUTHOR_NEG.

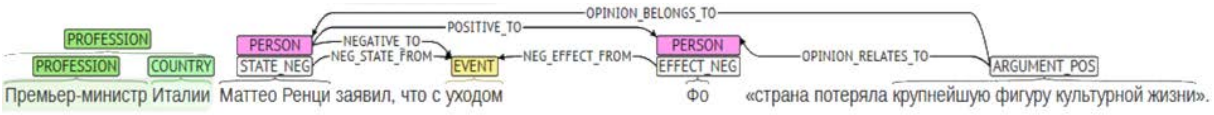


Figure 2: Example of sentiment annotation in sentence "Prime Minister of Italy Matteo Renzi said that with Fo's departure, "the country has lost a major figure in cultural life."" demonstrates positive attitude from Matteo Renzi towards Dario Fo and also negative emotions stemminng from his death.

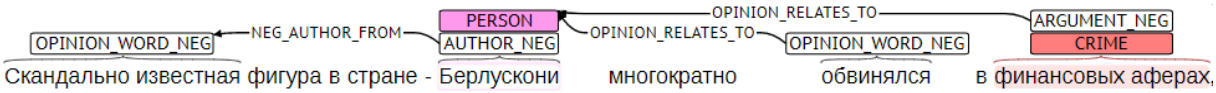


Figure 3: Example of sentiment annotation in sentence "Notorious figure in the country - Berlusconi has been repeatedly accused of financial fraud."" shows negative opinion of the author.

For the RuSentNE annotation, the BRAT annotation tool was used. The texts were annotated by both students and specialists in computational linguistics. All texts were then checked by a moderator, who discussed any identified issues' with annotators.

4 RuSentNE-2023 Dataset

The annotation of the RuSentNE corpus was partially used for the RuSentNE-2023 evaluation. Since RuSentNE contains 29 different classes of named entities, we selected those entities that are most frequent objects of the targeted sentiment in news texts. We selected the following classes of entities for the evaluation:

- PERSON — physical person regarded as an individual,
- ORGANIZATION — an organized group of people or company,
- COUNTRY — a nation or a body of land with one government,
- PROFESSION — jobs, positions in various organizations, and professional titles,
- NATIONALITY — nouns denoting country citizens and adjectives corresponding to nations in contexts different from authority-related.

The distribution of entity types in the training, validation, and test sets is presented in Table 1

Stage	Total	PERSON	ORG.	COUNTRY	PROF.	NATIONAL.
Train	6637	1934 (29%)	1487 (23%)	1274 (19%)	1666 (25%)	276 (4%)
Development	2845	857 (30%)	653 (23%)	686 (19%)	533 (24%)	116 (4%)
Final	1947	480 (25%)	484 (25%)	363 (19%)	510 (26%)	110 (5%)

Table 1: Distribution of entity types in training, validation and test sets

For RuSentNE-2023 evaluation devoted to targeted sentiment analysis, we used a subset of annotated entity tags and relations from the initial annotation of the RuSentNE corpus:

- AUTHOR_POS, AUTHOR_NEG tags describe the author's attitude towards the tagged entity,
- POSITIVE_TO, NEGATIVE_TO relations describe a relationship between two entities in a text,
- OPINION_RELATES_TO relation describes the relation from the opinion expressed in the text to the object of the opinion.
  - OPINION_WORD_NEG, ARGUMENT_NEG — negative to the object,

– `OPINION_WORD_POS`, `ARGUMENT_POS` — positive to the object

Entities without any sentiment annotations or relations directed towards them form the neutral class.

In particular, the following examples of targeted sentiment in the NEREL-2023 dataset are obtained from the annotation presented in Figures 1, 2, and 3 (Table 2).

Source	Entity	Sentiment	Source in RuSentNE
Figure 1	Samsung	negative	`NEGATIVE_TO`, `ARGUMENT_NEG`
Figure 1	Apple	negative	`NEGATIVE_TO`, `ARGUMENT_NEG`
Figure 2	Fo	positive	`POSITIVE_TO`, `ARGUMENT_POS`
Figure 2	Matteo Renzi	neutral	absense of annotation
Figure 3	Berlusconi	negative	`AUTHOR_NEG`, `ARGUMENT_NEG`

Table 2: Examples of sentiment labels in RuSentNE-2023 dataset

In the original RuSentNE corpus, sentiment-related relations can be annotated across sentences. In the RuSentNE-2023 evaluation, the targeted sentiment should be extracted from a single sentence. This implies that in a specific sentence, sentiment towards a target entity can be absent. Therefore, it was then necessary to ensure that the source of sentiment is located in the same sentence as the entity. If not, the sample was not included in the dataset. In the example below, the Tula transport prosecutor's office from the second sentence derives the positive annotation from the first one, which contradicts the relations from isolated second sentence:

- *«The Tula transport prosecutor's office **defended the rights** of workers. The Tula transport prosecutor's office **filed two lawsuits against Russian Railways** for the recovery of child care allowances».*

The pre-trained $\text{RuCoreNews}_{\text{small}}$ spaCy model[1] was utilized to segment texts into sentences.

The initial RuSentNE corpus contains so-called nested named entities, that is one entity can be annotated within another entity. Only an upper-level entity was selected for the RuSentNE-2023 dataset. An example of nested entities can be seen in Figure 2: entity *Prime-minister of Italy* contains entity *Italy*.

After the collection was formed, several post-processing steps were carried out. The same sentences that were included in the dataset multiple times according to different criteria were deduplicated. Examples with conflicting annotations were excluded. The minimum (40) and maximum (430) lengths of the text were chosen experimentally. Sentences excluded at this stage are presented below. The first sentence contains conflicting annotation for Iran (Tel Aviv vs Syria, Hezbollah and Hamas), while the following two sentences are too short:

- *«Tel Aviv believes that the price of the withdrawal of Israeli troops from the Golan Heights should be political concessions from Syria - the weakening of its ties with its strategic ally **Iran** and the cessation of support for the Lebanese resistance movement Hezbollah and the Palestinian Hamas».*
- *«The **shepherd** was not injured».*
- *«**Pirates** fought back».*

Table 3 presents the distribution of sentiment scores in the training, validation and test sets.

Type	Stage	Total	Positive	Negative	Neutral
train	Train	6637	856 (13%)	1007 (15%)	4774 (72%)
validation	Development	2845	362 (13%)	438 (15%)	2045 (72%)
test	Final	1947	269 (14%)	243 (12%)	1435 (74%)

Table 3: Distribution of sentiment scores in training, validation and test sets.

We estimated the inter-annotated agreement in the RuSentNE-2023 dataset using duplicating sentences from the initial dataset. Cohen's Kappa is calculated as 0.5 (moderate agreement). But the Cohen's Kappa can be unreliable in our case due to Kappa's sensitivity to class imbalance (Feinstein and Cicchetti, 1990). The percentage agreement between annotators is 84%.

[1] `https://spacy.io/models/ru#ru_core_news_sm`

5 Task Description

In the RuSentNE-2023 competition, named entities should be classified into one of three sentiment classes: positive, negative or neutral within the context of a single sentence. Each sentence is annotated as follows:

- `entity` — object of sentiment analysis
- `entity_tag` — tag of this object (see Section 4)
- `entity_pos_start_rel` — index of the initial character of the given entity
- `entity_pos_end_rel` — index of the next character after the last of the given entity
- `label` — sentiment label

Each entity has a three-scaled label. The following classes (labels) are used:

- Negative (-1)
- Neutral (0)
- Positive (1)

Participants are tasked with automatically annotating each test sentence with the appropriate sentiment label for a specified entity.

As the primary evaluation metric, we adopt the F1-PN-macro, which averages the F1-measures of the positive and negative classes. Additionally, we calculated the traditional F1-macro measure as a supplementary metric.

6 Results

Over 15 participants took part in development stage of the competition. Table 4 illustrates the results obtained by the competitors during this preliminary evaluation stage.

Participant	F1-PN-macro (rank)	F1-PN0-macro (rank)
mtsai	70.94 (1)	77.63 (1)
cookies	69.89 (2)	76.74 (2)
lsanochkin	68.11 (3)	75.36 (3)
Dmitry315	62.91 (4)	71.28 (4)
ryzhtus	62.34 (5)	70.57 (6)
mitrokosta	62.15 (6)	70.70 (5)
sag_m	61.35 (7)	69.73 (7)
shershulya	60.07 (8)	69.26 (8)
s231644	59.99 (9)	69.10 (9)
antongolubev	57.73 (10)	68.21 (10)
ild	57.46 (11)	67.52 (11)
GreatDispersion	57.00 (12)	65.38 (12)
abc111	55.25 (13)	65.00 (13)
baseline_model	44.32 (14)	57.89 (14)
postoevie	41.09 (15)	48.28 (16)
angyling	35.37 (16)	43.76 (17)
AlexSMSU	31.94 (17)	49.25 (15)

Table 4: Results of the Development evaluation stage, ordered by F1-PN-macro; participant nicknames with top-3 results are bolded; baseline_model results correspond to the application of the baseline model

The results of the final evaluation stage are illustrated in Table 5. Participants were not limited in the number o submissions due to the relatively high threshold provided for both evaluation stages with 1K submissions limit. During the final stage, those participants who did not exceed the baseline result chose not to include their submissions in the leaderboard. In the following, we provide brief descriptions of the methods adopted by participants during the final evaluation stage.

baseline_model. DeepPavlov (Burtsev et al., 2018) pre-trained conversational $\text{RuBERT}_{\text{base}}$[2] is used. The model is fine-tuned treating the targeted sentiment analysis task as a question-answering problem (Sun et al., 2019b).

Participant	F1-PN-macro (rank)	F1-PN0-macro (rank)
mtsai	66.67 (1)	74.11 (2)
cookies	66.64 (2)	74.29 (1)
lsanochkin	62.92 (3)	71.20 (3)
ChatGPT **	60.06 (-)	70.79 (-)
antongolubev	59.64 (4)	69.04 (4)
sag_m	59.33 (5)	68.71 (5)
mitrokosta	58.68 (6)	67.54 (6)
Dmitry315	53.60 (7)	62.92 (7)
ild *	53.20 (-)	63.78 (-)
abc111	49.98 (8)	61.32 (8)
Naumov_al	46.96 (9)	54.92 (10)
baseline_model	40.92 (10)	56.71 (9)

Table 5: Results of the Final evaluation stage, ordered by F1-PN-macro; participant nicknames with top-3 results are bolded; «*» corresponds to post-evaluation submissions, made rightafter the end of the final stage; baseline_model results correspond to the application of the baseline model; «**» represents zero-shot answers of ChatGPT application baseline

mtsai. The participant experimented with the following the following models: $\text{RuRoBERTa}_{\text{large}}$[3] , $\text{XLM-RoBERTa}_{\text{large}}$[4], RemBERT[5] (Chung et al., 2021). Two separate transformers-based models were adopted: one model was used for the original input, while the other was applied to the input with masked entities. The participant utilized the predefined «`[MASK]`» token for entities, identified with the Named Entity Recognition (NER) approach. Weights ranging from 0 to 1.0 were applied to the output, with the sentiment classes assigned a weight of 1.0 and the neutral class assigned a weight of 0.1. The participant also implemented threshold for neutral class: when the neutral class has the highest probability, but its value below threshold, they select the most probable class among "positive" and "negative". Participant employed an ensembling technique, which was based on a five-fold split of the dataset from the development stage. This technique involved training different transformers on various splits of data. In total, five models were used to ensemble the output.

cookies. The participant experiment with the following set of language models: language models $\text{RuBERT}_{\text{base}}$, $\text{RuBERT}_{\text{large}}$, $\text{RuRoBERTa}_{\text{large}}$. In terms of the input representation, various masks and markers for entity representation were considered. During the development phase, participant concluded that $\text{RuRoBERTa}_{\text{large}}$ illustrates the highest results across the other models of experiment set, as well as the absence of difference across different entity representation formats. For the final version of the input representation format, participant decided to mark the entity within the input sentence. The best submission for the development stage represents a system based on $\text{RuRoBERTa}_{\text{large}}$, fine-tuned using the inverse probability weighting technique. In the final evaluation stage, participant also employed ensemble learning and implemented automatic annotation of the development set.

lsanochkin. The participant used a language model with a prompting technique according to the following assumption: passing entity (e) jointly with the sentence[6] (T) in the following format: «`[CLS]` T `[SEP]` e». The model was fine-tuned on the input text format with transfromer-based classification pipeline (fully-connected layer on the top of a transformer model, cross-entropy loss). In terms

[2] `https://docs.deeppavlov.ai/en/master/features/models/bert.html`

[3] `https://huggingface.co/sberbank-ai/ruRoberta-large`

[4] `https://huggingface.co/xlm-roberta-large`

[5] `https://huggingface.co/google/rembert`

[6] like question answering or natural language inference prompt (Sun et al., 2019a)

of the embedding pooling for the sentiment class identification, the participant consider «[CLS]» token as a sentence representation.

antongolubev. The participant utilized the DeepPavlov RuBERT-base model fine-tuned in natural language inference (NLI) problem setting together with transfer learning approach. The model was pre-trained on targeted sentiment analysis data from SentiRuEval 2015-2016 evaluations (Loukachevitch et al., 2015) and automatically generated data from Russian news corpus (Golubev and Loukachevitch, 2021). In addition, the SMOTE (Chawla et al., 2002) augmentation approach for increasing of positive and negative classes was held.

sag_m. The participant's approach was based on the text-to-text generation approach using the $\text{ruT5}_{\text{large}}$[7] model. The generated output text is one of the possible sentiment labels for the analyzed named entity: "negative", "neutral" or "positive". The application process involved experimenting with several variants of data preparation for the model input and applying output token filtering to derive the final class label. The best results were achieved using additional data for training: the CABSAR dataset[8].

mitrokosta. The participant applied the RuBERT model. Entity token hidden states of the last layer were pooled. In addition, the hidden states of all layers were processed by convolutions and those that relate to the entity were averaged again. The latter yelded some improvements in quality.

ild. The participant adopts $\text{SBERT}_{\text{large}}$ language model, pre-trained initially on financial news dataset[9].

Dmitry315. The participant used the RuBERT model for fine-tuning on the provided data, with pooling for embeddings related to the entity.

Naumov_al. The participant considers the task as a multiclass-classification problem and experimented with Interactive Attention Network (IAN) (Ma et al., 2017). Numerous experiments were conducted using different word vector representations[10]. The participant experimented with several parameters including hidden_dim in the LSTM layers, the learning rate, and batch size. Alongside the competition dataset, the participant used CABSAR corpus[11], which contains Russian-language sentences from various sources: posts of the Live Journal social network, texts of the online news agency[12], and Twitter microblog posts.

We observe that in all approaches exceeding the baseline model results, participants used neural language models. Analyzing the results, it is worth to conclude the importance of the following findings, related to language model application (Vaswani et al., 2017):

1. a pretraining technique plays role: the contribution of RoBERTa results in higher performance rather than original BERT for the same sizes language models.
2. scale of the models is another direction: larger size of the model likely results in a higher baseline in the case of the fixed batch size and training evaluation methodology across all the model's sizes under consideration.
3. application of ensembling techniques allows gaining higher prediction results (the two first approaches in the final leaderboard).
4. task abstraction: this is a particular case when entities might be masked, marked or prompted which may assist with reaching the desired outcomes.

7 ChatGPT in RuSentNE-2023 evaluation

In (Zhang et al., 2022) authors report state-of-the-art results in the stance detection domain with zero-shot application of ChatGPT. The latter became a source of our inspiration to contribute with the related findings by analyzing responses for RuSentNE-2023 dataset.

We applied a conversational system that uses the GPT-3.5 model (Zhang et al., 2022) [13],which comes

[7] https://huggingface.co/ai-forever/ruT5-large

[8] https://github.com/sag111/CABSAR

[9] https://huggingface.co/chrommium/sbert_large-finetuned-sent_in_news_sents

[10] ELMo, $\text{RuBERT}_{\text{large}}$ and **$\text{XLM-RoBERTa}_{\text{large}}$** (the latter was used in final model), combined with the original IAN model

[11] https://github.com/sag111/CABSAR

[12] lenta.ru

[13] https://platform.openai.com/docs/models/gpt-3-5

with a pretrained state[14] corresponding to InstructGPT model. Access to this model is provided by the paid subscription service, ChatGPTPlus. The hidden state of the model was trained on text collections up to June 2021, and it supports up to 4,097 tokens inputs. The model is non-deterministic, implying that identical inputs can result in different outputs. We used a default temperature parameter. The sentences from the dataset, translated using the `googletrans` library[15], were input into the ChatGPT model and were also incorporated into a prompt formatted as follows:

*«What is the attitude of the sentence **[translated sentence]** to the target **[translated entity]**? Select one from «favor, against or neutral» and explain why».*

Regular expressions with later manual validation were used to analyse model responses. For several examples ($< 0.1\%$ of data), the entity was translated incorrectly, which led to the fact that the model denied entity to be in the sentence. In such cases, the neutral label was put. Due to the limit on the number of requests to ChatGPT model per hour, a shell script was written to send examples at a given frequency. The whole test dataset was processed in 55 hours.

The results of ChatGPT are included in Table 5. We can see that the model applied in zero-shot manner (without fine-tuning) took the fourth place in the evaluation reaching more than 60% F1-PN-macro.

8 Analysis of Examples

To analyze difficult cases, we extracted examples from the test set, which were erroneously classified by all (9) or almost all (8) participants. We encountered the following main problem cases (the examples below are translated from Russian):

1. Models fail to distinguish between the subject and the object of opinion. For example, in the sentence "In 2011, Azerbaijan will increase pressure on international organizations in connection with the Nagorno-Karabakh problem", eight models inaccurately predicted negative sentiment towards Azerbaijan. However, according to the ChatGPT explanation, which seems correct in this context: "The sentence simply states a fact about the intentions of the Azerbaijani government without expressing any positive or negative opinion about it".

A similar issue can be seen in the sentence: "In it, Crowley announced that the Obama administration intends to continue to cooperate with the Israeli politician". Here most models inaccurately predicted positive sentiment towards the Obama administration, while both the annotators and ChatGPT indicated neutral sentiment.

2. Models do not distinguish some sentiment adjectives (evident for a human reader, but ambiguous) and fail to identify sentiment directed towards the target entity. For example, in sentence "In the very first days of Moiseev's stay at the clinic of JSC "Medicina", the patient was examined by the country's **leading** neurosurgeon, who did not find any indication for surgery", nearly all models, including ChatGPT, failed to detect positive sentiment towards "neurosurgeon".

3. Models fail to identify sentiment that requires an understanding of the relationships between entities. For example, in sentence "On the first day of 2011, numerous messages appeared on the **Microsoft** forum from Hotmail users complaining about the disappearance of all read messages stored in their mailboxes". In this case, all models (including ChatGPT) did not detect the negative sentiment towards Microsoft, the owner of Hotmail service.

4. Models do not distinguish implicit sentiment. For example, in sentence "Over the years of her film career, she has appeared **in more than 30 films and television series**, including the sci-fi film Forbidden Planet and the detective series "Honey West", for which the actress **received a Golden Globe film award** and was **nominated for an Emmy Award.**" both the models and ChatGPT were unable to detect the positive sentiment toward the actress.

We also encountered several issues with human annotation within our dataset, particularly in sentences relating to sporting events. Some annotators labeled negative relationships between competing athletes but it seems in sports such relations should not be annotated. Additionally, discrepancies arose among

[14] `text-davinci-002-render-paid`, 2022/11/22

[15] `https://pypi.org/project/googletrans/`

annotators concerning the annotation of victories and defeats: while some treated such outcomes as factual information, others interpreted them as positive or negative sentiments towards the athletes. Similar variations were observed in the output from ChatGPT.

9 Conclusion

In this paper, we described the RuSentNE-2023 evaluation devoted to targeted sentiment analysis in Russian news texts. We think that targeted sentiment analysis in news texts has not been thoroughly explored, as prior research primarily focuses on sentiment within user reviews and social media content. The distinct characteristics of news texts include a significant proportion of neutral named entities, a substantial presence of implicit sentiment, and the occurrence of multiple named entities with varying sentiments within the same sentence.

The objective of the RuSentNE-2023 evaluation was to predict sentiment towards a named entity within a single sentence, effectively classifying the sentiment into one of three categories: positive, negative, or neutral. The dataset used for the RuSentNE-2023 evaluation is derived from the Russian News corpus, RuSentNE, which contains a comprehensive sentiment-related annotation. The corpus is marked with named entities and the sentiment directed towards these entities, alongside associated effects and emotional states.

The evaluation was organized using the CodaLab competition framework. The main evaluation measure was macro-averaged measure of positive and negative classes. There were 15 participants in the development stage of the evaluation, with 9 participants presenting their results in the final stage. The best results reached of 66% Macro F-measure for the positive and negative classes.

We additionally conducted experiments with ChatGPT on the test set of our evaluation, discovering that the zero-shot responses of ChatGPT reached 60% of the F-measure, which corresponds to the fourth-place in the evaluation. ChatGPT also provided comprehensive explanations of its conclusions. The outcomes are considerably notable, given that this was a zero-shot application.

Our future plans involve improving the annotation of the RuSentNE corpus, taking into account the results of the participants and the explanations provided by ChatGPT. Subsequently, we aim to conduct an evaluation focused on structured sentiment analysis, specifically tailored for the extraction of four-tuples (the subject of sentiment, object of sentiment, sentiment, and sentiment expression). Furthermore, it is important to eliminate sentence restrictions and advance towards classifying sentiment across sentences.

Acknowledgements

The work was supported by the Russian Science Foundation under Agreement No. 21-71-30003.

References

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019a. Improving relation extraction by pre-trained language representations. // *Proceedings of AKBC 2019*.

Christoph Alt, Marc Hübner, and Leonhard Hennig. 2019b. Improving relation extraction by pre-trained language representations. *arXiv preprint arXiv:1906.03088*.

Enrique Amigó, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten De Rijke, et al. 2012. Overview of replab 2012: Evaluating online reputation management systems. // *CLEF (online working notes/labs/workshop)*.

Enrique Amigó, Jorge Carrillo-de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Edgar Meij, Maarten de Rijke, and Damiano Spina. 2014. Overview of replab 2014: author profiling and reputation dimensions for online reputation management. // *Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings 5*, P 307–322. Springer.

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.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, Alexey Litinsky, Varvara Logacheva, Alexey Lymar, Valentin Malykh, Maxim Petrov, Vadim Polulyakh, Leonid Pugachev, Alexey Sorokin, Maria Vikhreva, and Marat Zaynutdinov. 2018. DeepPavlov: Open-source library for dialogue systems. *// Proceedings of ACL 2018, System Demonstrations*, P 122–127, Melbourne, Australia, July. Association for Computational Linguistics.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357.

Ilia Chetviorkin and Natalia Loukachevitch. 2013. Evaluating sentiment analysis systems in russian. *// Proceedings of the 4th biennial international workshop on Balto-Slavic natural language processing*, P 12–17.

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. *// 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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.

Alvan R Feinstein and Domenic V Cicchetti. 1990. High agreement but low kappa: I. the problems of two paradoxes. *Journal of clinical epidemiology*, 43(6):543–549.

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 and Natalia Loukachevitch. 2021. Use of augmentation and distant supervision for sentiment analysis in russian. *// Text, Speech, and Dialogue: 24th International Conference, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings 24*, P 184–196. Springer.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards target-dependent sentiment classification in news articles. *// Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16*, P 156–166. Springer.

Yangkun Lin, Chen Liang, Jing Xu, Chong Yang, and Yongliang Wang. 2022. Zhixiaobao at semeval-2022 task 10: Apporoaching structured sentiment with graph parsing. *// Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, P 1343–1348.

Natalia Loukachevitch and Anatolii Levchik. 2016. Creating a general russian sentiment lexicon. *// Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, P 1171–1176.

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.

NV Loukachevitch and N Rusnachenko. 2018. Extracting sentiment attitudes from analytical texts. *// Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 448–458.

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.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. 2021. Nerel: A russian dataset with nested named entities, relations and events. *// International Conference on Recent Advances in Natural Language Processing: Deep Learning for Natural Language Processing Methods and Applications, RANLP 2021*, P 876–885. Incoma Ltd.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *// Proceedings of the 26th International Joint Conference on Artificial Intelligence*, IJCAI'17, P 4068–4074. AAAI Press.

Gaku Morio, Hiroaki Ozaki, Atsuki Yamaguchi, and Yasuhiro Sogawa. 2022. Hitachi at SemEval-2022 task 10: Comparing graph- and Seq2Seq-based models highlights difficulty in structured sentiment analysis. // *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, P 1349–1359, Seattle, United States, July. Association for Computational Linguistics.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. // *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, P 1–18.

Maria Pontiki, Dimitrios Galanis, Harris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. // *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, P 486–495.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. // *ProWorkshop on Semantic Evaluation (SemEval-2016)*, P 19–30. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. // *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, P 502–518.

Taylor Shin, Yasaman Razeghi, Robert L. Logan IV, Eric Wallace, and Sameer Singh. 2020. AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 4222–4235, Online, November. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019a. Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. // *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 380–385, Minneapolis, Minnesota, June. Association for Computational Linguistics.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019b. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. // *Proceedings of NAACL-HLT*, P 380–385.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Biqing Zeng, Heng Yang, Ruyang Xu, Wu Zhou, and Xuli Han. 2019. Lcf: A local context focus mechanism for aspect-based sentiment classification. *Applied Sciences*, 9(16):3389.

Bowen Zhang, Daijun Ding, and Liwen Jing. 2022. How would stance detection techniques evolve after the launch of chatgpt? *arXiv preprint arXiv:2212.14548*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A robustly optimized BERT pre-training approach with post-training. // *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, P 1218–1227, Huhhot, China, August. Chinese Information Processing Society of China.

Frequency dynamics as a criterion for differentiating inflection and word formation (in relation to Russian aspectual pairs)

Elena V. Gorbova
independent researcher
elenagorbova12@gmail.com

Oksana Iu. Chuikova
Herzen State Pedagogical University of Russia
oxana.chuykova@gmail.com

Abstract

The paper reports the results of the critical evaluation of the quantitative approach to the distinction between inflection and word formation through the analysis of the trends in the frequency of word forms. The possibility of such analysis is provided by voluminous corpus data and tools for visualizing these trends. Both theoretical foundations of the proposed approach and the results of the pilot study of its applying to Russian aspectual triplets were considered. These cast doubt on the validity of distinguishing between inflection and word formation based on the trends in the frequency of word forms as a reliable tool used to reveal the unity or difference of lexical semantics and thus to define textual units as belonging to the same or different language units.

Keywords: Russian aspect; inflection; word formation, quantitative analysis; frequency, corpora
DOI: 10.28995/2075-7182-2023-22-142-160

Динамика частотности как критерий разграничения словоизменения и словообразования (применительно к видовой парности русского глагола)

Горбова Елена Викторовна
независимый исследователь
elenagorbova12@gmail.com

Чуйкова Оксана Юрьевна
РГПУ им. А. И. Герцена
oxana.chuykova@gmail.com

Аннотация

В статье представлены результаты критического осмысления количественного подхода к проведению границы между словоизменением и словообразованием через анализ динамики частотности употребления словоформ, возможность которого обеспечивается объемными корпусными данными и средствами их визуализации. Предлагается обсуждение как теоретических основ предложенного подхода, так и результатов пилотного исследования его применения к материалу русских видовых троек. Проведенный анализ позволяет усомниться в валидности разграничения словоизменения и словообразования через динамику частотности как действенного способа установления единства лексической семантики в качестве инструмента определения статуса текстовых единиц как вариантов одной языковой единицы или репрезентантов разных.

Ключевые слова: вид русского глагола; словоизменение; словообразование; количественный анализ; частотность, корпус

1 Вступительные замечания

Целью статьи является критическое обсуждение предложенного в [15, 16] «частотного подхода к лексической семантике», или «объективного численного подхода» [15: 73], как применительно к более общему вопросу различения словоизменения и словообразования, так и к его приложению к двум морфологическим типам видовой пары и проблеме трактовки грамматической категории

русского вида (аспекта)[1]. Отметим, что обозначенный подход основан, как указано в [16], на дистрибутивной гипотезе, изложенной в [13].

Далее работа выстроена следующим образом: раздел 2 представит частотный подход к лексической семантике и его критику с точки зрения теоретической лингвистики; в разделе 3 приведены случаи наличия несомненного (по общепризнанным лингвистическим критериям) статуса двух (или более) словоформ одной лексемы при значительно разнящейся частоте употребления словоформ в парадигме одной лексемы и случаи синхронного изменения частотности для словоформ разных лексем; в разделе 4 обсуждаются результаты частотного подхода к лексической семантике на материале видовых троек; в заключительном разделе 5 подведены итоги.

2 Частотный подход к лексической семантике: теоретические основы и его приложение к описанию категории вида русского глагола

Кратко изложим предложенный в [15, 16] подход. Основной задачей [16] является решение «исследовательского вопроса: в какой степени меняется семантика слов при суффиксальном и префиксальном способах образования аспектуальных пар?» [16: 1117]. При этом «проверяемая гипотеза» сформулирована так: «семантическая близость между глаголами в аспектуальных парах перфектив – вторичный имперфектив будет больше, чем между глаголами в парах базовый имперфектив – перфектив» [Ibid]. Постановка задачи и гипотеза базируются на: а) стремлении подвергнуть критическому анализу одну из теоретических моделей русского вида, а именно трактовку первого типа пар (с участием префигированных перфективов, противопоставленных симплексам-имперфективам) как словообразования и второго типа пар (единожды префигированного перфектива и образованного от него вторичного имперфектива) как словоизменения [16: 1116]; б) отказа от решения проблемы словоизменение vs. словообразование на основе обязательности и регулярности, поскольку регулярность отличается градуальностью (авторы ссылаются, в частности, на [11]), см. [16: 1116-1117]. В результате авторы [16] предлагают решать вопрос о границе между словоизменением и словообразованием[2] исключительно на основе идеи о единстве лексической семантики для словоформ одной лексемы, см. [16: 1118-1120], причем последнее устанавливается в ходе анализа изменения во времени частотности употребления словоформ, выявленного путем обращения к корпусу Google Books Ngram (GBN) с визуализацией посредством сервиса Ngram Viewer (https://books.google.com/ngrams/): частота употребления двух лексов с одинаковой лексической семантикой, находящихся в словоизменительном отношении, должна изменяться синхронно [16: 1120].

В [15] акценты чуть сдвинуты: здесь в фокусе внимания и критики находится операционный критерий установления видовой парности, известный как «критерий Маслова». Вместо него формулируется другой критерий видовой парности: «Глаголы из пары «первичный имперфектив – перфектив» имеют одинаковую лексическую семантику (т. е. образуют аспектуальную пару) тогда и только тогда, когда частотность их использования меняется синхронно – графики частотности имеют одинаковую форму» [15: 76]. Кроме визуального сравнения формы кривых на графике, построенном сервисом Ngram Viewer, применяется также их оценка через коэффициент Спирмена. Любопытно, что в [Ibid] авторы приводят три ограничения предложенного метода, и первым идет признание его приблизительности: «критерий не всегда дает точный ответ, т. е. он является не строгим, а приближенным» [Ibid], или градуальности. То есть то, на основании чего в [15, 16] отвергается общепринятый в современной морфологии подход к проблеме словоизменение vs. словообразование через регулярность и обязательность, имея в виду градуальность регулярности.

Заканчивая обзор предложенного в [15, 16] подхода к теоретическому описанию русского вида и критериям видовой парности, остановимся на результатах их исследования. С одной стороны, в [16] авторы приходят к выводу: «основной признак отнесения грамматической категории к

[1] Вполне солидаризируясь с позицией автора [2] относительно сомнительной валидности для лингвистики данных онлайн-инструментов автоматической обработки цифровых текстов, в том числе Google Books Ngram Viewer, мы будем вынуждены, вслед за авторами [15, 16] повторить их исследовательские приемы, чтобы выявить уязвимость предложенного способа решения ряда проблем теоретической лингвистики.

[2] Словоизменение и словообразование и в [15], [16], и в данной работе, понимаются вполне традиционно для отечественной лингвистики, как, напр., в [9], [12], [17], [7].

словоизменению или словообразованию – сохранение или не сохранение лексической семантики – не позволяет разграничить префиксальный и суффиксальный способы видообразования с точки зрения их грамматического статуса. Многолетнюю дискуссию о противопоставлении грамматического статуса аспектуальных пар, образованных префиксальным и суффиксальным способами, по нашему мнению, можно считать практически завершенной» [Ibid: 1127-1128]. С другой стороны, изложенные в [15: 78] результаты анализа графиков изменения частот употребления словоформ (инфинитивов) в парах «базовый глагол – естественный перфектив» и «базовый глагол – специализированный перфектив» (далее – приложение 1 и приложение 2), построенных сервисом Ngram Viewer, на наш взгляд, не дают оснований для таких выводов. Авторы отмечают, что «в приложении 1 из 101 пары в 85 случаях (85%) имеет место высокая (согласно шкале Чеддока) r > 0,7 корреляция. Среднее значение коэффициента корреляции 0,780. В приложении 2 из 101 пары в 57 случаях (57%) имеет место высокая r > 0,7 корреляция. Среднее значение коэффициента корреляции 0,607» [Ibid]. Ниже авторы отмечают: «вероятность получения ложноположительного решения (неаспектуальная пара показывает высокий коэффициент корреляции), существенно выше, чем ложноотрицательного (аспектуальная пара имеет низкий коэффициент корреляции)» [Ibid]. Однако это можно проинтерпретировать и иным образом: более чем в половине случаев выборки (57%) с парами типа «базовый глагол – специализированный перфектив» результат аналогичен подавляющему большинству (85%) случаев в выборке «базовый глагол – естественный перфектив». Это означает, что большинство случаев в каждом типе пар, постулируемых различными, однако идентичных по морфологической структуре (СВ2 ← НСВ1), отличаются высокой степенью семантической общности, а меньшая их часть (43% и 15% соответственно) показывают существенные семантические различия, отражаемые, по мысли авторов, в несинхронном изменении частотности. С нашей точки зрения, сомнительно, что такой результат исследования дает основания для приведенного выше вывода.

Обратимся к «дистрибутивной гипотезе» как основе предложенного «частотного подхода». В [13], той публикации, на которую ссылаются авторы обсуждаемого подхода как на теоретический источник, находим уточненную формулировку гипотезы: «A distributional model accumulated from co-occurrence information contains syntagmatic relations between words, while a distributional model accumulated from information about shared neighbors contains paradigmatic relations between words» [Ibid: 40]. Пафос [13] заключается как раз в том, чтобы показать, что дистрибутивная модель выстроена на теоретической основе структурализма: множественное цитирование З. Харриса, упоминание Л. Блумфильда, обращение к обоим как к последователям основоположника структурализма Ф. де Соссюра с его понятием значимости языковой единицы и двух видов отношений – синтагматических и парадигматических, что отражено и в приведенной формулировке гипотезы. Речь, следовательно, идет именно о той теоретической основе, которая дает нам используемый, в частности, в отечественной лингвистике подход к различению алло-единиц и «эмических единиц» (вариантов одной сущности и разных сущностей) через анализ дистрибуции и ее типов: дополнительной (непересекающейся) и свободного варьирования, с одной стороны, и контрастной, с другой, см. [9: 198-201]. Дистрибутивный анализ в конечном счете лежит в основе и решения проблемы словоизменение vs. словообразование через регулярность и обязательность, характерные для словоизменения и нехарактерные для словообразования (см. [11: 249-250, 283; 17: 25-26; 13: 51; 10: 326-332; 9: 198-201], т. е. присутствует в том подходе, который отвергается в [15, 16] как недостаточно эффективный из-за градуальности свойства регулярности (кстати, в [12] обязательность также градуальна). Итак, сомнительны как противопоставление дистрибутивной гипотезы в качестве альтернативы решению проблемы словоизменение vs. словообразование через свойства обязательности и регулярности, так и выводимая прямо из этой гипотезы (с отсылкой к [13]) идея об установлении единства лексической семантики на основе синхронности диахронического изменения частот употребления словоформ одной лексемы в качестве единственного критерия словоизменения, см. [16: 1118-1120].

3 Всегда ли словоформы одной лексемы демонстрируют синхронную частотность?

Остановимся на идее авторов [15, 16], согласно которой словоформы одной лексемы демонстрируют синхронную частотность своих вхождений, в то время как частоты вхождений различных

лексем, связанных словообразовательно, изменяются не синхронно. Данная идея подается авторами как сама собой разумеющаяся, без серьезной эмпирической проверки. Думается, что это утверждение представляет собой гипотезу, требующую доказательств. Наблюдаемая частотность вхождений словоформ в парадигмах словоизменительных категорий ставит высказанную идею под сомнение[3]. Приведем примеры.

NB! В первую очередь, следует отметить, что вводимое авторами [15, 16] ограничение по времени (в одном случае установлен временной диапазон с 1920 по 2019 гг., в другом – с 1950 по 2019 гг.) представляется недостаточно оправданным. С одной стороны, на аргумент, согласно которому это делается во избежание влияния старой орфографии, можно возразить, что если изменения в орфографии касаются рассматриваемых глаголов, то, как правило, они затрагивают оба, то есть существенно повлиять на синхронную частотность двух глаголов данный фактор вряд ли может (при этом в нескольких примерах ниже, где наблюдается резкое изменение тренда для определенной словоформы около 1920-го года, рассматривается суммарная частотность для двух вариантов написания). В то же время сокращение (до 100 или 70 лет) временного промежутка, на котором рассматривается динамика частотности глаголов, не позволяет выявить никаких значимых изменений в случае грамматических категорий, изменяющихся медленно. Поскольку было эмпирически установлено, что в русских текстах первой трети XIX в. наблюдаются большие и труднообъяснимые выбросы для любой словоформы, нами было принято решение рассматривать динамику частотности начиная с 1830 г.

1) Вряд ли вызывает сомнение принадлежность к одной словоизменительной парадигме глагольных форм, различающихся временем и лицом. Так, например, словоформы *был, буду* и *будет* относятся к одной парадигме, но их частотность в GBN меняется не синхронно, см. рис. 1.

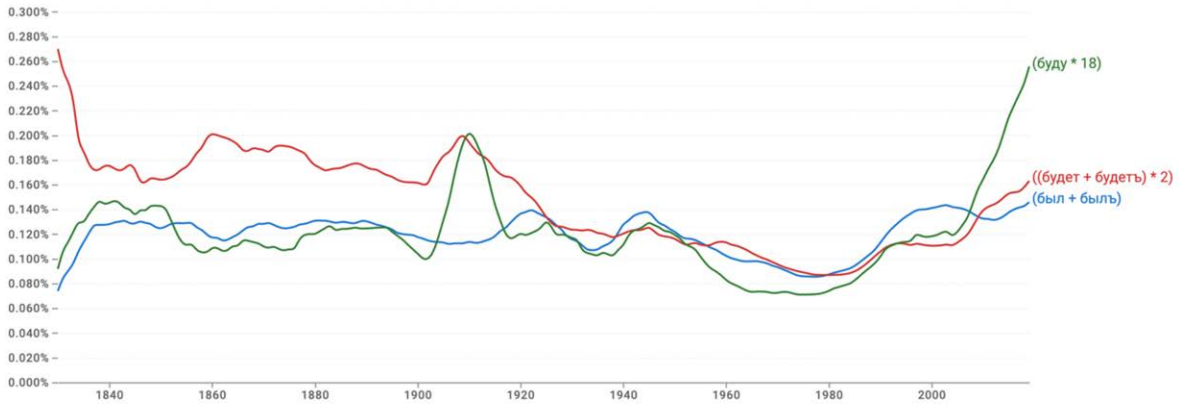


Рисунок 1: Графики частот *был – буду – будет*[4,5]

Для форм *будет* и *буду* на временном промежутке с 1830 по 2019 г. коэффициент корреляции будет невысоким (коэффициент Спирмена, r = 0,47; коэффициент Пирсона, r = 0,42)[6].

Отметим также, что утверждение о синхронном изменении частотности форм одного слова прямо противоречит тому, что известно о развитии граммем футурума и подтверждается на корпусном материале. Выражения со значением намерения, приобретая способность указывать на предсказание (но не наоборот), имеют тенденцию развиваться в граммемы с общим значением футурума (сохраняя способность к выражению значения намерения) см. [6: 310; 5: 106]. В [3] различные пути грамматикализации футурума имеют общую часть, завершаясь семантическим переходом INTENTION > FUTURE. Граммема футурума проходит стадию значения намерения — сначала говорящего, затем агенса высказывания. Далее намерение становится частью значения футурума, а значение предсказания развивается в результате переосмысления намерения третьего лица со стороны говорящего. Распределение форм 1-го и 3-го лица позволяет судить о

[3] Отметим, что статистической единицей в GBN (и в Ngram Viewer) является отдельная словоформа в ее конкретном графическом облике (при невозможности снятия омонимии и различения значений при полисемии).
[4] Степень сглаживания (“smoothong”) задается по умолчанию.
[5] Как и в [15], в случае различий в общей частотности словоформ, для наглядности и удобства сопоставления на графике частоты выравниваются путем кратного увеличения показателей низкочастотных единиц.
[6] Пример расчета коэффициентов корреляции приведен в Приложении Б.

способности граммемы к выражению значений намерения и предсказания и на этом основании судить об этапе развития, на котором находится граммема футурума, см. [6, 4].

На рис. 2 приведен аналогичный случай существенных различий в частотности употребления презентной и претеритальной формы английской проспективной конструкции *be going to+ Inf*.

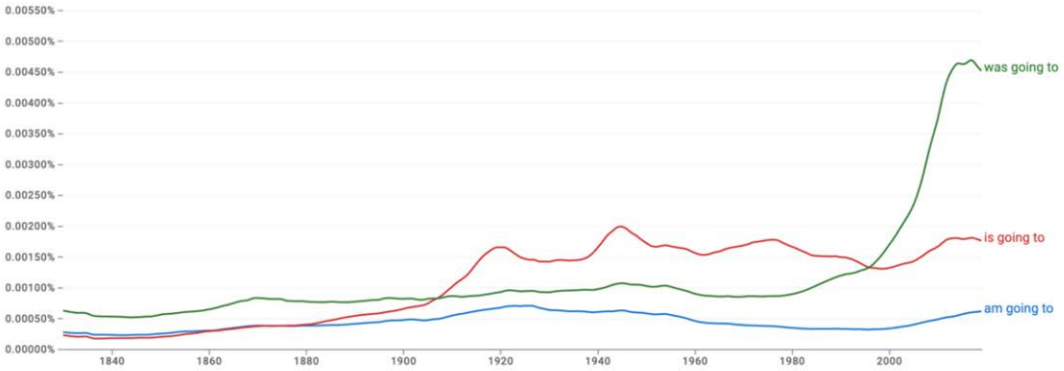


Рисунок 2: Графики частот *am – is – was going to+ Inf.*

Для форм *am going to* и *was going to* на временном промежутке 1830–2019 г. коэффициент корреляции средний или низкий (коэффициент Спирмена, r = 0,53; коэффициент Пирсона, r = 0,27).

2) С другой стороны, есть случаи, в которых высокий коэффициент корреляции наблюдается для глагольных единиц, которые не только не являются членами видовой пары (потенциально, словоформами одной лексемы), но и не связаны деривацией. Примерами служат ингестивные глаголы СВ *съесть* и *выпить* и делимитативы *поесть* и *попить,* относящиеся к одной семантической группе, но обозначающие различные ситуации и, несомненно, являющиеся разными лексемами.

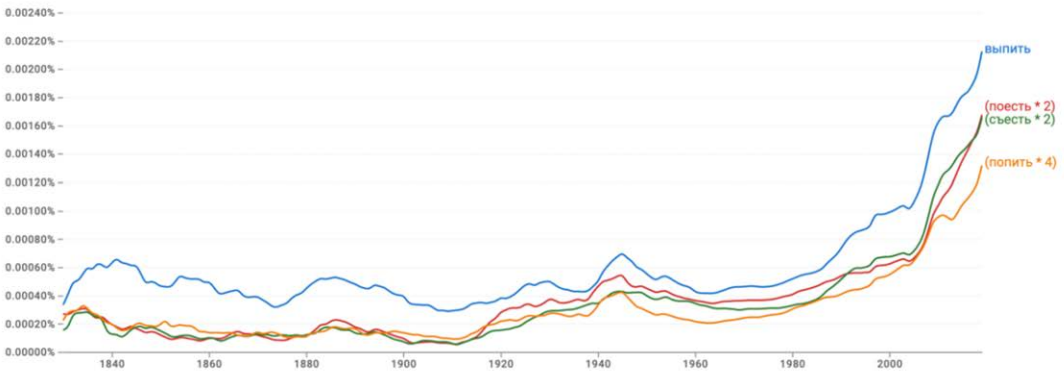


Рисунок 3: Графики частот *съесть – выпить – поесть – попить*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
выпить / попить	0.79	0.97
выпить / съесть	0.75	0.96
выпить / поесть	0.75	0.93
попить / съесть	0.93	0.98
попить / поесть	0.94	0.97
поесть / съесть	0.98	0.99

Таблица 1: Попарные коэффициенты корреляции для *съесть – выпить – поесть – попить*

Следует отметить, что довольно высокий коэффициент корреляции наблюдается для всех (!) парных комбинаций из четырех глаголов, в том числе для *попить*/*съесть* и *выпить*/*поесть*, не обнаруживающих единства ни корня (носителя лексического значения), ни префикса.

На рис. 4 и 5 приведены аналогичные случаи не из области глагольной морфологии: на рис. 4 на материале личных местоимений, на рис. 5 – на однокоренных существительных. В случае местоимений словоформы демонстрируют низкую корреляцию по частотности, а в случае разных, хотя и однокоренных, существительных – высокую (в терминах [15] – ложноотрицательный и ложноположительный результаты).

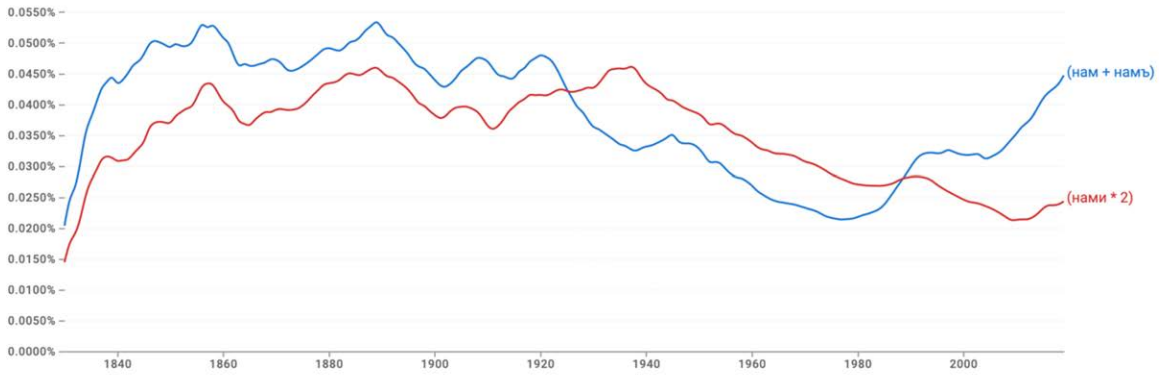


Рисунок 4: Графики частот местоименных форм *нам – нами*

Для форм *нам* и *нами* на промежутке 1830–2019 коэффициент корреляции невысокий (коэффициент Спирмена, r = 0,6; коэффициент Пирсона, r = 0,58).

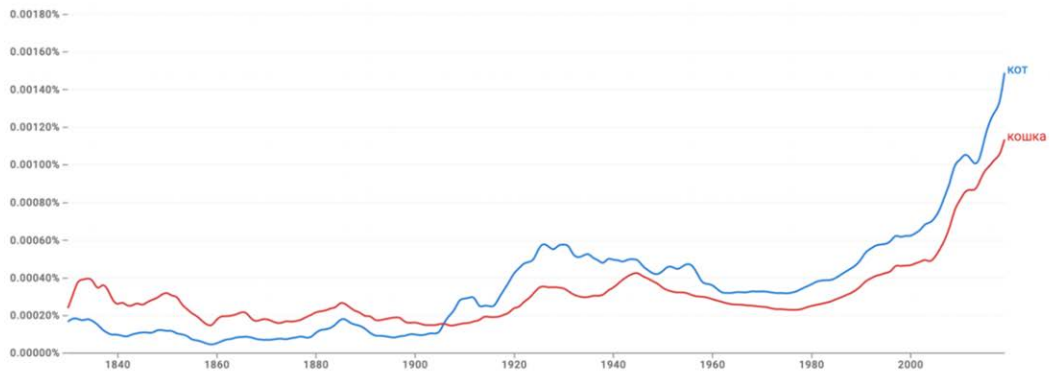


Рисунок 5: Графики частот существительных *кот – кошка*

Для словоформ *кот*[7] и *кошка* на промежутке 1830–2019 коэффициент корреляции высокий (коэффициент Спирмена, r = 0,82; коэффициент Пирсона, r = 0,9).

3) Авторы делают оговорку о возможности как ложноположительного («неаспектуальная пара показывает высокий коэффициент корреляции»), так и ложноотрицательного результата (аспектуальная пара имеет низкий коэффициент корреляции») [15: 78], при этом вероятность получения ложноположительного результата оказалась выше, чем ложноотрицательного. Ценность такого вывода была бы значительно выше при понимании того, каким образом ложноположительные и ложноотрицательные результаты распределены за пределами рассматриваемых пар. Наличие ложных результатов обоих типов подтверждается приведенными выше примерами. Однако отсутствие данных о системе не позволяет ответить на вопросы: 1) действительно ли распределение ложноположительных и ложноотрицательных результатов дает основание принимать гипотезу о синхронном изменении частотности при словоизменении; 2) можно ли говорить о том, что распределение ложноотрицательных и ложноположительных результатов в [15, 16] соотносимо с таковым в системе языка в целом.

[7] В [2] подмечена неожиданная картина частотности графического слова *кот* в текстах до 1920-х гг., в связи с чем его корреляция со словоформой *кошка* тем более необъяснима.

4 Дистрибутивная гипотеза и частотный подход к видовым тройкам

Еще один ракурс валидности обсуждаемого подхода к решению проблемы словоизменение vs. словообразование через оценку синхронности изменений частотности словоформ в диахронии на больших данных можно увидеть, применив предложенный подход к видовым, или морфологическим биимперфективным, тройкам типа СВ – НСВ1/НСВ2: *съесть – есть/съедать* [18: 235], в которых оба НСВ «претендуют на роль видового коррелята» к СВ [Ibid: 236][8,9]. Будем ориентироваться на именно так понимаемые видовые тройки, причем исключительно «образцовые» [Ibid: 241]: те, в которых оба НСВ, бесприставочный и приставочный (каждый из них реален, а не потенциален), претендуют на роль коррелята к приставочному СВ[10]. При этом все три единицы имеют общий корень и способны обозначать одну и ту же ситуацию, в силу чего характеризуются единством лексического значения.

В [8] представлено диахроническое исследование троек на материале НКРЯ (https://ruscorpora.ru/). Поскольку есть возможность опереться на результаты упомянутого исследования, из десяти троек в [8] отберем шесть – те, для которых частотность НСВ2 не исчезающе низка: *свариться — вариться / свариваться*[11] (также без *-ся: сварить — варить / сваривать*), *съесть — есть / съедать, оторвать — рвать / отрывать, пробить — бить / пробивать, сгореть — гореть / сгорать, сорвать — рвать / срывать*, добавив к ним тройки *разбить — бить / разбивать*, *разорвать — рвать / разрывать*, *намазать — мазать / намазывать, налить — лить / наливать*. Тем самым материалом настоящего исследования являются 11 видовых троек.

Рассмотрим коэффициенты корреляции для всех теоретически возможных парных комбинаций в тройках: СВ2/НСВ1, СВ2/НСВ2 и НСВ1/НСВ2. В табл. 2 и 3 приведены коэффициенты корреляции Спирмена для трех парных комбинаций (без учета порядка следования) в 11 рассматриваемых тройках. Соответствующие графики и расчеты коэффициентов корреляции Спирмена и Пирсона приведены в Приложении А.

Аспектуальная тройка	**СВ2 / НСВ1**	**СВ2 / НСВ2**	**НСВ1 / НСВ2**
съесть — есть / съедать	0.87	0.88	0.77
налить — лить / наливать	0.15	0.88	0.31
разорвать — рвать / разрывать	-0.28	0.88	-0.53
сорвать — рвать / срывать	0.69	0.83	0.84
намазать — мазать / намазывать	0.73	0.8	0.73
оторвать — рвать / отрывать	0.82	0.71	0.41
пробить — бить / пробивать	0.26	0.71	0.27
сгореть — гореть / сгорать	0.89	0.67	0.74
сварить — варить / сваривать	0.89	0.58	0.44
разбить — бить / разбивать	0.13	0.42	0.42
свариться — вариться / свариваться	0.38	0.16	-0.03
Среднее	**0.50**	**0.68**	**0.40**
Медиана	**0.69**	**0.71**	**0.42**

Таблица 2: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ2)

Коэффициент корреляции < 0.7 в парах СВ2 / НСВ2 наблюдается для троек, где НСВ2 низкочастотен (*сваривать(ся)*) и/или НСВ1 и НСВ2 проблемно взаимозаменяемы (*гореть / сгорать*), или же коэффициент низкий, но выше, чем для СВ2 / НСВ1 (*разбить — бить / разбивать*).

[8] Общая оценка видовых троек: это «неотъемлемая составляющая русской аспектуальной системы. <…> Тот факт, что многие русские приставочные глаголы сов. вида вступают в <…> корреляцию с двумя разными глаголами несов. вида, означает только то, что в этом участке системы, помимо собственно аспектуальной корреляции, в игре участвуют также парадигматические отношения лексической синонимии» [18: 247].

[9] В [14], более ранней работе, выполненной тем же исследовательским коллективом, что и в [15, 16], и с применением того же Ngram Viewer, рассматриваются видовые тройки (триплеты) с точки зрения их эволюции за два века. С выводом авторов о том, что «доля глаголов несовершенного вида уменьшается, а вторичные имперфективы вообще вымываются из языка» [14: 425], данные нашего исследования согласиться не позволяют.

[10] Вслед за [7] обозначим приставочный СВ как СВ2, считая первичным СВ, т. е. СВ1, перфектив-симплекс типа *дать*.

[11] Графическая подача и порядок следования членов тройки соответствует принятому в [18: 242-247].

Аспектуальная тройка	СВ2 / НСВ1	СВ2 / НСВ2	НСВ1 / НСВ2
сварить — варить / сваривать	0.89	0.58	0.44
сгореть — гореть / сгорать	0.89	0.67	0.74
съесть — есть / съедать	0.87	0.88	0.77
оторвать — рвать / отрывать	0.82	0.71	0.41
намазать — мазать / намазывать	0.73	0.8	0.73
сорвать — рвать / срывать	0.69	0.83	0.84
свариться — вариться / свариваться	0.38	0.16	-0.03
пробить — бить / пробивать	0.26	0.71	0.27
налить — лить / наливать?	0.15	0.88	0.31
разбить — бить / разбивать	0.13	0.42	0.42
разорвать — рвать / разрывать	-0.28	0.88	-0.53
Среднее	**0.50**	**0.68**	**0.40**
Медиана	**0.69**	**0.71**	**0.42**

Таблица 3: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ1)

Общие наблюдения:

1) по медианному уровню коэффициента корреляции Спирмена из трех возможных пар глаголов, формируемых на базе видовой тройки, наиболее высок уровень корреляции в СВ2 / НСВ2 (0,71), непосредственно за ней следует СВ2 / НСВ1 (0,69) и значительно отстает НСВ1 / НСВ2 (0,42); ту же тенденцию выявляет и среднее значение этого коэффициента, хотя оно менее устойчиво к выбросам: 0,68 – 0,50 и 0,40;
2) высокий (с коэффициентом Спирмена выше 0,7) уровень корреляции на большем количестве случаев (семь из 11) отмечен для СВ2 / НСВ2, для СВ2 / НСВ1 он отмечается в пяти случаях из 11; для НСВ1 / НСВ2 -- в трех случаях, еще в двух -- обратная корреляция;
3) обобщая предшествующие наблюдения, можно сделать вывод о том, что по итогам анализа диахронической частотности на 11 видовых тройках наибольшей согласованностью изменения таковой по обоим параметрам (медиана и среднее по выборке, а также количество случаев с высоким уровнем коэффициента Спирмена) характеризуется пара СВ2 / НСВ2 (две префиксальные формы), сразу за ней следует СВ2 / НСВ1, иную картину мы видим в случае НСВ1 / НСВ2.

5 Выводы и перспективы

Имея в виду риски экстраполяции выводов, сделанных на основе небольшого по охвату материала исследования, на всю языковую систему и не претендуя на окончательность формулировок, приведем несколько более общих соображений, вытекающих из проведенного исследования.

Главный вывод: отказ от широко применяемого в современной лингвистике подхода к решению проблемы словоизменение vs. словообразование через критерии обязательности и регулярности в пользу обращения к количественному подходу посредством учета диахронического изменения частотности словоформ как критерия единства лексического значения (в частности, с использованием GBN) не представляется оправданным в силу целого ряда причин.

I. Прежде всего, это сомнения в обоснованности предложенного подхода дистрибутивной гипотезой [13] и в его предпочтительности по сравнению с использованием критериев обязательности и регулярности, являющихся, в конечном итоге, производной от анализа типов дистрибуции рассматриваемой единицы.
II. Далее, это так называемые ложноположительные и ложноотрицательные результаты применения предложенного в [15, 16] подхода, т. е. наличие случаев, когда явно разные лексемы показывают высокий уровень коэффициента корреляции диахронической частотности (*кот* и *кошка*, ингестивные *съесть, выпить, поесть, попить*), и случаев низкого уровня корреляции словоформ одной лексемы (*был, буду, будет*; *нам* и *нами*; *am going to* и *was going to*).

III. Наконец, наше исследование в рамках предложенного в [15, 16] подхода на материале русских аспектуальных троек показало, что уровень корреляции частотности членов видовой тройки, самим фактом своего вхождения в нее характеризуемых высокой степенью близости лексической семантики, существенно различается при попарном разбиении тройки, см. наблюдения 1-2 выше. Полученные результаты подтверждают трактовку НСВ1 в видовой тройке как «джокера», «выполняющего чужие функции» [1: 106]. С учетом результатов исследования, позволим себе расширить это понятие и утверждать, что «джокером» симплекс НСВ1 выступает не только в случае замещения приставочного НСВ2 (без различий по виду), но и выступая аспектуальным партнером приставочного СВ2. Как показывают результаты исследования изменения частотности, во втором случае (НСВ1 как видовой партнер для СВ2) роль «джокера» симплекс выполняет значительно лучше, чем в первом (см. наблюдения в разделе 4 относительно двоек СВ2 / НСВ1 и НСВ1 / НСВ2), на чем и базируется понятие приставочной видовой пары.

Завершая подведение итогов применения анализа частотности членов видовой тройки в диахронии на материале GBN, отметим, что существенно различные результаты по тройкам (Приложение А) позволяют предположить, что симплексы НСВ1 с неодинаковой легкостью выступают в роли «джокера» как для СВ2, так и для НСВ2, что может быть обусловлено различными факторами, в том числе связанными с лексической семантикой симплекса, степенью его полисемичности, востребованностью в качестве синонима для приставочных СВ2 и НСВ2 в профессиональных языках [18: 248-257], явлением депрефиксации [Ibid: 274] и др. Все это подлежит дальнейшему изучению[12].

Условные обозначения

НСВ – несовершенный вид; СВ – совершенный вид; НСВ1 – симплекс НСВ (типа *лить*), СВ1 – симплекс СВ (типа *дать*); СВ2 – приставочный глагол СВ, дериват НСВ1 (типа *про-лить*); НСВ2 – т. наз. вторичный имперфектив, дериват СВ2, обладающий префиксом и суффиксом имперфективации (типа *про-ли-ва-ть*).

Благодарности

Авторы выражают благодарность Е. В. Еникеевой за помощь в работе с данными НКРЯ.

References

[1] Apresjan Yu. D. Interpretation of redundant aspectual paradigms in the defining dictionary // Apresjan Yu. D. Selected works. Vol. II: Integrated description of language and systematic lexicography [Izbrannye trudy. Vol. II: Integral'noe opisanie yazyka i sistemnaya leksikografiya]. — Moscow: Yazyki Russkoi Kul'tury, 1995. — P. 103–114.

[2] Belikov V. I. (2016), What and how can a linguist get from digitized texts? [Chto i kak mozhet poluchit' lingvist iz ocifrovannyh tekstov?], Siberian Journal of Philology [Sibirskij filologicheskij zhurnal], 3, pp. 17-34.

[3] Bybee J. L., Perkins R., Pagliuca W. (1994), The evolution of grammar: tense, aspect and modality in the languages of the world. Chicago: University of Chicago Press.

[4] Chuikova O. Iu. (2018), The Semantics of Future in Russian, English and Spanish: The Interaction of Temporality, Aspectuality and Modality. Candidate of Philology Thesis [Semantika budushchego vremeni v russkom, anglijskom i ispanskom yazykah (vzaimodejstvie temporal'nosti, aspektual'nosti i modal'nosti): diss. na soiskanie stepeni kand. filol. nauk]. St. Petersburg.

[5] Dahl Ö. (1985), Tense and Aspect Systems. — Oxford: Blackwell.

[6] Dahl Ö. (2000), The grammar of future time reference in European languages // Ö. Dahl (ed.). Tense and Aspect in the Languages of Europe. — Berlin, New York: Mouton de Gruyter. — P. 309–328.

[7] Gorbova E.V. (2017), Aspectual formation of Russian verbs: Inflection, derivation, or a set of quasigrammemes? ("Sore points" of Russian aspectology revisited) [Russkoe vidoobrazovanie: slovoizmenenie, slovoklassifikaciya ili nabor kvazigrammem? (eshche raz o bolevyh tochkah russkoj aspektologii)], Topics in the study of language [Voprosy yazykoznaniya], 1, pp. 24–52.

[12] Отметим также тот факт, что проведение аналогичного исследования тех же 11 аспектуальных троек на таком источнике языкового материала, как НКРЯ (см. Приложение В), дает при общем взгляде на средние величины сходные, однако разнящиеся относительно конкретных троек, результаты.

[8] Gorbova E.V. (2020), Aspectual triplets of the Russian verb in diachrony: Evidence from the Russian National Corpus [Vidovye trojki russkogo glagola v diahronii (na materiale NKRYa)], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2020” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2020”], pp. 321-347. DOI: 10.28995/2075-7182-2020-19-321-347
[9] Kasevich V.B. (2019), Problems of Semantics [Problemy semantiki]. — St. Petersburg: St. Petersburg University Press.
[10] Maslov Yu.S. (2004), Selected Works: Aspectology. General Linguistics [Izbrannye trudy: Aspektologiya. Obshchee yazykoznanie]. — Moscow, Yazyki slavyanskoi kul'tury Publ.
[11] Mel'chuk I.A. (1997), Course in General morphology [Kurs obshchej morfologii]. Vol. 1. — Moscow, Vienna: Yazyki russkoj kul'tury, Venskij slavisticheskij al'manah, Izdatel'skaya gruppa «Progress».
[12] Plungian V. A. (2011), Introduction to grammatical semantics: Grammatical meanings and grammatical systems of languages of the world [Vvedenie v grammaticheskuyu semantiku. Grammaticheskie znacheniya i grammaticheskie sistemy yazykov mira]. — Moscow: RGGU, 2011.
[13] Sahlgren, M. (2008), The Distributional Hypothesis. From context to meaning, Distributional models of the lexicon in linguistics and cognitive science [Special issue], Rivista di Linguistica, Vol 20(1), pp. 33–53.
[14] Solovyev V.D., Bochkarev V.V. (2015), Evolution of the aspectual triplets in Russian through the prism of Google Ngram [Evolucija aspectual’nyh tripletov v russkom jazyke cherez prizmu Google Ngram], Proceedings of the international conference «Corpus linguistics – 2015» [Trudy mezhdunarodnoj konferencii «Korpusnaja lingvistika – 2015»]. St. Petersburg, SPbSU [SPbGU], pp. 425-434.
[15] Solovyev V.D., Bochkarev V.V. (2022), The case for aspectual pairs reopened [Delo ob aspektual'nyh parah otkryvaetsya vnov'], Tomsk State University Journal of Philology [Vestnik Tomskogo gosudarstvennogo universiteta. Filologiya], Vol. 78, pp. 67–98. DOI: 10.17223/19986645/78/4
[16] Solovyev V., Bochkarev V., Bayrasheva V. (2022), Aspectual pairs: Prefix vs. sufix way o formation, Russian Journal of Linguistics, Vol. 26(4), pp. 1114–1135. DOI: 10.22363/2687-0088-27394.
[17] Zaliznyak A.A. (2002), «Russian nominal inflection» with selected works on Modern Russian and general linguistics [«Russkoe imennoe slovoizmenenie» s prilozheniem izbrannykh rabot po sovremennomu russkomu yazyku i obshchemu yazykoznaniyu]. Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul’tury], 2002.
[18] Zalizniak Anna A., Mikaelyan I.L., Shmelev A.D. (2015), Russian aspectology: In defense of the aspectual pair [Russkaya aspektologiya: v zashchitu vidovoi pary] — Moscow: Languages of Slavic Culture [Yazyki Slavyanskoi Kul’tury].

Приложение А. Графики и попарные коэффициенты корреляции в аспектуальных тройках

1) *свариться — вариться / свариваться*

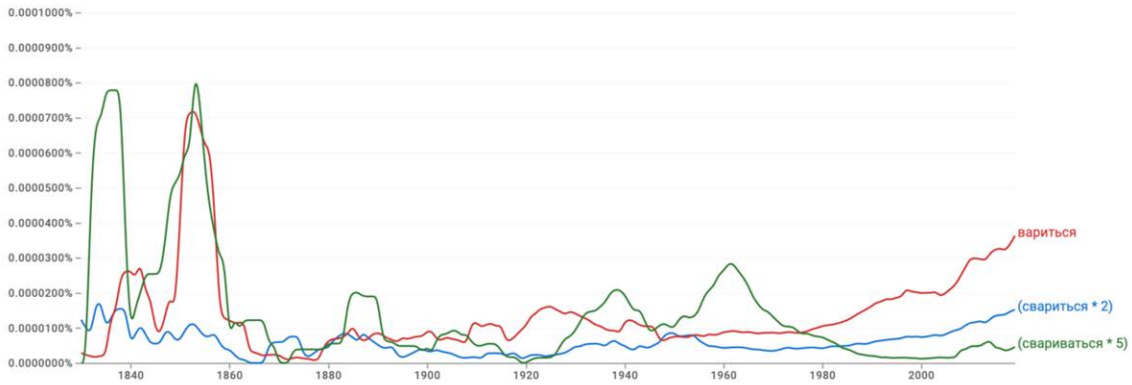


Рисунок А1: Графики частот *вариться – свариться – свариваться*

Пара	коэффициент Спирмена	коэффициент Пирсона
свариться / вариться	0.38	0.45
свариться / свариваться	0.16	0.43
вариться / свариваться	-0.03	0.4

Таблица А1: Попарные коэффициенты корреляции в тройке *свариться — вариться / свариваться*

2) *сварить — варить / сваривать*

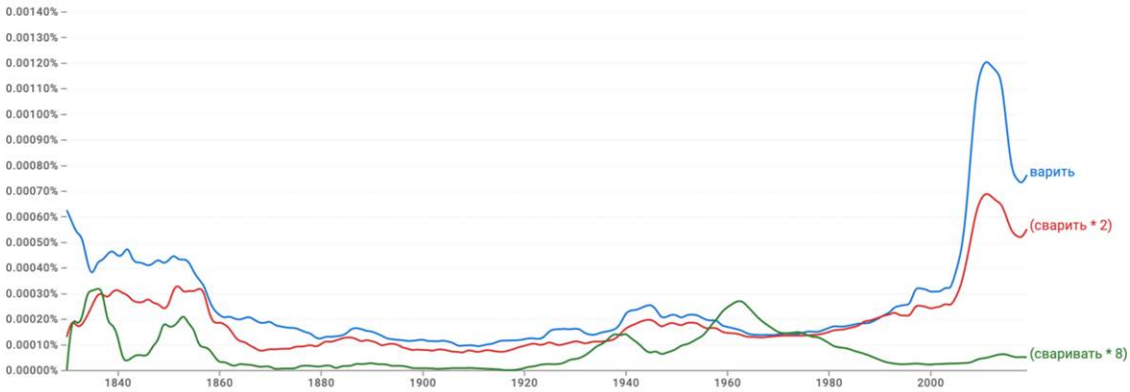


Рисунок А2: Графики частот *варить – сварить – сваривать*

Пара	коэффициент Спирмена	коэффициент Пирсона
сварить / варить	0.89	0.95
сварить / сваривать	0.58	0.14
варить / сваривать	0.44	0.11

Таблица А2: Попарные коэффициенты корреляции в тройке *сварить — варить / сваривать*

3) *съесть — есть / съедать*[13]

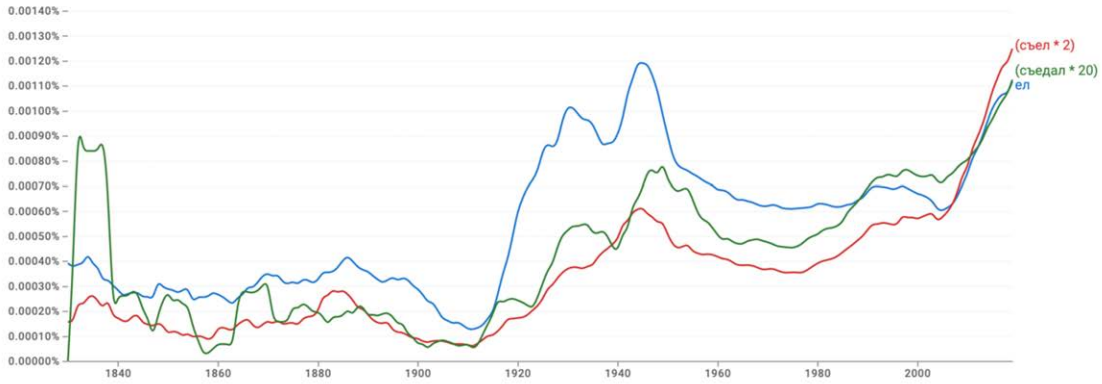


Рисунок А3: Графики частот *ел – съел – съедал*

Пара	коэффициент Спирмена	коэффициент Пирсона
съел / ел	0.87	0.8
съел / съедал	0.88	0.86
ел / съедал	0.77	0.74

Таблица А3: Попарные коэффициенты корреляции в тройке *съесть — есть / съедать*

4) *оторвать — рвать / отрывать*

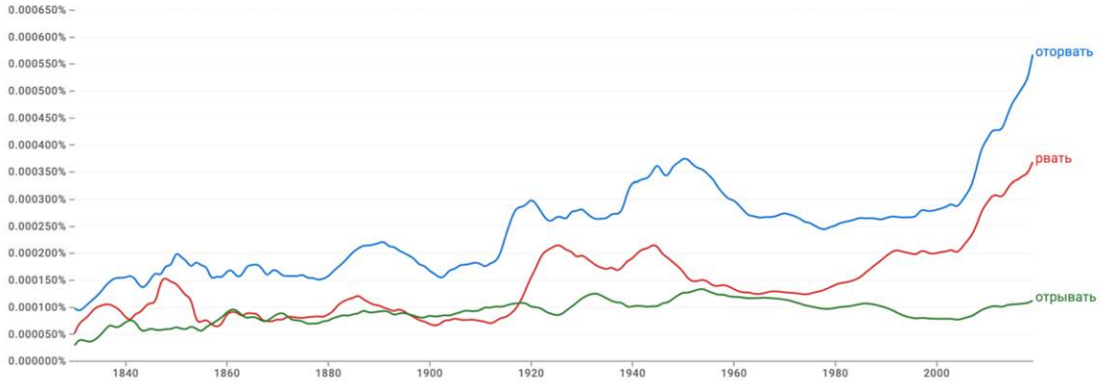


Рисунок А4: Графики частот *рвать – оторвать – отрывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
оторвать / рвать	0.82	0.88
оторвать / отрывать	0.71	0.67
рвать / отрывать	0.41	0.36

Таблица А4: Попарные коэффициенты корреляции в тройке *оторвать — рвать / отрывать*

[13] Для этой тройки рассмотрение инфинитивных форм приводит к некорректным результатам в связи с омонимией инфинитива *есть* и презентной формы глагола *быть*, поэтому исследование проведено на формах претерита.

5) *пробить — бить / пробивать*

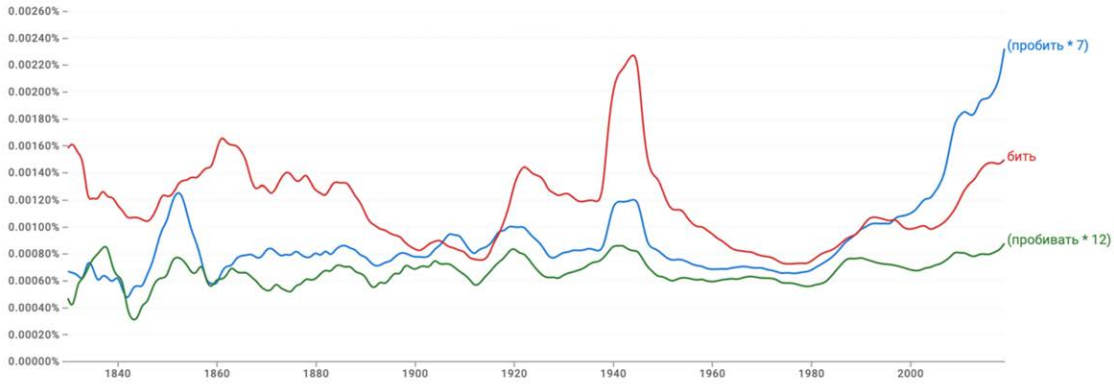


Рисунок А5: Графики частот *бить – пробить – пробивать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
пробить / бить	0.26	0.28
пробить / пробивать	0.71	0.62
бить / пробивать	0.27	0.36

Таблица А5: Попарные коэффициенты корреляции в тройке *пробить — бить / пробивать*

6) *сгореть — гореть / сгорать*

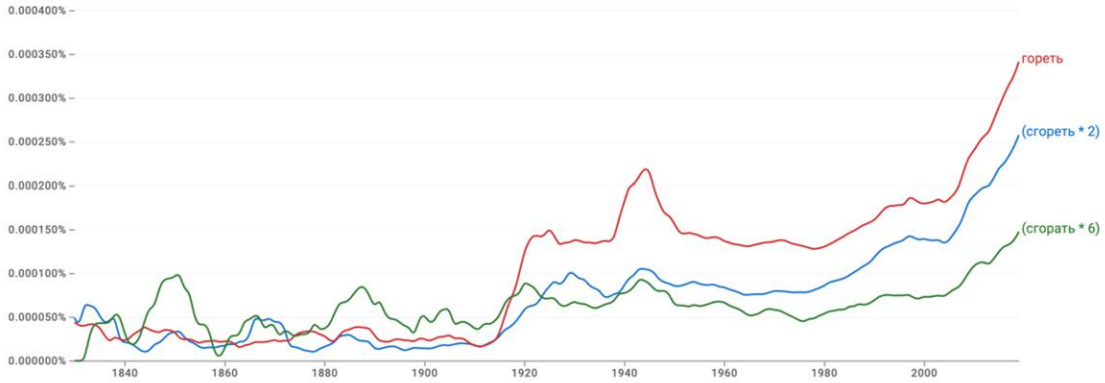


Рисунок А6: Графики частот *гореть – сгореть – сгорать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
сгореть / гореть	0.89	0.95
сгореть / сгорать	0.67	0.72
гореть / сгорать	0.74	0.74

Таблица А6: Попарные коэффициенты корреляции в тройке *сгореть — гореть / сгорать*

7) *сорвать — рвать / срывать*

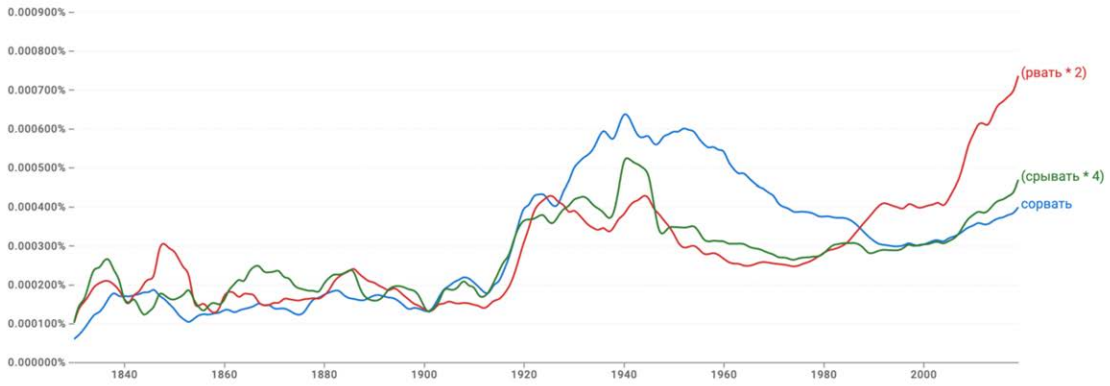


Рисунок А7: Графики частот *рвать – сорвать – срывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
сорвать / рвать	0.69	0.57
сорвать / срывать	0.83	0.85
рвать / срывать	0.84	0.79

Таблица А7: Попарные коэффициенты корреляции в тройке *сорвать — рвать / срывать*

8) *разорвать — рвать / разрывать*

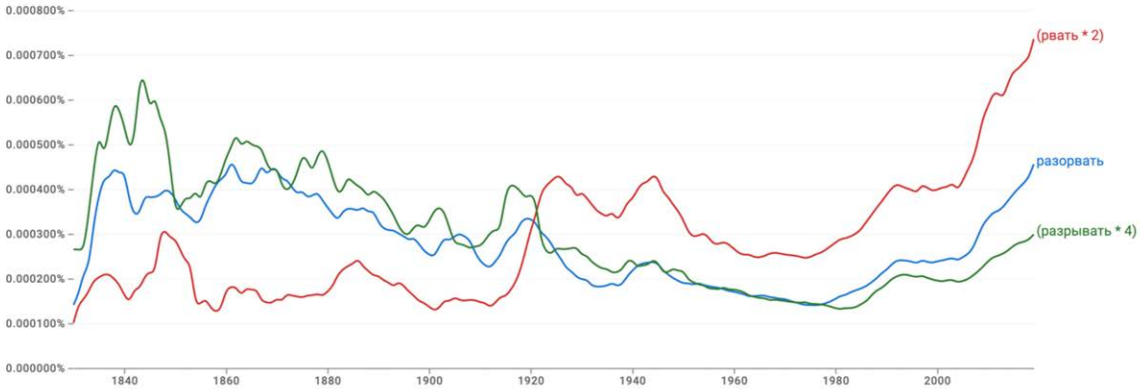


Рисунок А8: Графики частот *рвать – разорвать – разрывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
разорвать / рвать	-0.28	-0.13
разорвать / разрывать	0.88	0.85
рвать / разрывать	-0.53	-0.45

Таблица А8: Попарные коэффициенты корреляции в тройке *разорвать — рвать / разрывать*

9) *разбить — бить / разбивать*

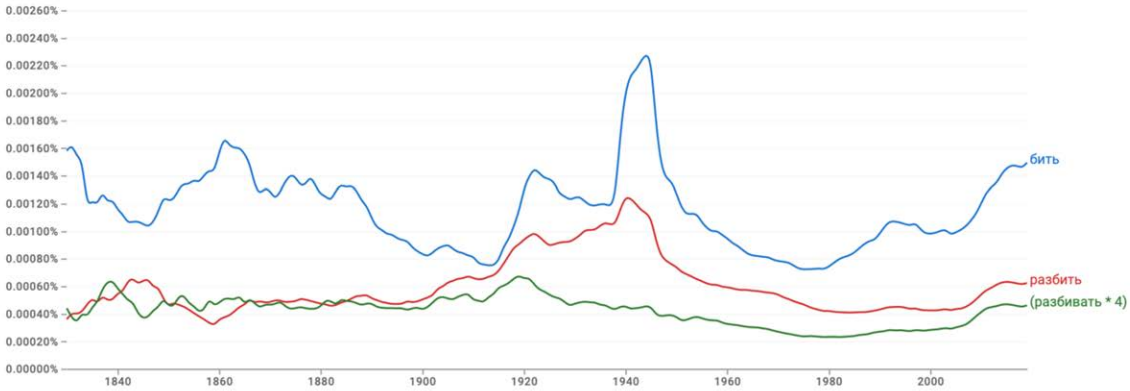


Рисунок А9: Графики частот *бить – разбить – разбивать*

Пара	коэффициент Спирмена	коэффициент Пирсона
разбить / бить	0.13	0.41
разбить / разбивать	0.42 (0.4199)	0.39 (0.3906)
бить / разбивать	0.42 (0.4181)	0.39 (0.3934)

Таблица А9: Попарные коэффициенты корреляции в тройке *разбить — бить / разбивать*

10) *намазать — мазать / намазывать*

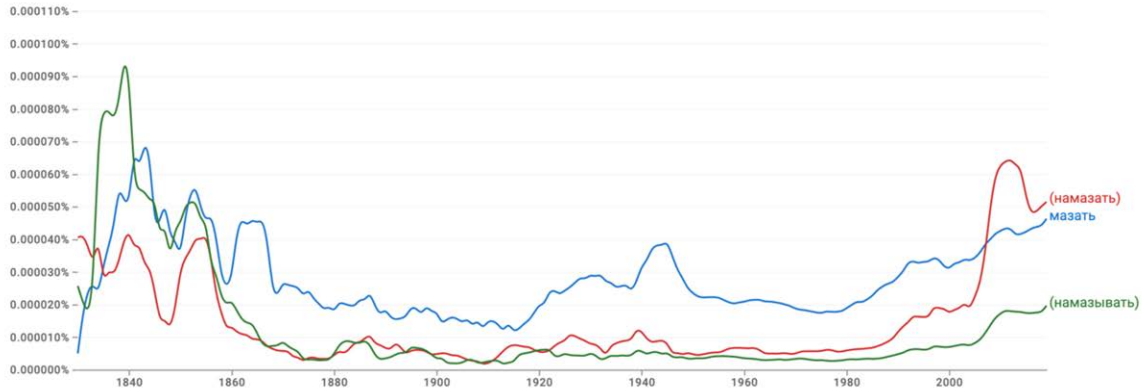


Рисунок А10: Графики частот *мазать – намазать – намазывать*

Пара	коэффициент Спирмена	коэффициент Пирсона
намазать / мазать	0.73	0.66
намазать / намазывать	0.80	0.60
мазать / намазывать	0.73	0.67

Таблица А10: Попарные коэффициенты корреляции в тройке *намазать — мазать / намазывать*

11) *лить — налить / наливать*

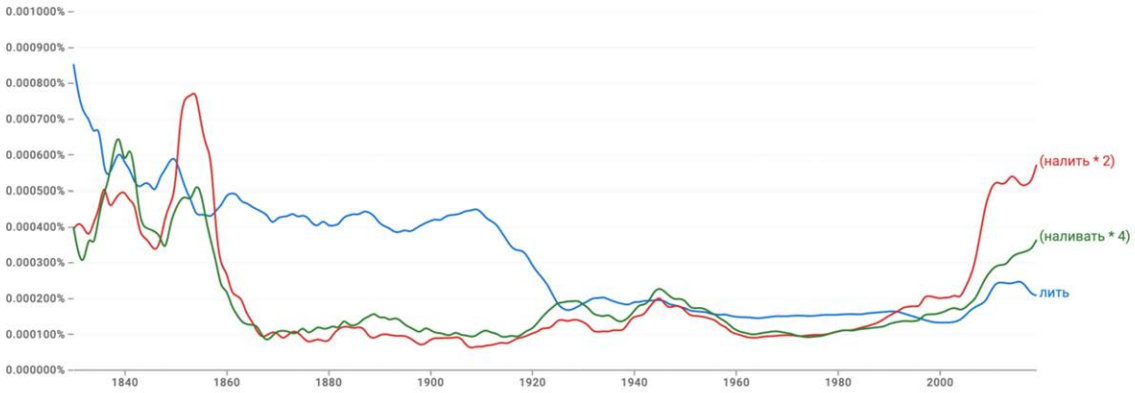


Рисунок А11: Графики частот *лить – налить – наливать*

Пара	**коэффициент Спирмена**	**коэффициент Пирсона**
налить / лить	0.15	0.36
налить / наливать	0.88	0.89
лить / наливать	0.31	0.48

Таблица А11: Попарные коэффициенты корреляции в тройке *налить — лить / наливать*

Приложение Б. Пример кода на Python для расчета коэффициентов корреляции

```
#https://www.geeksforgeeks.org/scrape-google-ngram-viewer-using-python/
import pandas as pd
import scipy.stats as stats
import requests
import urllib

def run_query(query, start_year=1830, end_year=2019,
              corpus='ru-2019', smoothing=3):
    query = urllib.parse.quote(query)
    url = (
        'https://books.google.com/ngrams/json?content=' +
        query + '&year_start=' + str(start_year) + '&year_end=' +
        str(end_year) + '&corpus=' + str(corpus) + '&smoothing=' +
        str(smoothing) + ''
    )

    response = requests.get(url)
    output = response.json()
    return_data = []
    if len(output) == 0:
        return "Нет данных для Ngram."
    else:
        for i in range(len(output)):
            return_data.append(output[i]['timeseries'])
    return return_data[0]

df = pd.DataFrame(
    {
        'год':list(range(1830,2020)),
        'варить':run_query('варить'),
        'сварить':run_query('сварить')
    }
)

df[['варить', 'сварить']] = (
    df[['варить', 'сварить']]
    .apply(lambda x: x*100)
) #переводим значения для всех столбцов, кроме "год", в проценты

# коэффициент корреляции Спирмена
rho, p = stats.spearmanr(df['сварить'], df['варить'])
print(rho, p)

#коэффициент корреляции Пирсона
df['сварить'].corr(df['варить'])
```

Приложение В. Данные дополнительного исследования аспектуальных троек, проведенного по НКРЯ (ruscorpora.ru/chart)

Аспектуальная тройка	СВ2/НСВ1	СВ2/НСВ2	НСВ1/НСВ2
оторвать – рвать/отрывать	0,71	0,59	0,59
съел – ел/съедал	0,71	0,49	0,21
сварить – варить/сваривать	0,70	0,26	0,49
намазать – мазать/намазывать	0,42	0,05	0,34
сгореть – гореть/сгорать	0,41	0,37	0,28
разбить – бить/разбивать	0,29	0,16	0,50
сорвать – рвать/срывать	0,28	0,50	0,51
свариться – вариться/свариваться	0,12	0,09	0,06
налить – лить/наливать	-0,19	0,15	0,08
пробить – бить/пробивать	-0,28	0,75	-0,40
разорвать – рвать/разрывать	-0,28	0,34	-0,23
Среднее	**0,26**	**0,34**	**0,22**
Медиана	**0,29**	**0,34**	**0,28**

Таблица В1: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ1)

Аспектуальная тройка	СВ2/НСВ1	СВ2/НСВ2	НСВ1/НСВ2
пробить – бить/пробивать	-0,28	0,75	-0,40
оторвать – рвать/отрывать	0,71	0,59	0,59
сорвать – рвать/срывать	0,28	0,50	0,51
съел – ел/съедал	0,71	0,49	0,21
сгореть – гореть/сгорать	0,41	0,37	0,28
разорвать – рвать/разрывать	-0,28	0,34	-0,23
сварить – варить/сваривать	0,70	0,26	0,49
разбить – бить/разбивать	0,29	0,16	0,50
налить – лить/наливать	-0,19	0,15	0,08
свариться – вариться/свариваться	0,12	0,09	0,06
намазать – мазать/намазывать	0,42	0,05	0,34
Среднее	**0,26**	**0,34**	**0,22**
Медиана	**0,29**	**0,34**	**0,28**

Таблица В2: Попарный коэффициент корреляции Спирмена в тройках (сортировка по убыванию в столбце СВ2 / НСВ2

Комментарий к таблицам В1 и В2

Общие средние результаты по Таблицам 2 и 3 основного текста статьи (результаты исследования по GBN) и по Таблицам В1 и В2 (результаты аналогичного исследования по НКРЯ) не совпадают количественно, однако совпадают по парам словоформ: самый высокий коэффициент Спирмена в среднем в парах СВ2 / НСВ2, далее следуют СВ2 / НСВ1, на последнем месте пары НСВ1 / НСВ2. При этом данные по отдельным тройкам, полученные по двум разным источникам материала (GBN и НКРЯ) существенно разнятся (ср. данные из Таблиц 2 и 3 и Таблиц В1 и В2).

		СВ2 / НСВ1	СВ2 / НСВ2	НСВ1 / НСВ2
GBN	среднее	0,50	0,68	0,40
	медиана	0,69	0,71	0,42
НКРЯ	среднее	0,26	0,34	0,22
	медиана	0,29	0,34	0,28

Таблица В3: Сравнение общих средних результатов коэффициентов корреляции Спирмена в тройках (по GBN и по НКРЯ)

Приложение Г. Пример кода на Python для расчета коэффициентов корреляции (НКРЯ)

```
import pandas as pd
import scipy.stats as stats
import requests
import urllib

def run_query_w_smoothing(query, start_year=1830, end_year=2020,
                          smoothing=3):
    query = urllib.parse.quote(query)
    url = (
        'https://processing.ruscorpora.ru/graphic.xml?env=alpha' +
        '&mode=graphic_main&mycorp=&mysent=&mysize=&mysentsize=' +
        '&mydocsize=&dpp=100&spp=&spd=1&text=lexform' +
        '&sort=i_year_created&g=i_year_created&lang=ru&nodia=1&req=' +
        query + ',&startyear=' + str(start_year) + '&endyear=' +
        str(end_year) + '&smoothing=' + str(smoothing) + '&format=json' +
        '&total=2&showChart=false&tableIsRender=false'
    )

    response = requests.get(url)
    output = response.json()
    year_freq_dict = {}
    for i in range(len(output['values'][0]['data'])):
        year_freq_dict[output['values'][0]['data'][i][0]]output[
            'values'][0]['data'][i][1]
    smoothed_data = []
    for year in range(start_year, end_year+1):
        start_smoothing = max(0, year - smoothing)
        end_smoothing = min(end_year, year + smoothing)
        freq_sum = 0
        for y in range(start_smoothing, end_smoothing+1):
            freq_sum += year_freq_dict.get(str(y), 0)
        smoothed_data.append(freq_sum / (2 * smoothing + 1))
    s = pd.Series(smoothed_data, index=range(start_year, end_year+1))
    return s

df = pd.DataFrame(
    {
        'варить': run_query_w_smoothing('варить'),
        'сварить': run_query_w_smoothing('сварить')
    }
)

# коэффициент корреляции Спирмена
rho, p = stats.spearmanr(df['сварить'], df['варить'])
print(rho, p)
#коэффициент корреляции Пирсона
df['сварить'].corr(df['варить'])
```

Computer-assisted detection of typologically relevant semantic shifts in world languages*

Ilya Gruntov
Institute of Linguistics
Bol.Kislovsky per. 1/12,
Moscow, 125009, Russia
altaica@narod.ru

Elisei Rykov
HSE University
20 Myasnitskaya ulitsa,
Moscow, 101000, Russia
esrykov@edu.hse.ru

Abstract

The paper contains the description of a semi-authomatic method for the detection of typologically relevant semantic shifts in the world's languages. The algorithm extracts colexified pairs of meanings from polysemous words in digitised bilingual dictionaries. A machine learning classifier helps to separate those semantic shifts that are relevant to the lexical typology. Clustering is applied to group similar pairs of meanings into semantic shifts.
Keywords: Lexical typology, semantic shifts, NLP, computational semantics, polysemy
DOI: 10.28995/2075-7182-2023-22-161-171

Автоматический поиск типологически релевантных семантических переходов в языках мира

Илья Грунтов
Институт языкознания РАН
Россия, г. Москва, 125009
Бол. Кисловский пер. 1/12
altaica@narod.ru

Елисей Рыков
Национальный исследовательский
университет «Высшая
школа экономики»
Россия, г. Москва, 101000
ул. Мясницкая, д. 20
esrykov@edu.hse.ru

Аннотация

Статья описывает полуавтоматический метод выявления типологически важных семантических переходов в языках мира. Алгоритм извлекает пары значений многозначных слов, встречающихся в двуязычных словарях. Машинно обученный классификатор позволяет определить степень релевантности перехода для лексической типологии. Пары значений объединяются в семантические переходы посредством кластеризации.
Ключевые слова: Лексическая типология, семантические переходы, АОТ, полисемия, вычислительная семантика

1 Introduction

The typology of semantic changes in the languages of the world is one of the areas of linguistics that can greatly benefit from computational methods. The current work is based on the datasets collected within the framework of the Database of Semantic Shifts in the Languages of the World (DSS, https://datsemshift.ru) which was founded in 2002. The main concept of the project has been described in (Zalizniak et al., 2012). The semantic shift is defined as a cognitive proximity between two meanings which are represented in different ways in the languages of the world, e.g. a relation between two meanings of a polysemic word, etymological cognates in related languages, semantic evolution at

*The paper is supported by the grant of Russian Science Foundation №22-18-00586, https://rscf.ru/project/22-18-00586/

different historical stages of a language, relations between the meaning of a loanword and a source word in two languages, the meanings of two morphological derivatives from a single root etc. See a more complete list in (Zalizniak, 2018). In the present paper, however, we restrict our subject to a specific subset of these possible types of semantic shifts, namely with a relation between two different meanings that can be represented by a colexification within a polysemic word of a given language. Such pairs of meanings are here called **realisations** of a semantic shift. For example, in at least 275 languages there are different realisations of the semantic shift "moon" - "month" [1]. Cf. also Table 1

Language	**Entry**	**Meaning 1**	**Meaning 2**
Ancient Hebrew	ḥōdeš	new moon ḥōdeš māḥār 'tomorrow is the new moon'	month ḥōdeš bəšānā 'one month in the year'
Swahili	mwezi	moon mwezi uliliwa na joka 'lunar eclipse'	month kila mwezi 'every month'
Adyghe	maze	moon мазэр къыкъокIыгъ 'the moon rose'	month январь мазэм 'in the month of January'
Tibetan	zla-ba	moon	month

Table 1: Several realisations of the semantic shift "moon" - "month" out of 275 present in the DSS

The purpose of this paper is to describe our attempt to automatically detect those semantic shifts that can be reproduced in different languages independently or through borrowing or loan translation. This may deepen our understanding of the cognitive mechanisms that give rise to polysemy. We have created a dataset of pairs of meanings representing a semantic shift extracted from the polysemous words of 75 languages, developed a machine learning classifier that defines the degree of typological relevance of a given pair of meanings, and performed clustering to group similar pairs of meanings from different languages. The result of our work is a pool of potential semantic shifts that can be further evaluated by a team of experts who can then incorporate them into the DSS.

The main source for our work are bilingual dictionaries "Language X -> Language Y". Usually language Y is one of the major world languages, such as English, Spanish, French, Russian etc. For these high-resource languages there are sophisticated language models, lexical databases of semantic relations and other linguistic tools. Thus, these languages are in fact meta-languages for describing meanings in the source Language X, and we can make cross-linguistic comparison of these data by applying NLP methods to these meta-languages. This approach therefore allows data from both low and high resource languages to be included in the common framework.

2 Methods and related works

There are many methods for detecting semantic shifts. One of them is the study of historical word corpora which contain documents from several hundred years, so that we can track changes in word meanings within a language over an extended time period. See e.g. (Hamilton et al., 2016), (Rodina and Kutuzov, 2020), (Fomin et al., 2019), (Kutuzov and Kuzmenko, 2017)), (Kutuzov et al., 2018) etc. Corpora analysis allows to find out that a given word has different meanings, but it is a highly difficult task to define strictly these meanings.

Another approach implies extraction data from etymological dictionaries and databases which usually contain cognate sets of diachronically related words with often different meanings in the modern languages. The etymological cognates are clear cases of semantic shift, but the main problem here is

[1] https://datsemshift.ru/shift0856

the robustness and reliability of the etymology. The shallower the language family, the more reliable the etymologies, but at deeper levels it becomes much more difficult to prove the validity of cognate relationships. See, e.g. the example of such diachronic studies in (Федотова, И.В., 2020).

Another approach detects the semantic changes arising within morphological derivation. See e.g. such works on Ukrainian (Melymuka et al., 2017) or Czech (Musil et al., 2019).

However, the present paper is concerned only with a particular subclass of semantic shifts, namely those that can be extracted from a dictionary as a separate pair of meanings of a polysemous word. The closest project to this task, apart from the DSS mentioned above, is another large project devoted to the aggregation of colexifications from various languages of the world called CLICKS (The Database of Cross-Linguistic Colexifications), the main principles of which are described in (Rzymski et al., 2020). CLICKS extracts meanings from the word-lists and maps them to concepts from the Concepticon catalogue (List et al., 2019) by semi-automatic fuzzy search.

Although we are aiming for a somewhat similar result, namely a database of semantic shifts, we took rather different approach to selecting and comparing meanings. According to our approach, not all pairs of meanings presented in the dictionaries should be included in the final database, but only those that are "non-trivial", "non-automatic" and thus most relevant for the typology of semantic changes.

There are certain criteria we use to define the degree of relevance of a semantic shift:

- The difference between meanings within a pair of meanings extracted from the dictionary should be "substantial". Of course, when one deals with semantic change, there is often a continuum between completely different and slightly different meanings. The degree of difference can vary greatly depending on the cultural and linguistic context of a given language. If a word ceased to mean 'yellow' and began to mean 'white', this would obviously be a major change, but if the meaning of a word changed from 'yellow' to 'light-yellow, pale', this would hardly be considered a major change by anyone. So it is not a distinct boundary between two classes of 'major change' vs. 'minor change', but a scale.
- The pair of meanings should not be related by regular metonymy. For example, "content" vs. "container", "author" vs. "author's work", etc.
- Neither meaning should be too vague.
- Differences in syntactic actant structure are not sufficient to turn a pair of meanings into a realisation of a suitable semantic shift. For example, we would not consider the semantic changes between *'to drink' (transitive) - 'to drink' (intransitive)* as a valid semantic shift (as in examples like: *He drinks water* vs *he has quit drinking*).
- The meanings within a pair should be connected immediately, i.e. if there is a minimal semantic shift between *foot* 'body part' and *foot* 'the lower part of the hill' and on the other hand there is another minimal semantic shift between *foot* 'body part' and *foot* 'unit of length', we would not postulate a semantic shift between the meanings *foot* 'lower part of the hill' and *foot* 'unit of length'.

To make these preliminary criteria more formal and quantifiable we use a number of factors to train the ML classifier to separate suitable semantic shifts from the others. The factors and the classifier itself are described in the "Machine learning classifier" section below.

3 Materials

We used data from 75 digitised dictionaries of the following languages:

- **Altaic:** Azeri, Buriad, Crimean Tatar, Dagur, Japanese, Khamnigan, Khalkha Mongolian, Korean, Nogai, Soyot, Tatar, Turkish
- **Austroasiatic:** Vietnamese
- **Austronesian:** Indonesian, Tagalog
- **Bantu:** Swahili, Zulu
- **Chukotko-Kamchatkan:** Koryak
- **Cushitic:** Somali
- **Indo-European:** Afrikaans, Ancient Greek, Armenian, Belarusian, Bulgarian, Czech, Danish, French, German, Greek, English, Icelandic, Italian, Kurdish, Latin, Latvian, Lithuanian, Norwe-

gian, Old English, Ossetian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Spanish, Swedish, Tajik, Tat, Ukrainian, Welsh
- **North-East Caucasian:** Avar, Bagvalal, Botlikh, Chamalal, Chechen, Lezgin, Rutul, Tabasaran
- **North-West Caucasian:** Adyghe, Abaza
- **Semitic:** Arabic, Hebrew
- **Sino-Tibetan:** Chinese
- **Uralic:** Estonian, Finnish, Hungarian, Komi Zyrian, Mari, Mokshan, Nganasan, Selkup, Yukaghir
- Basque (isolate), Esperanto (artificial).

These dictionaries vary greatly in size and in the percentage of polysemous words they contain, yielding from 500 pairs of meanings in Soyot up to 56000 in Azeri. In total we extracted from all these dictionaries a dataset consisting of 1,300,000 pairs of meanings each of them colexified in a lexical entry.

4 Pipeline

Our pipeline for detecting typologically relevant semantic shifts looks as follows[2].

1. Extraction of the polysemy from the digitised dictionaries by parsing.
2. Vectorization of all meanings of polysemous words using an encoder
3. Processing the data with a machine learning classifier that filters out all trivial or otherwise irrelevant cases.
4. Clustering of the filtered realisations from different languages into groups. These groups correspond to semantic shifts.
5. Postprocessing of the data by linguists.

In the following we will look at each stage in more detail.

5 Parsing

In the first stage we parse digitised dictionaries via Python scripts. The minimal entity we work with is a pair of meanings colexified within a lexical entry. In the simplest case the polysemic word has only two meanings. Often, however, the polysemic word has more than two meanings. It is rather impractical and far from linguistic reality to include all the possible combinations of meanings in our dataset. In our approach, to avoid combinatorial explosion, we assume that each meaning of a polysemic word is derived from its main meaning, which is defined by the dictionary authors as the first meaning. We know that this is an oversimplification derived from the simple model of radial polysemy (all secondary meanings are derived from the main one), while the real polysemy patterns may be different, e.g. chain polysemy (when the third meaning is derived from the second, not from the first). Perhaps, in further studies we will be able to overcome this limitation.

We extract polysemy from the dictionaries relying on the description of the meanings provided by the authors of the corresponding dictionaries. Sometimes, the meanings of a stem diverge so drastically that the authors of a dictionary decide to put them in separate entries as homonyms. In our approach, we did not distinguish between true homonyms which are different words that happen to be identical in form and false homonyms, which are the result of the evolution of the same word. We included such words in our data with the label “homonym”. Another possible decision would be to completely exclude such words from our data, but we supposed that recall was more important here than accuracy, and we do not want to lose the information of these “false” homonyms. The shift in meaning which is so strong that it gives life to two separate words synchronously is of great interest for semantic typology.

The result of this stage is a tsv file in which each line contains a lexical entry, a pair of meanings taken from the dictionary, the language name and dictionary metadata

6 Machine learning classifier

After parsing we obtained over 1,300,000 colexified pairs of meanings from 75 languages. However, not all of them represents a relevant semantic shift. Therefore, pruning is required to filter out irrelevant cases according to the criteria described above in Section 2.

[2]The data and scripts are available at `https://github.com/lmeribal/semantic-shifts`

To compare different classification approaches, we used the F1-measure. To train the classifier, we randomly selected a sample of pairs of meanings extracted from polysemous words and annotated it with the help of a team of linguistic experts[3]. The annotation was performed for two classes: positive class, when the given polysemy represents a semantic shift and vice versa. We considered a pair of meanings as marked for a certain class if at least 3 experts put it into that class. All the experts were experienced in semantic typology and have been working with the Database of semantic shifts for a considerable amount of time. As instruction for the markup they used the criteria for detecting relevance of a semantic shift presented in Section 2. Annotators inter-rater agreement according to Krippendorff's Alpha was 0.46. In total, over 22700 judgements were obtained for 2500 pairs of meanings. 1700 pairs become our train set. Validation set and test set consisted of 375 pairs each. The annotated sample from the dataset is shown in Table 2

language	entry	Meaning 1	Meaning 2	mark
Latvian	jēls	'сырой, сырое мясо'	'разг. непристойный, сальный'	1
Zulu	isi-bindi	'печень'	'смелость, храбрость'	1
Ancient Greek	σκολιός	'кривой, изогнутый'	'лукавый, коварный'	1
Welsh	marchnadaeth	'ware(s), merchandise'	'trade, traffic, commerce, business'	0
Lithuanian	dambra	'дудка, свирель'	'губная гармоника'	0
Azeri	şıltaq	'каприз'	'привередник, привередница'	0

Table 2: Sample of the annotated polysemy from the dictionaries

6.1 Methods

To find out which pairs of meaning represent a valid semantic shift relevant to the lexical typology, we tried several methods.

Cosine measure: As a baseline for comparing different approaches, we chose a method based on cosine distance alone. If the cosine distance between the embeddings of the definitions is greater than 0.5, this pair of meanings qualifies as a valid realisation of a semantic shift, and vice versa. Our dictionaries were usually bilingual translation dictionaries, not explanatory dictionaries. They contained a word in the original language and its translation in English, Russian, Spanish, etc. Since we needed to compare the translations of the original words, we looked for a multilingual embedding model that could compare sentences and noun phrases from different languages in the same embedding space. Therefore, we chose a Multilingual Universal Sentence Encoder (MUSE) [4] to obtain embeddings from dictionary definitions.

Feature-based classifier: The main source for feature engineering and feature selection was the dictionary definitions of the extracted meanings. We extracted features such as the number of words in these definitions and their lengths, the normalised Levenshtein distance between them and the cosine distances between their MUSE embeddings, the presence of common hyperonyms, synonyms, part of speech tags for the syntactic heads of the definitions, and so on. For a more detailed list, see the Table 4. We then trained a gradient boosting classifier on the data for 500 iterations with a depth parameter of 3. The algorithm chosen was Catboost (Dorogush et al., 2018).

Frozen LM fine-tuning: Previous work such as (Radford et al., 2017) shows that Language Models (LMs) perfectly solve downstream tasks related to language understanding, even when learning simple tasks such as next word or character prediction. Therefore, we assumed that Language Models already had the necessary knowledge about languages and their encoder embeddings could be applied to our task. Since we work with different high-resource languages, multilingual LMs such as BERT, RoBERTA, mt0 (Muennighoff et al., 2022) and FLAN-T5 (Chung et al., 2022) were used for training. As the mt0 and FLAN-T5 models are complete transformers, only frozen encoders were extracted from these models. To

[3] Annotation was performed by Ilya Gruntov, Sofia Durneva, Idaliya Fedotova, Viktoria Kaprielova, Veronika Kondratieva, Tatiana Mikhailova, Maria Orlova, Maksim Rousseau, Elisey Rykov, Anna Smirnitskaya, and Anna Zalizniak

[4] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

extract the embedding of the whole meaning, we applied mean pooling to the token embeddings. For all models, output embeddings for two meanings were obtained independently. Later, we concatenated the embeddings of two meanings and applied a classification head with binary output to these concatenated embeddings. In total, our model contained only 2050 trainable parameters. We also explored pre-trained language models of different sizes. During the training, we optimised Cross-Entropy Loss using the AdamW optimizer. Since our data is unbalanced (about 70% of the samples belong to the negative class), we passed class weights into the loss object. Negative class weight was calculated as the ratio of number of positive class samples to total samples, and vice versa. For each model, we trained the classifier for 50 epochs with a learning rate of 1e-3. For validation, we used the version of the classifier from the epoch with the lowest test loss value.

Multilingual Universal Sentence Encoder: In addition, we trained a classifier with the method described above, using embeddings from MUSE.

6.2 Evaluation

The evaluation of the classifiers is shown in the Table 3. For each method we calculated precision, recall and F1. ROC-AUC was only calculated for methods that were able to return probabilities of classes, so we didn't calculate it for our baseline.

The method based on a cosine measure does not achieve a high metric values. Thus, a noticeable difference between the definitions of the meanings of a polysemous word does not necessarily make it a valid semantic shift. BERT, RoBERTa and FLAN-T5 show similar results to the feature-based method. Classifier that accepts embeddings from the mt0-small model shows the best result and outperforms other methods. In addition, the method using embeddings from the MUSE shows ROC-AUC comparable to larger sizes of mt0.

Method		**Precision**	**Recall**	**F1**	**ROC-AUC**
Cosine measure		0.40	0.41	0.40	-
Feature-based		0.59	0.58	0.56	0.67
Multilingual Universal Sentence Encoder		0.65	0.63	0.62	0.71
Frozen LM fine-tuning	bert-base-multilingual-cased	0.62	0.60	0.59	0.64
	xlm-roberta-base	0.65	0.64	0.63	0.69
	xlm-roberta-large	0.63	0.62	0.61	0.66
	flan-t5-small	0.58	0.57	0.56	0.61
	flan-t5-base	0.61	0.61	0.61	0.65
	flan-t5-large	0.61	0.59	0.59	0.65
	mt0-small	**0.68**	**0.67**	**0.67**	**0.74**
	mt0-base	0.67	0.65	0.65	0.71
	mt0-large	0.65	0.64	0.63	0.71

Table 3: Performance of different classification models. For the Frozen LM fine-tuning method, the name of the pre-trained model from the HuggingFace is shown.

For the trained CatBoost classifier we additionally extracted importance of the input features. This data is presented in the Table 4. Cosine distance makes the greatest contribution, as well as the normalised Levenshtein distance.

To infer a mark on unmarked pairs, we used a classifier that accepts embeddings from the mt0-small, which had the highest metrics values. After classification, 800,000 pairs were marked as unsuitable, and 530,000 pairs were valid.

Feature	Importance
Cosine distance between definitions	23.23
Normalised Levenshtein distance between definitions	15.68
Normalised Levenshtein distance between hyperonyms	13.77
Cosine distance between hyperonyms	13.29
Cosine distance between synonyms	10.57
Common hyperonyms between syntactic heads of the definitions	6.12
Part of speech of the syntactic head of the first definition	4.37
Part of speech of the syntactic head of the second definition	4.21
Normalised Levenshtein distance between synonyms	3.77
Common parts of speech	3.14
Common synonyms between syntactic heads of the definitions	1.85

Table 4: Feature importance from the trained CatBoost classifier

7 Clustering

Each semantic shift consists of similar realisations from different world languages. One of our tasks is to group realisations into semantic shifts, which is similar to the clustering task in machine learning. So each realisation from a particular language would be a separate point within a cluster, while the cluster itself would represent a semantic shift as an abstract entity. We can therefore obtain embeddings of the realisations, and then apply any clustering algorithm to these embeddings to obtain clusters of the new semantic shifts.

Apart from that, there is another task when we already have semantic shifts, and our goal is to find new realisations for a given shift from our collection of realisations. For this we can use special algorithms that can be initialised with centroids, such as K-Means (Lloyd, 1982), and pass embedded shifts as cluster centroids. Alternatively, it is possible to cluster all realisations, and further match clusters with shift embeddings. If the distance between a cluster of realisations and a shift embedding is less than a certain threshold, we can say that this cluster corresponds to that shift. If there is no such shift for a given threshold, this may be the cluster of a new semantic shift. To speed up the matching process, efficient similarity search algorithms such as FAISS (Johnson et al., 2019) can be used.

Since each realisation is a pair of meanings, different methods can be used to obtain an embedding for the whole realisation:

- Sum of the embedding of meanings
- Average embedding of meanings
- Concatenation of embeddings of meanings
- Embedding of concatenation of meanings.

As a benchmark to test the quality of different clustering approaches we use the manually selected dataset of semantic shifts and their realisations from the DSS. Sample from the obtained dataset is shown in the Table 5. The dataset consisted of 7441 semantic shifts and 25407 realisations of these shifts from 1300 world languages.

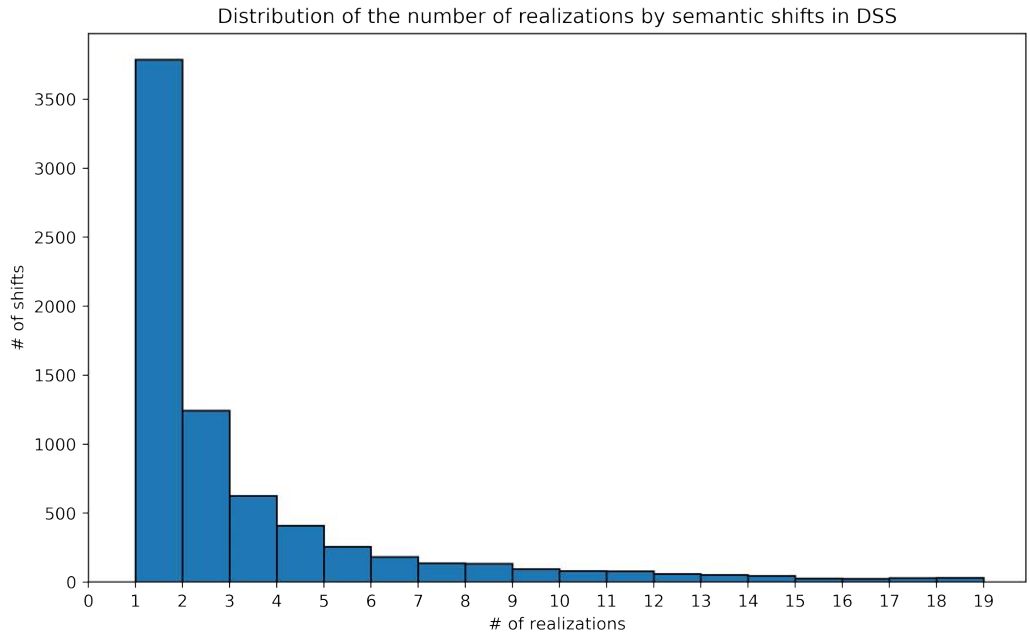


Figure 1: Distribution of the number of realisations by semantic shifts in DSS

ID	Shift ID	Language	Meaning 1	Meaning 2
19659	5218	Hungarian	head	chapter (of a book)
10706	3253	Thai	belly button	whirlpool
7336	1151	Swahili	to mix, to stir, to shake	to mix, to derange (plans etc)
1985	57	Latin	child, baby; boy	young servant, slave
8412	57	Fula	young boy	servant
11977	2798	Meadow Mari	interpreter	talkative person

Table 5: Sample of the clustering benchmark. If two realisations belong to one semantic shift their Shift Id would be the same

Figure 1 shows the distribution of the number of realisations by semantic shifts in DSS. It is noteworthy (and important for clustering) that most of the shifts have only one realisation. The Adjusted Rand Index (ARI) (Steinley, 2004) was chosen as a quality metric for clustering.

We tried different clustering approaches: K-Means, BIRCH (Zhang et al., 1996), DBSCAN (Ester et al., 1996). Since K-Means requires a number of clusters, a unique number of shifts was passed as a parameter, while we did not pass the number of clusters to other algorithms. If the number of clusters hasn't been passed to the BIRCH algorithm, it returns the subclusters, skipping the last clustering step. The DBSCAN algorithm itself determines the number of clusters using the threshold parameter, which we have left by default. Table 6 shows the performance of different clustering algorithms on the benchmark. The BIRCH algorithm outperforms K-Means and DBSCAN. In addition, we found that the sum of embeddings is the best of the methods observed to obtain the embedding of the whole realisation.

	K-Means	BIRCH	DBSCAN
Sum	0.48	**0.79**	0.73
Average	0.48	0.43	0.78
Embedding concat	0.37	0.76	0.59
String concat	0.44	0.62	0.75

Table 6: Performance of the different clustering approaches on benchmark

Below is the example of the cluster of realisations which corresponds to the semantic shift "fox" - "cunning person". Not all the clustered pairs of meanings are relevant to this semantic shift.

Language	Entry	Meaning 1	Meaning 2
Ancient Greek	ἀλώπηξ	‘лиса лисица’	‘лиса, разновидность акулы’
Avar	цер	‘лиса, лисица’	‘хитрец, пройдоха, лиса’
Bulgarian	лисица	‘лисица зверь’	‘перен. лиса, хитрец ’
Chinese	红狐 hónghú	‘рыжая лиса’	‘лиса обыкновенная’
Czech	liška	‘лисица, лиса’	‘старая лиса, плут’
Czech	lišák	‘лиса ’	‘старая лиса, хитрец, плут’
English	fox	‘лиса, лисица’	‘лиса, проныра, хитрец’
French	renard	‘лиса, лисица’	‘лиса, хитрец’
German	fuchs	‘лисица’	‘лиса, хитрец, пройдоха’
Italian	volpe	‘лиса, лисица’	‘лисий мех, лиса’
Korean	여우	‘лисица, лиса’	‘лиса’
Lezgian	сик1	‘лиса, лисица’	‘лиса, хитрец’
Mari	рывыж	‘лиса, лисица ’	‘лиса’
Polish	lis	‘лисица, лиса’	‘лис’
Slovak	lišiak	‘лисица, лиса, самец лис’	‘хитрец, плут, лиса’
Slovak	liška	‘лиса, лисица’	‘лиса’
Spanish	raposo	‘лисица’	‘хитрец, льстец, лиса’
Spanish	raposa	‘лисица, лиса’	‘хитрец, льстец, лиса’
Spanish	zorra	‘лиса, лисица’	‘лиса, шельма’
Spanish	zorrero	‘лисий’	‘королевский зверолов’
Spanish	zorro	‘лис, лиса’	‘лиса, лисий мех’
Turkish	tilki	‘лиса, лисица’	‘лиса, хитрец’
Ukrainian	лис	‘лисица, лиса, кобель, диал. лис’	‘перен. лиса’

Table 7: Cluster of realisations for the "fox" - "cunning person" semantic shift.

Meaning 1	Meaning 2	Number of languages
woman	wife	34
tooth	spike	29
long (size)	long (time)	23
to rip out	to take out	21
voice	sound	21
man	husband	21
head	classifier for round objects	21
grandmother	old woman	21
heel (anatomical)	heel (shoe)	20
Gossypium (plant)	cotton	20

Table 8: Top 10 largest clusters.

8 Validation and Postprocessing

The result is a pool of candidates that might qualify for relevant semantic shifts. In our framework we require a supervision from a linguistic team to exclude possible mistakes, true homonymy and possible inaccuracies of the result. The linguists review the data and include the best semantic shifts into the common database available online at `https://datsemshift.ru`. The manual approach that linguists used to apply previously implied looking through many polysemous words to find a suitable realisation of a semantic shift. Our approach allowed to "enrich" the raw linguistic material via filtering out of unsuitable pairs of meanings by means of a machine learning classifier. In order to quanitfy the value of our method of optimisation of linguistic work we ask the same team of experts who made an initial markup for the classifier to make judgements on a sample of the pairs of meanings which were considered as valid by our ML classifier. It turned out that when the linguists estimate a random sample of pairs of meanings they mark as "valid" only 30% of pairs. However, when they estimate a sample of pairs of meanings that received "valid" mark from the classifier, the approval rate increases up to 69%.

9 Conclusion

The method described above that processes dozens of dictionaries, extracts polysemic words, filters out typologically irrelevant cases and clustering similar pairs of meanings from various languages into semantic shifts is a valuable and powerful tool for detection of typologically relevant semantic shifts. It allows linguists to skip a lot of routine work of manually searching the dictionaries and looking for similar change of semantics in different languages. Thus, linguists can use this tool to make the conclusions about the cognitive mechanisms of the polysemy on the wide typological material.

Acknowledgements

We would like to thank Anna A. Zalizniak who read the preprint version of the paper and gave us her comments. Our colleagues Sofia Durneva, Idaliya Fedotova, Viktoria Kaprielova, Veronika Kondratieva, Tatiana Mikhailova, Maria Orlova, Maksim Rousseau, Anna Smirnitskaya, and Anna Zalizniak contributed substantially to the discussion of the criteria of a valid semantic shift. Alexander Gruntov helped a lot with preprocessing of the digitized dictionaries before parsing.

References

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Anna Veronika Dorogush, Vasily Ershov, and Andrey Gulin. 2018. Catboost: gradient boosting with categorical features support. *CoRR*, abs/1810.11363.

Федотова, И.В. 2020. Полисемия в списках самодийской базисной лексики и языковые контакты. *Урало-алтайские исследования*, 2(37):77–113.

Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. // *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, P 226–231. AAAI Press.

Vadim Fomin, Daria Bakshandaeva, Julia Rodina, and Andrey Kutuzov. 2019. Tracing cultural diachronic semantic shifts in russian using word embeddings: test sets and baselines. *ArXiv*, abs/1905.06837.

William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2016. Cultural Shift or Linguistic Drift? Comparing Two Computational Measures of Semantic Change. *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, 2016:2116–2121.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Andrey Kutuzov and Elizaveta Kuzmenko. 2017. Two centuries in two thousand words: Neural embedding models in detecting diachronic lexical changes. // *Quantitative approaches to the russian language*, P 95–112. Routledge.

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. 2018. Diachronic word embeddings and semantic shifts: a survey. *ArXiv*, abs/1806.03537.

JM List, C Rzymski, S Greenhill, N Schweikhard, K Pianykh, and R Forkel. 2019. Concepticon 2.2. *Jena, Germany: Max Planck Institute for the Science of Human History. See https://concepticon. clld. org*.

S. Lloyd. 1982. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28(2):129–137.

Mariia Melymuka, Gabriella Lapesa, Max Kisselew, and Sebastian Padó. 2017. Modeling derivational morphology in Ukrainian. // *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning.

Tomáš Musil, Jonáš Vidra, and David Mareček. 2019. Derivational morphological relations in word embeddings. // *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, P 173–180, Florence, Italy, August. Association for Computational Linguistics.

Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. 2017. Learning to generate reviews and discovering sentiment.

Julia Rodina and Andrey Kutuzov. 2020. RuSemShift: a dataset of historical lexical semantic change in Russian. // *International Conference on Computational Linguistics*.

Christoph Rzymski, Tiago Tresoldi, Simon J Greenhill, Mei-Shin Wu, Nathanael E Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A Bodt, Abbie Hantgan, Gereon A Kaiping, et al. 2020. The database of cross-linguistic colexifications, reproducible analysis of cross-linguistic polysemies. *Scientific data*, 7(1):13.

Douglas Steinley. 2004. Properties of the Hubert-Arabie Adjusted Rand Index. *Psychological methods*, 9:386–96, 09.

Anna A Zalizniak, Maria Bulakh, Dmitrij Ganenkov, Ilya Gruntov, Timur Maisak, and Maxim Russo. 2012. The catalogue of semantic shifts as a database for lexical semantic typology. *Linguistics*, 50(3):633–669.

Anna A. Zalizniak. 2018. The catalogue of semantic shifts: 20 years later. *Russian Journal of Linguistics*, 22/4.

Tian Zhang, Raghu Ramakrishnan, and Miron Livny. 1996. Birch: An efficient data clustering method for very large databases. *SIGMOD Rec.*, 25(2):103–114, jun.

Vague reference in expository discourse: multimodal regularities of speech and gesture

Olga Iriskhanova
Moscow State Linguistic University,
Institute of Linguistics RAS,
Moscow, Russia
oiriskhanova@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

Olga Agafonova
Moscow State Linguistic University,
Moscow, Russia
olga.agafonova92@gmail.com

Abstract

The paper looks into the vague reference expressed in speech and gesture distribution in expository discourse. The research data are the monologues of 19 participants with total length of 2 hours 38 minutes. In these monologues, the use of vague reference (expressed in placeholders and approximators, with total amount of 2528) and functional gesture types (deictic, representational, pragmatic and adaptors, with total amount of 2309) was explored, with the aim of identifying the regular patterns of speech and gesture distribution and co-occurrence. The multimodal regularities include 1) the proportional frequency of four gesture types use equal to 6.8 / 14.4 / 28.7 / 50.1, which manifests overall distribution of co-speech gesture in expository discourse, 2) the significant difference in co-speech gesture use with placeholders and approximators which manifests itself in the use of three gesture types, adaptors, representational and pragmatic gestures, 3) the individually maintained significant difference in co-speech gesture use with placeholders and approximators which manifests itself in adaptors. These regularities can serve as predictors for identifying the specifics of vague reference in multimodal expository discourse.

Keywords: expository discourse, multimodal behavior, co-speech gesture, vague reference, multimodal regularity

DOI: 10.28995/2075-7182-2023-22-172-180

Нечеткая референция в экспозиторном дискурсе: мультимодальные константы в речи и жестах

Ольга Ирисханова
Московский государственный
лингвистический университет,
Институт языкознания РАН,
Москва, Россия
oiriskhanova@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

Ольга Агафонова
Московский государственный
лингвистический университет,
Москва, Россия
olga.agafonova92@gmail.com

Аннотация

В статье исследуется категория нечеткой референции, реализуемая в мультимодальном поведении говорящего в речи и жестах в экспозиторном дискурсе. Материалом исследования являются записи монологов 19 участников общей продолжительностью 2 часа 38 минут. В ходе анализа устанавливаются особенности совместного использования речевых показателей нечеткой референции (заместителей и аппроксиматоров, общим количеством 2528) и функциональных типов жеста (дейктических, репрезентирующих, прагматических и адаптеров, общим количеством 2309). Цель исследования заключается в обнаружении мультимодальных констант в их распределении и в совместном использовании в дискурсе данного типа. К ним отнесены 1) относительное частотное распределение четырех типов жеста в пропорции 6.8 / 14.4 / 28.7 / 50.1, 2) наличие значимых различий в использовании жестов с заместителями и аппроксиматорами в отношении трех типов жестов, адаптеров, репрезентирующих и прагматических жестов, 3) наличие индивидуального варьирования в использовании адаптеров с заместителями и аппроксиматорами. Данные константы могут рассматриваться в качестве предикторов нечеткой референции в мультимодальном экспозиторном дискурсе.

Ключевые слова: экспозиторный дискурс, мультимодальное поведение, жест, сопровождающий речь, нечеткая референция, мультимодальные константы

1 Introduction

Exploring co-speech gestures as predictors of discourse types is an important task in multimodal studies. Methodologically, this idea is rooted in D. McNeill's theory of growth points which claims that "speech and gesture are co-expressive and opposed semiotically. Each has its own means of packaging the shared idea <…>" [1, p. 84]. In this study, it is the shared view of the discourse construal which according to McNeill, gives rise to growth points, or "the smallest package of gesture-speech unity" [ibid, p. 80]. Whereas speech and gesture have been commonly studied to explore single discourse construal effects in multiple studies, there is still scarce information on how co-speech gesturing contributes to discourse construal when discourse is viewed as a multi-function phenomenon.

In recent years, recognition of discourse structuring potential of co-speech gesturing has received special attention; however, methods and instruments of such analysis are now only developing. The best performing methods utilize the functional types of gesture [2; 3, 4], visuospatial virtual simulations of gesture [5; 6], and visuospatial types of gesture [7; 8; 9; 10]. The present study develops the functional approach to gesture analysis since it allows to explore both speech and gesture functions as interrelated in a specific discourse. We address the least studied discourse type, the expository discourse which explains or develops a topic and which maintains a focus on the relations between various phenomena [11; 12]. In contrast to other discourse types, for instance the descriptive discourse which has been explored in terms of functional co-speech gestures [2; 4, 13], we still know very little about speech and gesture distribution in expository discourse. The possible explanation for this is that as opposed to other discourse types, its speech characteristics which might have served to explore co-speech gesturing are less studied.

Consequently, the article develops a discourse functional approach to multimodal analysis of expository discourse. We seek to identify the regularities which appear in the speech and gesture distribution considering both overall data sample distribution and individual variance. The contributions of the current study include (i) establishing speech, gesture and co-speech gesture distribution in the compiled corpus of expository discourse; (ii) specifying the regularity patterns of multimodal behavior in expository discourse which can serve as predictors for the discourse type under consideration.

2 Theoretical framework

2.1 Vague reference in expository discourse. Placeholders and approximators

In expository discourse, the object of reference or the event is construed as having fuzzy boundaries; therefore, vague reference can serve as the key discourse characteristics of this discourse type. Vague reference can be viewed as a discourse category which directs the choice of the speaker towards a less distinct mode of referent or event construal [14]. Following V. Podlesskaya, vague reference results from the difficulties in speech generation in case direct reference seems problematic or undesirable [15]. We expect that vague reference will appear both in speech and in gesture since this discourse category can control both communicative modalities and serve as a growth point [1] in the choice of functional

discourse markers in speech and functional gesture types. This assumption is also cognitively rooted since fuzzy categorization of objects and events is a cognitive mechanism [16] which underlies vague reference and therefore can modulate multimodal behavior.

Most commonly, when exploring vague reference in speech, the works identify its two types of discourse markers, placeholders and approximators [15] which manifest two different speech functions. Placeholders are the discourse markers which are used instead of direct reference to objects, their properties, events and other speech patterns. Approximators are words and word combinations which accompany other (both direct and vague) means of reference. In this study, we adopt the vague reference typology of discourse markers developed and tested on a smaller data sample in the study of O. Iriskhanova and Yu. Abramova [14]. **Placeholders** include impersonal pronouns (*кто-то*, *где-нибудь*), shell-nouns [17] like *штука*, *объект*, *состояние*, nominalized adjectives (*хорошее, непонятное*, *большое, древнее*), metadiscourse markers (*вот так, как-то так, что-то в этом роде*). **Approximators** include hedges which make the statement sound less categorical (*как бы, что ли, ну в общем*), hedges pointing at personal opinion (*на мой взгляд, я думаю*), indefinite pronouns and particles accompanying nouns (*какой-то, чей-то*), modal adverbs and discourse markers (*вероятно, вряд ли*), deictic pronouns and adverbs (*тут, вот, этот*), metadiscourse accompanying comments (*в смысле там, скажем так, то есть*). Placeholders and adaptors frequently appear in clusters like *пламя это что-то скорее разгорающееся и большое* which includes three placeholders (impersonal pronoun *что-то*, and two nominalized units *разгорающееся* and *большое*), as well as an approximator (hedge *скорее*); or *мне кажется что чепуха это какое-то словесное понятие* which includes a placeholder (shell noun словесное понятие), and two approximators (a hedge pointing at personal opinion *мне кажется* and indefinite pronoun *какое-то*). Consequently, while presenting two ways of categorizing vague reference, placeholders and approximators do not constitute an opposition shaping the referent or event in their discourse construal. As known, multiple studies consider clustering patterns of discourse markers as a separate research task [7; 18], however in the compiled corpus these clusters display high variance which appears in the number and order of discourse markers presented within the clusters; therefore, the decision was adopted to consider the single uses of vague reference discourse markers accompanied by gestures. This approach commonly adopted for instance in [1; 2; 3; 9] allows to specify the use of gestures as contingent on each of the functional types of discourse markers and to further identify co-speech gesture distribution and their regularity patterns in the compiled corpus of expository discourse.

2.2 Functional types of gesture

In the study, we employ the functional gesture typology developed in the works of C. Müller, A. Cienki, and O. Iriskhanova [2; 19], who differentiate four basic **gesture types** with their further specification: deictic (Pointing, Touching gestures), representational (Holding, Molding, Acting, Embodying, Tracing gestures), pragmatic (Discourse structuring, Discourse representational, Discourse emphatic, Expressing attitude/evaluation, Contact establishing gestures), and adaptors (Self-adaptors, Object-adaptors). We expect that these gestures will manifest specific proportional distribution in expository discourse and that their distribution will be different with placeholders and adaptors since they clearly realize different discourse functions. Deictic gestures point at an object to foreground it [20]. Representational gestures can be described as gestures which stimulate the speech production process due to their iconicity, i.e., resemblance to some concepts in their physical/metaphorical properties [21; 22]. Pragmatic gestures include hand movements with different subfunctions and are primarily discourse related [23; 24]. Adaptors are used to reduce anxiety and cognitive load which helps to concentrate on the subject of speech [25; 26; 27]. The process of their identification in the recorded data involves: 1) visual analysis of gestures according to their visuospatial characteristics, 2) analysis of their functions in speech based on their semantics, dependent on the verbal context.

3 Experiment design

3.1 Participants and experiment procedure

19 participants (all students, aged 18-22) took part in the experiment. Their multimodal behavior was videorecorded with a frontal camera. The experiment mentors were seated in front of the participant, their role consisted in posing the questions which stimulated expository discourse in experiment participants. Each participant answered the same 10 questions which prompted to comment on the difference between 10 pairs of close synonyms, like *roar* and *howl (рык/вой)*, *line* and *lineament (линия/черта)*, *duty* and *obligation (обязанность/обязательство)*, *burden* and *load (бремя/ноша)*. The recorded corpus of speech and gesture manifestations is 2 hours 38 minutes long. The data were then analysed in ELAN software, where they were annotated in three layers: transcriptions, speech discourse markers, and gesture types. In Fig. 1 and 2 we present the annotation examples.

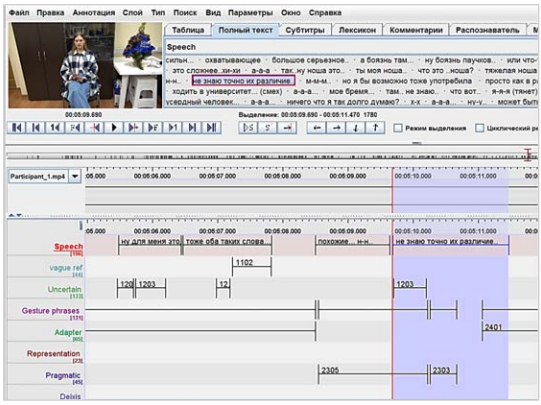


Figure 1: Approximator *не знаю* used with pragmatic gesture «не знаю точно их различие» (“don’t know exactly their difference”)

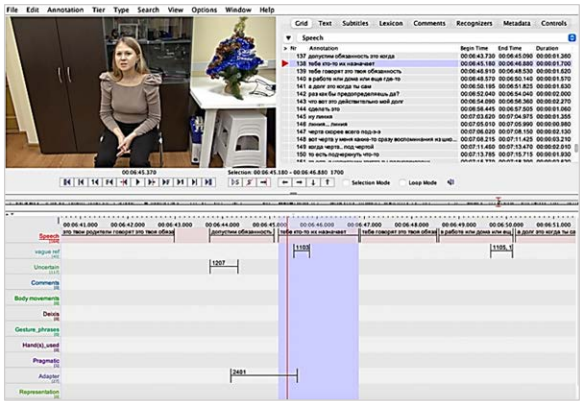


Figure 2: Placeholder *кто-то* used with self-adaptor «тебе кто-то их назначает» (“somebody assigns them to you”)

Figures 1 and 2 manifest that annotations allowed to synchronize the use of discourse markers and gestures. As seen in Figures 1 and 2, the discourse segment for speech analysis was the minimal discourse unit identified following both prosodic and syntactic criteria; most commonly it corresponds to a clause [28] manifested here in *не знаю точно их различие* (Figure 1) and in *тебе кто-то их назначает* (Figure 2). The discourse unit under consideration in Figure 1 displays three gesture uses which are pragmatic gestures (coded as 2305 and 2303) and one adaptor (coded as 2401); still the discourse marker of vague reference (coded as 1203) is synchronized only with the first pragmatic gesture (coded as 2305). In Figure 2 the discourse unit contains one discourse marker of vague reference (coded as 1103) which is synchronized with one gesture use of adaptor (coded as 2401). Two annotators decided on the choice of gesture types. In most cases it was a unanimous decision; in the cases when this decision was hampered by the presence of two possible types (or subtypes), we marked them as displaying both. Two annotators decided on the choice of discourse markers; since we had an inventory of markers, in very rare cases we had to discuss the choice.

The **data processing algorithm** included 4 steps described below.

Step 1. Analysis of frequency (activity) of two functional types of discourse markers, placeholders and approximators; and of four gesture types, deictic, representational, pragmatic and adaptors. At this step, we identify the proportional regularity of co-speech gesture use.

Step 2. Contingency tests with each function of discourse markers and each gesture type. This step helps determine whether there are specific gestures contingent with either type of discourse markers.

Step 3. Analyses of variance in speech and gesture in individual participants’ behavior. These analyses allow to qualify the differences as systemic or individual.

Step 4. Identifying the regularities in speech and gesture distribution and co-occurrence within the sample and in the individual behavior.

4 Results

4.1 Distribution of speech functions and gesture types

At Step 1 we explore frequency (activity) of two functional types of discourse markers, placeholders and approximators; and of four gesture types, deictic, representational, pragmatic and adaptors. The total number of placeholders and approximators in the compiled corpus is equal to 2528, and the total number of gestures (deictic, representational, pragmatic and adaptors) used as co-speech gestures is equal to 2309. The overall activity of speech and gesture in expository discourse is given in Table 1. Importantly, since there are cases of placeholders or approximators use not accompanied with gesture, the total number of placeholders and approximators is larger than the number of co-speech gestures.

	Deictic	**Representational**	**Pragmatic**	**Adaptors**		
Speech					**With gestures**	**Total**
Placeholders	58 (8.03)	140 (19.39)	241 (33.38)	283 (39.2)	722	768
Approximators	99 (6.24)	193 (12.16)	421 (26.53)	874 (55.07)	1587	1760
	157 (6.8)	333 (14.42)	662 (28.67)	1157 (50.11)		

Table 1: Speech and gesture frequency (Abs and (Rel))

Table 1 shows that adaptors prevail in the recorded corpus of expository discourse; still, pragmatic and representational gestures are also frequently observed. With the total number of gesture use, the proportional use of four gesture types in the sample is 6.8 / 14.4 / 28.7 / 50.1 (the mean values of gesture use in individual behavior are 9.24, 19.6, 38.94, 68.1), which can be considered a regularity of overall distribution of co-speech gesture in the expository discourse corpus.

The next question is whether different types of gestures are used with placeholders and adaptors. At Step 2 we conduct a series of contingency tests to identify the significance of differences in co-speech gesture frequency. With the number of gestures used with placeholders equal to 722 and the number of gestures used with approximators equal to 1587, the Chi-square contingency test did not show considerable differences in their distribution ($\chi2=0.478$, $p=0.49$). The results indicate that the data still manifest considerable uniformity and are more likely to be dependent on the discourse type rather than on the use of either of the two types of discourse markers. Meanwhile, we hypothesize that the use of single gesture types can manifest variance, since the proportional use of adaptors with placeholders vs. approximators is 39% and 55%, for deictic gestures it is 8% and 6%, for representational gestures – 19% and 12%, for pragmatic gestures – 33% and 27%. Four Chi-square contingency tests showed that the difference in the use of adaptors was highly significant with $\chi2=50.029$, $p< .001$; additionally, the differences in representational gestures with $\chi2=21.435$, $p< .001$, and pragmatic gestures with $\chi2=11.391$, $p< .001$ are also statistically significant. This means that there is a systemic difference in co-speech gesture use with placeholders and approximators, and it manifests itself in the use of three gesture types – adaptors, representational and pragmatic gestures. This difference can also serve as a multimodal regularity modulated by the discourse type.

However, we can expect that these differences are attributed to the individual variance in multimodal behavior. Therefore, at Step 3 we explore the variance in speech and gesture in individual participants' behavior in the recorded corpus. In Figures 3 and 4 we manifest the individual differences in speech and gesture distribution.

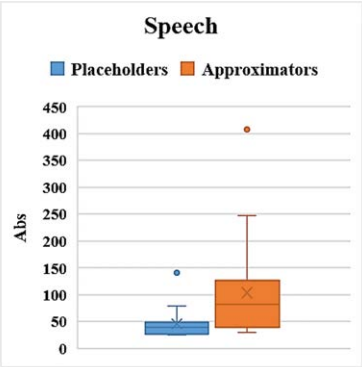


Figure 3: Box plot diagram of speech functions distribution

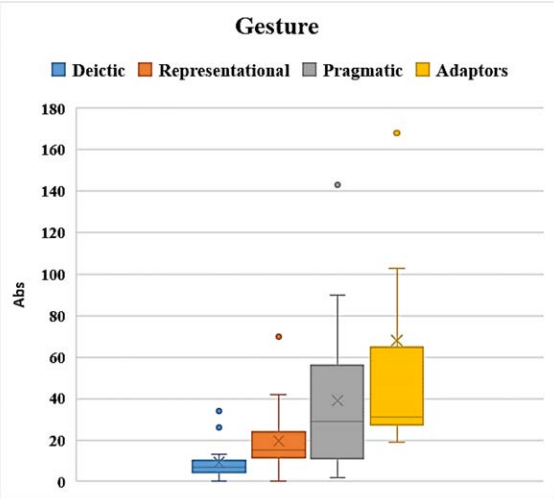


Figure 4: Box plot diagram of gesture types distribution

As we see from the diagrams, the data do not have the normal distribution (confirmed by Shapiro-Wilk tests, with $p<0.005$); therefore, we applied Repeated Measures ANOVA (Non-parametric) to determine the variance in individual use of the two functional types of discourse markers and gesture types. With $F(1) = 9.94$ and $p=0.002$ for the use of speech functions (placeholders and approximators) and $F(3) =35.8$ and $p < .001$, we can claim that the data manifest significant individual differences. However, these differences can occur either in all the gesture types or they can be attributed to a particular group of gestures. For this reason, we split the data describing the gesture use accompanying placeholders and approximators and analyze them separately. In Figure 5 the diagrams showing data distribution are presented.

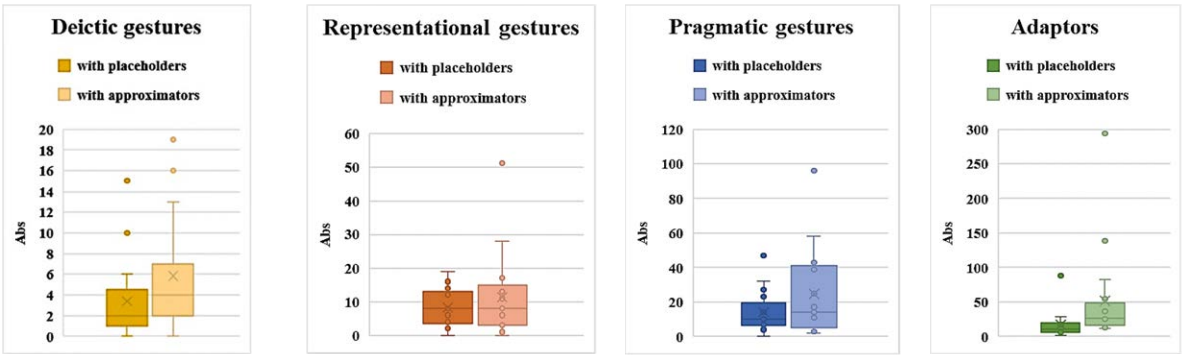


Figure 5: Box-plot diagrams with gesture distribution in individual multimodal behavior

The diagrams show that data distribution is not normal (confirmed by Shapiro-Wilk tests, with $p<0.005$). So, to determine the variance in individual use of the speech functions and gesture types we applied 4 Repeated Measures ANOVA (Non-parametric) tests.

In the use of deictic, representational and pragmatic gestures we did not observe statistically significant differences, with $F(1)=2.57$, $p=0.109$ for both deictic and representational gestures, and $F(1)=1.47$, $p=0.225$ for pragmatic gestures. Meanwhile, adaptors showed significant difference in individual use, with $F(1)=13.2$, $p < .001$. The results manifest that the differences in individual gesture use are mostly attributed to the use of adaptors, while other gesture types manifest relative uniformity. Therefore, the differences in the use of adaptors with placeholders and approximators in individual behavior serve as another multimodal regularity of the expository discourse corpus.

5 Discussion

In the present study, we expected to establish speech, gesture and co-speech gesture distribution in expressing vague reference in expository discourse, and to identify the regularity patterns of multimodal behavior which can serve as predictors for the vague reference in this discourse type.

In the recorded corpus (2 hours 38 minutes long) we identified several multimodal regularities in the use of placeholders and approximators as two speech functions of vague reference in expository discourse, and four gesture types, deictic, representational, pragmatic and adaptors. However, the regularities in individual behavior distribution appear most reliable.

The first regularity observed is **the proportional use of four gesture types**, which is 6.8 / 14.4 / 28.7 / 50.1. Although the proportional use of the gesture types cannot serve as a reliable regularity due to individual differences (see Figure 4), we can still claim that significant differences were observed only in the use of adaptors; therefore, deictic, representational and pragmatic gestures manifested common overall distribution.

The second regularity is the significant **difference in co-speech gesture use with placeholders and approximators**, which manifests itself in the use of three gesture types: adaptors (with $\chi 2=50.029$, $p< .001$), representational (with $\chi 2=21.435$, $p< .001$), and pragmatic gestures (with $\chi 2=11.391$, $p< .001$), with the last two types appearing more frequently with placeholders.

The third regularity is the **individually maintained significant difference** in co-speech gesture use with placeholders and approximators, which manifests itself in **adaptors** (with $F(1)=13.2$, $p < .001$).

The explanation of the results can be facilitated with the help of the discourse functions which gestures display. The frequent use of adaptors, which are the predominant type of gestures in expository discourse, as shown by the results obtained, proves that the speakers encounter difficulties in speech production and are forced to reduce anxiety and cognitive load to concentrate on the object of reference [25; 26]. This multimodal regularity is specific of expository discourse in contrast to descriptive or narrative discourse [13]. Meanwhile, the individual differences in the use of adaptors might indicate the difference in cognitive load/anxiety that every respondent experiences during the task, which supports the findings on individual variance in adaptors use resulting from the differences in perceived emotional stability and personality types [27].

The fact that pragmatic gestures display high frequency and high variance in their use with placeholders and approximators and more frequently accompany placeholders in the compiled corpus shows that they help the speaker to intensify or to formulate the idea of reference. Their high frequency may prove their multifunctionality in discourse which was described in [23; 24]. However, in this study we specified that this multifunctionality prevailed in case the speakers immediately construe the object of reference by using the means of vague reference rather than construe the discourse path towards an object or event while using approximators. We also found that the use of representational gestures shows variance, and representational gestures appear significantly more often with placeholders than with approximators, at least in the compiled corpus. These results conform to prior findings presented in [21; 22] which claim that representational gestures mostly display iconicity or resemblance to the objects or concepts in their properties. Still, relatively high frequency of their use with approximators may be explained by the fact that the preparation phase of gesture execution [23] is synchronized with the use of approximators.

These three multimodal regularities can be contrasted with the regularities observed in other discourse types, which will allow to identify their trans-discourse variance. Additionally, they can be used to explore the variance among different samples of expository discourse.

6 Final remarks

Overall, the study showed that there exist evident correspondences between the use of speech functions displaying vague reference and gesture types in expository discourse. The results prove that the category of vague reference habitually explored in speech is in fact a discourse structuring category which managers the choice of both speech functions and functional gesture types. Additionally, since these results were obtained via discourse functional approach to multimodality, the study also attests to the efficiency of this method in exploring multimodality in discourse.

The study is part of the project “Kinesic and vocal aspects of communication: parameters of variance” (FMNE-2022-0015) carried out at the Institute of Linguistics, Russian Academy of Sciences.

References

[1] 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.

[2] Cornelia Müller. Gestural Modes of Representation as techniques of depiction. *Body – Language – Communication: An international Handbook on Multimodality in Human Interaction*. (Handbooks of Linguistics and Communication Science). Berlin: De Gruyter Mouton, 2014. P. 1687–1701.

[3] Alan Cienki and Irene Mittelberg. Creativity in the forms and functions of spontaneous gesture with speech. In T. Veale, K. Feyaerts, C. Forceville (eds.) *The Agile Mind: A multi-disciplinary study of a multi-faceted phenomenon*. Berlin, De Gruyter Mouton, 2013. P. 231–252.

[4] Herbert H. Clark. Depicting as a Method of Communication. *Psychol. Rev.*, 123(3), 2016. P. 24–347.

[5] Stephen Kopp, Paul Tepper, Justine Cassell. Towards integrated microplanning of language and iconic gesture for multimodal output. *Proceedings of the International Conference on Multimodal Interfaces* (ICMI’04), 2004. P. 97–104.

[6] Dominique Boutet, Jean-François Jégo, Vincent Meyrueis. *POLIMOD Pipeline: tutorial: step-by-step tutorial: Motion Capture, Visualization & Data Analysis for gesture studies*. [Technical Report]. Université de Rouen, Université Paris 8, Moscow State Linguistic University, 2018.

[7] Andrej A. Kibrik and Vera I. Podlesskaya. *Night Dream Stories: A Corpus Study of spoken Russian discourse*. Moscow: Languages of Slavonic Culture, 2009.

[8] Elena A. Grishina. *Russian gesticulation as a linguistic phenomenon: corpus studies*. Moscow: YaSK, 2017.

[9] Jana Bressem, Silva H. Ladewig, Cornelia Müller. *Linguistic annotation system for gestures*. Berlin: De Gruyter Mouton, 2013.

[10] Alex Lascarides, Mathew Stone. A Formal Semantic Analysis of Gesture. *Journal of Semantics - J SEMANT*. 26. 2009. 10.1093/jos/ffp004

[11] Robert E. Longacre. *The grammar of discourse*. New York: Plenum, 1983.

[12] Ruth A. Berman, Bracha Nir-Sagiv. Comparing narrative and expository text construction across adolescence: A developmental paradox. *Discourse Processes*, 43(2), 2007. P. 79–120.

[13] Olga N. Prokofyeva, Olga K. Iriskhanova, Maria I. Kiose. Speech and gesture in descriptive discourse. In O. Iriskhanova (ed.) *Multimodal dimensions of discourse*. Moscow: YaSK, 2021. P. 63–109.

[14] Olga K. Iriskhanova, Yulia S. Abramova. Vague reference in metalinguistic tasks as a multimodal phenomenon. *Cognitive Studies of Language*, 4 (47), 2021. P. 233–244.

[15] Vera I. Podlesskaya. Vague names in Russian speech: a corpus study. *Computational Linguistics and Intellectual Technologies*, 12(19), 2013. P. 631–643.

[16] George Lakoff. Hedges. *Journal of Philosophical Logic*, 2, 1973. P. 458–508.

[17] Hans-Jeorg Schmid. *English Abstract Nouns as Conceptual Shells*. Berlin: Mouton de Gruyter, 2000.

[18] Nikolaj A. Korotaev, Vera I. Podlesskaya, Olga V. Fedorova. Disfluencies in Russian spoken monologues: a distributional analysis. *Computational Linguistics and Intellectual Technologies: Proceedings of the Annual International Conference “Dialogue”*, 19, 2020. P. 439–451.

[19] Olga K. Iriskhanova and Alan Cienki. The semiotics of gestures in cognitive linguistics: Contributions and challenges. *Issues in Cognitive Linguistics*, 4, 2018. P. 25–36.

[20] Kensy A. Cooperrider. The co-organization of demonstratives and pointing gestures. *Discourse Processes*, 53, 2015. P. 632–656.

[21] Sotaro Kita, Martha W. Alibali, Mingyuan Chu. How do gestures influence thinking and speaking? The gesture-for-conceptualization hypothesis. *Psychological review*, 124(3), 2017. P. 245–266.

[22] Martha W. Alibali, Amelia Yeo, Autumn B. Hostetter, Sotaro Kita. Representational gestures help speakers package information for speaking. *Why gesture? How the hands function in speaking, thinking and communicating*. Amsterdam: John Benjamins Publishing Company, 2017. P. 15–37.

[23] Adam Kendon. Pragmatic functions of gestures: Some observations on the history of their study and their nature. *Gesture*, 16 (2), 2017. P. 157–175.

[24] Renia Lopez-Ozieblo. Proposing a revised functional classification of pragmatic gestures. *Lingua*, 247, 2020. 102870.

[25] Paul Ekman, Wallace V. Friesen. The repertoire of non-verbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1, 1969. P. 49–98.

[26] Anna V. Leonteva. Contrastive analysis of adaptors in synchronic translation. *Vestnik of Moscow State Linguistic University. Humanities*, 1(869), 2023. P. 76–81.

[27] Michael Neff, Nicholas Toothman, Robeson Bowmani, Jean E. Fox Tree, Marilyn A. Walker. Don't Scratch! Self-adaptors Reflect Emotional Stability. In H.H. Vilhjálmsson, S. Kopp, S. Marsella, K.R. Thórisson (eds.) *Intelligent Virtual Agents. IVA 2011. Lecture Notes in Computer Science*, 6895, 2011. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-23974-8_43

[28] Olga V. Fedorova, Andrej A. Kibrik, Nikolaj A. Korotaev, Alla O. Litvinenko, Julia V. Nikolaeva. Temporal coordination between gestural and speech units in multimodal communication. *Computational Linguistics and Intellectual Technologies Proceedings of the Annual International Conference "Dialogue"*, 15(22), 2016. P. 159–170.

A new dataset for sentence-level complexity in Russian

Vladimir Ivanov
Kazan Federal University
Kazan, Russia;
Innopolis University,
Innopolis, Russia
v.ivanov@innopolis.ru

Elbayoumi Mohamed Gamal
Innopolis University
Innopolis, Russia
m.elbayoumi@innopolis.university

Abstract

Text complexity prediction is a well-studied task. Predicting complexity sentence-level has attracted less research interest in Russian. One possible application of sentence-level complexity prediction is more precise and fine-grained modeling of text complexity. In the paper we present a novel dataset with sentence-level annotation of complexity. The dataset is open and contains 1,200 Russian sentences extracted from SynTagRus treebank. Annotations were collected via Yandex Toloka platform using 7-point scale. The paper presents various linguistic features that can contribute to sentence complexity as well as a baseline linear model.

Keywords: sentence complexity, crowdsourcing, readability
DOI: 10.28995/2075-7182-2023-22-181-190

Набор данных с оценками сложности предложений на русском языке

Иванов Владимир
Казанский федеральный университет
г. Казань, Россия
Университет Иннополис
г. Иннополис, Россия
v.ivanov@innopolis.ru

Эльбайоуми Мохамед Гамаль
Университет Иннополис
г. Иннополис, Россия
m.elbayoumi@innopolis.university

Аннотация

Прогнозирование сложности текста — хорошо изученная задача. Предсказание уровня сложности отдельного предложения привлекает несколько меньший исследовательский интерес. В статье представлен новый набор данных с аннотацией сложности на уровне предложений. Набор данных открытый и содержит 1,200 предложений на русском языке, извлеченных из корпуса SynTagRus. Аннотации собирались через платформу Яндекс Толока по 7-бальной шкале. В статье представлены различные лингвистические признаки, которые могут быть использованы при оценке сложности предложений, а также предложена простая линейная модель.

Ключевые слова: сложность текста на уровне предложения, читабельность, краудсорсинг

1 Introduction

Text complexity prediction is a task studied at various levels of linguistic units ((Crossley et al., 2008; Collins-Thompson and Callan, 2005; Heilman et al., 2008; Shardlow et al., 2021; Shardlow et al., 2020)). The sentence-level complexity evaluation (SCE) subtask takes an intermediate position between the text fragment level (i.e., several coherent sentences) and the level of an individual word/phrase complexity prediction.

Recent works study sets of features that can be used in SCE, including lexical, syntactical features from the target sentence, and contextual features from surrounding sentences (Schumacher et al., 2016; Iavarone et al., 2021). One possible application of sentence-level complexity prediction is more precise modeling of text complexity beyond the passage-level. For longer texts readability measures such as Flesch-Kinkaid formula (Flesch, 1948) (as well as many others) make use of statistics and typically

provide a robust solution. However, statistics such as average sentence length and average word length tend to vary a lot when one analyze individual sentence which may produce less robust predictions. Therefore, in such cases a fine-grained model for sentence complexity prediction might be useful.

The SCE task presents issues, at the levels of interpretation of the model's results and feature selection. One of the state-of-the-art approaches is deep neural networks capable to explore a wide range of features and combine them in a hierarchical and non-linear manner. What is more, deep neural networks have been applied in SCE before. For instance, (Schicchi et al., 2020) evaluated the long short-term memory (LSTM) model with attention mechanism in a binary classification of Italian sentences.

Datasets with manual annotations of sentence complexity were created for a number of languages. (Brunato et al., 2018) present a detailed analysis of features that affect human perception of sentence complexity. The authors study the contribution of a set of lexical, morphosyntactic, and syntactic features. The most important features are sentence length, maximum dependency length in a dependency syntax tree, etc.; for sentences with the same length, the most important factors include average word length and lexical density. Analysis of text complexity in Russian academic text was performed in (Solovyev et al., 2018; Solnyshkina et al., 2018), where the main focus is modeling text complexity of a whole text or a passage.

In this paper, we address two issues. First, we present a new dataset with sentence-level complexity annotations on a scale from 1 to 7. The dataset contains 1,200 sentences with more than 23,000 individual complexity judgments. To the best of our knowledge, this is the first dataset of this kind in Russian. Second, we analyze several types of features and evaluate linear models for predicting sentence-level complexity.

The rest of the paper is organized as follows. Section 2 discusses the related work. Section 3 presents an approach to collect data. Section 4 describes the dataset and its features; Section 5 presents the experimental results on sentence complexity prediction.

2 Related Work

Here, we focus only on works that are closely related to the present study and consider sentence-level complexity datasets and evaluations. In (Inui and Yamamoto, 2001), authors study the relative complexity of sentences in the readability context for deaf people. Authors collected a corpus with pairs of sentences with paraphrases. Modeling complexity was targeted on the classification of paraphrases into three levels/classes ('left', 'right', 'same'). Inui and Yamamoto developed a rule-based method and compared it to the SVM classifier. Later, in (Vajjala and Meurers, 2014), authors evaluated an SVM classifier to predict relative complexity of pairs of complex and simplified sentences. The study (Maqsood et al., 2022) compares different algorithms for SCE in English dataset with seven categories. Classification of sentence difficulty in Arabic language is addressed in (Khallaf and Sharoff, 2021).

(Schumacher et al., 2016) studied models for predicting the reading difficulty of sentences, with and without the surrounding context. They binned sentences according to grade levels (e.g., a sentence from grade 1 was paired with sentences from grades 3-4, 5-6, 7-8, 9-10, 11-12). Authors studied lexical and grammatical features to train a logistic regression classifier and Bayesian ranker. These authors show that considering the context improves predicting sentence readability. The simplest model has only the AoA-based features, which allows to achieve higher score on the dataset. For Russian language sentence-level complexity prediction was addressed in (Ivanov, 2022), but that study used automatically generated complexity scores for sentences extracted from school textbooks.

(Brunato et al., 2018) applied crowdsourcing to model human perception of single-sentence difficulty in Italian and English. These authors investigate a wide set of linguistic features and their importance for human perception of sentence complexity. Brunato et al. analyzed few tens of features, such as 'char_tok' (average number of characters per word) and 'n_tokens' (average number of words per sentence). In their experiments, authors show that syntactic features can play important role in defining the sentence complexity, but 'char_tok' and 'n_tokens' features are always in the top important features as well. What is more, to explicitly control for sentence length, authors applied binning, i.e. sentences were grouped by length (e.g. 10, 15, 20, etc.) up to 35 tokens.

Finally, deep neural networks for sentence complexity classification were proposed in (Lo Bosco et al., 2021). Their model uses the TreeTagger to extract syntactic features, two LSTM layers, and a linear layer. The last layer outputs the probability of a sentence belonging to the easy or complex class. The experimental results show the increased approach effectiveness for both Italian and English, compared with several baselines such as Support Vector Machine, Gradient Boosting, and Random Forest.

3 Data Collection and Annotation

Our approach to dataset collection consists of two parts: selection of sentences and annotating them using the crowdsourcing platform (Yandex Toloka). We sampled sentences from the SynTagRus corpus. This Syntactically Tagged Russian text corpus contains more than 87,000 sentences (https://universaldependencies.org/treebanks/ru_syntagrus/index.html). The Universal Dependency version of SynTagRus is a comprehensive Russian dependency treebank that was developed by the Institute for Information Transmission Problems of the Russian Academy of Sciences (Lyashevkaya et al., 2016; Marneffe et al., 2021). It is a revision of the original SynTagRus treebank that uses the Universal Dependency annotation scheme. The Universal Dependency annotation scheme is a standard annotation scheme for dependency treebanks that is used in many different languages. The treebank covers a wide range of genres, including news articles, fiction books, and academic papers. It is annotated with a variety of linguistic features, including part-of-speech, morphology, syntax, and semantics. We chose the Universal Dependency version of SynTagRus because it is a high-quality treebank that covers a wide range of genres. We also believe that the linguistic features that are annotated in SynTagRus are relevant to the study of sentence-level complexity.

For extracting a sample of sentences for our dataset, we followed the methodology presented in (Brunato et al., 2018). Authors proposed reducing the influence of lexicon by pruning the sentences containing low-frequency lemmas using a lemma frequency list. In our study we use the frequency list developed by Sharov and Lyashevskaya (Lyashevskaya and S.A., 2009).

All the sentences contained in the SynTagRus corpus were grouped into 6 bins based on a different sentence length, i.e. 10, 15, 20, 25, 30, 35 tokens. Sentences in each subset were then ranked according to the average frequency of their lemmas. We extracted for each bin the first 200-top ranked sentences. Therefore, the dataset for annotation contains 1,200 sentences.

Assessments were collected via crowdsourcing of human judgments in the following way. Sentences were randomly shuffled and divided into task pages (one sentence per page). Each assessor should have passed a test for knowledge of Russian language. Out of all (approximately 10,000) such native speakers available at the Yandex Toloka platform, we admitted 30% of assessors with the highest score (according to the platform).

We used several mechanisms to ensure the quality of the data. First, each sentence was evaluated by multiple participants (each sentence got scores from ten assessors), which allowed us to calculate an average complexity score for each sentence and to estimate the level of agreement among the participants. Second, we used "gold standard" sentences in the task. These were sentences for which we already had reliable complexity ratings. The participants were not aware which sentences were the gold ones. Their ratings for these sentences were used to monitor their performance and to adjust their trust scores. If a participant consistently rated the gold standard sentences incorrectly, their future responses were given less weight in the final calculation of the sentence complexity scores.

Assessors were asked to read a sentence and rate how difficult it was on a 7-point scale where 1 means “very easy” and 7 “very difficult”. We chose a 7-point scale because we wanted to have a granular range of complexity ratings. We also wanted to avoid using a binary scale (e.g., easy vs. difficult), as we believe that sentence complexity is a spectrum. There are a number of theoretical bases for using a 7-point scale to measure sentence complexity. One theory is that sentence complexity is a continuous variable, rather than a discrete variable (Gernsbacher, 1999). This means that there are an infinite number of possible levels of complexity, rather than just a finite number of levels. Another theory is that sentence complexity is a multidimensional concept (Fletcher et al., 1986). This means that there are multiple factors that contribute to complexity, such as syntactic complexity, semantic complexity, and lexical complexity. A

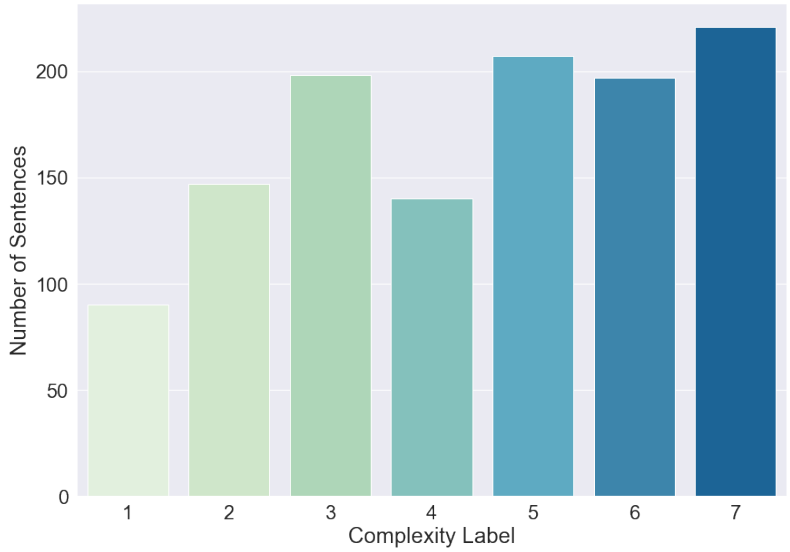


Figure 1: Distribution of scores in the corpus.

7-point scale can be used to capture these multiple factors. Last, but not least, a 7-point scale was applied in a similar previous work done for English and Italian (Brunato et al., 2018); which enables comparison between the datasets in future. In the following section we analyze the collected dataset.

4 Data Analysis

4.1 Analysis of annotations and agreement

The analysis of inter-annotator agreement is an important aspect of dataset validation, as it provides insights into the quality and reliability of the annotations. First, we make use of the Toloka's aggregation method (Dawid-Skene model) that provides a confidence score for each sentence. The mean score for each of seven label categories is above 99%. The distribution of aggregated labels are presented in Figure 1. One can see that overall the dataset is slightly imbalanced towards difficult sentences. The simplest score ('1') has only 90 examples.

Next, for each sentence we calculated the maximum number of assessors who agreed about some category for that sentence. On average, 4.3 assessors per sentence have agreed about a complexity label (with standard deviation of 1.2). Finally, in Figure 2 we plot the deviation of scores with respect to sentence length (bin). This plot clearly shows the correlation between sentence length and complexity score.

Our analysis suggests that the new dataset can provide a useful resource for studying sentence-level complexity in Russian, but caution should be exercised when interpreting the scores, especially for longer sentences. In the following subsection we analyze a set of features, including syntactical.

4.2 Exploring features and correlation

For our study, we extracted features that reflect various facets of sentence complexity, such as:

- **Average_path_length,** which represents the average dependency distance between words in a sentence. Dependency distance is defined as the number of words between two words that have a dependency relationship.
- **Maximum_path_length,** this feature represents the maximum dependency distance between words in a sentence.
- **Num_clauses,** represents the number of clauses in each sentence.
- **Num_phrases,** represents the number of phrases in each sentence.

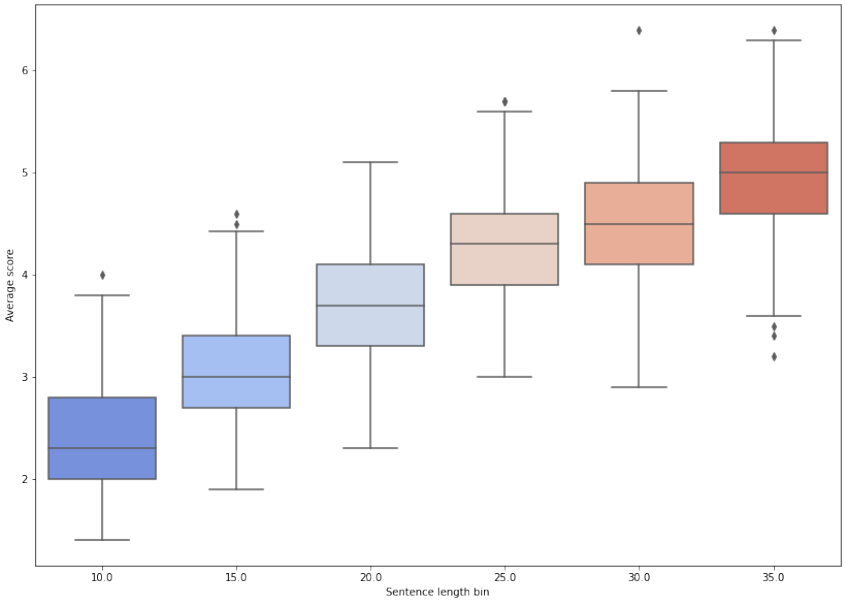


Figure 2: Distribution of complexity scores with respect to sentence length.

- **Num_subordinating_conjunctions** represents the number of subordinating conjunctions in each sentence. It is a measure of syntactic complexity and indicates the degree of subordination in the sentence.
- **Prop_nouns,** represents the proportion of nouns in each sentence based on their POS-tag.
- **Prop_verbs,** represents the proportion of verbs in each sentence.
- **Prop_adjectives,** represents the proportion of adjectives in each sentence.
- **Prop_pronouns,** represents the proportion of pronouns in each sentence.
- **Average_freq,** represents the average frequency of words in the sentence.
- **Avg_token_length,** represents average letters per word.
- **sen_len** is the sentence length measured in characters.

The following examples describe how these features can contribute to complexity.

Dependency distance is the length of the dependency path between a word and its head. A dependency path is the sequence of words that connect a word to its head. For example, in the sentence "The cat that sat on the mat was black," the dependency path between the word "black" and its head "cat" is "cat-sat-on-the-mat-black." The dependency distance between "black" and its head "cat" is 4. We found that dependency distance was positively correlated with sentence complexity. This means that sentences with longer dependency distances were more complex than sentences with shorter dependency distances. One reason why dependency distance is correlated with complexity is that it is a measure of the syntactic complexity of a sentence. Sentences with longer dependency distances have more complex syntax, which makes them more difficult to understand. This observation is supported by other studies of text complexity both at sentence level (Brunato et al., 2018) and at the passage level (Solovyev et al., 2023).

Number of Clauses, a clause is a group of words that has a subject and a verb. A sentence can have one or more clauses. For example, the sentence "The cat that sat on the mat was black" has two clauses: "The cat sat on the mat" and "The cat was black." We found that the number of clauses in a sentence was positively correlated with sentence complexity. This means that sentences with more clauses were more complex than sentences with fewer clauses. One reason why the number of clauses is correlated with complexity is that it is a measure of the semantic complexity of a sentence. Sentences with more clauses have more complex semantics, which makes them more difficult to understand.

Proportion of Nouns and Phrases, the proportion of nouns and phrases in a sentence is a measure of the lexical complexity of a sentence. Nouns and phrases are lexical items, which are words or groups of words that have meaning. We found that the proportion of nouns and phrases in a sentence was positively correlated with sentence complexity. This means that sentences with a higher proportion of nouns and phrases were more complex than sentences with a lower proportion of nouns and phrases. One reason why the proportion of nouns and phrases is correlated with complexity is that it is a measure of the vocabulary load of a sentence. Sentences with a higher proportion of nouns and phrases have a higher vocabulary load, which makes them more difficult to understand.

Table 1: Average values of linguistic features within different bins.

Feature Name	**L10**	**L15**	**L20**	**L25**	**L30**	**L35**
average_path_length	1.65	2.01	2.32	2.44	2.48	2.62
maximum_path_length	8.04	10.80	14.71	18.31	20.84	24.04
num_clauses	0.14	0.24	0.42	0.59	0.69	0.85
num_phrases	2.63	3.13	4.03	4.81	5.37	5.79
num_subord._conjunctions	0.15	0.22	0.33	0.47	0.46	0.67
prop_nouns	0.24	0.24	0.24	0.24	0.25	0.25
prop_verbs	0.13	0.11	0.11	0.11	0.10	0.10
prop_adjectives	0.08	0.08	0.10	0.10	0.10	0.10
prop_pronouns	0.02	0.02	0.02	0.03	0.03	0.02
avg_token_len	5.59	5.57	5.81	5.89	5.96	6.06
avg_freq	6333.51	5837.55	5548.08	5036.07	4687.93	4151.05
sen_len	55.90	83.53	116.18	147.33	178.76	212.18
score	2.43	3.09	3.73	4.27	4.52	4.94
std_score	1.05	1.11	1.22	1.28	1.36	1.31

Table 2: The correlation between linguistic features and sentence complexity within different bins.

Feature Name	**L10**	**L15**	**L20**	**L25**	**L30**	**L35**	**All**
average_path_length	0.15	0.02	0.12	-0.08	-0.10	-0.10	0.39
maximum_path_length	0.05	0.08	0.09	-0.09	-0.12	-0.15	0.58
num_clauses	0.07	0.13	0.08	-0.11	-0.00	-0.01	0.32
num_phrases	-0.14	-0.03	-0.05	-0.12	-0.04	-0.01	0.40
num_subordinating_conjunctions	0.09	0.01	0.03	-0.04	-0.07	-0.08	0.20
prop_nouns	-0.17	-0.01	0.00	0.20	0.16	0.15	0.08
prop_verbs	0.02	0.11	0.03	-0.16	-0.07	0.04	-0.10
prop_adjectives	0.02	-0.05	0.09	0.16	0.14	0.06	0.14
prop_pronouns	0.04	-0.03	0.09	-0.10	0.03	-0.04	0.02
avg_token_len	0.16	0.26	0.30	0.31	0.49	0.34	0.33
avg_freq	0.04	-0.06	-0.15	-0.04	-0.08	-0.19	-0.51
sen_len (in characters)	0.80	0.64	0.59	0.53	0.70	0.72	0.83

To analyze the correlation between linguistic features and sentence complexity, we first calculated the average complexity judgments for six bins of sentences with the same length (10, 15, 20, 25, 30, and 35 tokens). Pearson correlation coefficient is presented in Table 2. As anticipated, the feature with the strongest correlation to sentence complexity is sentence length (measured in characters).However, as indicated in Figure 3, exceptions exist where short sentences have high complexity scores and long sentences have low complexity scores.

Our analysis revealed that some features had a stronger correlation with sentence complexity than others. For example, we observed that the correlation coefficients for various features differ depending on the sentence length bin (see Table 2). Overall, features with the highest correlations are those related to

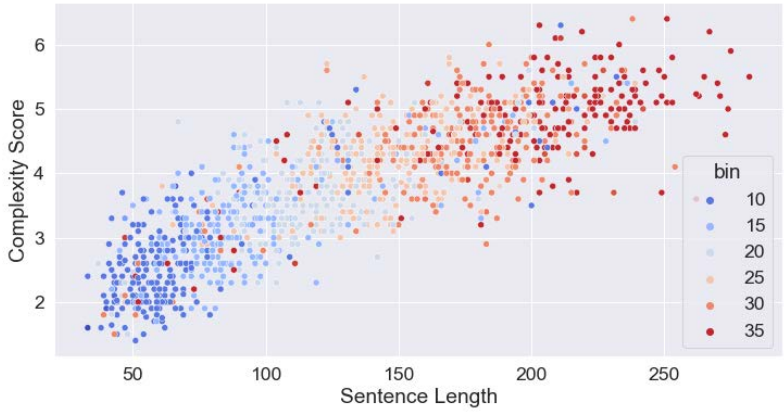


Figure 3: Complexity scores strongly correlate with sentence length. The plot also shows a substantial number of exceptions.

path length, proportions of nouns and phrases, and frequency as well as sentence/token lengths. These findings imply that specific linguistic features substantially influence sentence complexity in Russian. Our results provide insights into the linguistic factors that contribute to sentence-level complexity in Russian and highlight the importance of considering multiple features when assessing sentence complexity.

4.3 Comparison with English / Italian dataset

(Brunato et al., 2018) investigated the correlation between different linguistic features and human judgments of sentence difficulty, using Spearman's rank correlation coefficient. In contrast, our study explores the relationship between linguistic features and sentence complexity using Pearson correlation.

Comparing the findings of the two studies, some similarities can be observed. Both studies found that sentence length (in characters) has a strong positive correlation with sentence complexity. Additionally, the two studies identified similar linguistic features that are significantly correlated with sentence complexity, such as average token length and number of clauses.

In conclusion, while there are some differences in the correlation coefficients between the two studies, the overall findings suggest that certain linguistic features are consistently associated with sentence complexity.

5 Linear Regression Model for Sentence Complexity

Given the correlations coefficients (Table 2), we first train and evaluate a linear regression model. Feature selection shows that the best linear regression model can use three parameters, sentence length in characters (SLC), average path length (APL), and the number of clauses (NCL). The model presented below has MSE=0.32 (± 0.03), MAE=0.45 (± 0.02) and R^2 value of 0.71, while a model with a single parameter (SLC) has MSE=0.33 and MAE=0.46.

$$Compl.Score = -1.61 + 0.014 * SLC + 0.146 * APL + 0.057 * NCL$$

To confirm SLC is a strong predictor, we run two experiments. First, the Linear regression without the SLC parameter achieves only 0.61 (MSE). Second, we fine-tuned the pre-trained RuBERT model on 80% of the data. The performance of RuBERT is 0.54(MSE) and 0.57(MAE). It is worth noting that the linear model with three parameters systematically underestimates sentences with higher scores (close to 6) and overestimates the complexity of simple sentences with low scores (Fig. 4). Our analysis of such errors shows that the most errors are coming from relying on the SLC value. Therefore, we propose and evaluate models that make use of stratified by sentence length. To this end, we compare performance of

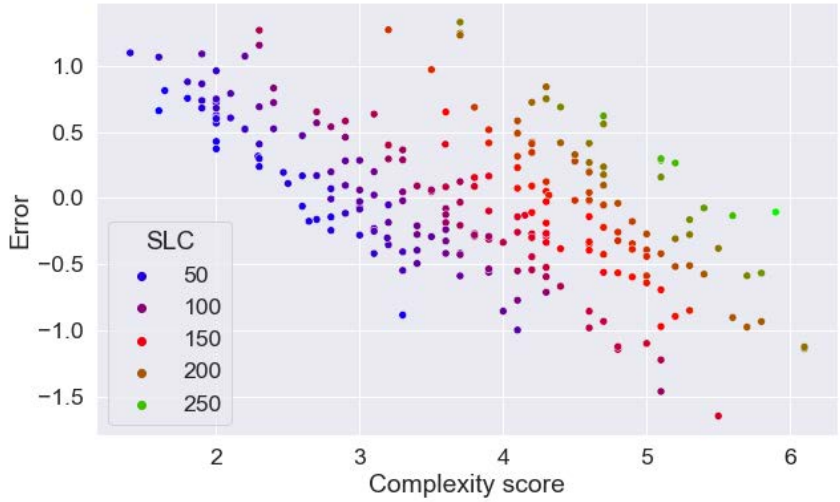


Figure 4: Negative correlation between linear model's errors and the true values of sentence complexity (Error = predicted value - true value).

linear models that either use or not use the SLC feature in each of the six bins. The results are provided in the Table 3.

Table 3: A comparison of models trained in a specific length range with and without sentence length parameter.

	with SLC		w/o SLC	
bin	MSE	MAE	MSE	MAE
L10	**0.254**	0.405	0.257	0.408
L15	0.272	0.417	**0.269**	0.414
L20	0.310	0.443	**0.293**	0.430
L25	0.295	0.438	**0.294**	0.435
L30	**0.295**	0.435	0.298	0.439
L35	0.313	0.413	**0.312**	0.421

6 Conclusion

In this paper, we present a dataset of 1,200 Russian sentences annotated for complexity, collected through crowdsourcing using the Yandex Toloka platform. The analysis of the dataset shows that it is slightly unbalanced towards difficult sentences, with a correlation between sentence length and complexity score. The paper also presents various linguistic features that contribute to sentence complexity in Russian, such as dependency distance, number of clauses and subordinating conjunctions, and proportion of nouns and phrases. The study found that certain features had a stronger correlation with sentence complexity than others. These findings provide insights into the linguistic factors that contribute to sentence-level complexity in Russian, and the dataset can be a useful resource for further research on this topic. The dataset is available at `https://zenodo.org/record/7937828#.ZGJEHC9ByZA`.

Acknowledgments

This work is funded by Russian Science Foundation, grant # 22-21-00334.

References

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? // *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, P 2690–2699, Brussels, Belgium, October-November. Association for Computational Linguistics.

Kevyn Collins-Thompson and Jamie Callan. 2005. Predicting reading difficulty with statistical language models. *Journal of the american society for information science and technology*, 56(13):1448–1462.

Scott A Crossley, Jerry Greenfield, and Danielle S McNamara. 2008. Assessing text readability using cognitively based indices. *Tesol Quarterly*, 42(3):475–493.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Garth Fletcher, Paula Danilovics, Guadalupe Fernandez, Dena Peterson, and Glenn Reeder. 1986. Attributional complexity. an individual differences measure. *Journal of Personality and Social Psychology*, 51:875–884, 10.

Morton Gernsbacher. 1999. Comprehension: A paradigm for cognition. *American Scientist*, 87, 11.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. // *Proceedings of the third workshop on innovative use of NLP for building educational applications*, P 71–79.

Benedetta Iavarone, Dominique Brunato, and Felice Dell'Orletta. 2021. Sentence complexity in context. // *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, P 186–199, Online, June. Association for Computational Linguistics.

Kentaro Inui and Satomi Yamamoto. 2001. Corpus-based acquisition of sentence readability ranking models for deaf people. // *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium, November 27-30, 2001, Hitotsubashi Memorial Hall, National Center of Sciences, Tokyo, Japan*, P 159–166.

Vladimir Ivanov. 2022. Sentence-level complexity in russian: An evaluation of bert and graph neural networks. *Frontiers in Artificial Intelligence*, 5.

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. // *Workshop on Arabic Natural Language Processing*.

Giosué Lo Bosco, Giovanni Pilato, and Daniele Schicchi. 2021. Deepeva: A deep neural network architecture for assessing sentence complexity in italian and english languages. *Array*, 12:100097.

Olga Lyashevkaya, Kira Droganova, Daniel Zeman, Maria Alexeeva, Tatiana Gavrilova, Nina Mustafina, and Elena Shakurova. 2016. Universal dependencies for russian: A new syntactic dependencies tagset. *SSRN Electronic Journal*, 01.

Olga Lyashevskaya and Sharov S.A. 2009. *Frequency dictionary of the modern Russian language (the Russian National Corpus)*. 01.

Shazia Maqsood, Abdul Shahid, Muhammad Tanvir Afzal, Muhammad Roman, Zahid Khan, Zubair Nawaz, and Muhammad Haris Aziz. 2022. Assessing english language sentences readability using machine learning models. *PeerJ Computer Science*, 7:e818.

Marie-Catherine Marneffe, Christopher Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal dependencies. *Computational Linguistics*, 47:1–52, 03.

Daniele Schicchi, Giovanni Pilato, and Giosué Lo Bosco. 2020. Deep neural attention-based model for the evaluation of italian sentences complexity. // *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, P 253–256. IEEE.

Elliot Schumacher, Maxine Eskenazi, Gwen Frishkoff, and Kevyn Collins-Thompson. 2016. Predicting the relative difficulty of single sentences with and without surrounding context. // *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, P 1871–1881, Austin, Texas, November. Association for Computational Linguistics.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. // *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, P 57–62, Marseille, France, May. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. *// Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, P 1–16, Online, August. Association for Computational Linguistics.

Marina Solnyshkina, Vladimir Ivanov, and Valery Solovyev. 2018. Readability formula for russian texts: a modified version. *// Advances in Computational Intelligence: 17th Mexican International Conference on Artificial Intelligence, MICAI 2018, Guadalajara, Mexico, October 22–27, 2018, Proceedings, Part II 17*, P 132–145. Springer.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of intelligent & fuzzy systems*, 34(5):3049–3058.

Valery Solovyev, Marina Solnyshkina, Vladimir Ivanov, and Svetlana Timoshenko. 2023. Complexity of russian academic texts as the function of syntactic parameters. *// Computational Linguistics and Intelligent Text Processing: 19th International Conference, CICLing 2018, Hanoi, Vietnam, March 18–24, 2018, Revised Selected Papers, Part I*, P 168–179. Springer.

Sowmya Vajjala and Detmar Meurers. 2014. Assessing the relative reading level of sentence pairs for text simplification. *// Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, P 288–297, Gothenburg, Sweden, April. Association for Computational Linguistics.

The problem of linguistic markup conversion: the transformation of the Compreno markup into the UD format

Alexandra Ivoylova
RSUH
Moscow, Russia
a.m.ivoylova@gmail.com

Darya Dyachkova
RSUH
Moscow, Russia
d.dyachkova@bk.ru

Maria Petrova
A4 Technology
Moscow, Russia
g-fox-ive@mail.ru

Mariia Michurina
RSUH
Moscow, Russia
marimitchurina@gmail.com

Abstract

The linguistic markup is an important NLP task. Currently, there are several popular formats of the markup (Universal Dependencies, Prague Dependencies, and so on), which are mostly focused on morphology and syntax. Full semantic markup can be found in the ABBYY Compreno model. However, the structure of the format differs significantly from the models mentioned above. In the given work, we convert the Compreno markup into the UD format, which is rather popular among NLP researchers, and enrich it with the semantical pattern.

Compreno and UD present morphology and syntax differently as far as tokenization, POS-tagging, ellipsis, coordination, and some other things are concerned, which makes the conversion of one format into another more complicated. Nevertheless, the conversion allowed us to create the UD-markup containing not only morpho-syntactic information but also the semantic one.

Keywords: Compreno, semantic markup, Universal Dependencies
DOI: 10.28995/2075-7182-2023-22-191-199

Проблемы конвертации лингвистической разметки: конвертация формата Compreno в UD-формат

Ивойлова А.М.
РГГУ
Москва, Россия
a.m.ivoylova@gmail.com

Дьячкова Д.С.
РГГУ
Москва, Россия
d.dyachkova@bk.ru

Петрова М.А.
A4 Technology
Москва, Россия
g-fox-ive@mail.ru

Мичурина М.А.
РГГУ
Москва, Россия
marimitchurina@gmail.com

Аннотация

Лингвистическая разметка является актуальной задачей NLP. В настоящее время существует несколько популярных форматов подобной разметки (Universal Dependencies, Prague Dependencies и др.), при этом в фокусе их внимания находятся, в первую очередь, морфология и синтаксис. Одним из немногих форматов, предлагающих не только морфо-синтаксическую, но и семантическую разметку, является формат ABBYY Compreno, однако в структурном отношении данный формат существенно отличается от указанных выше моделей. В предлагаемой работе делается попытка представить разметку Compreno в более привычном для пользователей формате UD и дополнить данный формат семантической разметкой.

Представление морфологии и синтаксиса в UD и Compreno имеет ряд значимых различий, касающихся, в числе прочего, токенизации, POS-tagging, эллипсиса, сочинения и других явлений, что создает определенные сложности при конвертации. Тем не менее, конвертация Compreno в UD позволила получить полную многоуровневую UD-разметку, содержащую как морфо-синтаксическую, так и семантическую информацию.

Ключевые слова: Compreno, семантическая разметка, Universal Dependencies

1 Introduction

Morphological, syntactic and semantic labelling is an essential part of natural language processing pipeline. A need for the universal multilanguage markup format has been acknowledged for a long time; one of the most known projects of creating such a format is the Universal Dependencies (UD) project (De Marneffe et al., 2006), although UD encompasses morphosyntax only.

As for the semantics, there is no markup standard so far which would be widely acknowledged. Currently, several projects deal with semantic labels, and some of them are meant for integral three-level labelling, for instance, Prague Dependencies (Hajic et al., 2001), or the ETAP system (Boguslavsky, 1999). Nevertheless, none of these projects provide both laconic and integral labelling format.

An attractive model in this respect seems the ABBYY Compreno model (Anisimovich et al., 2012; Petrova, 2014) which is capable to perform a full-scale morphosyntactic and semantic labelling. Its advantage is the ability to provide a complete well-structured semantic markup, which includes not only arguments, but also adjuncts, modifiers, and other dependencies. Besides, it has special means of handling non-tree links, such as ellipsis or dislocation. However, Compreno has its own drawbacks: first, the semantic part of the markup is too detailed which makes the markup too complicated; second, the formal structure of the markup format has some peculiarities.

Our primary goal is thereby to develop a new labelling standard that would benefit both from the conciseness of UD and the thoroughness of Compreno system. To achieve it, we decided to adopt the UD format for the morphosyntactic markup part and to enrich it with the simplified Compreno semantic markup. This task, in turn, demanded the conversion of the Compreno markup format into UD.

The elaboration of the integral markup standard and, especially, the semantic markup standard is a part of the Compreno-Based Linguistic Data (CoBaLD) Annotation Project which includes the creation of a fully-labelled Russian dataset [1] of approximately 400,000 tokens as well, containing news texts from the (now defunct) NewsRu.Com site. For more information on the standard and the dataset[2], see (Petrova et al., 2023).

At the first stage, the corpus was automatically annotated by the Compreno parser and checked manually by professional linguists.

At the second stage, the morphosyntactic part of the markup was automatically converted into the UD format. The conversion was partly checked as well. To evaluate the quality of the conversion, we checked abound 10% of the dataset. The percent of labels modified by different groups of annotators in manually checked automatic conversion varies from 5 to 10%, which means that the total quality of the conversion is close to 95%.

After the conversion, the UD markup was supplemented with the semantic pattern - word meanings for each token and the semantic relations between the constituents.

In the current paper, we focus on one important part of the work - the conversion of one format into another and the challenges we encountered solving this problem.

2 Related Work

The need to have a standardized format of natural text labelling (at first, POS-tagging) appeared when the first corpora were created. The pioneers of POS-tagging are the creators of the Brown corpus, the Lancaster/Oslo-Bergen corpus, the University of Pennsylvania corpus (UPenn) and others. The first language for labelling was English. A comprehensive table of rival English POS-tags can be found in (Atwell, 2008).

However, it turned out difficult to use the English POS-tags for other languages, so the attempts were made to create language-specific tagsets, such as (Bar-Haim et al., 2008) for Modern Hebrew or (Diab, 2007) for Arabic. Approximately at the same time, the UD project started. Its creators strove to develop a universal standard which could be applied to any language and which would combine both morphological and syntactic features.

[1] https://github.com/compreno-semantics/compreno-corpus

[2] The access to the dataset is provided according to the CC BY-NC 4.0 License which allows non-commercial use.

On the other side, semantic labelling formats were being developed as well, starting with the well-known Universal Networking Language (UNL) (Uchida and Zhu, 2001), and onto more recent projects like Universal Decompositional Semantics (UDS) (White et al., 2016) and Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013). The semantic projects, however, concentrated on the semantics itself. Some of them do not involve morphosyntactic parsing at all (the already mentioned UCCA and Abstract Meaning Representations (Banarescu et al., 2013) are the examples).

As for the Russian language, the only two projects aimed at the semantic parsing are the ETAP project and the above-mentioned Compreno project. For the moment, there are only a few Russian datasets labelled in UD (e.g., Taiga (Shavrina and Shapovalova, 2017) or SynTagRus (Boguslavsky, 2014; Droganova and Zeman, 2016; Droganova et al., 2018); as for the semantic labelling, no datasets are available.

3 Overview of the Format

Our markup format is derived from the UD format [3] and represents a format very similar to the well-known CONLL which is a syntactic parse tree with semantic data included. The main representation principles in UD are the following: a sentence is separated by a newline, any table row contains ten columns, and the columns include token ID, form, lemma, universal POS-tags and language-specific POS-tags, grammatical features, dependency head and dependency relation.

Unlike UD, the Compreno format represents sentences in a tree-like structure (see fig. 1):

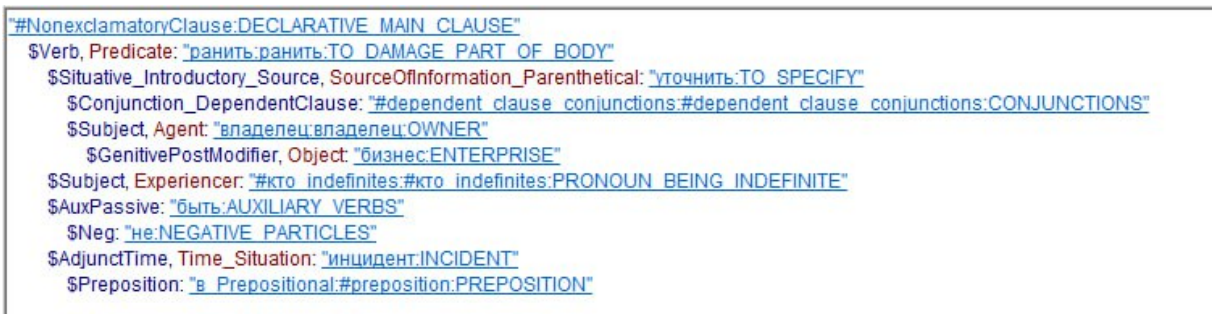


Figure 1: Compreno format: tree structure for *Как уточнил владелец бизнеса, никто не был ранен в инциденте.* 'As the business owner clarified, no one was injured in the incident.'

For our standard, we adopted the UD table format, but replaced the last two (usually empty, especially in the Russian UD corpora) columns with semantic slots and semantic classes taken from the Compreno format.

The Compreno model presents words in the form of a thesaurus-like semantic tree, which consists of universal semantic classes - semantic fields, filled with lexical contents in each language incorporated in the model. The total number of the classes is more than 200 000. For the current work, we used the simplified version of the hierarchy, cut to hyperonym classes only (about 1000 classes). For details, see (Petrova et al., 2023) and the relevant fragment of the hierarchy on Github[4].

Semantic slots, in turn, correspond to semantic roles, which define the semantic relations between the core and the dependent elements, including actants such as Agent or Experiencer, characteristics, adjuncts (time, condition, concession, etc.), and so on. Unlike syntactic roles, semantic ones can have different syntactic realizations, for instance, all bracketed constituents in "I will come [tomorrow]", "I will come [after sunset]", and "I will come [when the clock strikes twelve]" correspond to Time slot. Or, subject-Agent in active voice and by-Agent in passive voice correspond to one Agent slot. The list of the slots can be found on Github[5].

An example of the labelled text can be seen in fig. 2.

[3]The complete information on the UD tagset which was implemented here may be found at https://universaldependencies.org/).

[4]https://github.com/compreno-semantics

[5]https://github.com/compreno-semantics/compreno-corpus/blob/main/semantic_slots.xlsx

```
# text = Как уточнил владелец бизнеса , никто не был ранен в инциденте.
1   Как как SCONJ   _   _   2   mark    _   CONJUNCTIONS
2   уточнил уточнить    VERB    _   Aspect=Perf|Gender=Masc|Mood=Ind|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act  9   parataxis   Parenthetical   VERBAL_COMMUNICATION
3   владелец    владелец    NOUN    _   Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing   2   nsubj   Agent   HUMAN
4   бизнеса бизнес  NOUN    _   Animacy=Inan|Case=Gen|Gender=Masc|Number=Sing   3   nmod    Object_Situation    BUSINESS
5   ,   ,   PUNCT   _   _   2   punct   _   _
6   никто   никто   PRON    _   Animacy=Anim|Case=Nom|Gender=Masc|Number=Sing|Polarity=Neg  9   nsubj:pass  Experiencer BEING
7   не  не  PART    Polarity=Neg    8   advmod  _   PARTICLES
8   был быть    AUX _   Aspect=Imp|Gender=Masc|Number=Sing|Tense=Past|VerbForm=Fin|Voice=Act    9   aux:pass    _   AUXILIARY_VERBS
9   ранен   ранить VERB    _   Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Pass 0   root    Predicate   TO_DAMAGE_PART_OF_BODY
10  в   в   ADP _   _   11  case    _   PREPOSITION
11  инциденте   инцидент    NOUN    _   Animacy=Inan|Case=Loc|Gender=Masc|Number=Sing   9   obl Time    FACT_INCIDENT
12  .   .   PUNCT   _   _   9   punct   _   _
```

Figure 2: Our format: markup for *Как уточнил владелец бизнеса, никто не был ранен в инциденте.* 'As the business owner clarified, no one was injured in the incident.'

4 Compreno2UD Converter

The annotation of the dataset in such a format demanded the creation of the automatical converter which would transform Compreno morphosyntantic markup into UD.

The conversion program consists of several blocks. These blocks include original markup extraction, syntax and morphology conversion. The semantic layer is simply added over the resulting markup as it does not have to be converted.

The conversion pipeline is as follows:

- Labelled and manually checked texts are extracted from the Compreno system with the help of an API. On this stage, we get separate semantic and morphosyntactic data (morphological and syntactic labels are not divided technically);
- The extracted data is parsed and handed onto the syntactic module;
- Both the results of the syntax conversion and the original morphological data are passed to the morphological module, where tokenization and lemmatization issues are solved as well, and the results of both stages are merged;
- The semantic markup is merged with the results of the conversion.

The conversion of morphology and syntax can be performed in any order, so the reason for the syntax being converted first is purely technical.

Now let us consider the syntax and the morphology conversion in more detail, especially as far as the asymmetry between Compreno and UD is concerned.

5 Syntax

The description of the syntactic parsing in Compreno can be found in (Anisimovich et al., 2012). Shortly, the parser builds the dependency tree for each sentence, where each node is provided with the necessary grammatical features (both morphological and syntactic). Each dependency is marked with the surface slot (or syntactic role) such as Subject, Object_Direct, Object_Instrumental, and so on.

Compreno restores all elided nodes (such as copulas, for instance) and has a special set of labels for dislocation cases.

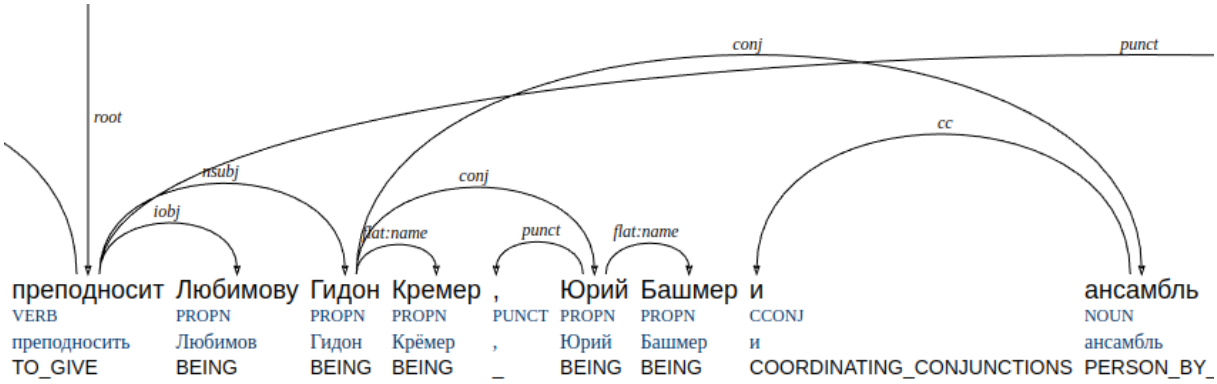
The process of syntax conversion is divided into two parts: the conversion of the heads and the conversion of the relations. Technically, the conversion of the heads must be done first, as the information about the heads is used during the relations conversion.

5.1 Dependency Heads

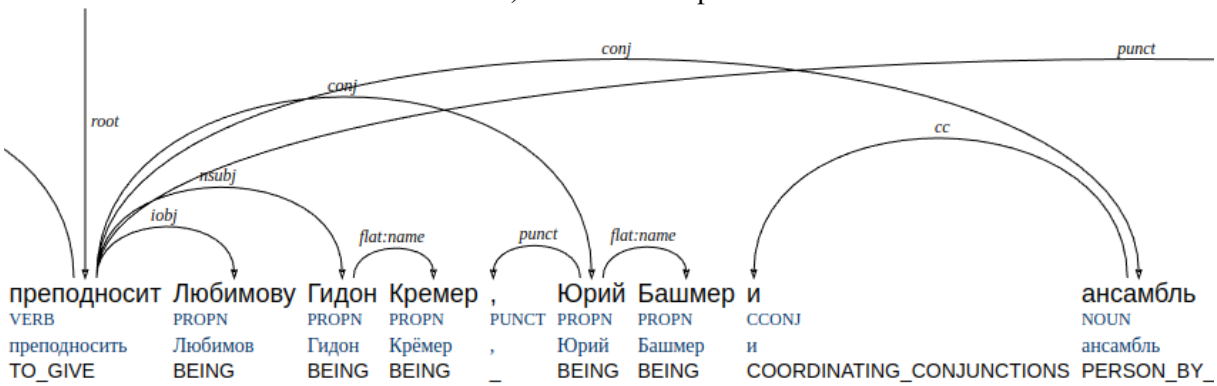
The conversion of the dependency heads from the Compreno format into the UD one seems quite straightforward to a large extent. However, there are several asymmetry cases dealing with ellipsis, coordination and movements. Moreover, the punctuation marks in Compreno are not regarded as separate nodes unlike it is in UD. Therefore, the main issues for the conversion were as follows:

- Punctuation marks had to get their dependency heads with the help of the rule-based algorithm which took into account the heads of the tokens on both sides of the mark in question;

- The cases of the elided heads were treated as close to the UD format as possible, with several rules according to the UD documentation on ellipsis. Nevertheless, this part of the conversion is prone to errors due to its rule-based nature;
- The copula in Compreno is the head of its clause, that is, in the sentence *Вася был студентом* 'Vasya was a student' the root is *был* 'was'. In UD, the root is the complement of the copula (*студентом* 'student');
- The preposition *согласно* 'according to' in Compreno is considered the head; in UD, it behaves like any other preposition, being dependent on its noun. Oddly enough, it is the only case of this kind we found in our data;
- The coordination is treated differently in UD and in Compreno - this divergence can be seen in fig. 3: as one can see, in UD, the coordinated elements depend on the first element of the coordination (*Гидон*), while in Compreno, all coordinated elements depend on the similar core (here - the verb *преподносит*). We adopted the UD concept.



a) The UD concept



b) The Compreno concept

Figure 3: Conjuncts representation (in UD style): a) the UD concept; b) the Compreno concept

For each case, we scripted the conversion to be as close to UD as possible.

Some of these differences were easy to eliminate, while the others involved a lot of discussions.

5.2 Dependency Relations

The dependency relations in UD cannot be treated as purely syntactic, as they often consider some semantic features. Compreno, on the other hand, has a strict distinction between syntax and semantics.

Therefore, the conversion of the Compreno format into UD included creating the set of rules which take into account various syntactic features, data on dependency heads and sometimes morphological and semantic categories. In most cases, it was not hard to align Compreno categories with UD dependency relation types.

The most difficult types for conversion appeared to be the following.

First of all, there are *obj*, *iobj* and *obl* relations in UD, and the distinctions between them are not purely syntactic: if there are two or more objects, one should choose the *obj* relation for the closest object and the *iobj* relation for the rest; the 'closeness' of the objects is hard to determine automatically.

Secondly, we could not truly define the *dislocated* tag, as there are no consistent features for it in Compreno, or they are difficult to derive.

We also did not implement the conversion rules for the *list*, *goeswith* and *reparandum* tags, as there were none in our dataset (typos in the data were corrected during manual semantic labelling).

6 Morphology

The morphology level is represented by POS-tags and grammatical features. As simple as it may seem, the attempts to develop a POS tagset for any language inevitably reveal some dubious areas. The same turned out to be true for the approach to the grammatical features. Both - the sets of POS and the sets of grammatical features do not coincide in the given formats. In general, we tried to follow the UD guidelines in most cases in order to be as consistent as possible with the format.

6.1 POS-tagging

Key differences between Compreno and UD in this respect are the following:

- There is no *Predicative* as a POS-tag in UD, but there is one in Compreno. This tag is assigned in cases like *Мне нужно идти* 'I must go'. Following the UD principles, we chose to convert the predicatives to adverbs, though it may be an arguable decision;
- There is no *Determiner* tag in Compreno, but there is one in UD. Taking into account that there is a closed set of tokens marked as determiners in UD (words like *мой, свой, этот, какой* 'mine', 'own', 'this', 'which'), we converted them using a list of tokens and a syntactic rule 'the head of the token in question must be a noun' (the rule helps to avoid placing the Determiner-tag in cases like *это было вчера* 'this happened yesterday');
- There is no POS-tag for proper nouns in Compreno, while there is one in UD. However, there are special grammatical features (grammemes) for proper nouns such as *Proper*, so UD POS-tags are set according to them;
- There are also inconsistencies with ordinal numerals (*одиннадцатый* 'eleventh'), which are tagged as numerals in Compreno, and as adjectives in UD (we convert them as adjectives);
- Some tokens in the Compreno format get a special POS-tag 'Invariable', which does not correspond to any of the UD tags; usually, these are discourse units and parenthetical constructions, for instance, *впрочем* 'however'. We created a special list of such tokens in order to process them according to UD.

6.2 Grammatical Features

We also encountered some asymmetry cases while mapping grammatical features. The information encoded in UD by one tag is sometimes distributed between several tags in Compreno, for example:

	UD	Compreno
Short forms	Variant=Short	ParticipleShortForm AdjectiveShortForm
Abbreviations	Abbr=Yes	Abbreviation Lex_Abbreviation Lex_KgSm Lex_LetterAbbreviation Lex_LetterDotAbbreviation

Furthermore, some grammatical categories are divided into morphological and syntactic ones in Compreno: for example, there are *Gender* and *SyntacticGender* tags for the grammatical category of gender. If some token has the *Gender=Common* tag, it means that its gender can change based on the context, that is, that the same token can function both as Feminine and Masculine. As a rule, these are surnames: *Шеварднадзе, Кириленко*, but the words like *убийца, камикадзе* also fall here, as well as some

foreign names: *Associated Press, УЕФА*. In this case, we use the information from the *SyntacticGender* tag.

Another example is the pronoun *себя* 'oneself' and alike. It does not have number and gender tags in most markups, but it gets them in Compreno in accordance with the semantic component as it inherits these categories from the controller of *себя*. In our resulting format, we decided to keep only information about case, as it is done in other UD-corpora.

6.3 Tokenization and Lemmatization

The conversion task also included the processing of lemmas and tokens since the principles of tokenization and in a lesser extent lemmatization are different in UD and in Compreno.

One of the prominent lemmatization differences is that the Compreno system puts verb lemmas in the perfect aspect, while in UD, verb lemma should have the same aspect as its form in a sentence. In order to comply with the UD format, we created a list of all verbs with both variants and restored their correct lemmas based on the aspect tag. This may be an arguable decision, as there are discussions on whether one should consider verb aspect an inflectional feature or not.

6.4 Re-tokenization

Technically, the most difficult part of the job was re-tokenizing sentences as tokenization rules for UD and Compreno differ significantly. For instance, the UD format implies that there cannot be a space inside a token, while Compreno treats many idiomatic and syntactically opaque expressions, such as кроме того 'besides', при этом 'moreover', and so on as a single unit.

To cope with this asymmetry in the conversion process, it was necessary not only to divide or split tokens in accordance with the UD standard, but also to decide what tags the parts of the split tokens should inherit. To divide and merge the tokens, we used the dictionary which was partly based on the list in the SynTagRus to UD conversion repository (257 tokens) and subsequently changed it and supplemented with new cases (now 389 tokens). Further, the list will be filled with all the tokens which include a space in the Compreno database. In this dictionary the head of a bigram is defined as a new head of a split token, and only this head inherits the semantic class label, while the others get none.

The splitting of foreign words and tokens like '*1990-1991*', company names, and time intervals demanded creating some rules as well. Such cases are innumerable and cannot be taken into account in any list, so we split them rule-based.

Another re-tokenization task was the merge of cases which could not be processed with the help of a list. The following token groups were merged:

- immutable parts of compound words (*авиа, фильмо, шведско*);
- model or product names (*Ту-34, Ил-76*);
- numerals with endings (*70-й, 19-го*).

Re-tokenization also invokes an issue of distributing semantic categories on the last stage of the markup conversion. When a token is split, its semantic slot and class are assigned to its part which would be the syntactic head in the split construction; the rest of the parts would get blanks. For example, the token *кроме того* 'besides' would be split into two parts, where *того* is the head of *кроме*. The semantic class DISCOURSIVE_UNITS would be assigned to *того*, as it is the class of the whole token *кроме того* in Compreno.

7 Compreno vs UD: challenges

As the conversion showed, there are two problems to work on further in more detail, both concern non-tree syntax.

First, UD does not restore the syntactic zeros, which leads to 'unnatural' dependencies. For example, in the phrase *Спортсменка, показав второй результат на первом участке, вылетела с трассы на втором* 'The athlete, having shown the second result in the first section, flew off the track in the second' the token *втором* 'second' depends on the verb and is marked with the *obl* relation, substituting its elided head. In the Compreno model, the elided head *участке* 'section' would be restored here, and

every constituent would get its correct tags. This difference was really hard to take into account during the conversion, so there must be inconsistencies in our dataset with such cases. As a consequence, the *orphan* relation was implemented only partially.

Second is the conversion of the constituents dislocation. For instance, let us take the sentence *Как подсказывает опыт, в классические шахматы лучшую игру демонстрируют сильнейшие шахматисты* 'As experience suggests, in classical chess the best play is demonstrated by the strongest chess players'. The correct head for the constituent *в шахматы* 'in chess' would be the head *игра* 'play' (and it is so in the Compreno format). Such information can be taken from the semantic structure of the sentence built by the parser, however, the current version of the converter does not process this information properly, and the head is assigned wrongly as *демонстрируют*. This problem is going to be solved by re-working the conversion script.

8 Further Developments

As a natural development of our work, we consider modifying our current markup format by adding the elided heads. This task will be probably tricky, as there is no satisfying concept for the labelling of the elided heads for now: it is difficult to include such nodes in the current CONLL format, because they do not have phonetically expressed forms.

As for the architecture of the converter, we will improve its work with the original parsed trees from Compreno in order to restore ellipsis and to label the dependency heads correctly in case of dislocation. Needless to say, we will focus on the correction of any bugs found in the current version of the converter.

9 Conclusion

It is commonly known that automatic conversion of any type of linguistic markup is a difficult task. In the current paper, we have shown the conversion process of the Compreno markup format into the UD morphosyntactic markup standard. The full description of the automatic conversion blocks - from tokenization to syntax - has been provided, with the focus on some fundamental differences and inconsistencies between the standards. The result of our work is a fully-labelled dataset for the Russian language which includes approximately 400,000 tokens. The dataset markup follows UD guidelines in the morphosyntactic part and is supplemented with the semantic pattern. Further work presupposes modifications in the syntax level such as restoring ellipsis.

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.

KV Anisimovich, K Yu Druzhkin, KA Zuev, FR Minlos, MA Petrova, and VP Selegei. 2012. Syntactic and semantic parser based on abbyy compreno linguistic technologies. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, P 91–103.

ES Atwell. 2008. Development of tag sets for part-of-speech tagging.

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.

Roy Bar-Haim, Khalil Sima'An, and Yoad Winter. 2008. Part-of-speech tagging of modern hebrew text. *Natural Language Engineering*, 14(2):223–251.

Igor Boguslavsky. 1999. Translation to and from russian: the etap system. // *EAMT Workshop: EU and the new languages*.

Igor Boguslavsky. 2014. Syntagrus–a deeply annotated corpus of russian. *Les émotions dans le discours-Emotions in Discourse*, P 367–380.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. // *Lrec*, volume 6, P 449–454.

Mona Diab. 2007. Improved arabic base phrase chunking with a new enriched pos tag set. // *Proceedings of the 2007 Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, P 89–96.

Kira Droganova and Daniel Zeman. 2016. Conversion of syntagrus (the russian dependency treebank) to universal dependencies. Technical report, Technical report, Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University.

Kira Droganova, Olga Lyashevskaya, and Daniel Zeman. 2018. Data conversion and consistency of monolingual corpora: Russian ud treebanks. // *Proceedings of the 17th international workshop on treebanks and linguistic theories (tlt 2018)*, number 155, P 53–66. Linköping University Electronic Press Linköping, Sweden.

Jan Hajic, Barbora Vidová-Hladká, and Petr Pajas. 2001. The prague dependency treebank: Annotation structure and support. // *Proceedings of the IRCS Workshop on Linguistic Databases*, P 105–114.

M Petrova, A Ivoylova, I Bayuk, D Dyachkova, and M Michurina. 2023. The CoBaLD Project: the creation and application of the full morpho-syntactic and semantic markup standard. // *International Conference on Computational Linguistics and Intellectual Technologies «Dialog»*.

MA Petrova. 2014. The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27(2):105–129.

Tatiana Shavrina and Olga Shapovalova. 2017. To the methodology of corpus construction for machine learning:“taiga” syntax tree corpus and parser. // *Proceedings of “CORPORA-2017” International Conference*, P 78–84.

Hiroshi Uchida and Meiying Zhu. 2001. The universal networking language beyond machine translation. // *International Symposium on Language in Cyberspace, Seoul*, P 26–27.

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.

Knowledge Transfer Between Tasks and Languages in the Multi-task Encoder-agnostic Transformer-based Models

Dmitry Karpov
MIPT
Dolgoprudny, Russia
dmitrii.a.karpov@phystech.edu

Vasily Konovalov
MIPT
Dolgoprudny, Russia
vasily.konovalov@phystech.edu

Аннотация

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by ∼0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform single-task ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings. The improvement can reach 4-5% if the Russian data are scarce enough. We have integrated these models to the DeepPavlov library and to the DREAM dialogue platform.

Keywords: multi-task, transformer, neural, dialog, emotion, sentiment, toxic, knowledge transfer, cross-lingual knowledge transfer, multi-task knowledge transfer, conversational tasks
DOI: 10.28995/2075-7182-2023-22-200-214

Перенос знаний между задачами и языками в многозадачных энкодер-агностичных моделях типа Трансформер

Дмитрий Карпов
Василий Коновалов
Московский физико-технический институт / Долгопрудный, Россия
dmitrii.a.karpov@phystech.edu vasily.konovalov@phystech.edu

Аннотация

В статье изучается перенос знаний в многозадачных энкодер-агностичных моделях типа Трансформер для пяти диалоговых задач – классификации эмоций, сентимента, токсичности, интентов и тематической классификации. В статье показано, что эти модели демонстрируют точность, отличающуюся от аналогичных однозадачных моделей примерно на 0.9%. Эти результаты верны для разных типов трансформеров. В то же время эти многозадачные модели имеют примерно на 0.1% больше параметров, чем любая из аналогичных однозадачных моделей. В статье также показывается, что начиная с определенного достаточно маленького размера набора данных, многозадачные модели начинают превосходить однозадачные модели, особенно на тех задачах, для которых меньше всего данных. Помимо этого, при обучении многоязычных моделей на русскоязычных данных, добавление англоязычных данных в обучающую выборку дополнительно улучшает результат многоязычных моделей в однозадачном и многозадачном режиме. Улучшение может достигать 4-5%, если русскоязычных данных достаточно мало. Эти модели также были интегрированы в библиотеку DeepPavlov и диалоговую платформу DREAM.

Ключевые слова: многозадачные, трансформер, нейросетевые, диалог, эмоции, тональность, токсичность, перенос знаний, межязыковой перенос знаний, многозадачный перенос знаний, разговорные задачи

1 Introduction

Transformer-based models, such as BERT, are widely used for text classification. The original article (Devlin et al., 2019) proposed the use of a separate BERT model for each task in multi-task benchmarks. Therefore, if several classification tasks need to be solved in parallel, several prediction models should be employed, which increases the demand for computational resources. One of the ways of tackling this problem is training one single model that can yield results for these tasks simultaneously.

Multi-task learning (MTL) is one of the transfer-learning techniques. It allows training one single model simultaneously for multiple related tasks so that the knowledge acquired in one task enhances another task's performance.

The real-world conditions require making a trade-off between the quality of neural models and their use of computational resources. Responding to this tradeoff necessitates the use of encoder-agnostic multitask transformer-based models, which allows quick replacement of the transformer backbone for different circumstances. The `transformers` (Wolf et al., 2020) library allows using different transformer-based models including distilled ones to save computational resources and speed up the inference time (Kolesnikova et al., 2022).

Our contributions are as follows:

1. We show how multi-task knowledge transfer occurs in the simple encoder-agnostic transformer-based models during training for multiple dialogue-related tasks.
2. We explore the effects of multi-lingual knowledge transfer in these models.
3. We implement these models in `DeepPavlov` framework.[1]

2 Related Work

Researchers have been conducting experiments with multi-task learning (MTL) for a long time (Caruana, 1997). Since the rise of neural networks, researchers have proposed a wide range of approaches to MTL, including cross-lingual word embeddings (Konovalov and Tumunbayarova, 2018). However, these methods did not develop further, as nowadays NLP is based on transformer-based models. Nevertheless, as transformer architectures come out quite often, this review mostly focuses on agnostic architectures, which work with all kinds of transformers, rather than transformer-specific architectures.

In some kinds (Karpov and Burtsev, 2021) of multi-task encoder-agnostic transformer-based architectures, every sample needs to be labeled or pseudo-labeled for all considered tasks. Even though this approach is successfully used in some dialogue systems (Kuratov et al., 2021), it lacks flexibility.

One of the most frequently used encoder-agnostic transformer-based architectures is (Liu et al., 2019). However, this architecture increases computational demands due to the specific stochastic attention layers for text classification.

The work (Asa Cooper Stickland, 2019) proposed different encoder-agnostic ways to work with BERT output in a multi-task setting.[2] One such way is to supplement the model with an extra BERT layer for each task. However, that way increases the number of required parameters for GLUE (Wang et al., 2018) by 67%, which is computationally heavy. Other encoder-agnostic approaches proposed in the same work worked no better than the plain use of *bert-base-uncased* output in the linear classifier in our experiments on GLUE.

Additionally, utilizing self-attention with a task-embedded module from the paper (Maziarka and Danel, 2021) instead of plain self-attention in the low-rank transformation did not yield improvements over the plain dense task-specific layers on top of BERT in our experiments. The task-embedded architecture presented in the same article is still not encoder-agnostic.

Another work (Huang et al., 2021) suggested a novel way to extract additional features from the BERT output – using lightweight convolutional ghost modules. Despite this approach being encoder-agnostic, utilizing attention with a ghost module in the low-rank transformation did not yield improvements over the plain dense task-specific layers above BERT in our experiments. This also holds for (Ali et al., 2021) architecture from computer vision.

[1] `https://github.com/deeppavlov/DeepPavlov/`

[2] Projective attention layers, presented in the same article as the superior result, are not encoder-agnostic.

At the same time, the performance of simple encoder-agnostic transformer-based models is still not fully explored. It is especially true for dialogue-specific datasets. Furthermore, the body of work lacks studies on the Russian language multi-task learning in general, and specifically on the dialogue tasks. Multilingual knowledge transfer in multi-task models for such tasks also remains unexplored. Our work is aimed to bridge this gap.

3 MTL Model Description

The MTL architecture we explore allows using different encoder-only Transformer architectures as a backbone. For our experiments, we utilized BERT-based models because they allow effective transfer learning (Chizhikova et al., 2023; Konovalov et al., 2020). However, the same approach can be applied to any Transformer-based model.

1. In the same way as in the original work (Devlin et al., 2019), we return the final hidden states for all tokens and apply the BERT pooling layer to them. Like in this article, we apply the pooler output.
2. Then, for every task, we apply the task-specific linear layer to the pooler output. The task-specific linear layer for every task type looks exactly like the linear fine-tuning layer for the single-task BERT models (see original article or Transformers (Wolf et al., 2020) manual).
3. Then we apply a loss function: mean squared error loss for the regression tasks, categorical cross-entropy loss for single-label tasks or binary cross-entropy loss for multi-label classification task. In this paper, we consider only the single-label classification.

The multi-task model in this setting requires almost no additional parameters and computational overhead, apart from the linear layers, so its simplicity singles it out. Also, the flexibility of this model allows using it with different kinds of backbones, which positively distinguishes it from (Asa Cooper Stickland, 2019).

For the distilBERT-like models, this multi-task model takes only 0.1% more parameters than single-task models. This computational overhead varies around this number, depending on the number of tasks, the number of classes, and the backbone model.

The encoder-agnostic multi-task transformer-based model is integrated into DeepPavlov (Burtsev et al., 2018). This implementation supports all Transformer backbones from the `AutoModel` class from HuggingFace. Our implementation is also successfully used in the Dream dialogue platform (Baymurzina et al., 2021).

4 Datasets

We explored the multi-task models' performance on the Russian and English datasets for five tasks, i.e. emotion classification, toxicity classification, sentiment classification, intent classification, and topic classification. For Russian and English data, the indexes of the same classes used by the models were also the same. We chose these tasks as they are pivotal for dialog systems (Kuratov et al., 2020). The datasets contain naturally occurring data, which are useful for dialogue systems development (Konovalov et al., 2016b; Konovalov et al., 2016a). Therefore, we consider these tasks to be conversational tasks.

4.1 Emotion Classification

We used the `go_emotions` dataset (Demszky et al., 2020) for emotion classification in English. This dataset consists of short comments from Reddit, such as *LOL. Super cute!* or *Yikes. I admire your patience*. We used Ekman-grouped emotions, grouping them into seven types, i.e *anger*, *fear*, *disgust*, *joy*, *surprise*, *sadness*, and *neutral*. After such grouping, we selected only single-label examples. There were 39,555 training examples of that kind. The train/test/validation split of this dataset was approximately 80/10/10.

For the same task in Russian, we used the `CEDR` dataset (Sboev et al., 2021). The dataset contains examples from different social sources: blogs, microblogs, and news. This dataset has five classes – *anger*, *fear*, *joy*, *surprise*, and *sadness* – but the samples from this dataset can belong to more than one single class or (unlike `go_emotions`) belong to no class. For example, the text *Надо утопать на встречу.* belongs to no class.

From this dataset, we selected only examples that belong to one single class or that have no class, labeling no-class examples as *neutral*. The class nomenclature of this dataset was almost the same as for the English dataset, except for the *disgust* class. Nonetheless, as *disgust* examples comprised less than 1.5% of the English training samples, it didn't impact knowledge transfer much.

The work (Sboev et al., 2021) provided only the train-test split of the CEDR dataset, which is 80/20. We singled out 12.5% of the training examples from CEDR as the validation set. The resulting dataset has 6,557 training samples.

4.2 Sentiment Classification

We used the DynaSent(r1) dataset (Potts et al., 2020) for sentiment classification in English. It contains naturally occurring sentences. i.e. *Need a cheap spatula?* We used only examples from the first round of the collection, to match the Russian data by difficulty. This single-label dataset with 80,488 training samples has three classes – positive, negative, and neutral. The dataset has 3,600 validation samples and the same number of test samples.

For the same task in Russian, we used the RuReviews dataset (Smetanin and Komarov, 2019). This three-class dataset consists of 90,000 product reviews from the "Women's Clothes and Accessories" category of a large Russian e-commerce website. As the considered product reviews already contain grades from the user, the authors of this dataset classified sentiment according to the grades. For example, the phrase *размер очень мал* was considered to be negative. We have chosen this dataset because it is open source and it has a relatively large size, even though it is domain-specific. As the train/validation/test split of this dataset was not provided, we used the same split as in the DynaSent(r1) dataset. After that, the training set had 82,610 training samples.

4.3 Toxicity Classification

For English toxic classification, we used the Wiki Talk dataset (Dixon et al., 2018). This Wikipedia comment dataset has two classes: toxic and not toxic. Unsurprisingly, the dataset contains vulgar slang. However, about 90% of examples from this dataset are not toxic, i.e. *Hi! so umm i guess yer incharge here hehehe. so wassup?*. This dataset has 127,656 training samples, 93,342 validation samples, and 31,915 testing samples. For Russian toxic classification, we used the RuToxic dataset (Dementieva et al., 2021). This two-class dataset was collected from Dvach, a Russian anonymous imageboard. This dataset originally has 163,187 samples. Among them, most of the samples are not toxic, e.g. *ещё бы. какой красавец.*. But obviously, some samples are toxic, e.g. *дворника тоже надо уничтожить!*. As the authors didn't provide the original split in their repository, we split this dataset in the same proportions as in the Wiki Talk dataset. After that, the training set had 93,342 training samples.

4.4 Intent Classification and Topic Classification

We used MASSIVE dataset (FitzGerald et al., 2022) for the intent classification for the Russian and English languages. The MASSIVE dataset for the English language contains the spoken utterances, which aim for the voice assistant, e.g. *play rock playlist*. All examples in this dataset were labeled and adapted simultaneously for 51 languages, including Russian.[3] This dataset has 11,514 train samples, 2,033 validation samples, and 2,974 test samples. Every sample belongs to one of 60 intent classes. This dataset is widely used for the conversational topic classification (Karpov and Burtsev, 2023).

We used the same dataset in the same way for topic classification as well, as this dataset is labeled by intent and by topic. Every sample from this dataset belongs to one of the 18 topic classes.

5 Experimental Setup

For all the experiments described in our work, the optimizer was AdamW (Kingma and Ba, 2015) with betas (0.9, 0.99), and the initial learning rate was 2e-5. We used average accuracies for all tasks as an early stop metric. The training had validation patience 3, and the learning rate was dropped by two times if the early stopping metric did not improve for two epochs.

[3]For example, the Russian dataset contains sample *расскажи новости russia today* instead of *stell me b. b. c. news*.

Table 1: Accuracy / F1-macro on the English data for the encoder-agnostic transformer-based model. English cased models trained on English data. Mode S stands for single-task, and mode M stands for multi-task.

Model	Mode	Average	Emotions 39.4k	Sentiment 80.5k	Toxic 127.6k	Intents 11.5k	Topics 11.5k	Batches seen
distilbert	S	**82.9/78.4**	**70.3/63.1**	74.7/74.3	**91.5/81.2**	**87.4/82.7**	**91.0/90.6**	11390
distilbert	M	82.1/77.2	67.7/60.7	**75.2/75.0**	90.6/79.8	86.3/80.4	90.8/90.1	14000
bert	S	**83.9/79.7**	**71.2/64.2**	76.1/75.8	**93.2/83.5**	**87.9/84.2**	**91.3/90.7**	9470
bert	M	83.0/78.4	69.0/63.1	**76.5/76.4**	91.4/80.8	87.1/81.2	91.2/90.6	11760

The training was usually completed in less than 10-15 epochs and never exceeded 25 epochs, even though the maximum number of epochs was set to 100.

We set the batch size to 160. We have also tried batch size 32, and the metrics for batch size 160 were just insignificantly better. However, the paper (Godbole et al., 2023) claims that this difference can be eliminated by better fine-tuning. Finally, we settled with batch size 160 because the computations with batch size 160 were performed several times faster.

In the preliminary multi-task experiments, apart from plain sampling (a sampling mode where the example sampling probability is proportional to the task size), we also tried annealed sampling (Asa Cooper Stickland, 2019) and uniform sampling (the same sampling probability for all tasks). We performed such experiments for Russian and English distilbert-like models, for Russian and English tasks. The results for these sampling modes did not bring out a noticeable improvement, thus we used only plain sampling.

We averaged all the experiment results by three runs.

6 Results and Analysis

We conducted experiments in mono-lingual mode with different transformer-based backbones to compare single-task and multi-task approaches. For the English-language tasks, we conducted the experiments for the backbones *bert-base-cased* (Devlin et al., 2019) (*bert*) and *distilbert-base-cased* (Sanh et al., 2019) (*distilbert*).

For the Russian-language tasks, we made experiments for the backbones *DeepPavlov/rubert-base-cased-conversational* (Kuratov and Arkhipov, 2019) (*rubert*) and *DeepPavlov/distilrubert-base-cased-conversational* (Kolesnikova et al., 2022) (*distilrubert*).

As distilled BERTs take 40% less memory than BERTs and are 60% faster, these experiments cover a variety of different model uses for different computational budgets and quality demands.

6.1 Single-task VS Multi-task: Backbones From Different Languages

In the first stage of the experiments, we compared the performance of our multi-task models to analogous single-task models with the same hyperparameters.

We present the results of the first stage of experiments in Tables 1-2. For every experiment, we provide accuracy / macro-averaged F1.

Overall, the performance of multi-task encoder-agnostic transformer-based models closely matches the performance of the analogous single-task models. This effect holds for the Russian language as well as for the English language.

While *distilbert* shows slightly worse metrics than *bert*, *distilrubert* even excels *rubert* on all but the largest tasks.

In the next experiments, we put the main focus on the distilbert-like models to speed up the computations.

Table 2: Accuracy / F1-macro on the Russian data for the encoder-agnostic transformer-based model. Russian cased models trained on Russian data. Mode S stands for single-task, and mode M stands for multi-task. RU means that models were trained and evaluated on Russian data, EN means that models were trained and evaluated on English data.

Model	Mode	Average	Emotions 6.5k	Sentiment 82.6k	Toxic 93.3k	Intents 11.5k	Topics 11.5k	Batches seen
distilrubert	S	**86.9/84.1**	**82.2/76.1**	**77.9/78.2**	**97.1/95.4**	**86.7/81.6**	90.4/89.5	8472
distilrubert	M	86.3/82.6	81.0/74.6	77.7/77.7	96.9/95.0	85.2/75.9	**90.7/89.9**	8540
rubert	S	**86.5/83.4**	**80.9/75.3**	**78.0/78.2**	**97.2/95.6**	**86.2/79.1**	90.0/89.0	7999
rubert	M	86.2/82.6	80.5/73.8	77.6/77.6	96.8/95.0	85.3/76.9	**90.5/89.8**	8113

Table 3: Accuracy / F1-macro on the Russian data for the encoder-agnostic transformer-based model. Multilingual cased models, batch size 160, plain sampling. Mode S stands for single-task, and mode M stands for multi-task. In the 'Training data' column, RU stands for the Russian language, 'RU+EN' means that Russian and English data are merged by task, and 'RU $\oplus$ EN' means that Russian and English tasks are treated as separate tasks.

Model	Training data	Mode	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
distilbert-mult	RU	S	**84.7/81.0**	77.4/69.1	**77.7/77.9**	**96.7/94.8**	**83.5/76.6**	88.1/86.9	10058
distilbert-mult	RU	M	84.3/80.2	**78.1/70.5**	76.8/76.7	96.5/94.4	81.9/72.3	**88.2/87.1**	9821
distilbert-mult	RU+EN	S	**85.2/81.8**	**78.9**/70.2	**77.4/77.3**	**96.8/94.9**	**84.7/79.1**	**88.4/87.4**	31843
distilbert-mult	RU+EN	M	84.5/81.1	77.9/**70.7**	76.6/76.7	96.5/94.5	82.9/76.5	**88.4**/87.2	17790
distilbert-mult	RU $\oplus$ EN	M	84.4/80.6	77.6/70.0	76.8/77.1	96.5/94.5	82.4/73.9	88.3/87.2	23688
bert-mult	RU	S	84.7/80.2	76.6/64.2	**77.8/78.2**	**96.9/95.1**	**83.9**/76.3	88.4/87.0	10884
bert-mult	RU	M	**84.8/81.4**	**78.4/71.4**	76.3/76.3	96.8/94.8	83.7/**76.6**	**89.0/87.8**	12810
bert-mult	RU+EN	S	**85.6/82.3**	78.9/70.1	**77.6/77.8**	**96.9/94.9**	**85.0/80.4**	**89.4/88.5**	23752
bert-mult	RU+EN	M	85.2/**82.3**	**79.2/72.7**	76.4/76.6	96.7/94.8	84.3/79.3	**89.4**/88.3	20755
bert-mult	RU $\oplus$ EN	M	85.0/81.6	78.3/71.4	77.1/77.0	96.7/94.7	84.0/76.7	89.1/88.0	22701

6.2 Multilingual Multi-task Backbones: Cross-lingual Training Impact

In the next stage of experiments, we have put the focus on multilingual knowledge transfer. To investigate this transfer, we utilized only multilingual backbones. In particular, we used *distilbert-base-multilingual-cased* and *bert-base-multilingual-cased*. In Table 3, we label them as *distilbert-mult* and *bert-mult*, respectively. Our main goals were:

- To compare the performance of the multi-task and single-task models with the multilingual backbones for the Russian language.
- To check how the performance of single-task models and the performance of multi-task models varies if we add the English data to them, and the data are merged by task (for every task, the model is trained on English+Russian training data and validated on Russian data).
- To check whether treating English-language tasks as separate tasks yields any improvements if we perform the validation on the Russian data.

As we see, the results of all settings are pretty similar: using Russian+English data puts us on the plateau, while improvements compared to using only Russian data are only moderate.

In the same setting, we also explored whether utilizing English-language tasks as separate tasks is more beneficial than merging Russian and English data by task. This approach did not prove to be any better and even brought out a small deterioration.

The impact of adding English data in case of having limited Russian data required additional investigation. We have researched this impact in the next series of experiments. In real-world conditions, we usually have a huge body of datasets for English data, but not nearly as much for Russian data. This

gives additional practical value to that experiments.

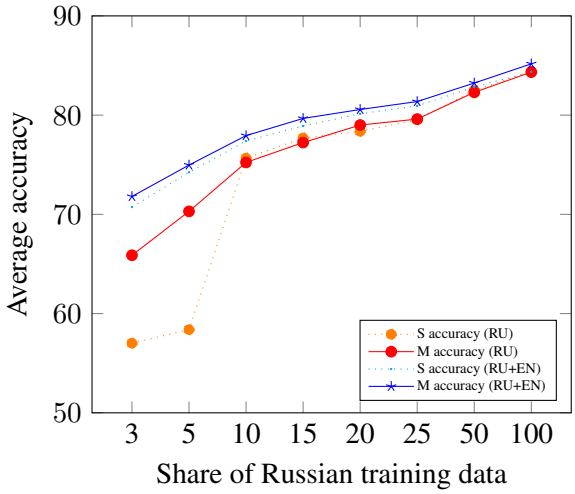
6.3 Impact of Adding the English Data

In this experiment series, we explored multi-task and single-task settings with Russian and English data merged by task. We studied how much the performance of *distilbert-base-multilingual-cased* (multi-task or single-task) improves when it is trained on some part of Russian train data if we add English training data to it and validate on the English validation data.

Specifically, we performed experiments for the following data shares: 0%, 3%, 5%, 15 %, 20%, 25%, 50%, and 100%. For 0%, we added to the table the model trained on English train data and validated on Russian validation data, and the model which is trained on English train data and validated on English validation data (but tested still on Russian test data). We restarted the experiments with three random seeds. For every series of experiments, we randomly shuffled the datasets and then selected all subsets at once, while the larger subsets contained all examples from the smaller subsets (like, 10% subset contains all examples from 5% and also from 3%)

We present the averaged results in Table 9, in Appendix. We averaged the results by three runs. For training on the 3-5% of the Russian data without the English data, we averaged the results by five runs due to the high variability of results. Additionally, we plot the results below, in Figure 1. The task-wise results for the experiments with data shares are also shown in Appendix, in Table 9.

We also note that in all the experiments from Table 9 where 100% share of the English data was used, we performed the experiments also with validation on the Russian data instead of the English data. That change did not impact the scores in any meaningful way (see Table 10).



RU share	S RU	M RU	S RU+EN	M RU+EN
3	57.0	**65.9**	**71.8**	70.7
5	58.4	**70.3**	**75.0**	74.2
10	**75.7**	75.2	**77.9**	77.4
15	**77.7**	77.2	**79.7**	78.9
20	78.4	**79.0**	**80.6**	80.1
25	79.5	**79.6**	**81.4**	80.9
50	**82.5**	82.3	**83.2**	82.8
100	**84.4**	84.3	**85.2**	84.4

Figure 1: Average accuracy on the Russian data for *distilbert-base-multilingual-cased*, batch size 160, plain sampling. 'S' stands for single-task mode, 'M' stands for multi-task mode, 'RU share' means the share of Russian training data, 'RU' means training only on the given percentage of Russian data, 'RU+EN' means training on the given percentage of Russian data with added full size English data. See Table 9 for more details.

For the Russian-only data, starting with a small enough percentage of the training data, the single-task metrics drop and become much lower than the multitask metrics. We do not see this effect for the Russian+English data, as in this case, even with a low share of Russian data, even single-task models still learn a much higher amount of knowledge from the English data.

7 Discussion

Multi-task encoder-agnostic transformer-based models almost match the single-task models by metrics on the dialogue tasks. The gap in average accuracy between the multi-task and single-task monolingual models is about 0.8-0.9% for the English language and about 0.3-0.6% for the Russian language. For the

multilingual models, the gap remains within the same limit, except for the *bert-base-multilingual-cased* trained only on Russian data, for which there is no gap.

We also show that if we train the multilingual model and have Russian and English data for the same tasks with the same classes, combining that into one task is slightly better than treating Russian and English tasks as separate tasks. We can explain it by the fact that while training multilingual models on merged data (see Table 3), the knowledge is transferred by the backbone and by the class-specific linear layers. At the same time, while training multilingual models on separate data, only knowledge transfer by the backbone takes place.

For the small-scale data, we can see that if we train the multilingual distilbert on small shares of Russian training data (2-5%), multi-task models outperform single-task models in the average accuracy. The Talbe 9 shows that this accuracy advantage increases while the dataset size decreases. For intent and topic datasets, this advantage disappears at 1,151 training samples. For the emotion dataset, surprisingly, this advantage holds with any dataset partition, possibly due to the effect of knowledge transfer from the sentiment task.

For experiments with adding English data, multi-task models showed no clear pattern of advantage over single-task ones. This fact also supports the hypothesis of the knowledge transfer dependency of the dataset size. If we added 100% of English training data, dataset sizes became too large for reaching the advantage from the multi-task knowledge transfer.

However, adding the English training data to the Russian training data improves the metrics on the Russian test set. The lower the size of the Russian training data we have, the more substantial the accuracy increase from adding the English data to the training sample. This accuracy gain can reach several percent if we have a limited amount of Russian training data (3-10% RU share in Table 9). This conclusion holds for multi-task and single-task models. The language of validation data (English or Russian) did not matter in our experiments.

The reason for metric improvement for the multilingual models by adding the English data is that while being pretrained on certain languages (in our cases - English and Russian), the models learn to represent the language-independent features of the examples. Therefore, while receiving Russian and English examples for the same tasks, the models fine-tune to the larger number of language-independent features and generalize more broadly, which helps to improve the results.

Our work did not cover the knowledge transfer to languages other than Russian. Also, we did not consider conditions under which multilingual models, with knowledge transferred from English data, excel analogous Russian-only models. We leave that for future work.

8 Conclusion

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by $\sim$0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform single-task ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings - up to 4-5% if the Russian data are scarce enough. We also have integrated these models into the DeepPavlov framework and into the DREAM library.

References

Alaaeldin Ali, Hugo Touvron, Mathilde Caron, Piotr Bojanowski, Matthijs Douze, Armand Joulin, Ivan Laptev, Natalia Neverova, Gabriel Synnaeve, Jakob Verbeek, and Herve Jegou. 2021. Xcit: Cross-covariance image transformers. // M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, *Advances in Neural Information Processing Systems*, volume 34, P 20014–20027. Curran Associates, Inc.

Iain Murray Asa Cooper Stickland. 2019. Bert and pals: Projected attention layers for efficient adaptation in multi-task learning. // *Proceedings of the 36th International Conference on Machine Learning*, volume 97, P 5986:5995.

Dilyara Baymurzina, Denis Kuznetsov, Dmitry Evseev, Dmitry Karpov, Alsu Sagirova, Anton Peganov, Fedor Ignatov, Elena Ermakova, Daniil Cherniavskii, Sergey Kumeyko, et al. 2021. Dream technical report for the alexa prize 4. *4th Proc. Alexa Prize*.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: An open source library for conversational ai. // *NIPS*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Anastasia Chizhikova, Vasily Konovalov, and Mikhail Burtsev. 2023. Multilingual case-insensitive named entity recognition. // Boris Kryzhanovsky, Witali Dunin-Barkowski, Vladimir Redko, and Yury Tiumentsev, *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, P 448–454, Cham. Springer International Publishing.

Daryna Dementieva, Daniil Moskovskiy, Varvara Logacheva, David Dale, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Russian toxicity dataset from 2ch.hk. dataset retrieved from `https://github.com/s-nlp/rudetoxifier`. *CoRR*, abs/2105.09052.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan S. Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547.

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.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification(paper for wiki talk dataset, cleaned version of the dataset retrieved from `https://huggingface.co/datasets/OxAISH-AL-LLM/wiki_toxic`). // *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, P 67–73, New York, NY, USA. Association for Computing Machinery.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2022. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado†. 2023. Tuning neural networks (google research github). `https://github.com/google-research/tuning_playbook`.

Zhiqi Huang, Lu Hou, Lifeng Shang, Xin Jiang, Xiao Chen, and Qun Liu. 2021. Ghostbert: Generate more features with cheap operations for bert. // *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, P 6512–6523, 01.

Dmitry Karpov and Michail Burtsev. 2021. Data pseudo-labeling while adapting bert for multitask approaches. // *Proceedings of the International Conference "Dialogue 2021"*.

Dmitry Karpov and Mikhail Burtsev. 2023. Monolingual and cross-lingual knowledge transfer for topic classification.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. // Yoshua Bengio and Yann LeCun, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Alina Kolesnikova, Yuri Kuratov, Vasily Konovalov, and Mikhail Burtsev. 2022. Knowledge distillation of russian language models with reduction of vocabulary.

VP Konovalov and ZB Tumunbayarova. 2018. Learning word embeddings for low resource languages: the case of buryat. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 331–341.

Vasily Konovalov, Ron Artstein, Oren Melamud, and Ido Dagan. 2016a. The negochat corpus of human-agent negotiation dialogues. // *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, P 3141–3145, Portorož, Slovenia, May. European Language Resources Association (ELRA).

Vasily Konovalov, Oren Melamud, Ron Artstein, and Ido Dagan. 2016b. Collecting Better Training Data using Biased Agent Policies in Negotiation Dialogues. // *Proceedings of WOCHAT, the Second Workshop on Chatbots and Conversational Agent Technologies*, Los Angeles, September. Zerotype.

Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, and Mikhail Burtsev. 2020. Exploring the bert cross-lingual transfer for reading comprehension. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis*, P 445–453.

Yuri Kuratov and Mikhail Y. Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *CoRR*, abs/1905.07213.

Yuri Kuratov, Idris Yusupov, Dilyara Baymurzina, Denis Kuznetsov, Daniil Cherniavskii, Alexander Dmitrievskiy, Elena Ermakova, Fedor Ignatov, Dmitry Karpov, Daniel Kornev, et al. 2020. Dream technical report for the alexa prize 2019. *Alexa Prize Proceedings*.

Y M Kuratov, I F Yusupov, D R Baymurzina, D P Kuznetsov, D V Cherniavskii, A Dmitrievskiy, E S Ermakova, F S Ignatov, D A Karpov, D A Kornev, T A Le, P Y Pugin, and M S Burtsev. 2021. Socialbot dream in alexa prize challenge 2019. *Proceedings of Moscow Institute of Physics and Technology*, 13(3):62–89.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-Task Deep Neural Networks for Natural Language Understanding. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 4487:4496.

Lukasz Maziarka and Tomasz Danel. 2021. Multitask learning using BERT with task-embedded attention. // *2021 International Joint Conference on Neural Networks (IJCNN)*, P 1–6. IEEE, 7.

Christopher Potts, Zhengxuan Wu, Atticus Geiger, and Douwe Kiela. 2020. Dynasent: A dynamic benchmark for sentiment analysis. *CoRR*, abs/2012.15349.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. *CoRR*, abs/1910.01108.

Alexander Sboev, Aleksandr Naumov, and Roman Rybka. 2021. Data-driven model for emotion detection in russian texts. *Procedia Computer Science*, 190:637–642.

Sergey Smetanin and Michail Komarov. 2019. Sentiment analysis of product reviews in russian using convolutional neural networks. // *2019 IEEE 21st Conference on Business Informatics (CBI)*, volume 01, P 482–486, July.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. // *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, P 353:355.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 38–45, Online, October. Association for Computational Linguistics.

9 Appendix

Table 4: Dataset sizes for the emotion classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	39555	4946	4968	6557	864	1862
joy	15216	1941	1863	1346	162	341
neutral	12823	1592	1606	2682	361	734
anger	4293	555	572	339	45	121
surprise	3858	459	488	491	77	165
sadness	2326	266	283	1207	158	368
fear	541	72	80	492	61	133
disgust	498	61	76	0	0	0

Table 5: Dataset sizes for the toxicity classification task, Russian and English data

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	127656	31915	63978	93342	23010	46835
not_toxic	114722	28624	57735	75452	18669	37659
toxic	12934	3291	6243	17890	4341	9176

Table 6: Dataset sizes for the sentiment classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	80488	3600	3600	82610	3695	3695
positive	21391	1200	1200	27570	1220	1210
neutral	45076	1200	1200	27531	1234	1235
negative	14021	1200	1200	27509	1241	1250

Table 7: Dataset sizes for the topic classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	11514	2033	2974	11514	2033	2974
calendar	1688	280	402	1688	280	402
play	1377	260	387	1377	260	387
qa	1183	214	288	1183	214	288
email	953	157	271	953	157	271
iot	769	118	220	769	118	220
general	652	122	189	652	122	189
weather	573	126	156	573	126	156
transport	571	110	124	571	110	124
lists	539	112	142	539	112	142
news	503	82	124	503	82	124
recommendation	433	69	94	433	69	94
datetime	402	73	103	402	73	103
social	391	68	106	391	68	106
alarm	390	64	96	390	64	96
music	332	56	81	332	56	81
audio	290	35	62	290	35	62
takeaway	257	44	57	257	44	57
cooking	211	43	72	211	43	72

Table 8: Dataset sizes for the intent classification task, Russian and English data.

data type	train	valid	test	train	valid	test
class	EN			RU		
Total	11514	2033	2974	11514	2033	2974
calendar_set	810	131	209	810	131	209
play_music	639	123	176	639	123	176
weather_query	573	126	156	573	126	156
calendar_query	566	102	126	566	102	126
general_quirky	555	105	169	555	105	169
qa_factoid	544	90	141	544	90	141
news_query	503	82	124	503	82	124
email_query	418	73	119	418	73	119
email_sendemail	354	63	114	354	63	114
datetime_query	350	64	88	350	64	88
calendar_remove	312	47	67	312	47	67
play_radio	283	46	72	283	46	72
social_post	283	50	81	283	50	81
qa_definition	267	55	57	267	55	57
transport_query	227	36	51	227	36	51
cooking_recipe	207	41	72	207	41	72
lists_query	198	50	51	198	50	51
play_podcasts	193	34	63	193	34	63
recommendation_events	190	26	43	190	26	43
alarm_set	182	31	41	182	31	41
lists_createoradd	177	25	39	177	25	39
recommendation_locations	173	31	31	173	31	31
lists_remove	164	37	52	164	37	52
music_query	154	30	35	154	30	35
iot_hue_lightoff	153	17	43	153	17	43
qa_stock	152	24	26	152	24	26
play_audiobook	150	35	41	150	35	41
qa_currency	142	32	39	142	32	39
takeaway_order	135	20	22	135	20	22
alarm_query	130	19	34	130	19	34
transport_ticket	127	25	35	127	25	35
email_querycontact	127	16	26	127	16	26
iot_hue_lightchange	125	22	36	125	22	36
iot_coffee	124	14	36	124	14	36
takeaway_query	122	24	35	122	24	35
transport_traffic	117	22	15	117	22	15
music_likeness	113	16	36	113	16	36
play_game	112	22	35	112	22	35
audio_volume_mute	110	15	32	110	15	32
audio_volume_up	110	12	13	110	12	13
social_query	108	18	25	108	18	25
transport_taxi	100	27	23	100	27	23
iot_cleaning	93	19	26	93	19	26
alarm_remove	78	14	21	78	14	21
qa_maths	78	13	25	78	13	25
iot_hue_lightdim	76	17	21	76	17	21
iot_hue_lightup	76	12	27	76	12	27
general_joke	72	15	19	72	15	19
recommendation_movies	70	12	20	70	12	20
email_addcontact	54	5	12	54	5	12
datetime_convert	52	9	15	52	9	15
iot_wemo_off	52	5	18	52	5	18
audio_volume_down	52	8	11	52	8	11
music_settings	51	8	6	51	8	6
iot_wemo_on	48	7	10	48	7	10
general_greet	25	2	1	25	2	1
iot_hue_lighton	22	5	3	22	5	3
audio_volume_other	18	0	6	18	0	6
music_dislikeness	14	2	4	14	2	4
cooking_query	4	2	0	4	2	0

Mode	RU share	EN share	Validated on	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
S	100%	100%	EN	85.2/82.1	79.0/70.8	77.2/77.4	96.5/94.5	84.5/80.6	88.4/87.4	15946
M	100%	100%	EN	84.4/80.9	77.2/70.5	75.8/75.8	96.4/94.4	83.5/76.3	88.9/87.8	20737
S	50%	100%	EN	83.2/79.5	75.6/65.8	75.6/75.7	96.1/93.9	82.2/76.5	86.8/85.5	16672
M	50%	100%	EN	82.8/78.1	76.2/64.5	74.0/73.4	95.9/93.5	80.9/72.7	87.2/86.1	19336
S	25%	100%	EN	81.4/76.7	73.7/61.4	73.7/73.9	95.5/92.7	78.8/71.9	85.1/83.6	16589
M	25%	100%	EN	80.9/76.4	73.1/63.9	73.7/73.7	95.1/92.2	77.5/68.1	85.3/83.9	16665
S	20%	100%	EN	80.6/76.0	71.8/60.3	74.0/74.0	95.1/92.1	78.0/71.1	83.9/82.4	12951
M	20%	100%	EN	80.1/75.0	71.9/61.2	73.5/73.5	94.9/91.9	76.1/65.5	84.2/82.8	17429
S	15%	100%	EN	79.7/74.7	70.8/57.8	72.6/72.7	94.6/91.3	77.3/70.1	83.1/81.6	13037
M	15%	100%	EN	78.9/73.5	70.0/58.4	71.9/71.5	94.5/91.2	74.7/65.0	83.5/81.8	15599
S	10%	100%	EN	77.9/72.0	68.3/52.1	72.3/72.7	93.9/90.0	73.9/65.8	81.2/79.4	13545
M	10%	100%	EN	77.4/70.9	67.9/51.2	71.7/71.7	93.7/90.1	72.3/61.5	81.6/79.9	14471
S	5%	100%	EN	75.0/67.9	64.1/45.0	70.2/70.4	92.7/87.8	69.9/60.5	77.9/75.8	12567
M	5%	100%	EN	74.2/66.3	63.4/41.2	70.1/70.2	92.3/87.6	67.6/56.6	77.7/75.9	12779
S	3%	100%	EN	71.8/64.6	59.1/38.8	68.4/68.6	91.0/85.6	65.9/57.4	74.6/72.6	12065
M	3%	100%	EN	70.7/64.2	58.5/44.8	67.8/67.7	90.9/85.5	62.4/51.3	74.0/71.6	14896
S	0%	100%	EN	52.4/42.0	48.3/26.6	43.8/43.1	80.0/58.6	37.5/31.5	52.3/50.4	15469
M	0%	100%	EN	51.0/41.5	42.6/23.8	45.4/42.8	78.6/61.6	38.0/30.6	50.0/48.4	14000
S	100%	0%	RU	84.4/80.4	76.5/66.5	77.2/77.3	96.7/94.7	83.5/76.4	88.2/87.0	11199
M	100%	0%	RU	84.3/80.4	77.9/70.4	76.4/76.5	96.5/94.4	82.3/73.3	88.4/87.4	11956
S	50%	0%	RU	82.5/78.0	74.0/63.2	76.4/76.4	96.1/93.8	80.0/71.9	86.1/84.8	5878
M	50%	0%	RU	82.3/78.0	75.0/66.5	74.6/74.7	96.0/93.7	79.5/69.8	86.4/85.2	8090
S	25%	0%	RU	79.5/72.5	67.0/45.0	75.1/75.4	95.4/92.5	76.6/68.1	83.6/81.5	3496
M	25%	0%	RU	79.6/74.3	72.3/62.1	72.7/72.8	95.3/92.6	73.8/61.5	83.7/82.1	5830
S	20%	0%	RU	78.4/70.3	64.3/36.2	74.4/74.7	95.1/92.0	75.3/67.5	82.9/81.0	2796
M	20%	0%	RU	79.0/74.2	71.4/61.4	73.0/73.2	95.0/91.9	73.5/63.8	82.3/80.8	5773
S	15%	0%	RU	77.7/70.1	66.1/44.5	74.0/74.0	94.8/91.5	72.2/61.2	81.6/79.5	1997
M	15%	0%	RU	77.2/71.3	70.7/59.6	71.7/72.0	94.6/91.4	68.6/54.9	80.6/78.7	5320
S	10%	0%	RU	75.7/67.1	64.5/41.1	73.3/73.5	93.9/90.0	67.7/54.7	78.8/76.2	1469
M	10%	0%	RU	75.2/68.2	68.7/55.3	71.5/71.7	94.0/90.2	64.0/48.4	77.8/75.5	2836
S	5%	0%	RU	58.4/47.9	48.3/20.3	71.0/71.1	92.7/87.9	29.9/18.2	50.1/41.8	739
M	5%	0%	RU	70.3/61.6	64.8/48.3	70.1/70.3	92.6/88.0	53.0/35.0	71.2/66.3	2095
S	3%	0%	RU	57.0/45.2	49.1/20.5	69.5/69.6	91.5/85.8	38.9/24.7	36.2/25.6	521
M	3%	0%	RU	65.9/55.1	62.6/41.3	69.0/69.2	91.2/85.6	42.6/24.2	63.9/55.1	1132

Table 9: Impact of small-scale training and adding parts of Russian data to the English data. Accuracy / F1-macro on the Russian data for *distilbert-base-multilingual-cased*, batch size 160, plain sampling. Mode S stands for singletask, mode M stands for multitask, RU share is the share of samples from every train Russian dataset, and EN share is the share of samples from every train English dataset. Averaged by 3-5 runs.

Table 10: Accuracy / f1 macro on the Russian data for the transformer-agnostic *distilbert-base-multilingual-cased*, batch size 160, plain sampling. Mode S stands for singletask, mode M stands for multitask, Impact of small-scale training and adding parts of Russian data to the English data. RU share is the share of samples from every train Russian dataset, and EN share is the share of samples from every train English dataset. Averaged by three runs. Comparison of validation on Russian and English data.

Mode	RU share	EN share	Validate on	Average	Emotions	Sentiment	Toxic	Intents	Topics	Batches seen
S	100%	100%	EN	85.2/82.1	79.0/70.8	77.2/77.4	96.5/94.5	84.5/80.6	88.4/87.4	15946
S	100%	100%	RU	85.3/81.9	79.2/71.4	77.2/77.3	96.7/94.7	84.6/78.2	88.6/87.7	29204
M	100%	100%	EN	84.4/80.9	77.2/70.5	75.8/75.8	96.4/94.4	83.5/76.3	88.9/87.8	20737
M	100%	100%	RU	84.4/80.7	77.6/69.8	76.8/76.9	96.6/94.6	82.4/74.5	88.8/87.8	21726
S	50%	100%	EN	83.2/79.5	75.6/65.8	75.6/75.7	96.1/93.9	82.2/76.5	86.8/85.5	16672
S	50%	100%	RU	83.5/79.6	76.7/67.6	76.1/76.2	96.2/93.9	81.7/74.9	86.7/85.4	17882
M	50%	100%	EN	82.8/78.1	76.2/64.5	74.0/73.4	95.9/93.5	80.9/72.7	87.2/86.1	19336
M	50%	100%	RU	82.7/78.6	75.5/66.3	74.5/74.7	96.0/93.6	80.7/72.8	86.8/85.8	23203
S	25%	100%	EN	81.4/76.7	73.7/61.4	73.7/73.9	95.5/92.7	78.8/71.9	85.1/83.6	16589
S	25%	100%	RU	81.8/77.3	74.5/63.4	74.6/74.9	95.4/92.6	79.1/71.7	85.1/83.7	15304
M	25%	100%	EN	80.9/76.4	73.1/63.9	73.7/73.7	95.1/92.2	77.5/68.1	85.3/83.9	16665
M	25%	100%	RU	81.0/76.6	73.3/63.8	73.5/73.8	95.0/92.2	78.1/69.5	85.1/83.9	19329
S	20%	100%	EN	80.6/76.0	71.8/60.3	74.0/74.0	95.1/92.1	78.0/71.1	83.9/82.4	12951
S	20%	100%	RU	81.0/76.3	73.2/62.3	74.5/74.6	95.2/92.2	77.6/69.6	84.4/83.0	15798
M	20%	100%	EN	80.1/75.0	71.9/61.2	73.5/73.5	94.9/91.9	76.1/65.5	84.2/82.8	17429
M	20%	100%	RU	80.3/75.3	72.3/61.5	73.9/74.1	94.9/92.0	76.1/66.1	84.5/83.1	14847
S	15%	100%	EN	79.7/74.7	70.8/57.8	72.6/72.7	94.6/91.3	77.3/70.1	83.1/81.6	13037
S	15%	100%	RU	80.0/75.4	71.5/60.5	73.7/73.9	94.8/91.6	76.6/69.2	83.3/81.7	18014
M	15%	100%	EN	78.9/73.5	70.0/58.4	71.9/71.5	94.5/91.2	74.7/65.0	83.5/81.8	15599
M	15%	100%	RU	79.0/73.1	69.8/54.1	71.7/71.5	94.5/91.3	75.5/66.6	83.5/82.1	17471
S	10%	100%	EN	77.9/72.0	68.3/52.1	72.3/72.7	93.9/90.0	73.9/65.8	81.2/79.4	13545
S	10%	100%	RU	78.2/72.0	69.4/50.5	72.1/72.4	94.3/90.6	74.4/66.8	81.0/79.5	17812
M	10%	100%	EN	77.4/70.9	67.9/51.2	71.7/71.7	93.7/90.1	72.3/61.5	81.6/79.9	14471
M	10%	100%	RU	77.3/71.2	67.8/54.3	72.0/72.1	93.4/89.7	71.4/59.7	81.7/79.9	13267
S	5%	100%	EN	75.0/67.9	64.1/45.0	70.2/70.4	92.7/87.8	69.9/60.5	77.9/75.8	12567
S	5%	100%	RU	75.2/68.7	64.8/47.9	70.5/70.7	93.0/88.4	69.5/59.9	78.1/76.4	16024
M	5%	100%	EN	74.2/66.3	63.4/41.2	70.1/70.2	92.3/87.6	67.6/56.6	77.7/75.9	12779
M	5%	100%	RU	73.6/66.3	61.3/44.7	70.1/70.1	92.4/87.6	66.1/52.8	78.0/76.1	11618
S	3%	100%	EN	71.8/64.6	59.1/38.8	68.4/68.6	91.0/85.6	65.9/57.4	74.6/72.6	12065
S	3%	100%	RU	72.1/64.8	59.9/40.1	68.7/69.0	91.8/86.5	65.5/55.9	74.6/72.5	12298
M	3%	100%	EN	70.7/64.2	58.5/44.8	67.8/67.7	90.9/85.5	62.4/51.3	74.0/71.6	14896
M	3%	100%	RU	70.7/63.0	58.7/39.5	67.8/67.2	90.6/85.2	62.1/50.8	74.2/72.1	14323

Attention-based estimation of topic model quality

Veronika Kataeva
ITMO University
St Petersburg, Russia
kataevaveronika@niuitmo.ru

Maria Khodorchenko
ITMO University
St Petersburg, Russia
mkhodorchenko@niuitmo.ru

Abstract

Topic modeling is an essential instrument for exploring and uncovering latent patterns in unstructured textual data, that allows researchers and analysts to extract valuable understanding of a particular domain. Nonetheless, topic modeling lacks consensus on the matter of its evaluation. The estimation of obtained insightful topics is complicated by several obstacles, the majority of which are summarized by the absence of a unified system of metrics, the one-sidedness of evaluation, and the lack of generalization. Despite various approaches proposed in the literature, there is still no consensus on the aspects of effective examination of topic quality. In this research paper, we address this problem and propose a novel framework for evaluating topic modeling results based on the notion of attention mechanism and Layer-wise Relevance Propagation as tools for discovering the dependencies between text tokens. One of our proposed metrics achieved a 0.71 Pearson correlation and 0.74 ϕ_K correlation with human assessment. Additionally, our score variant outperforms other metrics on the challenging Amazon Fine Food Reviews dataset, suggesting its ability to capture contextual information in shorter texts.

Keywords: Topic modeling, evaluation metrics, language models, attention mechanism, Layer-wise Relevance Propagation

DOI: 10.28995/2075-7182-2023-22-215-224

Оценка качества тематических моделей на основе механизма внимания

Аннотация

Тематическое моделирование является важным инструментом для исследования и выявления скрытых закономерностей в неструктурированных текстовых данных, что позволяет исследователям и аналитикам извлекать ценную информацию о какой-либо конкретной области. Тем не менее, тематическое моделирование не имеет единого мнения по вопросу его оценки. Оценивание полученных тем осложняется несколькими препятствиями, большинство из которых сводится к отсутствию единой системы метрик, односторонности оценки и недостаточной обобщаемости. Несмотря на различные подходы, предложенные в литературе, до сих пор нет единого мнения об аспектах эффективной и качественной экспертизы полученных тем. В данной исследовательской работе мы рассматриваем эту проблему и предлагаем новую систему оценки результатов тематического моделирования, основанную на понятии механизма внимания и послойного распространения релевантности как инструментов для обнаружения зависимостей между текстовыми токенами. Одна из предложенных нами метрик достигла корреляции Пирсона 0,71 и корреляции ϕ_K 0,74 с сравнении с оценками человека. Кроме того, наш вариант метрики превосходит другие методы оценивания на сложном наборе данных Amazon Fine Food Reviews, что свидетельствует о его способности фиксировать контекстную информацию в более коротких текстах.

Ключевые слова: Тематическое моделирование, метрики оценки, языковые модели, механизм внимания, послойное распространение релевантности

1 Introduction

The tremendous growth of digital information in recent years has evoked an increasing need to effectively process and analyze an enormous amount of text in short time. As far as most of the information is not labeled and markup with assessors takes resources and time, there is a clear tendency to utilize such data with unsupervised methods.

Topic modeling has emerged as an essential instrument for identifying semantically related sets of words that holistically encapsulate the underlying information in the document collection. The topic model receives a corpus of text and outputs the topic distribution for each document and the word distribution for each topic. While such approaches as Latent Dirichlet Allocation (LDA) highly rely on the prior which significantly constrains the possible solutions, others, like Additive Regularization (ARTM) are much more flexible and thus demand to be tuned for each of the input text corpora (Bulatov et al., 2020; Khodorchenko et al., 2022b).

Nonetheless, the estimation of resulting topic models is complicated due to several obstructions, primarily caused by a lack of a unified system of metrics. Across various papers, researchers conduct their experiments differently and employ a variety of metrics, hence intricating the comparison of performances (Abdelrazek et al., 2023). Furthermore, Doogan and Buntine substantiate the need for new evaluation measures, as the new models may be incompatible with older metrics. Another negative aspect of evaluation is the absence of generalization in experimental settings. This problem is exacerbated by the non-availability of benchmark datasets, compared to, for example, classification tasks (Doogan and Buntine, 2021). Furthermore, the metrics may reflect only a particular side of the produced model quality (Hoyle et al., 2021). Additionaly, best evaluation metrics can differ from dataset to dataset (Khodorchenko et al., 2022a). The suboptimal decision of the best topic model may cause an inaccurate representation of data and, therefore, its biased understanding. Thorough and comprehensive manual control of topic modeling outputs is still required, and it is critical for obtaining unbiased and high-quality results (Rüdiger et al., 2022). Various metrics emerged in attempts of evaluating topic quality, such as Normalized Pair-wise Mutual Information (NPMI), Perplexity, Topic Switch Percent (SwitchP), Coherence, Topic Significance, etc. However, they are not capable of closing the gap in evaluation.

As well as most of the existing topic model quality estimation scores do not fully encounter the context and rely mostly on statistics of the corpora at hand, in this paper we're addressing the power of language models. Transformer (Vaswani et al., 2017) models specifically have proven their effectiveness for numerous natural language understanding tasks, making them state-of-the-art architecture. One of the essential features of the models is attention mechanisms, which facilitate selective focus on certain parts of the input when making predictions, as well as allowing to identify the relationship of input with itself.

In this study, we propose to calculate the frequency and strength of the relationships between pairs of words in the topics with regard to attention scores. Each topic consists of words that are present in the text corpus, and a fine-tuned language model, having a deep understanding of the structure and semantics of data it has been trained on, can detect latent associations and their strength between them.

Our main contributions can be summarized as follows: (1) a novel approach for topic model quality estimation based on attention extraction with Layer-wise Relevance Propagation (LRP) mechanism, (2) an analysis of various ways to utilize attention information for the presented task, (3) a comparative study of correlations between different metrics with human evaluation to justify the proposed approach.

2 Related Work

2.1 Topic Modeling

Topic modeling has a long development story and includes wide range of models with the goal to extract latent component of the corpora which defines the topic starting from matrix factorization approaches (NMF (Févotte and Idier, 2011), SeaNMF (Shi et al., 2018)) to neural-based models (Card et al., 2018; Bai et al., 2018). The task, in general, can be formulated as follows:

$$p(w|d) = \sum_{t \in T} p(w|t)p(t|d) = \sum_{t \in T} \phi_{wt}\theta_{td}, d \in D, w \in W \tag{1}$$

where ϕ is a matrix of token probabilities in the topic, θ is a matrix of topic probabilities in the documents, D is a collection of documents, W is a finite set of vocabulary tokens, and T is a set of topics.

Among probabilistic models LDA approach (Blei et al., 2003) is being used as a solid baseline for modeling purposes despite a range of criticism which include weekly explainable Dirichlet prior and difficulties in inference adaptation to domain-specific corpora, though they can be enhanced in terms of parameter learning (Deeva et al., 2023).

In recent years, active development of neural topic models results in a wide range of new models (Card et al., 2018; Tian et al., 2020). At the same time, such models are vulnerable to overfitting and thus demand a carefully designed quality metric and loss function.

A semi-probabilistic additive regularization approach (Vorontsov and Potapenko, 2014) is one of the most flexible in terms of domain-specific models creation, as it allows combining regularizers to produce models with specific characteristics. Still, while providing such wide tools for model designing, it significantly increases the number of hyperparameters to be tuned.

2.2 Evaluation Metrics

Throughout the development of topic modeling, a range of automated metrics have been developed to quantify the performance of topic models.

The earliest and later heavily criticized (Hoyle et al., 2021; Doogan and Buntine, 2021) for low correlation with human assessments and unreliability are perplexity (Blei et al., 2003) which measures predictive likelihood of document given topic matrix and hyperparameters and topic coherence (Newman et al., 2010) that is based on pairwise words concurrences in the corpora. One of the prominent variations of coherence is NPMI has shown a substantial correlation with human judgment on word relatedness in previous studies. It compares joint probability of words to the probability of them occurring independently.

One of the recent approaches to assessing topic quality is the Topic significance (Lund et al., 2017). The metric considers the entire topic-word distribution, unlike the coherence measure. SwitchP (Lund et al., 2019) estimates local topic quality, that regards to the quality of a topic within a specific document. SwitchP demonstrates a higher positive correlation with human judgments in comparison to coherence (Lund et al., 2019; Rezaee and Ferraro, 2020).

While coherence is viewed as a measure of topic interpretability, there are introduced several attempts to evaluate other topic qualities, such as topic stability (Xing and Paul, 2018) and topic diversity (Dieng et al., 2020).

It is essential to note that some researchers apply combination of metrics (Dieng et al., 2020) or view topic modeling as classification or clustering (Harrando et al., 2021).

The scope of this paper is to study the usefulness of neural networks for topic quality assessment.

3 Attention-based topic model evaluation

The proposed attention-based topic evaluation consists of several steps, which include 1) language model fine-tuning to acquire the connections in input text; 2) performing layer-wise propagation to understand which heads and words connections are important; 3) identifying co-dependency value of individual pairs of tokens in topic and averaging the values; 4) calculating the final model quality by averaging scores from step 3 for all topics.

The first stage of our research involves fine-tuning the language model BERT (Devlin et al., 2018) to solve text classification tasks. One of the key features of the model is the multi-head attention mechanism. Learned in parallel, multiple attention heads produce versatile representations that provide various aspects of the input (Vaswani et al., 2017). For our experiments, we choose BERT and RuBERT (Kuratov and Arkhipov, 2019), BERT adaptation for the Russian language (further both referred as BERT). Both models consist of 12 Transformer blocks, where each layer has 12 heads.

As far as the original architecture of the BERT model appears as a black box, several approaches attempt to explain the decision-making behind the model predictions by addressing attention mechanisms. This work employs these methods to trace which interconnections between tokens BERT may detect: Layer-wise propagation (LRP) (Bach et al., 2015); Improved LRP (Chefer et al., 2020); raw output attentions.

Output layer attentions in Transformer-based architectures are calculated with multi-head attention mechanism which concatenates the results from all layer heads (eq. 2-4).

$$A(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

$$head_i = A(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

$$Multihead(Q, K, V) = Concat_i(head_i)W^o \tag{4}$$

where Q - query matrix, K - key, V - value, d_k - key dimensionality, W_i and W^o are parameter matrices.

LRP algorithm intends to examine the individual contribution of input to model output by propagating the relevance of the output back through the network layers consecutively to the input, using the same set of weights that have been used to compute the output (Voita et al., 2019).

Propagating relevance scores at a given layer are calculated as:

$$R_{i\leftarrow k}^{(l,l+1)} = R_k^{(l+1)} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}, \tag{5}$$

where k, i are neurons from layer $l+1$, and proceeding layer l, provided that i has a forward connection to k, w is weights, and a is an output from an activation function.

In Improved LRP, the local relevances are assigned based on the Deep Taylor Decomposition (Montavon et al., 2015), a method based on the Taylor series. This propagation involves an advanced approach to operating with matrix multiplication and skip connections.

The output of the method is defined through the weighted attention relevance:

$$\bar{A}^{(b)} = I + E_h(\nabla A_{(b)} \odot R^{n_b}) \tag{6}$$

$$C = \bar{A}^{(1)} \cdot \bar{A}^{(2)} \cdot ... \cdot \bar{A}^{(B)}, \tag{7}$$

where $\odot$ is Hadamard product, E_h is the mean across attention heads, $A^{(b)}$ is attention map of block b and $\nabla A_{(b)}$ is its gradients, R^{n_b} is layer's relevance.

Since our focus is solely on the relevance of the heads, we do not continue propagating down to the input variables and instead stop at desired self-attention layer, producing relevance matrices.

The next step is dedicated to determining BERT attention heads that carry non-trivial information. Since multi-head attention block learns different representations, each attention function may dissimilarly contribute to forming a prediction.

We use head confidence to handle raw output attentions, which is proportional to the average highest attention weight assigned across all instances of the evaluation dataset, excluding the end-of-sentence token. LRP-based techniques estimate head importance by calculating head relevance as the sum of the neurons' relevances within the head, normalized across all heads within a layer. The final relevance of a head is the mean relevance in the evaluation dataset (Voita et al., 2019).

The task of topic model quality estimation in case of attention-based approach can be generalized as an average quality of the resulting topics

$$Q^b = \frac{1}{T} \sum_t^T \sum_s^S \sum_h^H \sum_{i,j,i\neq j}^N a_{tshij}^b, \tag{8}$$

where T is the overall number of topics, S is overall number of texts, H is overall number of confident/important heads, N - amount of tokens in text and size of attention matrix, i is the i-th token of n-th text that attends to j-th token, $a^b \in \{a^{Attn\_sum}, a^{Imp\_LRP\_sum}, a^{LRP\_sum}, a^{Count}\}$ is an element of one of attention/relevance matrix.

Depending on the type of a^b we defined 4 alternative quality functions:

1. Q^{Attn_sum} denoting *Attention sum*, which derives information from output attentions matrix (eq. 4) from head with maximum relevance according calculated as $\frac{1}{n}max_i^L(head_i^l)$, where n is amount of layers,
2. $Q^{Imp\_LRP_sum}$ denoting *Improved LRP sum*, which identifies important heads by relevance matrices obtained through propagation by employing Improved LRP approach (eq. 6-7),
3. Q^{LRP_sum} denoting *LRP sum*, which uses LRP matrices as a source of information on tokens interconnections according to eq. 5,
4. Q^{Count} denoting *Count*, which uses binary matrix ($a_{ij} = 1$ if $a_{ij} > 0$ obtained from relevance matrix) that shows any present co-dependent token pairs.

4 Experimental study

4.1 Datasets

In this work, we use three datasets in Russian and English languages:

1. 20 Newsgroups dataset (Lang, 1995), a collection of news posts that covers 20 various topics including sports, religion, science, and politics.
2. Lenta.ru dataset, which comprises Russian news from an electronic resource spanning 20 years.
3. Amazon Fine Food Reviews (Amazon Reviews) dataset (McAuley and Leskovec, 2013), which consists of short reviews on food categories gathered over 10 years.
4. The dataset with evaluation of topics (Khodorchenko et al., 2022a) contains automatic and human scores for a variety of sampled topics outputted by 100 variously configured ARTM models with different amounts of topics built on the datasets 1-3 from this list. To measure the quality of topics, they were presented as tasks in Toloka (Tol, 2023) crowdsourcing platform interface. The assessors were asked whether a common topic for the presented word set is distinguishable. If the answer is positive, they are asked to name the topic and identify irrelevant to the topic words. Each of the topics was evaluated by several assessors to fit into weighted categories: a score of 2 for *yes*, 1 for *rather yes*, -1 for *rather no*, and -2 for *no*, with a score of 0 awarded in cases of inhomogeneous evaluations of human assessors. To compute correlation coefficients between human and automated model quality, we average the quality scores for each topic, enabling a comparison with human decisions.

Each dataset (1-3 from datasets list) is reduced to contain approximately 10 000 samples in order to diminish computational costs. Finally, the texts are pre-processed by removing any HTML tags, punctuation, links, tags, digits, and stop words, as well as by being lemmatized and filtered out to contain more than five tokens.

Both 20 Newsgroups and Lenta.ru contain pre-defined topic-related labels, whereas, for Amazon Reviews, we establish the labels by K-means clustering on reduced via Truncated SVD TF-IDF vectors with the number of clusters equal to 20.

We fine-tuned BERT instances on the classification task for each of the datasets. Details on hyperparameters of BERT are presented in Table 1. Models achieved F1-scores of 0.8545, 0.8056, and 0.8933 correspondingly, denoting sufficient understanding of texts BERT models have learned.

Input data	Model	Max len	Batch size	Learning rate	# of epochs
20 Newsgroups	$BERT_{BASE}$	128	8	5e-5	5
Lenta.ru	$RuBERT_{BASE}$	256	8	1e-5	5
Amazon Reviews	$BERT_{BASE}$	256	8	1e-5	5

Table 1: Hyperparameters of fine-tuned models.

4.2 Attention-based metrics performance

To understand the effectiveness of the developed metrics, we conduct a correlation analysis, using Pearson's r to detect linear relationships and ϕ_K to trace non-linear ones.

Firstly, correlation is measured between the human assessments and proposed metrics. Specifically, we measure dependencies for the quality values of distinct topics. The results of our experiments indicate that the Count metric of interconnections between tokens exhibits a greater degree of correlation than other attention-based methods. However, it is characterized by larger fluctuations of coefficient values across different datasets. In contrast, two LRP-based approaches demonstrate significantly greater stability.

Secondly, we measure the correlation between human assessments and model scores (see Table 2) calculated as the mean quality of all topics within each model.

In assessing the performance of attention-based metrics at the model level, our examination has revealed that the Improved LRP approach demonstrates higher correlation values than other techniques. Nonetheless, the LRP approach remains a viable and competitive option, displaying significant performance on the 20 Newsgroups dataset.

Dataset	Corr.	Attn sum	Imp. LRP sum	LRP sum	Count
20 Newsgroups	r	0.74	0.51	**0.78**	0.73
	ϕ_K	0.68	0.64	**0.86**	0.61
Lenta.ru	r	0.65	**0.79**	0.65	0.20
	ϕ_K	**0.80**	0.63	0.69	0.75
Amazon Reviews	r	0.65	**0.75**	0.69	0.61
	ϕ_K	0.61	**0.73**	0.69	0.66

Table 2: The correlation between human assessments of model quality and scores from developed automated metrics, as measured by Pearson's r and ϕ_K coefficients.

Figure 1 shows the propagated relevance matrix derived via applying Improved LRP to Amazon Reviews text. One of the topics obtained during topic modeling is *dog treat chew toy bone teeth ball puppy training liver piece bread pet play vet*, which was unanimously voted by assessors as a good topic. As we can see, the explainability method can identify significant relationships between words in the text *dog*, *treat*, and *pet*, which are part of the aforementioned topic as well and therefore contribute to a higher automated quality score during the proposed attention-based metrics calculations.

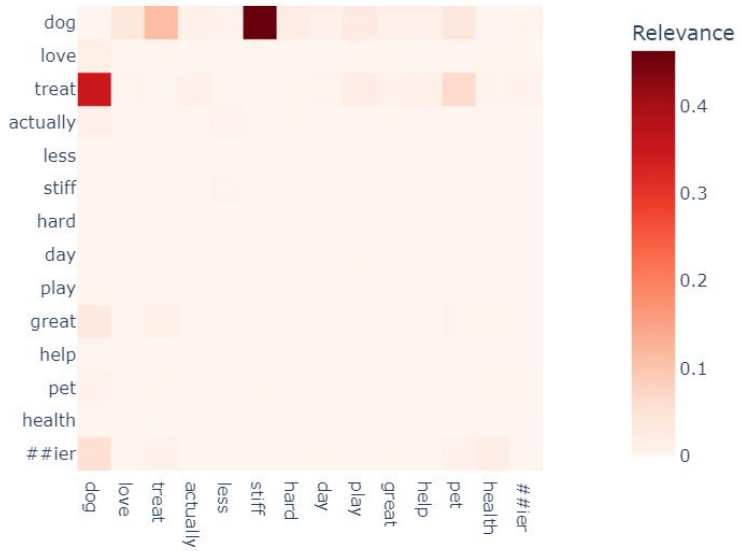


Figure 1: Relevance matrix of Amazon Reviews text derived using Imporved LRP. By employing this matrix with topic word set *dog treat chew toy bone teeth ball puppy training liver piece bread pet play vet*, we notice the substantial dependencies between topic and text words *dog*, *treat*, and *pet*.

Examples of obtained top-10 most probable words per topics from Lenta.ru dataset and corresponding scoring are presented in Table 3.

Quality	Score	Topic	Attn sum	Imp. LRP sum $\times 1e6$	LRP sum $\times 1e4$	Count
High	2	уголовный статья убийство обвинение следствие срок расследование преступление прокуратура комитет	349.94	650	790	11944
	1	олимпийский япония японский спортсмен спорт олимпиада сочи алкоголь спортивный клуб	31.82	62	2.3	2500
	2	продажа строительство квартира стоимость метр продавать квадратны жилье площадь сделка	207.06	240	490	7628
Low	-2	святой медиа сенат австрия действительность уверять алиев добывать окончательно разрушение	2	0.024	0.12	36
	-2	относиться певица опрос потеря свидетельствовать опрашивать сегодняшний рождаться рождение треть	5.94	37	2.2	400
	-1	деньги знать пытаться жить узнавать удаваться говорить решать вернуть помогать	43.79	17	-4	3192

Table 3: Examples of obtained high- and low-quality top-10 most probable words per topics obtained from Lenta.ru dataset with corresponding human labeling and received with proposed metrics scores.

Figure 2 illustrates how individual topics are scored by different metrics. Improved LRP is showing the best ability in distinguishing between high and low quality topics. It should be also noted that different automatic metrics make mistakes on different examples, so it is potentially possible to make an ensemble approach to better approximate human labelling.

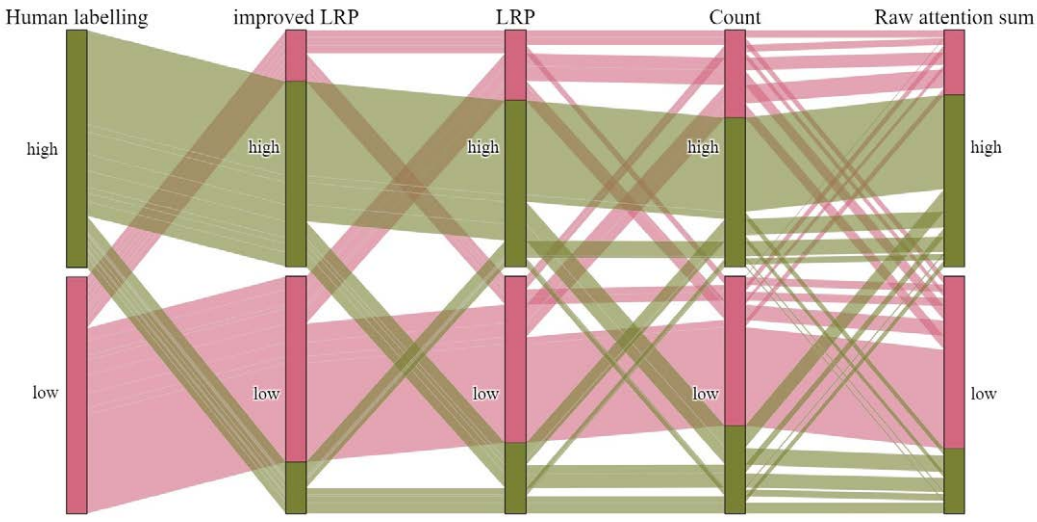


Figure 2: Quality comparison of individual topics. This illustration shows a scoring for each individual sentence. For “Human labelling” scores -1 and -2 were combined to “low” category and +1 and +2 - to “high” category. We also balanced amount of “high” and “low” scores by human labelling with random sampling. For other metrics intervals were divided by median value. “Improved LRP” shows better abilities in high and low quality topics.

4.3 Model-level correlation with human assessment comparison for automatic metrics

To estimate the general performance of the proposed metric, we compare the best attention-based variants (LRP and Improved LRP), and other commonly used metrics in both academia and applied settings, based on their ability to approximate human judgment. The results are presented in Table 4.

Dataset	20 Newsgroups		Lenta.ru		Amazon Reviews		Average	
Correlation	r	ϕ_K	r	ϕ_K	r	ϕ_K	r	ϕ_K
Improved LRP sum (our)	0.51	0.64	**0.79**	0.63	**0.75**	**0.73**	0.68	0.67
LRP sum (our)	0.78	0.86	0.65	0.69	0.69	0.69	**0.71**	**0.74**
NPMI	**0.86**	0.72	0.75	0.78	0.43	0.52	0.68	0.67
Perplexity	0.28	0.71	-0.43	0.75	0.49	0.58	0.4	0.68
Background Tokens Ratio	-0.22	0.58	0.73	0.8	-0.58	0.47	0.51	0.62
Avg SwitchP	-0.75	0.68	-0.2	0.7	-0.73	0.67	0.56	0.68
Coherence	0.71	0.71	0.75	0.8	0.53	0.69	0.66	0.73
Contrast	0.67	**0.94**	0.59	**0.87**	0.51	0.39	0.59	0.73
Purity	0.73	0.71	0.19	0.77	0.35	0.0	0.42	0.49
Kernel Size	0.65	0.64	0.24	0.59	0.44	0.62	0.44	0.62
Topic Significance Avg	0.29	0.72	0.19	0.67	0.25	0.54	0.25	0.64

Table 4: The Pearson's r and ϕ_K correlation coefficients between human assessments of model quality and scores produced by automated metrics, including our novel attention-based approaches and widely-used automated metrics. Best scores are indicated by bold text, while worst results are underlined. Average is calculated as an average correlation strength without sign.

One of the most stable results according to an average of the scores is demonstrated by LRP sum metric for both of the correlations. In this case, attention-based metric is showing good performance regardless of the dataset. At the same time, Improved LRP sum shows superior performance on Lenta.ru (linear correlation) and Amazon reviews (both correlations) while being worse on average. Proposed attention-based scores in general indicate good linear and non-linear correlations with human assessment.

Considering the results, our findings demonstrate the lack of consensus in the observed results, highlighting the existence of varying degrees of linear and non-linear correlation with human judgment across the different metrics evaluated. However, proposed LRP Sum metric can be used as a good metric for topics on average and Improved LRP version – for shorter text cases.

5 Conclusion and Future Work

In this paper, we presented an attention-based method to evaluate topic model quality. Results indicate that the proposed utilization of LRP approach to extract and summarize the interconnections between words in the topics based on fine-tuned BERT architecture is showing a better quality compared to other existing metrics, reaching 0.71 Person and 0.74 ϕ_K correlations with human assessment for LRP sum score. At the same time, Improved LRP sum score variant is revealing superior quality on the most difficult for other metrics dataset – Amazon Reviews, indicating its ability to catch more context-based information in case of shorter texts.

In future work, we are going to conduct experiments in the setting where BERT fine-tuning is done on the task of masked language model task to omit the necessity of label creation in case of their absence. We will also investigate ways to speed up LRP computations to insert the proposed scores into a topic models tuning framework.

Acknowledgements

This research is financially supported by the Foundation for National Technology Initiative's Projects Support as a part of the roadmap implementation for the development of the high-tech field of Artificial Intelligence for the period up to 2030.

References

Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems*, 112:102–131.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLOS ONE*, 10(7):1–46, 07.

Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. 2018. Neural relational topic models for scientific article analysis. // *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, CIKM '18, P 27–36, New York, NY, USA. Association for Computing Machinery.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022, mar.

Victor Bulatov, Vasiliy Alekseev, Konstantin Vorontsov, Darya Polyudova, Eugenia Veselova, Alexey Goncharov, and Evgeny Egorov. 2020. TopicNet: Making additive regularisation for topic modelling accessible. // *Proceedings of the Twelfth Language Resources and Evaluation Conference*, P 6745–6752, Marseille, France, May. European Language Resources Association.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. // *ACL*.

Hila Chefer, Shir Gur, and Lior Wolf. 2020. Transformer interpretability beyond attention visualization. *Computing Research Repository*, arXiv:2012.09838.

Irina Deeva, Anna Bubnova, and Anna V. Kalyuzhnaya. 2023. Advanced approach for distributions parameters learning in bayesian networks with gaussian mixture models and discriminative models. *Mathematics*, 11(2).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *Computing Research Repository*, arXiv:1810.04805.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. // *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 3824–3848. Association for Computational Linguistics, June.

Cédric Févotte and Jérôme Idier. 2011. Algorithms for nonnegative matrix factorization with the -divergence. *Neural Computation*, 23(9):2421–2456.

Ismail Harrando, Pasquale Lisena, and Raphael Troncy. 2021. Apples to apples: A systematic evaluation of topic models. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, P 483–493. INCOMA Ltd., September.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Lee Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. // *Proceedings of 35th Conference on Neural Information Processing Systems (NeurIPS 2021)*.

Maria Khodorchenko, Nikolay Butakov, and Denis Nasonov. 2022a. Towards better evaluation of topic model quality. // *2022 32nd Conference of Open Innovations Association (FRUCT)*, P 128–134.

Maria Khodorchenko, Nikolay Butakov, Timur Sokhin, and Sergey Teryoshkin. 2022b. Surrogate-based optimization of learning strategies for additively regularized topic models. *Logic Journal of the IGPL*, 02. jzac019.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *ArXiv*, abs/1905.07213.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. // Armand Prieditis and Stuart Russell, *Proceedings of the 12th International Conference on Machine Learning*, P 331–339, San Francisco (CA). Morgan Kaufmann.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: a multiword anchor approach for interactive topic modeling. // *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 896–905, Vancouver, Canada, July. Association for Computational Linguistics.

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 788–796, Florence, Italy, July. Association for Computational Linguistics.

Julian John McAuley and Jure Leskovec. 2013. From amateurs to connoisseurs: Modeling the evolution of user expertise through online reviews. // *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, P 897–908, New York, NY, USA. Association for Computing Machinery.

Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2015. Explaining nonlinear classification decisions with deep taylor decomposition. *Computing Research Repository*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. // *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, P 100–108, USA. Association for Computational Linguistics.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. // *Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 2020)*. Curran Associates Inc.

Matthias Rüdiger, David Antons, Amol M. Joshi, and Torsten-Oliver Salge. 2022. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLOS ONE*, 17:1–25, 04.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K. Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. // *Proceedings of the 2018 World Wide Web Conference*, WWW '18, P 1105–1114, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Runzhi Tian, Yongyi Mao, and Richong Zhang. 2020. Learning VAE-LDA models with rounded reparameterization trick. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 1315–1325, Online, November. Association for Computational Linguistics.

2023. Toloka ai: Powering data-centric ai. `https://toloka.ai/`.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, arXiv:1706.03762.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 5797–5808, Florence, Italy, July. Association for Computational Linguistics.

Konstantin Vorontsov and Anna Potapenko. 2014. Additive regularization of topic models. *Machine Learning*, 101:1–21, 12.

Linzi Xing and Michael Paul. 2018. Diagnosing and improving topic models by analyzing posterior variability. // *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, volume 32, P 6005–6012, apr.

Foregrounding and accessibility effects in the gaze behavior of the readers with different cognitive style

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

Anna Izmalkova
Higher School of Economics,
Moscow State Linguistic University,
Moscow, Russia
mayoran@mail.ru

Sergey Makeev
Lomonosov Moscow State University
Moscow, Russia
sergeymak98@gmail.com

Abstract

This paper explores accessibility effects in the gaze behavior of readers with different cognitive style, impulsive and reflective, as mediated by graphological and linguistic foregrounding in the discursive acts in 126 areas of interest (AOIs). The study exploits 1890 gaze behavior probes available at open access Multimodal corpus of oculographic reactions MultiCORText. We identified that while graphological foregrounding makes initial or final components of discursive act more accessible for the impulsive readers, reflective readers also observe the components within the act. Linguistic foregrounding produces higher access with impulsive readers in case the linguistic form is visually focalized (phonological foregrounding and parallel structures); meanwhile, with reflective readers this is the information density appearing in elliptical and one-component sentences which maintains higher access.

Keywords: foregrounding, graphological foregrounding, linguistic foregrounding, accessibility, gaze behavior, cognitive style, impulsivity / reflectivity

DOI: 10.28995/2075-7182-2023-22-225-232

Выдвижение и доступность информации в глазодвигательном поведении читателей с разным когнитивным стилем

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

Анна Измалкова
Московский государственный
лингвистический университет,
Москва, Россия
mayoran@mail.ru

Сергей Макеев
Московский государственный
университет им. М.В. Ломоносова
sergeymak98@gmail.com

Аннотация

В работе исследуются особенности доступа к информации через анализ глазодвигательного поведения читателей с разным когнитивным стилем, импульсивных и рефлективных, под влиянием семантического

выдвижения, графологического и лингвистического, в дискурсивных актах (в 126 зонах интереса). Материалом анализа являются 1890 проб глазодвигательного поведения, размещенных в Мультимодальном корпусе глазодвигательных реакций MultiCORText. Установлено, что графологическое выдвижение инициальных или финальных компонентов дискурсивного акта облегчает доступ к информации для импульсивных читателей; рефлективные читатели обращают внимание и на срединные компоненты дискурсивного акта. Лингвистическое выдвижение, которое делает информацию более доступной для импульсивного читателя, проявляется в ее представлении с помощью определенных формально-языковых средств – фонологических средств и параллельных конструкций. В то же время рефлективных читателей привлекает информация более высокой плотности, которая проявляется в эллиптических и однокомпонентных предложениях как средствах синтаксического выдвижения.

Ключевые слова: выдвижение, графологическое выдвижение, лингвистическое выдвижение, доступность, глазодвигательное поведение, когнитивный стиль, импульсивность / рефлективность

1 Introduction

The present study addresses the research problem of information accessibility in reading attested via gaze behavior of readers. One of the best explored factors mediating accessibility is the information foregrounding (salience, prominence, focalization) which is commonly studied via foregrounding cues or primes. As known, the effects of various types of foregrounding cues have been identified, with syntactic priming, lexical priming, visual cuing, perceptual priming, event orientation cueing. However, other factors apart from foregrounding type may contribute to accessibility effects. In the study, via the readers' gaze behavior we explore the effects of the two factors, the type of foregrounding and the cognitive style of the readers as potentially significant for information accessibility among the readers. The research exploits the data available at the Multimodal corpus of oculographic reactions MultiCORText which is a pilot open-access search corpus of gaze behavior contingent on the text semantic parameters.

2 Theoretical framework

The study is built on two theoretical frameworks to foregrounding effects. While foregrounding is commonly viewed as a construal operation stimulating mental structures activation realized through selected semiotic means, it can be addressed either as an information production operation or the operation which stimulates information accessibility. In the first case, foregrounding becomes the key interest for linguistic studies. In the second case, it is explored as a counterpart of information accessibility for readers displaying different cognitive skills; therefore, it is mainly the cognitive psychological object of research. Although cognitive psychology has developed its methods of studying foregrounding and accessibility effects, the considerable experience of linguistics in analyzing foregrounding can help specify these effects. Still, linguistics will also benefit from these studies as it receives an instrument of ranging linguistic foregrounding effects in terms of their perception.

2.1 Foregrounding in linguistics

Exploring foregrounding in cognitive linguistics is aimed at scaling its effects in information construal. Foregrounding in text can be attributed to either activating non-verbal information in event construal [26; 27], or activating linguistic information on the text structure [7; 9; 18; 25]. In the present study we address the second stance and consider it as a linguistic operation of directing attention towards definite language structures and their semantics. For instance, in [18] foregrounding in syntax is viewed via Newness expressed in End-Focus in English, whereas in Russian [15] it can appear both in sentence initial and final Rheme as well as in Theme-New in *Истории о своем доме и жизни / старик / не ощутил* (trans. *the stories of his house and life / the old man / did not feel*), in Complex Rheme in *Послышались шаги и веселый говор* (word by word trans. – *were heard the steps and gaily talk;* trans. – *We heard the steps and gaily talk*). Olga Iriskhanova [24] lists the linguistic means of foregrounding in all language levels. For instance, foregrounding in syntax is expressed via sentence-final position for Neutral Focus and sentence-initial position for Contrastive Focus, parallel structures, one-component and elliptical sentences. In morphology it appears in the use of proper names, superlative adjectives, verbs in the perfective aspect, verbs in the historical present. In the lexical level, it is expressed via tropes, expressive means, codes switching. Additionally, graphological foregrounding in letter capitalization, the use of punctuation marks, etc. also contributes to focalizing information [24]. This typology of graphological and linguistic foregrounding cues will serve to analyse foregrounding effects in the present study.

2.2 Foregrounding and accessibility in cognitive psychology

The second framework was developed in cognitive and experimental psychology and is aimed at scaling the accessibility effects produced by information foregrounding. Whereas several linguistic approaches do not differentiate between accessibility and foregrounding [1], in experimental studies accessibility is considered as the operation of obtaining the access to information in the text. The studies mostly name two psychological processes which contribute to information accessibility, which are attention control [6] and working memory activation [4]. Accessibility is vastly explored via foregrounding cues [8; 12, 16], where the latter produce different activation effects.

One of the most efficient methods of exploring accessibility modulated by linguistic foregrounding cues is via gaze behavior [3; 13]. In the study, accessibility is accessed via higher gaze costs or higher values of gaze metrics applied in attesting gaze behavior [14; 17; 21]. For instance, in [17] it was shown that extremely long fixations (>1000ms) provide evidence of information processing difficulties. Therefore, identifying longer gaze duration or higher fixation number can signal about higher information accessibility produced by the foregrounding cues in the text areas of interest (AOIs) under scrutiny.

However, we can hardly expect that accessibility is equally and solely dependent on foregrounding cues. First, they can produce different accessibility effects as shown in [3; 5; 13; 20]. For instance, M. Reingold and K. Rayner show that longer fixations appear on the words given in bold [22]. Second, other factors can affect information accessibility. For instance, in [24; 28] the readers' cognitive style explored via impulsivity and reflectivity was reported to have affected the reaction time and the number of errors made. In [10] it was found that impulsive readers' gaze behavior as opposed to the gaze behavior of reflective readers was more affected by graphological foregrounding cues, which was attributed to the differences in their attention types, bottom-up and top-down attention, consistent with the notion of the impulsive-reflective cognitive style [23]. Overall, the impulsive and reflective styles are treated as a "property of the cognitive system that combines individuals' decision-making time and their performance in problem-solving situations, which involve a high degree of uncertainty" [24: 451].

Therefore, the study seeks to explore the accessibility effects of foregrounding as modulated by two factors, the type of the foregrounding cue (graphological and linguistic, with further specification) and the cognitive style (reflective and impulsive) of the reader.

3 Methods and procedure

3.1 Experiment design

Stimuli. The stimuli were 5 one-page drama texts (authored by L. Petrushevskaya, L. Razumovskaya, A. Arbuzov, A. Vampilov, A. Chekhov) which involved 126 AOIs corresponding to discursive acts or the acts performing "responses and interpretations from an external world" [19]. With each discursive act representing an act of instruction, order, command, recommendation, prayer, plea, etc., we identified three basic formal types of discursive acts in our stimuli: 1) a clause (Она сидит у пианино), 2) two clauses representing one discursive act (Что за ребенок, что за ребенок золотой?), 3) a clause with discourse markers (А в четверг – ну, ей-богу, ну, клянусь – сидел в кресле). Although there were more than 126 discursive acts in the 5 stimuli, we had to choose only the ones which were located in one line since the transfer from one line to another would require higher gaze costs [21].

The AOIs were annotated using the semantic protocol incorporated into the Multimodal corpus of oculographic reactions MultiCORText which is an open access database (https://multicortext.linguanet.ru/) that allows parametric search using both semantic parameters and gaze metrics. Graphological foregrounding implied the use of italics, bold type, brackets, letters/words capitalization, full stop, comma, exclamatory/question mark, hyphen, etc. Linguistic foregrounding was annotated in all language levels, in phonological level (onomatopoeia, alliteration, etc.), lexical level (proper name, superlative degree of adjective/adverb, perfective/imperfective verb form, etc.), phraseological and syntactic level (phraseological units and set phrases, elliptical sentences, parallel constructions, etc.). Below, we present several examples of annotation:

In (AOI 38) / Заждалась вас, радость моя, светик…/ we mark the graphological cues, here first letter capitalization, comma, dots (suspension points). Linguistic foregrounding cues appear in perfective aspect of a verb, nonce word, expressive means, one-component sentence, parallel constructions.

In (AOI 60) / Варя подбрасывает карты, Михалев отбивается. / we identify the graphological cues which are first letter capitalization, full stop, comma. Linguistic foregrounding is realised through proper name, parallel constructions.

In (AOI 120) / *Она говорит тихо даже в минуты волнения.)* / we mark italics, brackets, first letter capitalization, full-stop, proper name, expressive means.

We presume that the effects of conventionality appearing in the more frequent use of several foregrounding means in all levels might cause the differences in the gaze behavior; still in the experiment two participants' groups were exposed to the same stimuli, which allows to disregard it in contrastive analysis.

Experiment procedure. The experiment was a two-step procedure. At the first step, the psychological test to identify impulsivity / reflectivity score was conducted. At the second step, the eye tracking experiment was carried out. 16 (15) subjects (students, age range 20-26, mean age 22) participated in the study.

Impulsivity / reflectivity score was measured with traditional Familiar Figures Test (MFFT) [11]. In the test, the subjects are expected to find a match for a target image among eight variants. Impulsivity / reflectivity assessment is carried out considering 1) latency (time taken to respond) and 2) accuracy (number of mistakes) score; consequently, the subjects are classified as impulsive if they manifest short latency and low accuracy, and reflective if they manifest long latency and high accuracy. As known, gaze patterns of impulsive and reflective subjects vary in visual search task [2]; therefore, we hypothesized that significant distinctions in gaze behavior would be observed for the reading task as well. MFFT allowed to identify two subject groups: more reflective and more impulsive participants.

During the eye tracking experiment, the eye tracker SMI Red-x binocular system, frequency = 60 Hz, accuracy = 0.4°, head movement 40x20 cm, operating distance = 60-80 cm, was applied. 126 AOIs gaze data were further analyzed in BeGaze 3.0 software. We received 1890 probes which were later subjected to analysis. Since there were two subject groups (more reflective and more impulsive), the probes were analyzed in 2 data sets with each data set annotated for presence or absence of 28 parameters of graphological and linguistic foregrounding. In the experiment, 3 gaze metrics were considered: First Fixation duration, Max Fixation duration, and Average Fixation duration in AOIs; they were selected following the gaze behavior studies employing text stimuli with AOIs [21; 14].

JAMOVI software was applied to explore gaze behavior variance and identify the degree of accessibility. Kruskal-Wallis One-way ANOVA (non-parametric) tests preceded by Shapiro-Wilk normality tests were performed to identify whether there are significant distinctions in gaze behavior of reflective and impulsive readers. The tests were used to estimate how the means of a dependent variable (First Fixation duration, Max Fixation duration, Average Fixation duration) change according to the 2-level independent variable, the presence or absence of each of the 28 foregrounding cues in two participant groups. We then scaled the Kruskal-Wallis χ^2-values of foregrounding cues (considering only the cases with significant p-values) as mediated by impulsive and reflective participants.

4 Results

4.1 Gaze metrics

MFFT [11] conducted at the first step of the experiment allowed to assess the time taken to respond (T) and the number of mistakes (MN) made by 16 participants (the gaze results of one participant were further discarded due to calibration problems). Two subject clusters were identified, 9 impulsive subjects (T = 370.3 s, MN =10.7) and 7 reflective subjects (T = 756.7 s, MN = 4.9). The gaze results of 15 participants were subjected to analysis. In Table 1 we show the gaze metrics (First Fixation duration, Max Fixation duration, Average Fixation duration) extracted from MultiCORText, which were further split into 2 data sets, for impulsive and reflective readers.

	First Fixation duration	Max Fixation duration	Average Fixation duration
N	700 / 637	700 / 637	700 / 637
Mean	169 / 164	210 / 200	165 / 161
Standard deviation	65.6 / 63.2	87.6 / 85	49.5 / 48.2
Variance	4309 / 3991	7675 / 7228	2447 / 2327

Table 1: Gaze metrics of impulsive / reflective readers

Descriptive statistics shows that although the differences in Mean values are not high, in all cases the values are lower for reflective readers. To identify the effects of foregrounding onto accessibility, individual gaze probes (700 for impulsive readers and 637 for reflective readers) were subjected to analysis. Since the gaze data do not have normal distribution (Shapiro-Wilk test with $p < .001$ proves it), a series of Kruskal-Wallis One-way ANOVA non-parametric tests (Kruskal-Wallis H tests) was conducted. Since 28 foregrounding cues were explored, 56 tests were carried out (in JAMOVI software).

4.2 Graphological foregrounding

15 Kruskal-Wallis H tests in each of the two groups were carried out with graphological foregrounding cues, 1a) no graphic foregrounding, 1b) italics, 1c) bold type, 1d) spacing, 1e) brackets, inverted commas, 1f) first letter capitalization, 1g) words capitalization 1h) non-standard graphology, 1i) tabulation, 1j) no orthographic foregrounding, 1k) full stop, 1l) comma (commas), 1m) exclamatory/question mark, 1n) dots, colon, 1o) hyphen. Both impulsive and reflective readers were affected by graphological foregrounding; still, we observed several differences.

Importantly, in the group of impulsive readers, only Max Fixation duration was modulated by foregrounding. We identified 5 foregrounding cues which produced significant effects onto the gaze behavior, which are first letter capitalization (Kruskal-Wallis χ^2 (1, 699) =5, p=0.025), tabulation (χ^2 (1, 699) = 6.13, p=0.013), full stop (χ^2 (1, 699) = 6.14, p=0.013), comma (χ^2 (1, 699) = 9, p=0.003), dots, colon (χ^2 (1, 699) = 6.85, p=0.009). The results show that impulsive readers were mostly affected by initial or final discursive act foregrounding cues, like in / Лицо ее выражает глубокое горе. / with both first letter capitalization and full stop, / Пауза. / with tabulation, first letter capitalization and full stop. This means that initial or final discursive act foregrounding cues make information more accessible for the impulsive readers.

In the group of reflective readers, both First Fixation duration and Max Fixation duration were affected. First Fixation duration was modulated by italics (χ^2 (1, 636) = 3.66, p=0.056) and brackets, inverted commas (χ^2 (1, 636) = 4.85, p=0.028). Max Fixation duration was modulated by bold type (χ^2 (1, 636) = 3.69, p=0.055), first letter capitalization (χ^2 (1, 636) = 7.44, p=0.006), words capitalization (χ^2 (1, 636) = 6.78, p=0.009), tabulation (χ^2 (1, 636) = 6.41, p=0.011), comma (χ^2 (1, 636) = 4.69, p=0.03). The results manifest that reflective readers observe the foregrounded information which appears both in the initial and final position of the discursive act and in within the discursive act like in / МИХАЛЕВ. / with words capitalization, / Я обомлела, когда вошла. / with comma. We can deduce that reflective readers develop a better access to any component of a discursive act than impulsive readers who mostly observe its beginning and its end.

4.3 Linguistic foregrounding

13 Kruskal-Wallis H tests in each of the two groups were carried out with linguistic foregrounding cues, 2a) phonological foregrounding (onomatopoeia, alliteration, etc.), 2b) proper name, 2c) superlative degree of adjective/adverb, 2d) perfective verb form, 2e) present tense verb manifesting past action, 2f) nonce-word, 1g) repetition of a word or word combination, 1h) code shifting, 1i) expressives and tropes, 1j) phraseological units and set phrases, 1k) elliptical or one-component sentence, 1l) sentence-final position for neutral syntactic focus, and sentence-initial position for contrastive syntactic focus, 1m) parallel constructions. In both groups, only Max Fixation duration was modulated by linguistic foregrounding cues.

Both impulsive and reflective readers showed higher gaze costs with parallel constructions (χ^2 (1, 699) =7.11, p=0.008 with impulsive readers, and χ^2 (1, 636) = 5.49, p=0.019 with reflective readers). This means that repeatability of linguistic structures attracts attention and consequently provides better access to information. Impulsive readers were also affected by phonological foregrounding (onomatopoeia, alliteration, etc.) (χ^2 (1, 699) =4.91, p=0.027). Reflective readers had higher gaze costs with elliptical or one-component sentence (χ^2 (1, 636) = 7.66, p=0.006). We assume that different reasons may cause these accessibility effects. Phonological foregrounding in a written text is expressed via graphic means mostly displayed in repetition of letters or combination of letters like in / Суббота, суббота – хороший вечерок. / where there is the repetition of letters and root morphemes or in the onomatopoetic words like in / Снова взрыв веселья. / where the combination of letters -в-з-р-в-

indicates an onomatopoetic sound combination. In both cases this type of foregrounding implies that a graphic form is focalized since it differs from other graphic forms. With reflective readers the situation is different. Higher gaze costs which appear in elliptical or one-component sentences indicate that these AOIs attract more attention because (in most cases) the propositional information is not distributed among the subject and the predicate but is packed within one syntactic unit, consequently it requires higher gaze costs to unpack it. Therefore, higher density of information can increase its accessibility for reflective readers. Seen in this way, the higher access of the reflective readers to parallel structures may be also explained by higher density of information produced by replicating either nominative structures like in / Заждалась вас, радость моя, светик… / or predicate ones like in / Добрый он, хороший. /.

5 Discussion

In this section we will present the foregrounding and accessibility effects on a systemic basis, for impulsive and reflective readers separately.

5.1 Foregrounding and accessibility with impulsive readers

The results have shown that only Max Fixation duration (out of 3 gaze metrics tested) was modulated by foregrounding cues. The results are in line with current trends, indicating that central tendency measures of fixation duration alone are not sufficient for eye movement analysis of information processing [17]. As shown by S. Negi and R. Mitra, extremely long fixations (>1000ms) contribute negatively to learning; therefore, Max fixation metric can provide evidence of information processing difficulties.

Major effects were observed with graphological foregrounding, which attests to the results of M. Reingold and K. Rayner [22] regarding longer fixation duration on the words in boldface. Still, we specified the foregrounding cues which produced higher gaze costs: first letter capitalization, tabulation, full stop, comma, dots, colon. We hypothesized that initial or final discursive act foregrounding cues make information more accessible for the impulsive readers, which conforms to the results received in the previous research [10]. Linguistic foregrounding does not produce such significant effects. Still, parallel constructions and phonological foregrounding (onomatopoeia, alliteration, etc.) appeared to attract higher attention. We assumed that these effects also account for graphic focalization, since in both cases we observe repeatability of linguistic structures. Consequently, this is the foregrounding by means of visually focal information which becomes more accessible for impulsive readers.

5.2 Foregrounding and accessibility with reflective readers

We found that both Max Fixation duration and First Fixation Duration displayed variance modulated by foregrounding effects; however, First Fixation Duration was affected only in two cases of graphological foregrounding. The results show that two metrics can be applicable to assess information processing difficulty, with initial processing manifesting in First Fixation Duration and processing difficulty manifesting in Max Fixation Duration.

Similarly with impulsive readers, reflectives were mostly attracted by graphological foregrounding; however, we observed that they had high access to both the information which appeared in the initial and final position of the discursive act and within the discursive act. The results indicate that their attention is guided by the position of foregrounding features in AOI. Importantly, linguistic foregrounding which produced higher access (parallel constructions and elliptical or one-component sentences) relates to the type which accounts for higher information density. This means that reflective readers tend to demonstrate top-down attention, which is consistent with the notion of the impulsive-reflective cognitive style [23]. The findings also prove that the developed semantic protocol specifying foregrounding cues following [7; 9; 18; 26; 27] is an efficient instrument to explore information accessibility.

6 Final remarks

In the paper, we explored the interrelation of foregrounding and information accessibility in reading. The study showed that information accessibility is maintained by at least two factors, different types of foregrounding (here – graphological and linguistic) and the readers' cognitive style (here – impulsive and reflective). The results help predict possible clines for impulsive and reflective readers attributed to

the differences in their attention types, bottom-up and top-down. Importantly, the results allow to range linguistic foregrounding effects in terms of their perception. Additionally, the study paves the way for developing a synergetic approach to information foregrounding and accessibility, which will make both linguistics and cognitive psychology benefit from it.

Acknowledgements

This research is supported by the Russian Science Foundation, project No. 22-28-01754 “Cognitive load economy in media texts interpretation: Multimodal Corpus of Oculographic Reactions MultiCOR”.

References

[1] Mira Ariel. Accessibility theory: an overview. In: Ted Sanders, Joost Schilperoord, & Wilbert Spooren (eds.) *Text representation: linguistic and psychological aspects*. Amsterdam: John Benjamins, 2001. P. 29–87.

[2] Irina V. Blinnikova, Anna I. Izmalkova. Modeling search in web environment: the analysis of eye movement measures and patterns. *Intelligent Decision Technologies 2017.* P. 297–307.

[3] Tatiana V. Chernigoskaya, Tatiana E. Petrova. *The gaze of Shredinger’s cat: identifying gaze metrics in psycholinguistic studies*. Saint-Petersburg, 2018.

[4] Nelson Cowan. *Attention and Memory: An Integrated Framework*. Oxford: Oxford University Press, 1995.

[5] Yulia Esaulova, Martina Penke, Sarah Dolscheid. Referent Cueing, Position, and Animacy as Accessibility Factors in Visually Situated Sentence Production. *Frontiers in Psychology*, 2020. Aug 27;11:2111. Accessed at: 10.3389/fpsyg.2020.02111.

[6] Maria Falikman. Perception and Attention Research in Russia: Traditions and State of the Art. *Journal of Russian and East European Psychology*, 49(5), 2011. P. 3–9.

[7] Talmy Givón. Beyond foreground and background. In Russell S. Tomlin (ed.) *Coherence and grounding in discourse*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 1987. P. 175–168.

[8] Shir Givoni and Rachel Giora. Salience and Defaultness. In: Frank Liedtke and Astrid Tuchen (eds.) *Handbuch Pragmatik.* Stuttgart: J.B. Metzler, 2018. P. 207–213.

[9] Olga Iriskhanova. *Games of focus in language*. Moscow: YaSK, 2014.

[10] Anna I. Izmalkova, Anastasia A. Rzheshevskaya. Graphological and semantic foregrounding as affecting gaze and speech of impulsive and reflective readers. *Languages and Modalities*, 2, 2022. P. 19–26.

[11] Jerome Kagan. Reflection-impulsivity: The generality and dynamics of conceptual tempo. *Journal of abnor mal psychology*, 71 (1), 1966. P. 17–24.

[12] Andrey A. Kibrik. Reference and working memory. In: Karen van Hoek, Andrey A. Kibrik, Leo Noordman *Discourse studies in cognitive linguistics*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 2009. P. 29–52.

[13] Maria I. Kiose, Anastasia A. Rzheshevskaya, Anna I. Izmalkova. Gaze behavior in single-page monomodal and cross-modal switches as affected by Event construal. *Computational Linguistics and Intellectual Technologies*. Papers from the Annual International Conference “Dialogue”, 21 (Supplementary volume), 2022. P. 1078–1088.

[14] Reinhold Kliegl, Ellen Grabner, Martin Rolfs, and Ralf Engbert. Length, frequency, and predictability effects of words on eye movements in reading. *European Journal of Cognitive Psychology*, 16, 2004. P. 262–284.

[15] Irina I. Kovtunova. *Modern Russian language. Word order and actual division of the sentence*. Moscow: Prosveschenije, 1976.

[16] Andriy Myachikov, Simon Garrod, Christoph Scheepers. Attention and memory play different roles in syntactic choice during sentence production. *Discourse processes*, 55(2), 2018. P. 218–229.

[17] Shivsevak Negi, Ritayan Mitra. Fixation duration and the learning process: An eye tracking study with subtitled videos. *Journal of Eye Movement Research*, 13.6, 2020. P. 1–15.

[18] Jan-Ola Östman and Tuija Virtanen. Theme, Comment, and Newness as Figures in Information Structuring. In: Karen van Hoek, Andrey A. Kibrik, Leo Noordman *Discourse studies in cognitive linguistics*. Amsterdam, Philadelphia: John Benjamins Publishing Company, 2009. P. 91–110.

[19] Robert Perinbanayagam. *Discursive acts. Language, Signs, and Selves*. Routledge, New York, USA, 2011.

[20] Mikhail Pokhoday, Yuri Shtyrov, Andriy Myachykov. Effects of Visual Priming and Event Orientation on Word Order Choice in Russian Sentence Production. *Frontiers in Psychology,* 10: 1661. PMID 31481907.

[21] Keith Rayner. Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, 124 (3), 1998. P. 372–422.
[22] Eyal Reingold, Keith Rayner. Examining the word identification stages hypothesized by the EZ Reader model. *Psychological Science,* 17(9), 2006). P. 742–746.
[23] Richard Riding, Stephen Rayner. *Cognitive styles and learning strategies: Understanding style differences in learning and behavior*. London: Routledge, 2013.
[24] Paulette Rozencwajg & Denis Corroyer. Cognitive Processes in the Reflective-Impulsive Cognitive Style. *The Journal of genetic psychology*, 166, 2005. P. 451–463.
[25] Russell S. Tomlin. Linguistic reflections of cognitive events. In: Russell S. Tomlin (ed.) *Coherence and Grounding in Discourse.* Amsterdam, Philadelphia: John Benjamins Publishing Company, 1987. P. 455–480.
[26] Arie Verhagen. Construal and perspectivization. In: Dirk Geeraerts and Hubert Cuyckens (eds.) *The Oxford Handbook of Cognitive Linguistics*. Oxford: Oxford University Press, 2007. P. 48–81.
[27] Brita Wårwik. What is foregrounded in narratives? Hypotheses for the cognitive basis of foregrounding. In Tujia Virtanen (ed.). *Approaches to Cognition through Text and Discourse*. Berlin, New York: Mouton de Gruyter, 2004. P. 99–122.
[28] Li-fang Zhang, Robert J. Sternberg, & Stephen Rayner. (eds.) *Handbook of Intellectual Styles: Preferences in cognition, learning, and thinking*. New York: Springer Publishing Co, 2012.

Towards a Russian Multimedia Politeness Corpus

Ksenia Klokova
MIPT
Moscow, Russia
klokova.ks@mipt.ru

Maxim Krongauz
HSE
Moscow, Russia
mkrongauz@hse.ru

Valery Shulginov
MIPT, HSE
Moscow, Russia
shulginov.va@mipt.ru

Tatiana Yudina
MIPT
Moscow, Russia
yudina.tatiana.a@mipt.ru

Abstract

Communication involves an exchange of information as well as the use of linguistic means to begin, sustain, and end conversations. Politeness is seen as one of the major language tools that facilitate smooth communication. In English, politeness has been an area of great interest in pragmatics, with various theories and corpus annotation approaches used to understand the relationship between politeness and social categories like power and gender, and to build Natural Language Processing applications. In Russian linguistics, politeness research has largely focused on lexical markers and speech strategies. This paper introduces the ongoing work on the development of the Russian Multimedia Politeness Corpus and discusses an annotation framework for oral communicative interaction, with an emphasis on adapting politeness theories for discourse annotation. The proposed approach lies in the identification of frames that encompass contextual information and the selection of relevant spatial, social, and relational features for the markup. The frames are then used to describe standard situations, which are marked by typical intentions and politeness formulae and paraverbal markers.

Keywords: politeness, discourse, multimedia corpora, speech act, sociolinguistics
DOI: 10.28995/2075-7182-2023-22-233-244

Мультимедийный корпус вежливости русского языка

Ксения Клокова
МФТИ
Москва, Россия
klokova.ks@mipt.ru

Максим Кронгауз
НИУ ВШЭ
г. Москва, Россия
mkrongauz@hse.ru

Валерий Шульгинов
МФТИ, НИУ ВШЭ
Москва, Россия
shulginov.va@mipt.ru

Татьяна Юдина
МФТИ
Москва, Россия
yudina.tatiana.a@mipt.ru

Аннотация

Коммуникация включает в себя обмен информацией, а также использование языковых средств с целью начала, поддержания и завершения разговора. Вежливость рассматривается как один из основных языковых инструментов, сглаживающих общение. В английском языке вежливость представляет большой интерес для прагматики, различные теории и подходы к аннотации корпусов используются для понимания взаимосвязи между вежливостью и социальными категориями, такими как власть и гендер, а также для создания приложений для обработки естественного языка. В традиции русской лингвистики исследования вежливости в основном сосредоточены на лексических маркерах и речевых стратегиях. В данной работе представляется текущая работа по разработке Мультимедийного корпуса вежливости русского языка и обсуждается структура

аннотации для устного коммуникативного взаимодействия с акцентом на адаптацию теорий вежливости для разметки дискурса. Предлагаемый подход заключается в идентификации фреймов, которые охватывают контекстную информацию, и выборе соответствующих пространственных, социальных и реляционных характеристик для разметки. Затем фреймы используются для описания стандартных ситуаций, которые отмечаются типичными речевыми намерениями и формулами вежливости, а также паравербальными маркерами.

Ключевые слова: вежливость, дискурс, мультимедийный корпус, речевой акт, социолингвистика

1 Introduction

Communication involves not only the exchange of information between speakers, but also the use of a whole set of linguistic means to begin, sustain, and end conversations. These means belong to the language, but are largely determined by social and cultural preconditions, and the strategies of their use are usually denoted by the term *politeness*. Politeness in English has long been a subject of great interest in pragmatics, from classical politeness theories, where politeness is seen as a result of rational communicative behavior (Lakoff, 1973; Leech, 1983; Brown and Levinson, 1987) to discursive ones focused on the analysis of the dialogue interaction and its evaluations by the participants of communication (Eelen, 2001; Watts, 2003; Ogiermann, 2009).

Findings from these theories have been used for annotation of politeness corpora which served various purposes: from the analysis of the relationship between politeness and such social categories as power, status and gender (Danescu-Niculescu-Mizil et al., 2013) to the NLP applications of adjusting the degree of politeness in written communication (Madaan et al., 2020) and creating a politeness adaptive dialogue system (Mishra et al., 2022). Recently, more and more interest has been drawn to the research of how politeness affects users' perceptions in chat-bots (Liebrecht et al., 2020; Rana et al., 2021; Shan et al., 2022).

Research of politeness in Russian linguistics has mostly been concerned with the use of lexical markers, which constitute a system of stable communicative formulae for establishing contact and maintaining communication in a chosen tone (Krongauz, 2004; Formanovskaja, 2002). In a broader context, there have been works on speech strategies in interactions between interlocutors in situations of persuasion and provocation (Issers, 2009), functions of imperative verb forms in a situation of request or prompt (Paducheva, 2010), and extralinguistic conditions of spontaneous speech interactions (Zemskaja et al., 1981). Most of these studies describe qualitative traits of communication while quantitative studies have been confined to the tasks of automatic detection of speech aggression and detoxification of online communication (Dementieva et al., 2021; Dementieva et al., 2022). The aim of our project is to develop the annotation framework for oral communicative interaction that on one hand takes into account approaches from linguistic politeness theories and on the other could be used for qualitative research and the NLP applications.

Since the corpus is centered around oral communication, data for annotation are taken from excerpts of modern Russian movies, series, talk shows and phone conversations. One of our current goals has been the adaptation of politeness theories for practical discourse annotation. In addition, since politeness is acquired through the process of socialization and therefore is influenced by a large number of contextual and social factors, the second goal was to select appropriate factors to include in the markup.

The paper is structured as follows. First, we provide an overview of existing politeness theories and politeness-related corpora for English and Russian. Then we focus on defining the scope of the discourse unit that should be taken for annotation. The last section is devoted to the description of contextual and social features that are included in the markup. In the last section we present some of the high level statistics of the current corpus.

2 Related work

2.1 Politeness theories

Politeness has been one of the major research fields in pragmatics since its scientific establishment in the last century. The first wave of research took Grice's Cooperative principle (Grice, 1975) and its assump-

tion of rational communicative behavior as a starting point and proposed general politeness principles that serve a purpose of smoothing friction in interpersonal interaction (Lakoff, 1973; Leech, 1983). Following the same rational principle, the seminal Universal Politeness Theory (Brown and Levinson, 1987) considers politeness as measures to preserve the social face (of a speaker and that of an addressee) by mitigating face-threatening speech acts (FTAs) and describes a number of strategies a speaker could use when performing an FTA. The social context is included in the speaker's account of how threatening a given speech act is. According to the theory, the speaker can assess the weight of the threat by taking into account the social distance between them and the addressee, the power (status) distance, and the level of imposition.

Despite still being popular among the researchers, the Universal Politeness Theory was vigorously criticized for the universality of norms it claimed, its focus on individual speech acts, and static social context. The reliance on individual speech acts makes it impossible to account for the politeness in communicative situations – the interlocutors' reactions and responses, as well structure of the dialogue. As a result, the second wave of politeness research was centered around the context of a particular interaction and the interpretation of speech activity by the communicants themselves (Ide, 1989; Werkhofer, 1992; Marriott, 1993; Spencer-Oatey, 2000; Eelen, 2001; Watts, 2003; Ogiermann, 2009). In these discursive theories norms were considered changeable and constituted by interpersonal interactions. The main challenge of the discursive approach lies in rejection of any generalizations and consideration of all possible contextual and personal features that could influence the politeness markers. The third wave of politeness research aimed at overcoming the shortcomings of both classical and discursive theories by integrating them in one another. Its main interest lies in oral communication and the analysis of discourse pieces that can be generalized (Arundale, 2013; Haugh, 2021; Kádár and Haugh, 2013; Terkourafi, 2005b).

The Russian tradition generally distinguishes between politeness and etiquette. Karasik posits that politeness is concerned with the manifestation of respect for another person, whereas etiquette constitutes a set of rules that govern human social behavior (Karasik, 2002). He additionally argues that etiquette research covers not only verbal units, but non-linguistic means of communication and the determination of the parameters of etiquette social variability. Formanovskaja agrees and makes further distinction that speech etiquette constitutes a socially defined and national-specific set of rules that govern speech behavior in interactions, conditioned on social roles of the communicants and their relations (Formanovskaja, 2002). For her, politeness draws from Leech's Politeness principle, whereas etiquette manifests in a separate class of etiquette speech acts. Common broad and repetitive situations in which speech etiquette is analyzed are similar to the ones explored in politeness theories discussed above: getting someone's attention, greeting, acquaintance, farewell, apology, compliment, consolation, condolence, etc. (Prokhorov and Stepin, 2006). In similar fashion, Larina discusses the difference between politeness in Russian and English, mainly leaning on the classical politeness theories (Larina, 2009). Since research of politeness and speech etiquette are much less formalized in Russian linguistic tradition and the distinction between them for most part is not well defined, we prefer using the term *politeness*.

2.2 Politeness and multimedia corpora

Most of the studies related to the development of the politeness corpora are focused on English. In the English-speaking environment, there is a tendency to refine existing well-known corpora to solve problems related to polite communication. One of the examples is the Enron corpus (Klimt and Yang, 2004), which consists of e-mail correspondence of an American corporation (1.39 million sentences). In its polite version (Madaan et al., 2020) the sentences were first automatically scored for the degree of politeness using the pre-existing classifier (Niu and Bansal, 2018) and then the top-scoring ones were used as training data for the politeness transfer task. Another example is the large MultiDoGO dialogue dataset (Peskov et al., 2019), which contains conversations between an agent and a customer in several domains (airline, fast food, finance, insurance, media, and software). For the polite version, each utterance was annotated with one of four fine-grained politeness classes to be used in a politeness adaptive dialogue system (Mishra et al., 2022).

In an earlier work on politeness corpora the area of interest were requests (Danescu-Niculescu-Mizil

et al., 2013) as one type of speech acts that pose threats to the addressee's social face according to the Universal Politeness Theory. For this study a portion of data from Wikipedia and Stack Exchange requests was annotated with domain-independent lexical and syntactical features (e.g. polite formulae like please and hedges), as well as politeness scores obtained from crowdsourced annotation. It was then used to train a politeness classifier for automatic labeling of the remaining data. Preliminary analysis of the relation between the degree of politeness and social features showed its variation conditioned on power, status, gender.

In linguistic studies of politeness, the British National Corpus [1], both text and audio, is also often used (McEnery et al., 2002; Deutschmann, 2006; Vizcaíno, 2007). Additionally, there are a number of works on polite communication that use a range of multimedia corpora: political comics corpora (Abdel-Raheem, 2021), the Santa Barbara series corpora (Brown, 2014) and studies on TV charity commercials (Pennock-Speck and del Saz-Rubio, 2013). These studies show that corpora annotated with paraverbal features are important for a deeper study of politeness.

Although there are no corpora annotated for politeness in Russian language, several corpora within toxic communication research exist. Namely, the Russian Language Toxic Comments dataset (Belchikov, 2019), the Toxic Russian Comments corpus (Semiletov, 2020) and the RuToxic data corpus (Dementieva et al., 2021), which is based on the former two. These corpora contain comments from the social networks Odnoklassniki, Pikabu and the Dvach forum. Studies in sociolinguistics often use corpora of oral communication (Sherstinova, 2009; Bogdanova-Beglarian et al., 2016; Cui, 2019). The most representative corpus of oral communication in Russian is the ORD ("One Speaker's Day"), which contains 240 hours of recordings of everyday telephone conversations (Asinovsky et al., 2009; Sherstinova, 2009).

3 Basic data unit for annotation

One of the most important tasks in the development of an oral politeness corpus is a delineation of a minimal discourse unit that should be considered for annotation. Approach that should be chosen to resolve this task depends on the theoretical politeness concepts. If we adhere to the assumption of rational communicative behavior (theories based on Grice's principle), it should be the elementary discursive unit (EDU) as proposed by Kibrik and Podlesskaja (Kibrik and Podlesskaja, 2009). EDU is a quantum of oral discourse, the minimum step by which the speaker moves the discourse forward. It correlates with the notion of the speech act in the classical politeness theories, with the former being determined to a large extent by prosodic features, the latter being mostly defined by the speaker's intention.

However, reducing a politeness to an individual EDU, as well as speech act, does not allow for the analysis of composite polite utterances or a combination of formulae. If we consider, for example, a composite greeting: "Hello, hope you are doing well today." – it is clear that it consists of two speech acts (greeting and expressive), however, it could be considered as one long greeting.

On the other hand, the discursive approach can be considered, in which politeness is based on an understanding of discourse as a dialogic transaction between the participants (Watts, 2003; Locher and Watts, 2005). Thus, marking a statement as polite or impolite is possible only in the case of a match/mismatch between the addressee's expectations and the signals provided by the speaker (Jary, 1998). In data annotation for politeness, it is necessary to consider the broader communicative context, which includes the discursive interaction of the speaker and listener, their social and cultural characteristics, as well as standard communicative expectations and responsibilities.

To designate this level of discourse analysis, we specify the term *frame*, by which Terkurafi understands "immediately observable, indispensable extra linguistic information about a situation" that is "summarised together with information about the appropriate linguistic politeness marker(s)" (Terkourafi, 2005a). Thus, the frame contains prerequisites for the use of politeness markers and social parameters that regulate the choice of linguistic means. At the current stage of our project, we distinguish the following types of prerequisites: the beginning or the end of communicative contact, compensation for causing damage or inconvenience and doing good.

[1] http://www.natcorp.ox.ac.uk/

The frames contain *standard situations* which are marked with an occurrence of typical intentions of the EDU. A *standard situation* then is a special case of a frame realization in which ritualized politeness formulae (and non-verbal markers) manifest themselves (in greetings, acquaintances, farewells, apologies and gratitude). Thus, the structure of our data suggests the descriptions of intentions and formulae that determine the type of polite interaction, as well as broad social and communicative context that reflects discursive approaches to politeness.

4 Context features

Broad context

As it is evident from the previous sections, context and individual characteristics of the interlocutors influence the way politeness is manifested. In order to give the users of the Russian Multimedia Politeness Corpus an opportunity to conduct research on its various aspects, we chose to annotate macro-level dialogues along with the social context of a frame in which a standard situation takes place. In the following we describe different levels of context that are annotated and illustrate them on an excerpt from a Russian movie called "Exercises in beauty" (*Uprazhneniya v prekrasnom*, 2011):

(1) Hotel employee: Ah, Evgeny Sanych, good afternoon. Do you remember, you have stayed here two years ago?
Evgeny nods
(2) Hotel employee [hands over keys]: Here is your suite, second floor.
Evgeny nods
*(3) *Evgeny starts leaving but the hotel employee blocks his way**
(4) Hotel employee [holds out a piece of paper and a pen]: Can I also ask for an autograph for my niece Liza? And here is the pen... It doesn't write... [holds out another pen] To Lizunchik from...
Evgeny signs the paper shoves it back (5) and leaves (6)

We first specify the place and time of the discourse fragment in a separate entity called an episode. In our example those would be a hotel lobby and the action takes place after the main characters arrive in a new town.

The next level of annotation is to identify the frames that are present in the fragment. As mentioned in Section 3, currently there are four types of prerequisites for the usage of politeness markers and these are used to mark the frame type. In our example the frames are the following: the beginning of the conversation (1), doing good (2), compensation for causing inconvenience (3), doing good (5), the end of the conversation (6). Additional frame – request (4) – is not included in the annotation yet.

On the lower annotation level we enter the frames themselves, which are confined to specific timestamps and the dialogue boundaries. Here we annotate the interlocutors (Evgeny and the hotel employee). Further, it is determined whether the frames are realized and, if so, they are annotated with the standard situation type. In our case these are just greeting (the beginning of the conversation frame) and gratitude (the doing good frame (2)). The frames (3), (5) and (6) are not annotated with standard situations since there are no verbal or paraverbal markers that label typical communicative behavior. The standard situations are then annotated with timestamps, text, presence of non-verbal markers and address terms (*Evgeny Sanych* as in (1)).

Basic social features of interlocutors

According to our approach, the choice of politeness strategies and markers is influenced by social and cultural factors that are specific to different interaction situations. We have identified two types of features: basic social features and relation features. Basic social features are constant characteristics of the communicators that influence their speech behavior. These features are fixed and do not change during the course of the interaction. Relation features, on the other hand, are characteristics that become important during the interaction between communicators. These characteristics are often not fixed and may change depending on the context of the interaction.

It is commonly agreed in psychological and sociolinguistic research that various psychological and social characteristics of individuals are marked in speech in general and in how politeness is manifested as well. Among those are age (Helfrich, 1979; Bella, 2009), gender (Smith, 1979; Holmes, 1995; Mills, 2003) and belonging to a particular social group (Brown and Levinson, 1979; Mahmud, 2013).

All of these features are annotated in our corpus, with the social group reflected in profession and education if available. Both of these features can assume one value from an open list which is populated throughout the annotation process. If there is no information about a character's age, an age bin is assigned instead. It is also worth noting that although social status of a character is not annotated separately, it can be inferred from the combination of age, profession and education. In our running example, Evgeny would be assigned the following values: male, 40 years old, adult, actor, Russian Institute of Theatre Arts. Similar for the hotel employee, however, for him exact age and education features are missing: male, young man, hotel employee.

Relation features

Following the inclusion of the status variables in the Universal Politeness Theory (Brown and Levinson, 1987, p. 74), we annotate symmetric relations of social distance between the interlocutors: degree of familiarity in the frame (strangers, little-known to each other and acquaintances) and relationships (co-workers, friends, spouses, relatives, etc.). In the same way asymmetric relations are taken into account in the form of hierarchy. This includes two features: one interlocutor's position relative to the other (higher, lower or equal) and the specification of such a position (age difference, status, rank, etc.).

In our example, Evgeny and the hotel employee are little-known to each other and do not have any relationship. In the given context Evgeny's social role (famous actor) is hierarchically higher in relation to the hotel employee. This is partially confirmed by the observed verbal behavior (Evgeny does not use politeness formulae where those would be expected) and the fact the hotel employee asks for his autograph. Thus, in the annotation the specification of the hierarchy for these two characters would be Evgeny's higher position due to his status.

Gestures and address terms

Gumperz termed the social and cultural factors which facilitate politeness in interaction as "signaling mechanisms" (Gumperz, 1982, p. 16). These mechanisms are often not consciously used by the participants, making them useful resources for analyzing the results of communicative exchanges. Important signaling mechanisms we focus on include nonverbal cues (e.g. gestures) and politeness features not tied to a particular frame (such as addressing and pronoun switching at the future stages).

Gestures can either support the corresponding speech act or alter its meaning. Indeed, if we consider a polite utterance thank you accompanied by a rude gesture, the overall intention would not be an expression of gratitude. Furthermore, the communicative act can be performed in multiple modalities and the speaker generally is not restricted to choose a verbal strategy (e.g. to nod instead of saying *hello*) (Arndt and Janney, 1985; Ambady et al., 1996).

Address terms (names, titles), on the other hand, are usually used as supportive mechanisms to either show respect or confirm the existing relationship, as in our running example. They could also serve a double function – for example, as an attention getter or as a greeting. Furthermore, address terms are frequently used to model politeness (Voigt et al., 2017; Yeomans et al., 2018).

As mentioned above, there is one address term in our example – in the frame (1), the standard situation of greeting. Furthermore, in the standard situations of both greeting and gratitude Evgeny's response is a nod which would be annotated as a gesture in the corresponding situations.

Figure 1 illustrates the overall conceptual schema for the proposed annotation, including social features and general metadata.

5 Preliminary annotation results

In this section we present the results of annotation that have been achieved and the current corpus volume. The data for annotation came from fifteen movies and series released after 2000. At the first stage of the project it was decided to focus on fiction media which closely reflect modern life and communication.

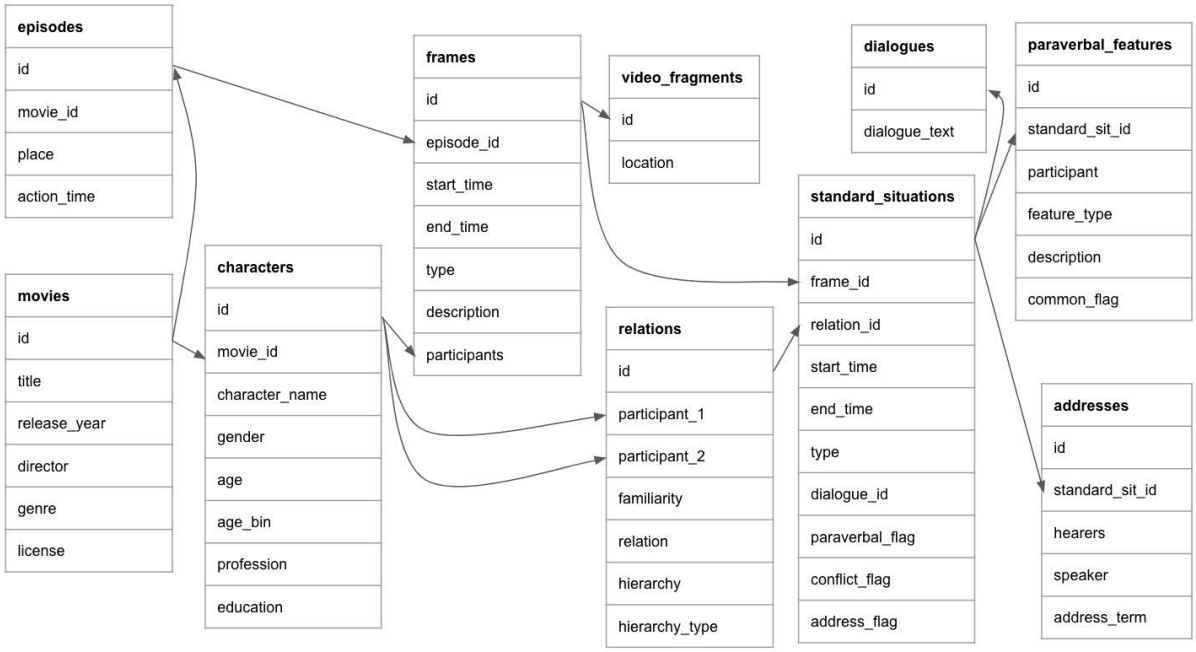


Figure 1: Conceptual schema for the proposed annotation.

The full list of the movies is presented in Table 1, the current version of the corpus includes 3,000 potential frames and 525 frames that have been verified by mark-up match among annotators.

Table 1 presents statistics on the verified annotated data. The standard situations in which frames are realized are shown in Figure 2. The largest share is occupied by greetings, which correlates with the distribution of frames, since greetings are included at the beginning of communication. This can be explained by the peculiarities of the media taken for marking, since in films the situations of meetings and acquaintances are more significant for the narrative.

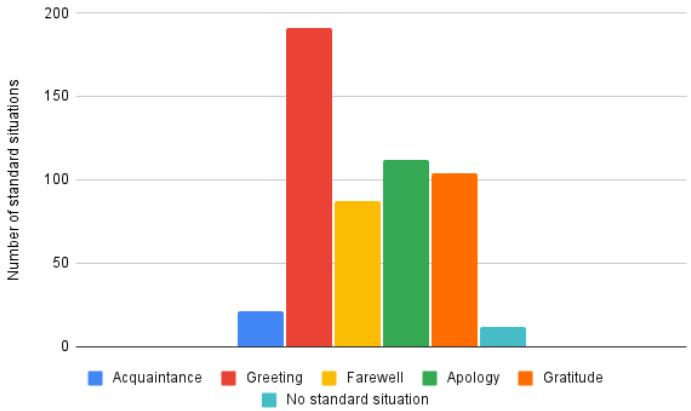
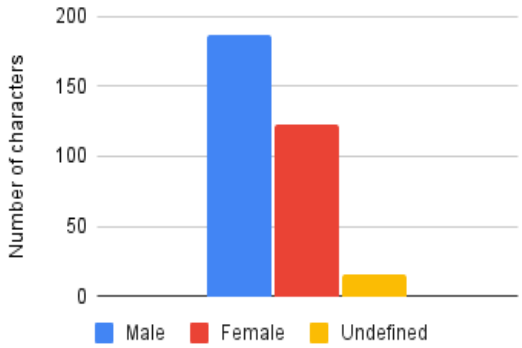


Figure 2: Distribution of the standard situation types. *No standard situation* means that the corresponding frame was not realized.

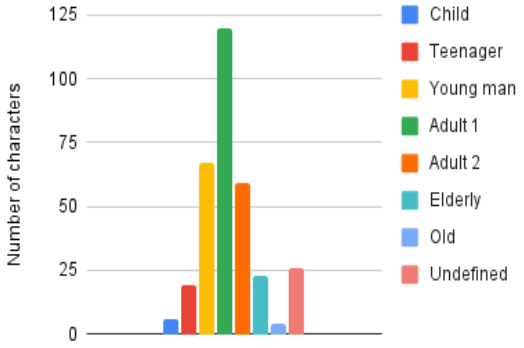
At the level of the participant feature labeling Figure 3 (Subfigures 3a, 3b, 3c, we can note some skew, which is manifested in the gender composition of the interactors (a greater number of men); age characteristics (Adult 1 dominates (range of 35-40 years old)); as well as types of familiarity degree, where communication between acquaintances occurs more often. *Undefined* means a mixed group of different genders, ages or familiarity degree to the speaker.

Movie	Number of frames	Min duration	Max duration	Total
Arrhythmia (2017)	34	0:00:01	0:00:24	0:03:30
The geographer drank away the globe (2013)	36	0:00:02	0:01:12	0:06:29
Fool (2014)	30	0:00:02	0:00:48	0:05:47
Speakerphone (2018)	34	0:00:01	0:02:28	0:04:42
Exercises in beauty (2011)	39	0:00:01	0:00:36	0:05:58
Inadequate people (2010)	31	0:00:02	0:01:57	0:09:29
Stories (2012)	53	0:00:02	0:02:52	0:12:00
Radio day (2008)	36	0:00:01	0:00:25	0:03:44
The stroll (2003)	29	0:00:02	0:02:36	0:09:21
Major (series, 2004)	30	0:00:02	0:00:24	0:03:45
Peter FM (2006)	30	0:00:02	0:00:43	0:04:35
What Men Talk About (2010)	37	0:00:01	0:00:39	0:04:02
Last minister (series, 2020)	47	0:00:01	0:00:25	0:07:55
Election day (2007)	41	0:00:02	0:00:38	0:05:57
This is what's happening to me (2012)	18	0:00:16	0:03:03	0:19:49

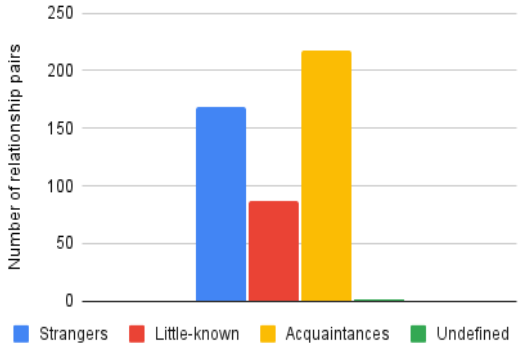
Table 1: Number of frames and general durations of video fragments per annotated movie



(a) Gender distribution



(b) Age distribution



(c) Familiarity degree distribution

Figure 3: Statistics on characters' features

6 Conclusion and future work

Politeness research is a well-established scientific area in English, however relatively little attention has been paid to systematic studies of politeness in Russian. With growing interest from both sociolinguists and NLP practitioners in how politeness in oral communication can be modeled, which strategies and formulae are used and how they can be integrated in natural language processing applications, it seems natural to attempt to construct a language resource that could be used for research in these areas. In this paper we introduced the ongoing work on such an attempt – the Russian Multimedia Politeness Corpus.

One of the main intricacies in the construction of the corpus is delineation of the discourse piece that should be annotated. Classical politeness theories operate on the single speech act level, whereas discursive approaches tend to consider much longer sequences and dismiss any effort for generalizations. Therefore, our suggested approach is based on frames, which encompass extra linguistic information about the communicative situation and can consist of one or several standard situations in which conventionalized politeness manifests itself. Currently, we annotate the frames that are concerned with the beginning or the end of communicative contact, compensation for causing damage or inconvenience and doing good. Then, the corresponding standard situations are greetings, acquaintances, farewells, apologies and gratitude.

Being a part of oral speech, the choice of politeness strategies and formulae are too influenced by a great variety of contextual factors. The one included in annotation can be split into three groups: spacial (place and time of the action), social features of the interlocutors (age, age bin, gender, profession and education) and relational features between the interlocutors (degree of familiarity, relationship, hierarchy). Additionally, we annotate gestures as they can align with, alter the meaning of or replace actual utterances, and address terms for they play an important role in the confirmation of the existing relationship and have several pragmatic functions.

Currently the corpus consists of 3,000 potential frames and 525 frames that have been verified. The descriptive statistics of the latter data shows skews in the distributions of social features, as well as frames and standard situations. If the asymmetry of social features should be balanced at the future stages, the distribution of frame types might be explained by the media chosen for annotation or be representative of Russian oral communication in general.

Acknowledgements

The research is fulfilled within the framework and with the support of the strategic academic leadership program "Priority 2030" of the Moscow Institute of Physics and Technology (National Research University).

References

Ahmed Abdel-Raheem. 2021. Multimodal metaphor and (im)politeness in political cartoons: A sociocognitive approach. *Journal of Pragmatics*, 185:54–72.

Nalini Ambady, Jasook Koo, et al. 1996. More than words: Linguistic and nonlinguistic politeness in two cultures. *Journal of Personality and Social Psychology*, 70(5):996—-1011.

Horst Arndt and Richard W. Janney. 1985. Politeness revisited: Cross-modal supportive strategies. *IRAL: International Review of Applied Linguistics in Language Teaching*, 23(4):281–300.

Robert B. Arundale. 2013. Conceptualizing 'interaction' in interpersonal pragmatics: Implications for understanding and research. *Journal of Pragmatics*, 58:12–26.

Alexander Asinovsky, Natalia Bogdanova, et al. 2009. The ord speech corpus of russian everyday communication "one speaker's day": Creation principles and annotation. *// International Conference on Text, Speech and Dialogue*, P 250–257, Pilsen, Czech Republic.

Anatolij Belchikov. 2019. Russian language toxic comments. Accessed on March 10, 2023.

Spyridoula Bella. 2009. Invitations and politeness in greek: The age variable. *Journal of Politeness Research: Language, Behaviour, Culture*, 5(2):243–271.

Natalia Bogdanova-Beglarian, Tatiana Sherstinova, et al. 2016. An exploratory study on sociolinguistic variation of russian everyday speech. // *International Conference on Speech and Computer*, P 100–107, Budapest, Hungary.

Penelope Brown and Stephen C. Levinson. 1979. Social structure, groups and interaction. // Scherer K. R. and Giles H., *Social Markers in Speech*, P 291–342. Cambridge University Press.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language*. Cambridge University Press, Cambridge, UK.

Lynnelle Rhinier Brown. 2014. Requesting the context: A context analysis of let statement and if statement requests and commands in the santa barbara corpus of spoken american english. *Rice Working Papers in Linguistics*, 5:100–107.

Lili Cui. 2019. The refusal speech act in russian everyday speech (considering the social relations aspect). *Russkaja Rech'*, 5:79–92.

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, et al. 2013. A computational approach to politeness with application to social factors. // *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 250–259, Sofia, Bulgaria.

Daryna Dementieva, Daniil Moskovskiy, et al. 2021. Methods for detoxification of texts for the russian language. *Multimodal Technologies and Interaction*, 5(9):54.

Daryna Dementieva, Daniil Moskovskiy, et al. 2022. Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora. // *Proceedings of the Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*, P 114–131, Online.

Mats Deutschmann. 2006. Social variation in the use of apology formulae in the british national corpus. // Renouf A. and Kehoe A., *The Changing Face of Corpus Linguistics, Language and Computers*, volume 55, P 205–221. Brill.

Gino Eelen. 2001. *A Critique of Politeness Theories*. St. Jerome Publishing, Manchester, UK.

Natal'ja Formanovskaja. 2002. *Rechevoje obschenie: kommunikativno-pragmaticheskij podkhod*. Russkij jazyk, Moskva.

Paul H. Grice. 1975. Logic and conversation. *Syntax and Semantics: Speech Acts*, 3:41–58.

John J. Gumperz. 1982. *Discourse Strategies*. Cambridge University Press, Cambridge, UK.

Michael Haugh. 2021. Discourse and politeness. // Hyland K., Paltridge B., and Wong L., *The Bloomsbury Handbook of Discourse Analysis*, P 219–232. Bloomsbury Academic.

Hede Helfrich. 1979. Age markers in speech. // Scherer K. R. and Giles H., *Social Markers in Speech*, P 63–108. Cambridge University Press.

Janet Holmes. 1995. *Women, Men and Politeness*. Routledge, London, UK.

Sachiko Ide. 1989. Formal forms and discernment: Two neglected aspects of universals of linguistic politeness. *Multilingua*, 8(2-3):223–248.

Oksana Issers. 2009. Strategija rechevoj provokatsii v publichnom dialoge. // Pichkhadze A. and Derzhavina E., *Russkij jazyk v nauchnom osveschenii*, volume 2, P 92–104. Jazyki slavyanskoj kul'tury, Moskva.

Mark Jary. 1998. Relevance theory and the communication of politeness. *Journal of Pragmatics*, 30(1):1–19.

Vladimir Karasik. 2002. *Jazyk sotsial'nogo statusa*. Gnosis, Moskva.

Andrej Kibrik and Vera Podlesskaja. 2009. *Rasskazy o snovedenijakh. Korpusnoje issledovanije ustnogo russkogo diskursa*. Jazyki slavyanskoj kul'tury, Moskva.

Bryan Klimt and Yiming Yang. 2004. Introducing the enron corpus. // *Proceedings of International Conference on Email and Anti-Spam*, Mountain View, CA.

Maxim Krongauz. 2004. Russian oral politeness in the new century. *Russian Linguistics*, 28:163–187.

Dániel Z. Kádár and Michael Haugh. 2013. *Understanding politeness*. Cambridge University Press, Cambridge, UK.

Robin Lakoff. 1973. The logic of politeness of minding your ps and qs. // *Papers from the Ninth Regional Meeting of the Chicago Linquistic Society*, P 292–305.

Tatiana Larina. 2009. *Kategorija vezhlivosti i stil' kommunikatsii: sopostavlenije anglijskikh i russkikh lingvokul'turnykh traditsij*. Jazyki slavyanskoj kul'tury, Moskva.

Geoffrey N. Leech. 1983. *Principles of Pragmatics*. Longman, London, UK.

Christine Liebrecht, Lena Sander, and Charlotte van Hooijdonk. 2020. Too informal? how a chatbot communication style affects brand attitude and quality of interaction. *Journal of Politeness Research*, 12604(1):16–31.

Miriam A. Locher and Richard J. Watts. 2005. Politeness theory and relational work. *Journal of Politeness Research*, 1(1):9–33.

Aman Madaan, Amrith Setlur, et al. 2020. Politeness transfer: A tag and generate approach. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 1869–1881, Online.

Murni Mahmud. 2013. The roles of social status, age, gender, familiarity, and situation in being polite for bugis society. *Asian Social Science*, 9(5):58–72.

Helen Marriott. 1993. Politeness phenomena in japanese intercultural business communication. *Intercultural Communication Studies*, 3(1):15–38.

Tony McEnery, Paul Baker, and Christine Cheepen. 2002. Lexis, indirectness and politeness in operator calls. // *New Frontiers of Corpus Research*, P 53–69, Rodopi, Amsterdam.

Sara Mills. 2003. *Gender and politeness*. Cambridge University Press, Cambridge, UK.

Kshitij Mishra, Mauajama Firdaus, and Asif Ekbal. 2022. Please be polite: Towards building a politeness adaptive dialogue system for goal-oriented conversations. *Neurocomputing*, 494:242–254.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6:373–389.

Eva Ogiermann. 2009. Politeness and in-directness across cultures: A comparison of english, german, polish and russian requests. *Journal of Politeness Research*, 5:189–216.

Elena Paducheva. 2010. *Semanticheskije issledovanija: Semantika vremeni i vida v russkom jazyke. Semantika narrativa*. Jazyki slavyanskoj kul'tury, Moskva.

Barry Pennock-Speck and Milagros M. del Saz-Rubio. 2013. A multimodal analysis of facework strategies in a corpus of charity ads on british television. *Journal of Pragmatics*, 49(1):38–56.

Denis Peskov, Nancy Clarke, et al. 2019. Multi-domain goal-oriented dialogues (multidogo): Strategies toward curating and annotating large scale dialogue data. // *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 4526–4536, Hong Kong, China.

Jurij Prokhorov and Iosif Stepin. 2006. *Russkije: kommunikativnoe povedenie*. Flinta, Nauka, Moskva.

Kanishk Rana, Rahul Madaan, and Jainendra Shukla. 2021. Effect of polite triggers in chatbot conversations on user experience across gender, age, and personality. // *30th IEEE International Conference on Robot & Human Interactive Communication (RO-MAN)*, P 813–819, Vancouver, Canada.

Alexander Semiletov. 2020. Toxic russian comments. Accessed on March 10, 2023.

Yi Shan, Meng Ji, et al. 2022. Language use in conversational agent-based health communication: Systematic review. *Journal of Medical Internet Research*, 24(7).

Tatiana Sherstinova. 2009. The structure of the ord speech corpus of russian everyday communication. // *Text, Speech and Dialogue*, P 258–265, Pilsen, Czech Republic.

Philip M. Smith. 1979. Social structure, groups and interaction. // Scherer K. R. and Giles H., *Sex markers in speech*, P 109–146. Cambridge University Press.

Helen Spencer-Oatey. 2000. Rapport management: A framework for analysis. // Spencer-Oatey H., *Culturally Speaking: Managing Rapport through Talk across Cultures*, P 11–46. Continuum, London, UK.

Marina Terkourafi. 2005a. An argument for a frame-based approach to politeness : Evidence from the use of the imperative in cypriot greek. *Pragmatics and beyond. New series*, 139:99–116.

Marina Terkourafi. 2005b. Beyond the micro-level in politeness research. *Journal of Politeness Research*, 1(2):237–262.

María José García Vizcaíno. 2007. Using oral corpora in contrastive studies of linguistic politeness. *Corpus Linguistics Beyond the Word, Language and Computers*, 60:117–142.

Rob Voigt, Nicholas P. Camp, et al. 2017. Language from police body camera footage shows racial disparities in officer respect. // *Proceedings of the National Academy of Sciences*, volume 114, P 6521–6526.

Richard J. Watts. 2003. *Politeness*. Cambridge University Press, Cambridge, UK.

Konrad T. Werkhofer. 1992. Traditional and modern review: The social constitution and the power of politeness. // R. J. Watts, Ide S., and Ehlich K., *Politeness in Language: Studies in its History, Theory and Practice*, P 155–200. Mouton de Gruyter, Berlin, Boston.

Michael Yeomans, Alejandro Kantor, and Dustin Tingley. 2018. The politeness package: Detecting politeness in natural language. *The R Journal*, 10(2):489–502.

Elena Zemskaja, Margarita Kitajgorodskaja, and Evgenij Schirjaev. 1981. *Russkaja razgovornaja rech': oschije voprosy. Slovoobrazovanije. Sintaksis*. Nauka, Moskva.

An experimental study of argument extraction from presuppositional clauses in Russian

Mikhail Knyazev
Institute for Linguistic Studies, Russian Academy of Sciences, Saint Petersburg, Russia;
HSE University, Saint Petersburg, Russia;
Lomonosov Moscow State University, Moscow, Russia;
misha.knjazev@gmail.com

Abstract

The paper discusses two acceptability rating studies testing wh-interrogative and relative extractions of arguments from *čto*-clauses of presuppositional predicates like *žalet'* 'regret', as contrasted with nonpresuppositional predicates like *nadejat'sja* 'hope' and nominalized (*to čto*) clauses. The results show a difference in extraction between bare and nominalized clauses but no difference between presuppositional and nonpresuppositional clauses, raising potential doubts about the analysis of presuppositional clauses as DPs with a silent D.

Keywords: syntactic islands, presuppositional clauses, nominalized clauses, Russian, experimental study
DOI: 10.28995/2075-7182-2023-22-245-253

Экспериментальное исследование выдвижения аргументов из пресуппозициональных клауз в русском языке

Михаил Князев
Институт лингвистических исследований РАН, Санкт-Петербург, Россия;
НИУ ВШЭ, Санкт-Петербург, Россия;
МГУ им. М. В. Ломоносова, Москва, Россия
misha.knjazev@gmail.com

Аннотация

В статье обсуждаются два эксперимента на оценку приемлемости, проверяющие выдвижение аргументного вопросительного слова и относительного местоимения из клауз со *что* при пресуппозициональных предикатах типа *жалеть* в свавнении с непресуппозициональными предикатами типа *надеяться*, а также номинализованными клаузами с *то, что*. Результаты показывают различие между выдвижением из простых и номинализованных клауз при отсутствии различий между пресуппозициональными и непресуппозициональными клаузами, создавая потенциальную проблему для анализа пресуппозициональных клауз как DP-проекций с нулевым D.

Ключевые слова: синтаксические острова, пресуппозициональные клаузы, номинализованные клаузы, русский язык, экспериментальное исследование

1 Introduction

In Russian, complement clauses can be bare or nominalized, when preceded by the demonstrative *to* 'that' (Kobozeva, 2013, a.o.). It is usually assumed that nominalized CPs, as in (1a), are (strong) islands, whereas bare CPs, as in (1b), generally allow extraction, although it is considered marked for indicative (*čto*) clauses (Khomitsevich, 2007; Morgunova, 2021b).

(1) a. **Komu*$_1$ *Lena nadeetsja na* [$_{DP}$ *to čto pomožet* t_1 *s kvartiroj*]*?*
to whom Lena hopes on that.ACC that will help with apartment
Intended: 'Who does Lena hope that she will help with the apartment?'

b. *??Komu*$_1$ *Lena nadeetsja* [$_{CP}$ *čto pomožet* t_1 *s kvartiroj*]*?*
to whom Lena hopes that will help with apartment
Intended: 'Who does Lena hope that she will help with the apartment?'

Extraction may also depend on the lexical semantic class of the predicate. Thus, complement clauses of *presuppositional* predicates, including cognitive and emotive factives like 'remember' or 'regret', are assumed to be more difficult to extract from, compared to *nonpresuppositional* predicates, including nonfactives like 'say' or 'think'. Because this contrast is much stronger for adjunct compared to argument (object) extractions, presuppositional clauses are usually considered *weak islands* (Hegarty, 1992; Basse, 2008, a.o.). An influential account of presuppositional islands (Kastner, 2015) (see also (Honcoop, 1998)) explains them by analyzing presuppositional clauses as DPs (cf. (Kiparsky and Kiparsky, 1970)) headed by a silent definite determiner creating a barrier for extraction (in contrast to nonpresuppositional clauses analyzed as bare CPs). The crucial assumption of this account is that when D merges with a CP it creates a *weak* island.[1] However, there is also a prominent view that definite or presuppositional DPs create a *strong* island (Davies and Dubinsky, 2003; Sichel, 2018, a.o.), leading to uncertainty as to the validity of Kastner's silent D analysis, at least for English (cf. (Haegeman, 2012; Djärv, 2019)).

The main goal of this paper is to experimentally investigate the contrast in extraction from presuppositional and nonpresuppositional clauses in Russian in order to examine the predictions of Kastner's silent D analysis, which was recently adopted to presuppositional *čto* clauses in (Knyazev, 2022) (based on independent considerations).[2] The present paper looks only at *argument* extractions and thus provides a test of the *strong island version* of the silent D analysis, i.e. testing the latter under the assumption that (definite) D is an absolute barrier for extraction in Russian, whether in general (Pereltsvaig, 2007; Lyutikova, 2010) or specifically when it merges with a CP (Bondarenko, 2022). With this qualification, the silent D analysis predicts a contrast in (argument) extraction between presuppositional *čto* clauses, as in (2b), and nonpresuppositional clauses in (1b). It further predicts that extraction from presuppositional clauses should not differ from the corresponding extraction from *to čto* clauses, as in (2a).

(2) a. **Komu*$_1$ *Vasya žaleet o tom, čto odolžil den'gi* t_1*?*
to whom Vasya regrets about that.PREP that lent money
Intended: 'Who does Vasya regret that he has lent the money to?'

b. **Komu*$_1$ *Vasya žaleet* [$_{PP}$ $\emptyset_P$ [$_{DP}$ $\emptyset_D$ [$_{CP}$ *čto odolžil den'gi* t_1]]]*?*
to whom Vasya regrets that lent money
Intended: 'Who does Vasya regret that he has lent the money to?'

Because extraction from presuppositional vs. nonpresuppositional CPs to my knowledge has not been experimentally tested in Russian, the paper also aims to clarify the empirical picture in this regard.

The paper also examines whether nominalized (*to čto*) clauses are indeed strong islands, which to my knowledge also has not been shown experimentally. In particular, it is important to control for the acceptability of *to čto* clauses in the baseline (no extraction) condition since they may be independently degraded for some verbs (Kobozeva, 2013). To address this confound, the paper focuses on *oblique/PP-taking* predicates like *nadejat'sja* 'hope' and *žalet'* 'regret' (cf. (1)–(2)), for which *to čto* clauses are systematically allowed. Two additional questions are also addressed: is there is a difference in extraction between *čto*- and *čtoby*-clauses? and between wh-interrogative and relative clause dependencies?

Two acceptability rating studies were conducted, one with wh-interrogative extractions (Section 2) and the other with relativization (Section 3). The results confirm the view that nominalized CPs uniformly block extraction. At the same time, they do not show a contrast between presuppositional and nonpresuppositional CPs, contrary to the silent D view and in line with the null hypothesis, according to which presuppositional clauses are bare CPs (Bondarenko, 2022). An alternative interpretation of the results in terms of the weak island version of the silent D analysis is also discussed (Section 4).

[1] More precisely, silent D is assumed to merge directly with a CP creating a weak island, whereas overt D is assumed to merge with (possibly null) N + CP creating a strong (complex NP) island (Kastner, 2015, p.168).

[2] Island data are not discussed in (Knyazev, 2022).

2 Experiment 1

2.1 Design, Materials and Procedure

The experiment had a 2 × 2 × 2 design, with factors: (i) predicate class (presuppositional vs. nonpresuppositional); (ii) presence/absence of extraction; and (iii) complement type (*čto*- vs. *to čto*-clause). The 4 conditions with extraction were shown in (1)–(2); the baseline/no extraction conditions are given in (3).[3]

(3) a. *Lena nadeetsja (na to), čto pomožet Vane s kvartiroj.*
Lena hopes on that.ACC that will help to Vanya with apartment
'Lena hopes that she will help Vanya with the apartment.'

b. *Vasya žaleet (o tom), čto odolžil den'gi Andreju.*
Vasya regrets about that.PREP that lent money to Andrey
'Vasya regrets that he lent money to Andrey.'

4 verbs were used in each class, given in (4) (with subcategorization). The nonpresuppositional class had 4 nonfactive belief/speech predicates; the presuppositional class had 3 emotive factive predicates (*žalet'* 'regret', *rad* 'glad', *gordit'sja* 'proud') and 1 communicative factive *priznat'sja* 'confess'.[4]

(4) a. nonpresuppositional: *nadejat'sja* (*na* ACC) 'hope', *uveren* (*v* PREP) 'certain', *namekat'* (*na* ACC) 'hint', *xvastat'sja* (INS) 'boast'
b. presuppositional: *žalet'* (*o* PREP) 'regret', *rad* (DAT) 'glad', *gordit'sja* (INS) 'proud', *priznat'sja* (*v* PREP) 'confess'

With each predicate, 4 lexically matched sets, crossing extraction and complement type, were created. The 32 experimental sentences were distributed among 4 lists in a Latin Square design (i.e. participants saw each predicate in 1 of the 4 conditions). There were 19 fillers (including practice items, 9 unacceptable and 9 acceptable); the unacceptable fillers contained 4 complex NP violations and 5 selectional violations; 1 sentence contained extraction from the complement of *dumat'* 'think', used as a baseline.

The task was to rate the naturalness of the sentences on a 1–7 scale. The experiment was hosted on PCIbex Farm (`https://farm.pcibex.net/`) and was completed by 45 people.

2.2 Analysis and Predictions

Data from 44 participants who complied with the task were analyzed. A linear mixed effects model was fitted to z-score transformed data, as implemented by the lmerTest package for R. Predicate class (with nonpresuppositional as baseline), complement type (with bare as baseline) and extraction were entered as fixed effects and a maximum random effects structure that allowed for convergence was used.

The silent D analysis predicts an interaction between all 3 factors such that with nonpresuppositional predicates extractions from nominalized CPs should be less acceptable compared to extractions from bare CPs (relative to the baseline condition), whereas with presuppositional predicates there should be no difference between extractions from nominalized and bare CPs. By contrast, the alternative analysis, whereby presuppositional clauses are bare CPs, predicts an interaction only between complement type and extraction (for both predicate classes). Both analyses also predict the main effect of extraction such that extraction from bare CPs should be less acceptable compared to the baseline, due to the markedness of extractions from *čto*-clauses in Russian (Khomitsevich, 2007; Morgunova, 2021b).

The predictions of the analyses can be visualized by plotting for each predicate class an interaction plot with the mean ratings for the 4 conditions (with extraction plotted on the x-axis and complement type represented by line type). The silent D analysis predicts non-parallel lines for nonpresuppositional predicates, with a steeper slope for the line corresponding to nominalized CPs but parallel lines for presuppositional predicates, whereas the CP analysis predicts non-parallel lines for both predicate classes.

[3] All sentences involved extraction of accusative or dative objects.

[4] No independent tests for presuppositionality of the predicates were done for this (and the next) experiment; the classification relied on usual treatments of their translational equivalents in the literature, e.g. it matches (Anand et al., 2019), except that *priznat'sja* 'confess' was analyzed as a (semi-)factive ((Sheehan and Hinzen, 2011)). On problems with classification of factive predicates see (Degen and Tonhauser, 2022).

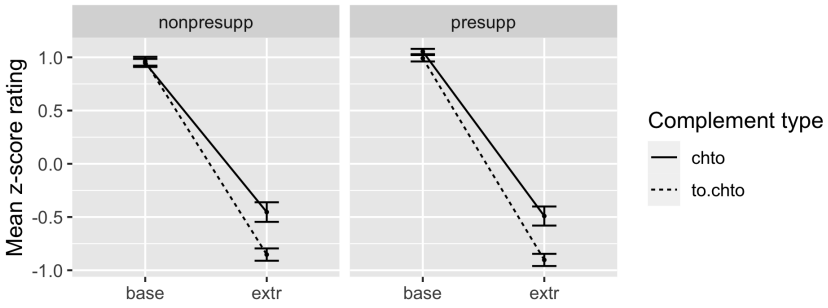


Figure 1: Condition means of Experiment 1.

2.3 Results and Discussion

The results are summarized in Figure 1. As can be seen, the plots are in line with the CP analysis. This was confirmed statistically. The model with item as a random effect showed the effect of extraction (Estimate = –1.39, SE = 0.08, $p < 0.001$) such that extractions from bare CPs were mildly unacceptable, with the ratings of –0.45/–0.49, similar to extractions with *dumat'* 'think' in the fillers (–0.42), confirming the view that *čto*-clauses are not fully transparent (Morgunova, 2021b).[5]

The model also showed an interaction between extraction and complement type (Estimate = –0.4, SE = 0.01, $p < 0.001$) such that the decrease in acceptability due to extraction was stronger for nominalized compared to bare CPs. Extractions from nominalized CPs had the ratings of –0.85/–0.9, similar to complex NP (–1.04) and selectional violations (–0.89), confirming their status as strong islands.

Other effects were not significant, including crucially the 3-way interaction ($p = 0.85$), suggesting that there was no contrast in argument extraction between presuppositional and nonpresuppositional CPs, contrary to the silent D hypothesis. This is further supported by the fact that the interaction effect for individual predicates (measured by DD-scores) did not pattern according to presuppositionality, e.g. the DD-scores for presuppositional predicates *gordit'sja* 'be proud' (0.60) and *rad* 'glad' were higher than for the nonpresuppositional predicate *nadejat'sja* 'hope' (0.18).[6] To summarize, the results are consistent with the bare CP view but do not provide support for the silent D analysis. (For an alternative interpretation in terms of the weak island version of the latter analysis see Section 4.)

3 Experiment 2

3.1 Design

The experiment was similar to Experiment 1 but tested extractions of the relative pronoun (*kotoryj* 'which'), which may lead to weaker (compared to wh-interrogative extractions) or no island effects with some island types (Sprouse et al., 2016) (cf. (Morgunova, 2021a, p.54–55)). In addition, it also tested *čtoby*-clauses, which are considered more transparent for extraction (Demina, 2021). The experiment had a $3 \times 2 \times 2$ design, as in (5)–(6), which was similar to Experiment 1, except that predicate class had 3 levels: *čto*-nonpresuppositional, *čto*-presuppositional and *čtoby*.[7]

(5) a. *Akcii, kotorye on byl uveren (v tom), čto budut aktivno pokupat',*
shares which.PL.ACC he was certain in that.PREP that will actively buy
neožidanno ruxnuli.
unexpectedly crashed
'Shares that he was certain that people would actively buy unexpectedly crashed.'

[5] "Estimate" refers to the estimated coefficient, or slope, of a predictor in the model; "SE" refers to the standard error of the estimate; "p" refers to the p-value for a coefficient estimate (using Satterthwaite approximation of degrees of freedom).

[6] DD-scores were calculated using the formula $DD = (mean_{chto[extr]} - mean_{to.chto[extr]}) - (mean_{chto[base]} - mean_{to.chto[base]})$ (Sprouse et al., 2016).

[7] All sentences involved extraction of *kotoryj* 'which' from the accusative object position.

b. *Kniga, kotoruju on gordilsja (tem), čto napisal v soavtorstve s nobelevskim*
'book which.SG.ACC he was proud that.INS that wrote in coauthorship with Nobel
laureatom, ne imela uspexa.
laureate not had success.'
'The book which he was proud that he wrote with a Nobel laureate was not successful.'

c. *Stat'ja, kotoruju on nastaival (na tom), čtoby studenty pročitali, okazalas'*
article which.SG.ACC he insisted on that.PREP that.SUBJ students read turned out
nedostupna dlja skačivanija.
unavailable for download
'The article that he insisted that students should read was not available for downloading.'

(6) a. *On byl uveren (v tom), čto èti akcii budut aktivno pokupat'.*
he was certain in that.PREP that these shares will actively buy
'He was certain that people will actively buy these shares.'

b. *On gordilsja (tem), čto napisal knigu v soavtorstve s nobelevskim laureatom.*
he was proud that.INS that wrote book in coaathorship with Nobel laureate
'He was proud that he wrote a book with a Nobel laureate'.

c. *On nastaival (na tom), čtoby studenty pročitali ètu stat'ju.*
he insisted on that.PREP that.SUBJ students read this article
'He insisted that students should read this article'.

3.2 Materials and Procedure

12 predicates, as in (7), were tested, including 4 from Experiment 1.[8] The nonpresuppositional class (with *čto*) had 4 nonfactive predicates. The presuppositional class had 3 emotive factives *žalet'* 'regret', *gordit'sja* 'proud' and *udivlën* 'surprised' and 1 response-stance verb *soglasit'sja* 'agree'.[9]

(7) a. *čto*-nonpresuppositional: *uveren* (*v* PREP) 'certain', *namekat'* (*na* ACC) 'hint', *nastaivat'* (*na* PREP) 'insist', *mečtat'* (*o* PREP) 'dream'

b. *čto*-presuppositional: *žalet'* (*o* PREP) 'regret', *gordit'sja* (INS) 'proud', *udivlën* (DAT) 'surprised', *soglasit'sja* (*s* INS) 'agree'

c. *čtoby*: *nastaivat'* (*na* PREP) 'insist', *mečtat'* (*o* PREP) 'dream', *stremi'sja* (*k* DAT) 'strive', *sledit'* (*za* INS) 'see to (it)'

As in Experiment 1, with each predicate, 4 sentence sets were constructed, distributed among 4 lists. There were 18 filler sentences (including 2 practice items): 10 acceptable (6 without extraction and 4 with relative extractions with *sčitat'* 'believe', *predpolagat'* 'suppose', *xotet'* 'want' and *prosit'* 'ask', used as baselines) and 8 unacceptable (2 with complex NP and 6 with selectional violations).

The procedure was as in Experiment 1. The experiment was completed by 49 people.

3.3 Analysis

5 participants (who rated complex NP violations higher than acceptable extractions from *čtoby*-clauses) were excluded. The analysis was similar to Experiment 1, except that predicate class was coded using 2 contrasts (for an easier comparison with Experiment 1): (A) *čto* vs. *čtoby*; and (B) nonpresuppositional vs. presuppositional (for the *čto* classes).

As in Experiment 1, the silent D analysis predicts a 3-way interaction involving contrast B. By contrast, the CP analysis predicts only a two-way interaction between extraction and complement type.

Both analyses also predict a 3-way interaction with contrast A such that extractions from bare *čtoby*-clauses should be more acceptable compared to bare *čto*-clauses (relative to the baseline), whereas extractions from nominalized *čto*- and *čtoby*-clauses should be equally unacceptable.

[8] Two predicates in the *čto*-nonpresuppositional and *čtoby* class, i.e. *nastaivat'* 'insist' and *mečtat'* 'dream', coincided in order to test the effect of *čtoby* directly.

[9] Response-stance predicates (Cattel, 1978) are classified as presuppositional, along with factives (Hegarty, 1992, a.o.). Accordingly, they receive a silent D analysis in (Kastner, 2015).

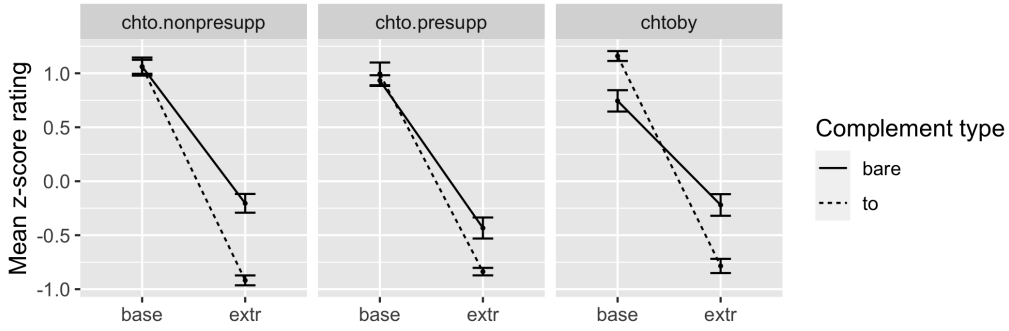


Figure 2: Condition means of Experiment 2.

3.4 Results and Discussion

The results are summarized in Figure 2. The model with item, subject and by-item complement type as random effects showed the effect of extraction (Estimate = –1.21, SE = 0.06, $p < 0.001$) and an interaction between extraction and complement type (Estimate = –0.73, SE = 0.09, $p < 0.001$) such that extractions from nominalized CPs were less acceptable compared to bare CPs. Although this interaction is visually larger for the nonpresuppositional (*čto*) class than for the presuppositional class, as expected on the silent D analysis, the 3-way interaction with contrast B was not significant (Estimate = –0.11, SE = 0.11, p = 0.31). Also, the by-predicate DD-scores did not consistently differ according to presuppositionality, e.g. the DD-scores for *žalet'* 'regret' (1.30) were higher than the DD-scores for all nonpresuppositional verbs (0.31–1.19). Thus, like in Experiment 1, the results suggest that extractions from presuppositional *čto*-clauses (–0.43) do not significantly differ from exractions from nonpresuppositional clauses (–0.20) but at the same time are significantly more acceptable than extraction from *to čto* clauses (–0.84). This is in line with the CP analysis and contrary to the silent D view.

The results also showed a 3-way interaction with contrast A (Estimate = –0.14, SE = 0.06, p = 0.02).[10] This interaction is best interpreted by fitting separate models for bare and nominalized CPs (with subject and item as random effects). The model for bare CPs showed the effect of extraction (Estimate = –1.26, SE = 0.07, $p < 0.001$), the effect of contrast A (Estimate = –0.10, SE = 0.04, p = 0.04) and an interaction (Estimate = 0.15, SE = 0.05, p = 0.002), such that although *čtoby*-clauses were rated as lower than *čto*-clauses in the baseline condition this difference disappeared in the extraction condition, suggesting that extractions from *čtoby*-clauses are more acceptable than extractions from *čto*-clauses *relative to the baseline* (cf. the steeper slope of the solid line in the leftmost panels in Figure 2), in accordance with the literature (Khomitsevich, 2007; Demina, 2021).[11] By contrast, the model for nominalized CPs showed only the effect of extraction (Estimate = –1.92, SE = 0.05, $p < 0.001$), confirming that nominalized CPs are strong islands, which is further supported by the fact that extractions from nominalized CPs had the mean ratings ranging from –0.92 to –0.78, close to complex NP violations (–1.01).

Finally, there was no clear difference between relative and wh-interrogative extractions.

4 General Discussion and Conclusion

What can we conclude from these results? The fact that argument extractions from presuppositional *čto*-clauses were only mildly unacceptable (in contrast to severely degraded extractions from *to čto*-clauses) and did not differ from extractions from nonpresuppositional clauses is inconsistent with the (strong

[10] The model also showed the effect of *to* (Estimate = 0.16, SE = 0.07, p = 0.03), the effect of contrast A (Estimate = –0.09, SE = 0.03, p = 0.01), as well as its interaction with extraction (Estimate = 0.13, SE = 0.04, p = 0.002) and with *to* (Estimate = 0.13, SE = 0.05, p = 0.02). Other effects were not significant.

[11] Interestingly, extractions from *čtoby*- and (nonpresuppositional) *čto*-clauses did not differ in *absolute* terms (–0.22 and –0.20), although the corresponding contrast did show up in the fillers (0.07 and –0.46).

island version of the) silent D analysis of presuppositional clauses in (2)b), as proposed in (Knyazev, 2022), following (Kastner, 2015). Instead, it supports the null hypothesis view that both presuppositional and nonpresuppositional clauses are bare CPs (Bondarenko, 2022, p.338–340) (see also footnote 14).

As mentioned in Section 1, an alternative way to interpret the results is to assume the weak island version of the silent D analysis, i.e. that D creates only a weak island when it merges with a CP, as suggested in (Kastner, 2015). On this view, we should not expect a contrast between presuppositional and nonpresuppositional clauses, assuming that weak islands do not block argument extractions.

There are two main issues with this alternative. First, it has been proposed that *čto*-clauses are *generally* weak islands (Bailyn, 2020), providing a potential account of the fact in Russian extractions even from nonpresuppositional clauses are marked (Morgunova, 2021b), as we also saw in the experiments.[12] Yet, if weak islandhood is to be explained by merging of a (definite) D on top of a CP, then *both* presuppositional and nonpresuppositional clauses should have a silent D, unlike in (Kastner, 2015).[13]

The second, and more important, issue is that Kastner's view that silent D + CP creates a weak island depends on his assumption that *overt* D creates a strong island by virtue of having the structure with a null N (D + N + CP), as in complex NP island (see footnote 1). However, there is convincing evidence that overtly nominalized (*to čto*) clauses in Russian have the structure D + CP, with no null N (Knyazev, 2022; Bondarenko, 2022). Yet, if silent D is associated with the same structure as overt D, we should normally expect it to similarly create a strong island, contrary to Kastner's view.

This is indeed what (Bondarenko, 2022, p.328) proposes, deriving the strong islandhood of D + CP from Anti-Locality (see references therein). Evidence for this view comes from verbs like *ob"jasnjat'* 'explain', *kommentirovat'* 'comment' and others, which are ambiguous between the presuppositional ('CP = fact explained/commented on') and the nonpresuppositional reading ('CP = content of explanation/comment'). Bondarenko argues that the presuppositional reading corresponds to the structure with a (possibly silent) D, whereas the nonpresuppositional reading corresponds to bare CP. Crucially, the presuppositional reading categorically blocks extraction regardless of the overtness of D, as in (8), supporting the view that D + CP creates a strong island.[14]

(8) **Kogo*$_1$ *Lena argumentirovala* [$\emptyset_{\text{D}}$ / *to čto Zenit legko odoleet* t_1]*?*
who.ACC Lena argued that.ACC that Zenit easily will win
'Who did Lena argue (for the position) that Zenit will easily defeat?' (Bondarenko, 2022, p. 326–327)

A potential objection to this argument is that overtness of D may sometimes matter for islandhood, e.g. in the case of subjunctive clauses with factive verbs under negation, where extraction is blocked only by overt but crucially *not* silent D, as in (9) (Bondarenko, 2022, p.329), suggesting that non-overtness of D may obviate Anti-Locality (Erlewine, 2016).[15] Something similar might be going on with presuppositional clauses studied in this paper.

(9) *Kogo*$_1$ *Katja ne pomnit (*takogo / *togo), čtoby Ira priglašala* t_1*?*
who.ACC Katya not remembers such.GEN that.GEN that.SUBJ Ira invited
'Who does Katya not remember Ira inviting?' (Bondarenko, 2022, p. 329)

To conclude, while the present experimental results do not necessarily falsify the silent D analysis of presuppositional clauses, they provide no specific evidence for it. Thus, to the extent that the burden of proof is on the proponents of silent D, the CP analysis seems preferable. However, further data, particularly on adjunct extractions, are ultimately needed to decide between the two alternatives.

[12] But see (Demina, 2021), which did not find a contrast between argument and adjunct extractions from *čto*-clauses in experimental data (as would be expected on their weak island status).

[13] Incidentally, this (across-the-board) version of the silent D analysis was proposed in (Knyazev, 2016).

[14] On Bondarenko's view, the D + CP structure depends on whether the clause is a true argument (as opposed to a modifier, corresponding to bare CP), rather than to presuppositionality per se. This allows her to maintain the view that presuppositional clauses of oblique/PP-taking verbs like *žalet'* 'regret'/*gordit'sja'sja* 'be proud' are bare CPs required by her treatment of silent D is restricted to the *accusative* position—provided they can be analyzed as modifiers (Bondarenko, 2022, p.338–340). While she does not discuss extraction with the latter predicates, the present results can be taken to support the CP analysis for them.

[15] Such clauses are assumed to be DPs based on independent semantic considerations (Bondarenko, 2022).

Acknowledgements

This research is supported by Russian Science Foundation, RSF project 22-18-00037 realized at Lomonosov Moscow State University, https://rscf.ru/en/project/22-18-00037/, whose support is gratefully acknowledged. I also thank three anonymous reviewers for their helpful comments.

References

Pranav Anand, Jane Grimshaw, and Valentine Hacquard. 2019. Sentence embedding predicates, factivity and subjects. // C Condoravdi and Tracy Holloway King, *Tokens of meaning: Papers in honor of Lauri Karttunen*, P 279–308.

John Frederick Bailyn. 2020. The scrambling paradox. *Linguistic Inquiry*, 51(4):635–669.

Galen Basse. 2008. Factive complements as defective phases. // *Proceedings of the 27th West Coast Conference on Formal Linguistics*, P 54–62. Somerville, MA: Cascadilla Proceedings Project Massachussetts.

Tatiana Igorevna Bondarenko. 2022. *Anatomy of an attitude*. Ph.D. thesis, MIT.

Ray Cattel. 1978. On the source of interrogative adverbs. *Language*, 54:61–77.

William Davies and Stanley Dubinsky. 2003. On extraction from NPs. *Natural Language & Linguistic Theory*, 21:1–37.

Judith Degen and Judith Tonhauser. 2022. Are there factive predicates? An empirical investigation. *Language*, 98(3):552–591.

Julia Demina. 2021. Asimmetrii vydviženija argumentov i ad"junktov. // E. Lyutikova and A. Gerasimova, *Russkie ostrova v svete èksperimental'nyx dannyx. Kollektivnaja monografija*. Buki-Vedi, Moscow.

Kajsa Djärv. 2019. *Factive and assertive attitude reports*. Ph.D. thesis, University of Pennsylvania.

Michael Yoshitaka Erlewine. 2016. Anti-locality and optimality in Kaqchikel Agent Focus. *Natural Language & Linguistic Theory*, 34:429–479.

Liliane Haegeman. 2012. *Adverbial Clauses, Main Clause Phenomena, and Composition of the Left Periphery: The Cartography of Syntactic Structures*. Oxford: Oxford University Press.

Michael Vincent Hegarty. 1992. *Adjunct extraction and chain configurations*. Ph.D. thesis, MIT.

Martin Honcoop. 1998. *Dynamic excursions on weak islands*. Ph.D. thesis, Leiden University.

Itamar Kastner. 2015. Factivity mirrors interpretation: The selectional requirements of presuppositional verbs. *Lingua*, 164:156–188.

Olga Khomitsevich. 2007. *Dependencies across Phases: From sequence of tense to restrictions on movement*. Ph.D. thesis, Utrecht University.

Carol Kiparsky and Paul Kiparsky. 1970. Fact. // Manfred Bierwisch and Karl Erich Heidolph, *Progress in Linguistics*, P 143–173. The Hague: Mouton.

Mikhail Knyazev. 2016. *Licensing clausal complements: The case of Russian čto-clauses*. Ph.D. thesis, Utrecht University.

Mikhail Knyazev. 2022. Spanning complement-taking verbs and spanning complementizers: On the realization of presuppositional clauses. *Journal of Linguistics*, 58(4):887–899.

Irina M. Kobozeva. 2013. Uslovija upotreblenija «to» pered pridatočnym iz"jasnitel'nym s sojuzom «čto». // Inkova O., *Du mot au texte. Études slavo-romanes*. Bern: Peter Lang.

Ekaterina Lyutikova. 2010. K voprosu o kategorial'nom statuse imennyx grupp v russkom jazyke. *Vestnik MGU, ser. 9 (Filologija)*, (6):36–76.

Ekaterina Morgunova. 2021a. Ostrovnye konstrukcii v russkom jazyke. // E. Lyutikova and A. Gerasimova, *Russkie ostrova v svete èksperimental'nyx dannyx. Kollektivnaja monografija*. Buki-Vedi, Moscow.

Ekaterina Morgunova. 2021b. Ostrovnye svojstva pridatočnyx iz"jasnitel'nyx s sojuzom čto. // E. Lyutikova and A. Gerasimova, *Russkie ostrova v svete èksperimental'nyx dannyx. Kollektivnaja monografija*. Buki-Vedi, Moscow.

Asya Pereltsvaig. 2007. The universality of DP: A view from Russian. *Studia linguistica*, 61(1):59–94.

Michelle Sheehan and Wolfram Hinzen. 2011. Moving towards the edge. *Linguistic Analysis*, 37(3-4):405–458.

Ivy Sichel. 2018. Anatomy of a counterexample: Extraction from relative clauses. *Linguistic Inquiry*, 49(2):335–378.

Jon Sprouse, Ivano Caponigro, Ciro Greco, and Carlo Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. *Natural Language & Linguistic Theory*, 34:307–344.

Collaborative constructions in Russian conversations: A multichannel perspective

Korotaev N. A.
Institute of Linguistics RAS / Moscow
Russian State University for the Humanities / Moscow
n_korotaev@hotmail.com

Abstract

The talk provides a multichannel description of how interlocutors co-construct utterances in conversation. Using data from the "Russian Pears Chats & Stories", I propose for a tripartite sequential scheme of collaborative constructions. When the scheme is fully realized, its first step not only includes the initial component of the construction, but also presupposes that the first participant makes a request for a co-operative action; the final component of the construction is provided by the second participant during the second step; while the third step consists of the first participant's reaction. On each step, the participants combine vocal and non-vocal resources to achieve their goals. In some cases, non-vocal phenomena provide an essential clue to what is actually happening during co-construction, including whether the participants act in a truly co-operative manner. I distinguish between three types of communicative patterns that may take place during co-construction: "Requested Cooperation", "Unplanned Cooperation", and "Non-realized Interaction". The data suggest that these types can be influenced by the way the knowledge of the discussed events is distributed among the participants.

Keywords: shared syntax; multichannel communication; conversation
DOI: 10.28995/2075-7182-2023-22-254-266

Мультиканальное взаимодействие при совместном построении синтаксических конструкций в диалоге

Коротаев Н. А.
Институт языкознания РАН / Москва
РГГУ / Москва
n_korotaev@hotmail.com

Аннотация

В докладе представлен мультиканальный подход к анализу случаев совместного построения синтаксических конструкций в диалоге. На материале корпуса «Рассказы и разговоры о грушах» предлагается общая схема коммуникативного обмена, содержащая три шага: запрос на совместное действие со стороны первого участника, включающий в себя начальный компонент конструкции; завершение конструкции, реализуемое вторым участником; реакцию первого участника на действия второго участника. На каждом шаге, помимо собственно речевой составляющей, участники также опираются на невокальные ресурсы, учет которых в ряде случаев позволяет точнее определить характер коммуникативной ситуации — в том числе, насколько успешно участники сотрудничают, осуществляя совместное действие. В зависимости от степени полноты реализации общей схемы выделяется три типа коммуникативных ситуаций, приводящих к возникновению совместного построения. Высказывается предположение, что на относительную частотность этих типов может оказывать влияние распределение между участниками доступа к содержанию обсуждаемых в диалоге событий.

Ключевые слова: совместный синтаксис; мультиканальная коммуникация; диалог

1 К постановке задачи

Совместное построение (*co-construction;* в более узком смысле — *collaborative completion*) — диалогическое явление, заключающееся в том, что один из участников разговора начинает реализацию некоторой синтаксической структуры, а другой ее заканчивает ([Ono, Thompson 1996; Helasvuo 2004] и др.). Наиболее естественная среда возникновения совместных построений — это неподготовленный устный диалог. Иллюстрацией этого явления может служить пример (1). Начальный компонент — подлежащее и часть глагольной группы — реализует одна говорящая, после чего вступает второй говорящий, который достраивает имеющуюся структуру до законченного сложноподчиненного предложения.

(1) Pears04: «шляпу уронил»[1]

1054.26	R-vE260	**А-а /шляпу он уронил-л %**
1055.78	C-vE233	**% ещё \до того как \падал.**
1057.58		(0.14)
1057.17	R-vE261	\–А-а!

Речевые характеристики совместного построения изучены достаточно подробно, в первую очередь, в рамках Анализа бытового диалога. Благодаря исследованиям, проведенным, в частности, в [Sacks 1992; Szczepek 2000a, b; Lerner 2004; Clancy, McCarthy 2014], к настоящему моменту накоплен значительный массив сведений о синтаксической, просодической и коммуникативной организации совместного построения. Полезный обзор основных англоязычных работ представлен в [King 2018].

Фундаментальное исследование совместного построения в русском языке было предпринято в работе [Гренобль 2008]. На материале записей радиопередач Л. Гренобль обосновала разграничение случаев совместного построения на расширения и завершения. Это противопоставление основано на введенном еще в классической статье [Sacks et al. 1974] понятии точки перехода (*transition relevance place*) — месте, в котором завершается текущая реплика и может произойти смена говорящего. При расширении второй участник совместного построения вступает после точки перехода, при завершении — еще до нее; именно случай завершения представлен в примере (1) выше. Еще одним важным результатом Гренобль стал вывод о неравномерной частотности случаев совместного построения: в одних диалогах они встречаются достаточно часто, в других — редко или вообще никогда. В число факторов, влияющих на частотность совместных построений, входят индивидуальные особенности говорящих. Так, в [Оленикова, Федорова 2020] было показано, что совместные построения в целом более частотны в диалогах с заикающимися.

В то же время значительно меньше известно о том, как, реализуя совместное построение, говорящие используют неречевые коммуникативные ресурсы: жесты, направления взгляда, кивки и проч. Этот вопрос включает анализ совместного построения в контекст мультимодальных (или мультиканальных) исследований — направления, рассматривающего речевое (вокальное) и неречевое (невокальное) поведение как (относительно) равноправные

[1] Все примеры в тексте доклада приводятся в транскрипционном формате, используемом в корпусе «Рассказы и разговоры о грушах» [Kibrik et al. 2020] и частично модифицированном для целей настоящего исследования. В заголовке примера указывается кодовый номер записи и короткое неформальное обозначение коммуникативного эпизода. Нумерованные строки в транскрипте соответствуют элементарным дискурсивным единицам (ЭДЕ). Литеры N, C и R в номере ЭДЕ указывают на закрепленные роли участников (подробнее см. раздел 2 ниже); слева от номера указывается время начала произнесения ЭДЕ. В скобках приводится продолжительность абсолютных и заполненных пауз; для обозначения наложения реплик используются квадратные скобки и горизонтальное выравнивание. При помощи слешей и стрелок размечены движения частоты основного тона в акцентированных словоформах. Кроме того, компоненты совместно реализуемых конструкций дополнительно выделены полужирным шрифтом; все реплики первого участника набраны синим цветом, все реплики второго участника — красным цветом. Символ % в конце строки указывает на завершение первого компонента конструкции при наличии запроса на совместное действие (см. ниже разделы 3 и 4.1); этот же символ в начале строки маркирует начало финального компонента конструкции, произносимого вторым участником.

составляющие при реализации коммуникативного замысла [Müller et al. eds. 2013/2014; Church et al. 2017; Кибрик 2018]. Мультиканальная перспектива совместного построения рассматривается, в частности, в работах [Bolden 2003; Hutchins, Nomura 2011; Iwasaki 2011; Kalkoff, Dressel 2019; Song, Vukadinovich 2021]: в них анализируются отдельные образцы невокальных явлений, значимых при совместном построении, и их связь с вокальными. Для русского языка подробный анализ совместного построения в мультиканальной перспективе, насколько мне известно, пока не проводился. В настоящем докладе я надеюсь частично восполнить этот пробел, привлекая материал корпуса «Рассказы и разговоры о грушах». Основная задача исследования — выполнить описание случаев совместного построения в русском диалоге, адекватно учитывающее вклад невокальных каналов коммуникации, и предложить классификацию типов мультиканального взаимодействия, приводящего к возникновению совместно реализованных конструкций[2].

Структура работы такова. В разделе 2 будет кратко представлен используемый корпус и обоснован его выбор в качестве материала исследования. В разделе 3 будет приведена общая схема коммуникативного взаимодействия, релевантная при описании конкретных случаев совместного построения. В разделе 4 будут рассмотрены типы совместного построения, обнаруженные в исследованном материале. Раздел 5 содержит обсуждение полученных результатов; раздел 6 — заключение.

2 Материал

Корпус «Рассказы и разговоры о грушах» (далее — RUPEX; https://multidiscourse.ru/) — это аннотированная коллекция аудио- и видеозаписей, организованных согласно общему дизайну. В каждой записи задействованы три активных участника: Рассказчик (в транскриптах приводимых в тексте примеров и на скриншотах обозначается при помощи литеры N), Комментатор (C) и Пересказчик (R). До начала записи N и C смотрят «Фильм о грушах» [Chafe ed. 1980], после чего к ним присоединяется R — и все трое, следуя полученным инструкциям, последовательно реализуют три этапа: рассказ (N подробно рассказывает содержание фильма, обращаясь к R), разговор (все трое обсуждают содержание фильма) и пересказ (R, основываясь на информации, полученной на предыдущих двух этапах, рассказывает содержание фильма четвертому участнику, который не смотрел фильм и не присутствовал до этого в комнате). Подробнее о содержательных и технических характеристиках записей корпуса см. [Kibrik, Fedorova 2018]. RUPEX уже неоднократно использовался для проведения исследований устной речи и мультиканального поведения; см., в частности, [Litvinenko et al. 2018; Кибрик и др. 2019; Korotaev et al. 2020].

В настоящем исследовании анализируются этапы разговора восьми записей корпуса. Ценность этого материала для изучения совместного построения определяется, на мой взгляд, двумя соображениями. Во-первых, единый дизайн записей позволяет сопоставлять то, как разные участники реализуют сходные коммуникативные задачи — в том числе, насколько часто они склонны прибегать к совместному построению. Во-вторых, опять-таки благодаря общему дизайну записей, можно проследить, каким образом на характер совместного построения влияет распределение между участниками доступа к содержанию обсуждаемого в диалогах фильма. Этот фактор представляется достаточно важным, поскольку в имеющихся исследованиях преимущественно анализируется материал, однородный с этой точки зрения. Например, в [Kalkoff, Dressel 2019], где показано, как носители испанского языка координируют свои вокальные и невокальные действия, завершая реплики друг друга в контексте совместного рассказа о личном опыте, у участников, очевидно, есть равноправный доступ к содержанию историй. Противоположная ситуация представлена в работе [Hutchins, Nomura 2011], в которой рассматриваются диалоги между пилотами, проходящими обучение на авиасимуляторе, и инструкторами: здесь участники обладают заведомо неравноправным доступом к обсуждаемой информации. В свою очередь, в диалогических этапах записей RUPEX представлены обе эти

[2] Подчеркну, что, согласно приведенным выше определениям, ситуация совместного построения задается исключительно формальными свойствами *языковых выражений*, т. е. исходной точкой для анализа все же остается речевая составляющая мультиканального взаимодействия. Это необходимо иметь в виду при обсуждении полученных результатов, см. раздел 5.

возможности: когда в совместном построении задействованы Рассказчик и Комментатор, речь идет о равноправном доступе (оба участника одновременно смотрели фильм); когда же одним из участников, реализующих совместное построение, выступает Пересказчик, второй участник обычно лучше осведомлен о содержании фильма.

При анализе обнаруженных в корпусе случаев совместного построения я частично опирался на ранее уже выполненную разметку речи [Kibrik et al. 2020], мануальной жестикуляции [Litvinenko et al. 2018] и глазодвигательного поведения [Fedorova 2021] участников. Однако интерпретация зафиксированных в разметке жестов и фиксаций взгляда по большей части была выполнена отдельно — преимущественно на перцептивных основаниях. Безусловно, такой подход страдает от недостатка объективных критериев, поэтому далее речь пойдет лишь о качественных наблюдениях, а не о количественных результатах.

3 Общая схема коммуникативного обмена

Осуществляя совместное построение, участники диалога демонстрируют сложным образом скоординированное мультиканальное поведение. Для обобщенного описания этого поведения я предлагаю использовать схему коммуникативного обмена, представленную на рис. 1. Важной особенностью этой схемы является то, что в ней, как это предлагается и в ряде других работ (см., в частности, [Mondada 1999; Lerner 2004]), выделяется не два, а три дискурсивных шага. Помимо двух очевидных шагов (соответствующих начальному и финальному компонентам совместно реализуемой конструкции), значимым для интерпретации общего характера коммуникативной ситуации является и третий шаг — реакция участника 1 на вклад участника 2. Три шага реализуются последовательно, при этом характер временны́х отношений между ними не является жестким: шаги могут быть разделены паузами, следовать встык или накладываться друг на друга.

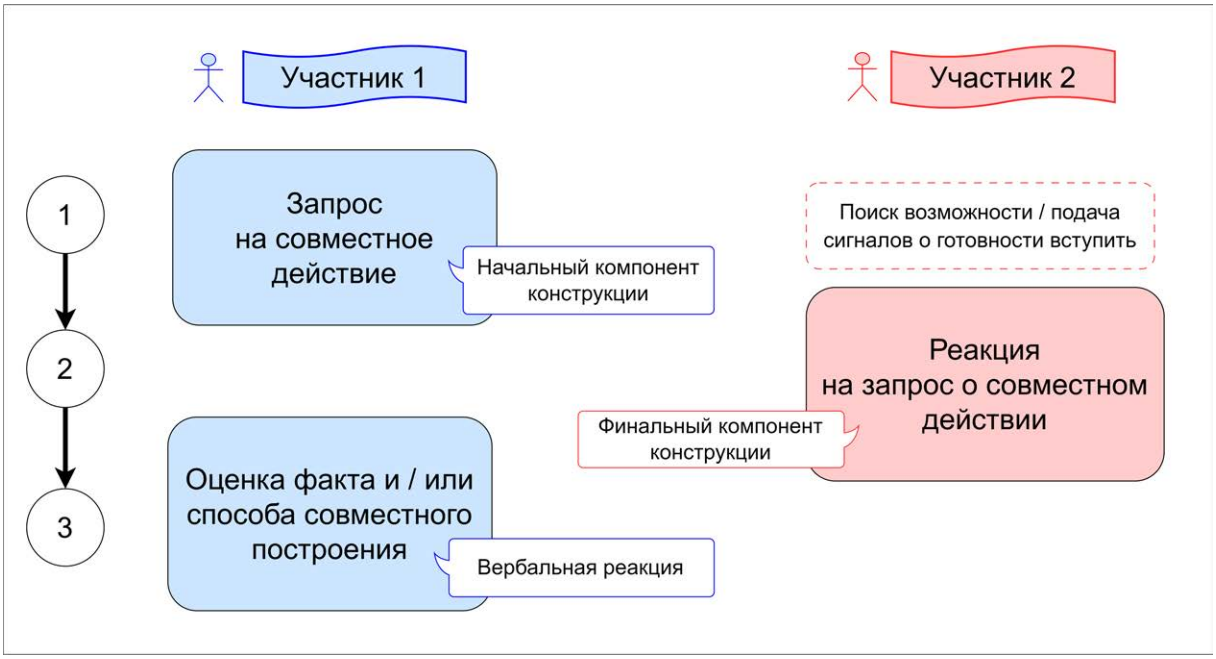


Рисунок 1. Схема мультиканального коммуникативного обмена при совместном построении. Закрашенными прямоугольниками обозначены коммуникативные действия, незакрашенными выносками — их речевые реализации

Реализацию схемы, представленной на рис. 1, можно проследить на примере (2).

(2) Pears37: «этот мужчина»

370.97	C-vE075	и (0.12) проходили как раз –↓мимо-о этого /дерева,
371.20	C-vE076	(**Где был-л** (ə 0.27) (ш 0.20) (0.29) [(0.21) этот \мужчина,
374.55	N-vE217	**% [\мужчина!**
375.50	N-vE218	[\Да!,
375.51	C-vE077	[\да.
375.82	N-vE219	да [–да.
376.02	C-vE078	[\Груши который собирал.)
377.12	C-vN019	(ц 0.28)
377.40	C-vE079	он /спустился,

Как видно из транскрипта, совместное построение происходит в строках C-vE076 и N-vE217. Общий контекст данного коммуникативного эпизода таков: Пересказчица, выясняя необходимые подробности сюжета, просит Комментатора и Рассказчицу уточнить, куда именно направились персонажи фильма. При этом она предполагает, что они двигались в сторону дерева, и рассчитывает получить подтверждение или опровержение своей гипотезы. Первой на этот запрос реагирует Комментатор. Она поворачивает голову в сторону Пересказчицы, устанавливает с ней зрительный контакт и начинает описывать релевантную последовательность событий. В свою очередь Рассказчица внимательно смотрит на Комментатора. Данная диспозиция видна на рис. 2а, на котором зафиксирован момент начала произнесения Комментатором словоформы *был*.

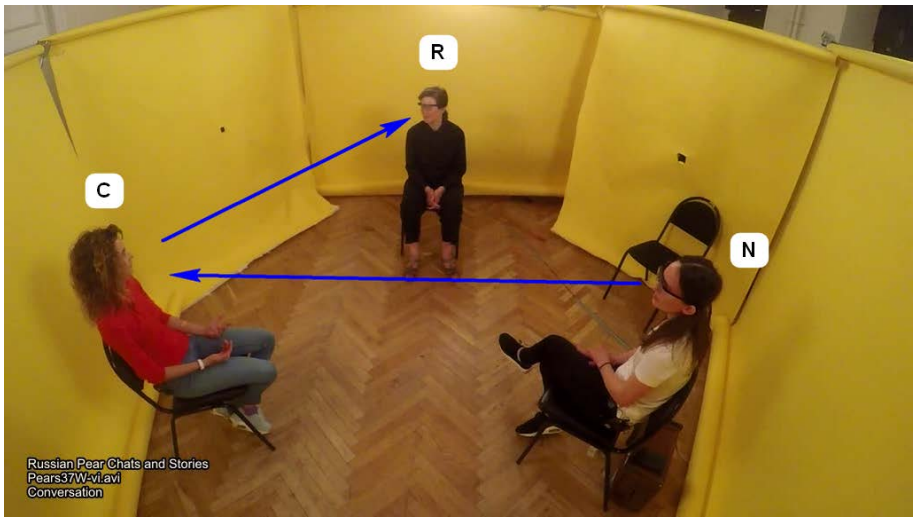


Рисунок 2а. Направление взгляда Комментатора (C) и Рассказчицы (N) на момент начала произнесения Комментатором словоформы *был* в строке C-vE076 фрагмента (2)

Примерно в этот же момент Комментатор начинает испытывать затруднения при реализации своего замысла. Это проявляется в серии хезитационных речевых сигналов: словоформа *был* произносится с нефонологическим удлинением финального согласного, после чего следуют заполненные паузы хезитации. Одновременно Комментатор выполняет поисковый жест кистью левой руки и переводит взгляд с Пересказчицы на Рассказчицу, формируя таким образом невокальный запрос на совместное действие; см. рис. 2б. Эта последовательность событий соответствует ***первому шагу*** общей схемы совместного построения[3].

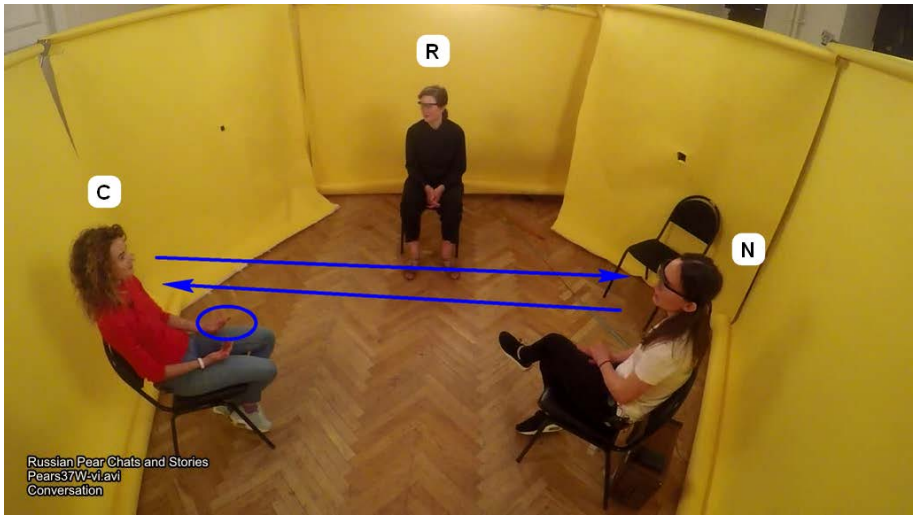


Рисунок 2б. Направление взгляда Комментатора (C) и Рассказчицы (N); мануальная жестикуляция Комментатора во время хезитационной паузы в строке C-vE076 фрагмента (2)

На ***втором шаге*** Рассказчица, реагируя на запрос Комментатора, предлагает свой вариант завершения начатой клаузы. Свой вклад она сопровождает кивками головой, призванными, по всей видимости, дать понять, что она в целом согласна с линией повествования, осуществляемой Комментатором. Примечательно, что одновременно с этим Комментатор, преодолев свои затруднения, также завершает начатую ей конструкцию, причем практически таким же образом, что и Рассказчица. При этом, несмотря на возникающее наложение, участницы не проявляют сколько-либо конкурентного поведения. Напротив, они продолжают смотреть друг на друга и подавать взаимные сигналы подтверждения. Из этих сигналов состоит ***третий шаг*** коммуникативного обмена, в рамках которого участницы реализуют однотипные подтверждающие реплики. Произнося свое *да*, Комментатор выражает согласие как с самим фактом совместного построения (которое она оценивает как полезную помощь со стороны Рассказчицы), так и со способом завершения, выбранным собеседницей. Попутно она завершает хезитационную жестикуляцию и возвращает руки в нейтральное положение на коленях. Расценив, что коммуникативный обмен с Рассказчицей завершен, Комментатор вновь переводит взгляд на Пересказчицу и продолжает текущее описание; см. рис. 2в и строку C-vE078 транскрипта.

[3] Ключевая роль смены направления взгляда при инициировании совместного построения, в частости, отмечена в [Bolden 2003: 203-208]: автор показывает, что перевод взгляда с окружающей обстановки на собеседника регулярно считывается как приглашение завершить начатую реплику. Подробнее о распределении зрительного внимания на материале RUPEX см. [Федорова 2021].

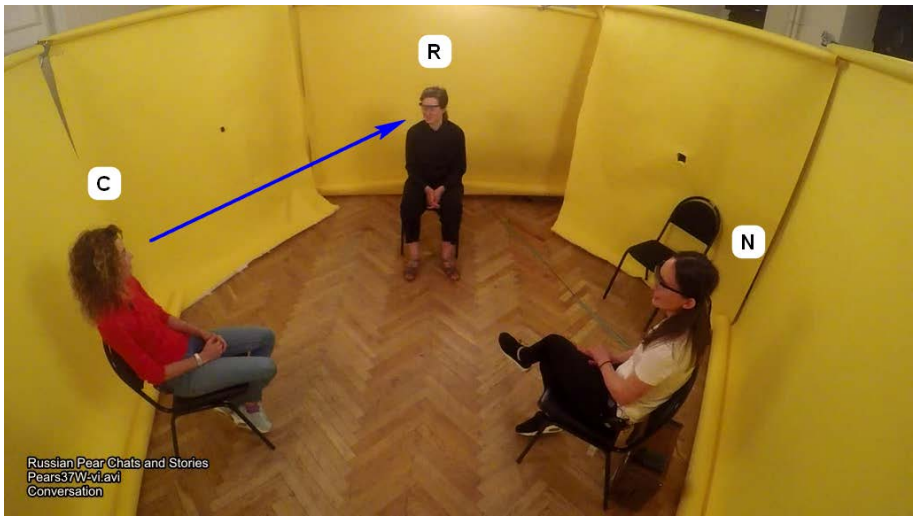


Рисунок 2в. Направление взгляда Комментатора (C) во время произнесения словоформы *который* в строке C-vE078 фрагмента (2)

4 Типы взаимодействия при совместном построении

Как следует из приведенного выше анализа примера (2), при совместном построении участники диалога могут координировать свои усилия, пошагово реализуя согласованные вокальные и невокальные действия. Именно такое согласованное поведение зафиксировано в схеме коммуникативного обмена на рис. 1. Однако далеко не во всех случаях совместного построения эта схема реализуется в полной мере: некоторые ее элементы могут опускаться, что приводит к ощутимому изменению общего характера коммуникативной ситуации. Степень полноты реализации общей схемы можно положить в основу классификации типов мультиканального взаимодействия при совместном построении. В данном разделе будут кратко охарактеризованы три таких типа, достаточно надежно выделяемые на материале RUPEX.

4.1 «Запрошенное сотрудничество»

При «Запрошенном сотрудничестве» участники полноценно реализуют все три шага общей схемы. Как это может происходить, уже было подробно продемонстрировано для фрагмента (2). Отмечу, что в этом фрагменте задействованы две участницы (Комментатор и Рассказчица), обладающие равным доступом к содержанию обсуждаемого фильма. Основанием для запроса на совместное действие здесь становится локальная неуверенность одной из них в способе описания известной ей ситуации. Другая стандартная для этого типа конфигурация возникает, когда первым участником является Пересказчик, полагающий, что второй участник (Рассказчик или Комментатор) обладает бóльшим доступом к содержанию фильма. В таких случаях совместное построение функционально сближается с вопросно-ответными парами: начальный компонент конструкции содержит в себе не просто приглашение к совместному действию, но и запрос на получение информации.

Именно так обстоит дело в примере (1), приведенном выше в разделе 1. Инициатором коммуникативного обмена здесь выступает Пересказчица, которая рассчитывает уточнить последовательность происходящих в фильме событий. Для этого она произносит первый компонент конструкции с заметным восходящим (тематическим) акцентом на словоформе *шляпу*, удлиняет финальный согласный глагольной словоформы *уронил* и после этого отчетливо прерывает вокализацию. Перерыву в вокализации сопутствует и замершая жестикуляция: Пересказчица поднимает правую руку и удерживает ее в этом маркированном положении на протяжении всего коммуникативного обмена. Для описания первого шага также существенно, что, формируя запрос на совместное действие, Пересказчица переводит взгляд на Рассказчицу, которую она, видимо, считает более «авторитетной» участницей записи; см. рис. 3а.

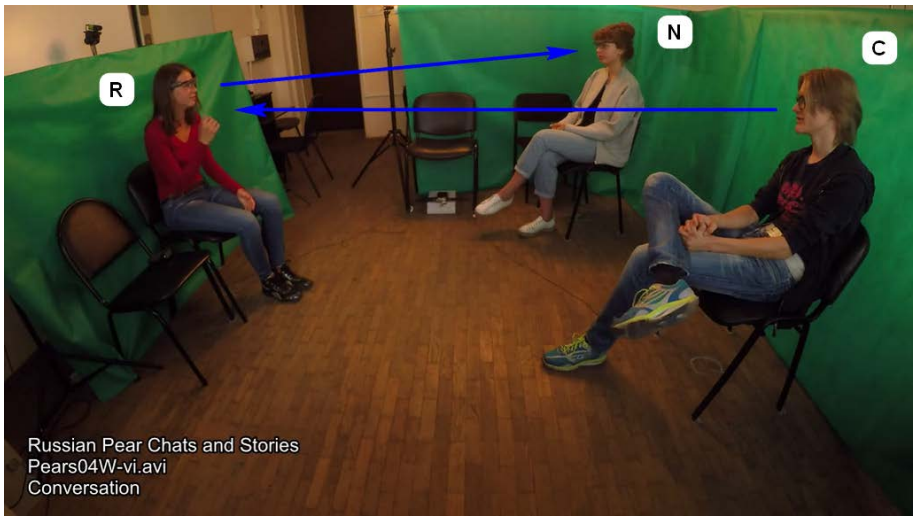


Рисунок 3а. Направление взгляда Пересказчицы (R) и Комментатора (C) в момент произнесения Пересказчицей слова *уронил* в строке R-vE260 фрагмента (1)

Однако ответственность за реализацию второго шага принимает на себя Комментатор, во время первого шага продолжающий смотреть на Пересказчицу. Именно он достраивает конструкцию до полной сложноподчиненной клаузы, и как только он начинает это делать, Пересказчица переводит взгляд на него, по-прежнему не опуская руку; см. левый кадр на рис. 3б. Получив запрошенную информацию, Пересказчица подает вокальный сигнал принятия и опускает руку, обозначая таким образом завершение коммуникативного обмена; см. правый кадр на рис. 3б.

Рисунок 3б. Невокальные действия Пересказчицы (R) во фрагменте (1): на левом кадре — направление взгляда при произнесении Комментатором (C) словоформы *падал* (строка C-vE233); на правом кадре — движение руки вниз при произнесении Пересказчицей строки R-vE261

4.2 «Незапланированное сотрудничество»

Если в примерах (1) и (2) первый участник при помощи вокальных и невокальных сигналов формирует запрос на совместное действие, то в некоторых других случаях второй участник может выступить со своим продолжением начатой конструкции и без такого запроса. Реакция первого участника на подобное развитие событий может быть разной. В том случае если он в итоге оценивает действие второго участника как допустимое или даже желательное, реализуется тип «Незапланированное сотрудничество». Стандартное распределение ролей при таком типе взаимодействия — это равноправный доступ к содержанию фильма (участвуют Рассказчик и Комментатор) или меньший доступ у второго участника (Пересказчика). Образец второй из этих двух конфигураций представлен во фрагменте (3).

(3) Pears23: «третья корзина»

517.20	N-vE262	**Когда мужчина /спустился,**
518.33	N-vE263	[увидел одну /полную,
518.35	R-vE075	**% [он \третью напол= ~**
519.33	N-vE264	(ʔ 0.11) ʔа /другую \пустую.
520.31	N-vE265	/\Нет,
520.56	N-vE266	он не \успел наполнить,
521.66	N-vE267	он просто /спустился, ≈≈

Рассказчица передает последовательность событий, обращаясь к Пересказчику. Несмотря на то что она смотрит на своего собеседника, это сложно интерпретировать как запрос на совместное действие. Произнеся препозитивное придаточное в строке N-vE262, она сразу же приступает к реализации запланированной главной клаузы (строки N-vE263–264). При этом она сопровождает речь серией изобразительных жестов. Тем временем Пересказчик, улучив подходящую возможность, вступает со своим вариантом продолжения конструкции. С функциональной точки зрения он высказывает догадку, рассчитывая, что Рассказчица оценит ее правильность. Это действие он подкрепляет прагматическим жестом ладони, направленным в сторону Рассказчицы; см. рис. 4.

Рисунок 4. Невокальные действия Пересказчика (R) и Рассказчицы (N) во время произнесения строк R-vE075 и N-vE263 фрагмента (3). Пересказчик реализует прагматический жест, Рассказчица продолжает ранее начатую серию изобразительных жестов

В строке N-vE264 Рассказчица завершает реализацию своего исходного плана и уже после этого реагирует на догадку Пересказчика. Таким образом, действие Пересказчика, которое не было обусловлено запросом со стороны Рассказчицы, в итоге все же учитывается Рассказчицей как релевантное для текущего отрезка диалога.

4.3 «Неслучившееся взаимодействие»

Вступление участника 2 без запроса может вызвать и менее кооперативную реакцию со стороны участника 1. В частности, видимой реакции может не последовать вовсе. В таких случаях можно говорить о ситуации «Неслучившегося взаимодействия». В RUPEX этот тип преимущественно встречается, когда Рассказчик и Комментатор параллельно обращаются к Пересказчику. Так,

в примере (4) Рассказчица, обращаясь к Пересказчику, передает последовательность событий, приведших к краже груш.

(4) Pears16: «багажники»

На первом шаге она описывает устройство велосипеда одного из персонажей и в какой-то

748.30	N-vE378	Потом такой /о-опс!,
749.24	N-vE379	/берёт,
749.82	N-vN035	(ц 0.46)
750.28	N-vE380	(ʔ 0.34) и \ставит вот ==
751.54	N-vE381	/знаешь как[ʔ],
752.11		(0.47)
752.58	N-vE382	**\раньше были на задних /сиденьях —**
754.58		(0.14)
754.72	N-vE383	(\велосипедов),
755.37	C-vE140	**% [\багажники.**
755.42	N-vE384	— [\такие вот || (0.24) [\сидушки.
756.28	C-vE141	[Только /тут с-с= [(0.27) \=переди багажник.
756.83	N-vE385	[Вот \такие || \металлические как бы.

момент сталкивается с трудностями при упоминании багажника, который располагался на этом велосипеде не сзади (что для нее более привычно), а спереди. Комментатор в это время смотрит на Рассказчицу, но Рассказчица не подает каких-либо сигналов, которые можно было бы интерпретировать как запрос на совместное действие; см. левый кадр на рис. 5. На втором шаге Комментатор поворачивается к Пересказчику и, воспользовавшись тем, что Рассказчица временно отложила завершение клаузы, начатой в строке N-vE382, предлагает свое завершение — *багажники*. Рассказчица не обращает внимания на действие Комментатора и продолжает свое развитие. Обе они при этом смотрят на Пересказчика и активно жестикулируют; см. правый кадр на рис. 5. Это параллельное, конкурентное действие продолжается также на протяжении еще некоторого времени, за которое обе участницы успевают произнести еще по одной ЭДЕ; какого-либо дополнительного взаимодействия, обусловленного совместным построением, между ними не происходит.

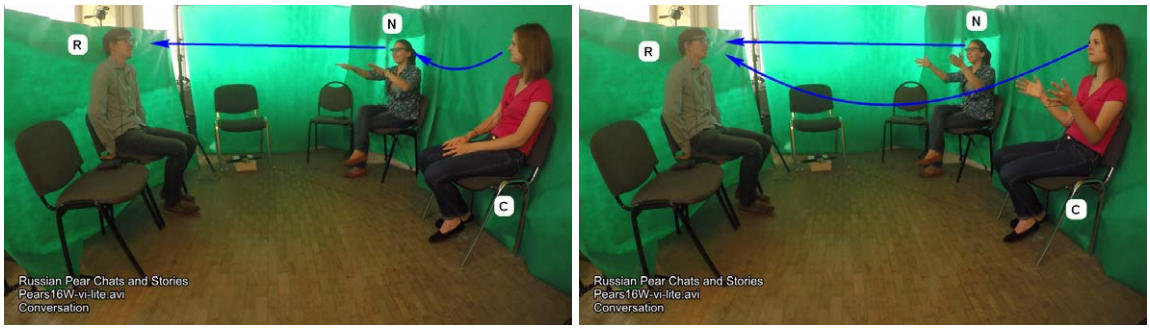


Рисунок 5. Невокальные действия Рассказчицы (N) и Комментатора (C) во фрагменте (4): слева — во время произнесения Рассказчицей словоформы *сиденьях* (строка N-vE382); справа — во время параллельного произнесения строк N-vE384 и C-vE140

5 Обсуждение

Интерпретируя полученные результаты, хочется отметить два пункта. С одной стороны, анализ случаев совместного построения достаточно убедительно демонстрирует, что учет невокальных действий участников диалога необходим для адекватного описания стоящих перед ними коммуникативных задач и способов их реализации. Так, во фрагменте (2) запрос на совместное действие со стороны участницы 1 выражен не столько посредством речевой хезитации (в конце концов, участница в результате показывает, что способна самостоятельно справиться с речевым затруднением), сколько за счет перевода взгляда с одной собеседницы на другую. И наоборот, тот факт, что обе участницы совместного построения во фрагменте (4) независимо друг от друга обращены к третьей участнице и с точки зрения глазодвигательной активности, и с точки зрения мануальной жестикуляции, ярче всего указывает на некооперативный характер этого коммуникативного эпизода. При этом речевое воплощение этих двух совместных построений обнаруживает обманчивое сходство: в обоих случаях участник 1 достраивает начатую конструкцию параллельно с участником 2.

Как отмечается в [Кибрик 2018: 71], для мультиканального дискурса характерна кофункциональность различных коммуникативных средств: при реализации своего замысла адресанту сообщения часто не важно, при помощи какого именно канала это происходит. В случае совместного построения уместно говорить не столько о кофункциональности, сколько о разнесении функций различных каналов. Поскольку техника совместного построения требует от участников при использовании вокальных ресурсов опираться на достаточно жесткую общую синтаксическую рамку [Гренобль 2008: 34-35], то ряд важных коммуникативных задач — в первую очередь, реализация запроса на совместное действие и сигнализирование о готовности принять в нем участие — становятся «естественной» сферой применения невокальных ресурсов. Далее возникает вопрос, является ли такое разнесение функций исключительной особенностью ситуации совместного построения, обусловленной ее собственно языковыми (синтаксическими) характеристиками, или же это проявление более общей диалогической тенденции. Для решения этого вопроса требуется привлечь материал других диалогических контекстов: вопросно-ответных обменов, споров и проч.

С другой стороны, различия между рассмотренными типами совместных построений можно соотнести с различиями в степени скоординированности усилий участников по выполнению совместного действия. Наряду со случаями, характеризующимися высокой степенью кооперативности на каждом шаге (примеры (1), (2)), мы можем наблюдать и ситуации, в которых участники либо согласуют свои действия «на лету» (3), либо вовсе не стремятся к видимой координации своих усилий (4). Таким образом, можно считать, что часто высказываемая в литературе идея о совместном построении как признаке кооперативного поведения в диалоге [Sacks 1992: 147; Szczepek 2000b; Kalkoff, Dressel 2019] нуждается в уточнении.

6 Заключение

Итак, в докладе представлен мультиканальный подход к описанию случаев совместного построения синтаксических конструкций в неподготовленном устном диалоге на русском языке. На материале диалогических этапов записей корпуса «Рассказы и разговоры о грушах» удается показать, ка́к участники, с одной стороны координируют vs. не координируют свои усилия, с другой стороны, распределяют имеющиеся в их распоряжении вокальные и невокальные ресурсы при реализации своих коммуникативных задач. С достаточной регулярностью в корпусе фиксируются три типа мультиканального взаимодействия, сопутствующего возникновению совместных построений. Различия между ними обусловлены тем, какие шаги из общей схемы коммуникативного обмена воплощаются в конкретном случае совместного построения. Согласно предварительным наблюдениям, на тип взаимодействия может влиять характер распределения между участниками доступа к содержанию обсуждаемого фильма. Этот качественный вывод еще нуждается в количественном подтверждении, возможном при анализе бóльшего объема данных. Другое потенциальное направление дальнейших исследований — это сопоставление выделенных типов мультиканального взаимодействия с прочими формальными параметрами варьирования при совместном построении: видом синтаксической границы между компонентами конструкции, наличием пауз / наложений, способами интонационного оформления.

Благодарности

Исследование выполнено в рамках проекта Минобрнауки «Кинетические и вокальные аспекты коммуникации: параметры варьирования». Автор также выражает благодарность анонимным рецензентам, высказавшим важные замечания к нефинальной версии текста.

Литература

[1] Bolden Galina. Multiple modalities in collaborative turn sequences // Gesture. — 2003. — Vol. 2. — P. 187–212. https://doi.org/10.1075/gest.3.2.04bol
[2] Chafe Wallace (ed.). The Pear Stories: Cognitive, cultural, and linguistic aspects of narrative production. — Norwood, NJ: Ablex, 1980.
[3] Church R. Breckinridge, Alibali Martha W., Kelly Spencer D. (eds.). Why Gesture? How the hands function in speaking, thinking and communicating. — John Benjamins, 2017.
[4] Clancy Brian, McCarthy Michael. Co-constructed turn-taking // Karin Aijmer, Christoph Rühlemann (eds.) Corpus pragmatics: A handbook. — Cambridge: Cambridge University Press, 2014. — P. 430–453.
[5] 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. — P. 213–226. https://doi.org/10.28995/2075-7182-2021-20-213-226
[6] Grenoble Lenore. Syntax and co-constructed turns in Russian dialogue [Sintaksis I sovmestnoe postroenie repliki v russkom dialoge] // Issues in Linguistics [Voprosy jazykoznanija]. — 2008. — Vol. 1. — P. 25–36.
[7] Helasvuo Marja-Liisa. Shared syntax: the grammar of co-constructions // Journal of Pragmatics. — 2004. — Vol. 36 (8). — P. 1315–1336.
[8] Hutchins Edwin, Nomura Saeko. Collaborative construction of multimodal utterances // J. Streeck, C. Goodwin, C. LeBaron (eds.) Embodied interaction: Language and body in the material world. — Cambridge; New York: Cambridge University Press, 2011. — P. 29–43.
[9] Iwasaki Shimako. The multimodal mechanics of collaborative unit construction in Japanese conversation // J. Streeck, C. Goodwin, C. LeBaron (eds.) Embodied interaction: Language and body in the material world. — Cambridge; New York: Cambridge University Press, 2011. — P. 106–120.
[10] Kalkhoff Alexander Teixeira, Dressel Dennis. Co-constructing utterances in face-to-face-interaction: A multimodal analysis of collaborative completions in spoken Spanish // Social Interaction. Video-Based Studies of Human Sociality. — 2019. — Vol. 2(2). https://doi.org/10.7146/si.v2i2.116021
[11] Kibrik Andrej A. Russian multichannel discourse. Part I. Setting up the problem [Russkij mul'tikanal'nyj diskurs. Čast' I. Postanovka problemy] // Psychological Journal [Psixologičeskij žurnal]. — 2018. — Vol. 39 (1). — P. 70–80.
[12] 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
[13] Kibrik Andrej A., Korotaev Nikolay A., Fedorova Olga V., Evdokimova Alexandra A. Unified multichannel annotation: A tool for analysing natural communication [Edinaja mul’tikanal’naja annotacija kak instrument analiza estestvennoj kommunikacii] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2019). Issue 18. — 2019. — P. 265–280.
[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] King Allie H. Collaborative completions in everyday interaction: A literature review // Studies in Applied Linguistics & TESOL. — 2018. — Vol. 18 (2). — P. 1–14.
[16] Korotaev Nikolay A., Podlesskaya Vera I., Smirnova Katerina V., Fedorova Olga V. Disfluencies in Russian spoken monologue: A distributional analysis // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference “Dialogue” (2020). Issue 19. — Moscow: RGGU, 2020. — P. 439–451. https://doi.org/10.28995/2075-7182-2020-19-454-466
[17] Lerner Gene H. Collaborative turn sequences // Gene H. Lerner (ed.) Conversation analysis: Studies from the first generation. — Philadelphia: John Benjamins, 2004. — P. 225–256. https://doi.org/10.1075/pbns.125.12ler
[18] 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.
[19] Mondada Lorenza. L'organisation séquentielle des ressources linguistiques dans l'élaboration collective des descriptions // Langage et Société. — 1999. — Vol. 89 (1). — P. 9–36. https://doi.org/10.3406/lsoc.1999.2882

[20] Müller Cornelia, Cienki Alan, Fricke Ellen et al. (eds.). Body – Language – Communication: An international handbook on multimodality in human interaction. — Berlin: De Gruyter Mouton, 2013/2014. 2 vols. https://doi.org/10.1515/9783110261318
[21] Olenikova A. V., Fedorova Olga V. Co-constructed syntactic units in dialogues with people who stutter [Sovmestnyj sintaksis v dialogax s zaikajuščimisja] // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue" (2020). Issue 19. — Moscow: RGGU, 2020. — P. 581–590. https://doi.org/10.28995/2075-7182-2020-19-596-605
[22] 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
[23] Sacks Harvey. Lectures on Conversation. — Oxford: Basil Blackwell, 1992. — Vol. 1.
[24] 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
[25] Song Zixuan, Vukadinovich Stefana. Collaborative construction of turn constructional units in responsive positions of question-answer sequences in Mandarin conversation // Chinese Language and Discourse. — 2021. — Vol. 12 (1). — P. 84–108. https://doi.org/10.1075/cld.00038.son
[26] Szczepek Beatrice. Formal aspects of collaborative productions in English conversation // InLiSt (Interaction and Linguistic Structures). — 2000. — Vol. 17. — P. 1–35.
[27] Szczepek Beatrice. Functional aspects of collaborative productions in English conversation // InLiSt (Interaction and Linguistic Structures). — 2000. — Vol. 21. — P. 1–36.

Fact-checking benchmark for the Russian Large Language Models

Anastasia Kozlova
SberDevices
anastasi2510@gmail.com

Denis Shevelev
SberDevices
reddraner@gmail.com

Alena Fenogenova
SberDevices
alenusch@gmail.com

Abstract

Modern text-generative language models are rapidly developing. They produce text of high quality and are used in many real-world applications. However, they still have several limitations, for instance, the length of the context, degeneration processes, lack of logical structure, and facts consistency. In this work, we focus on the fact-checking problem applied to the output of the generative models on classical downstream tasks, such as paraphrasing, summarization, text style transfer, etc. We define the task of internal fact-checking, set the criteria for factual consistency, and present the novel dataset for this task for the Russian language. The benchmark for internal fact-checking and several baselines are also provided. We research data augmentation approaches to extend the training set and compare classification methods on different augmented data sets.

Keywords: fact-checking, factual consistency, lm, nlg, text generation
DOI: 10.28995/2075-7182-2023-22-267-277

Факт-чекинг для улучшения языковых моделей на русском языке

Анастасия Козлова
SberDevices
anastasi2510@gmail.com

Денис Шевелев
SberDevices
reddraner@gmail.com

Алена Феногенова
SberDevices
alenush93@gmail.com

Аннотация

Генеративные языковые модели сейчас стремительно развиваются и используются повсеместно. Однако, у них всё ещё есть ряд лимитов, и упущений, таких как ширина контекста, склонности к галлюцинациям и дегенерациям, логические связи, и изменения фактической информации. В данной работе мы рассматриваем задачу проверки фактов для непосредственно выхода генеративных моделей в классических генеративных задачах, таких как: парафраз, суммаризация, перенос стиля и подобных. В данной работе мы определяем задачу и критерии внутреннего факт-чекинга, впервые представляем новый русскоязычный датасет для этой задачи, а также набор тестов для оценки моделей и их сравнения с базовыми решениями. Мы также рассмотрели несколько методов аугментации данных для тренировочного сета и провели сравнительный анализ методов на разных наборах данных.

Ключевые слова: факт-чекинг, консистентность фактов, большие языковые модели, автоматическая генерация текста

1 Introduction

Large language models are fast developing and excel at producing text. The interest in language models continues to grow as such models are used to solve various downstream tasks, such as paraphrasing, summarization, style transfer, etc. Plenty of these tasks can be defined as generating the text based on some source text, where the model generates new original text, preserving the same sense. For such generative models, one of the main requirements for generated texts is factual correctness and consistency of text with the source data.

Despite progress in the quality of language models and the growth of scientific research in this field, texts generated using language models may contain inaccuracies, hallucinations (Zhou et al.,

2020)(Bender et al., 2021), and misinformations (Kryściński et al., 2019a). Automatic fact-checking can serve as an effective means of identifying inconsistencies in generated text, thereby enhancing the quality and reliability of the output. The significance of factual accuracy cannot be overstated, particularly in the context of news and medical articles, legal documents, and other socially consequential texts. At the same time, an automatic fact-checker can provide a more time-efficient solution to the problem of inaccurate information than manual fact-checking, making it available to a broader group of people. Thus, automatic fact-checking plays a vital role in improving the accuracy and consistency of information, helping to overcome the problem of false or misleading information.

Existing approaches to fact-checking are based on consistency testing of statements against evidence (Thorne et al., 2018a)(Mesgar et al., 2020) but do not consider the original information's completeness. For generative downstream tasks, preserving the consistency and completeness of the data is essential. Thus, the fact-checking systems may also be used as an essential tool for the evaluation of the large language models (Tam et al., 2022), (Chaudhury et al., 2022).

This work focuses on the internal fact-checking task as a fact-preservation problem and defines its criteria. In this paper, we present a new dataset and the factual verification benchmark[1] for the Russian language. The dataset contains tagged examples labeled *consistent* and *inconsistent*; for inconsistent examples, ranges containing violations of facts in the source text and statements are also presented. Various sources were used for data collection, such as texts obtained by the paraphrasing task and summarization data, translations from English to Russian of existing datasets for fact-checking, and text argumentation. We use the obtained dataset to fine-tune and evaluate models, such as ruBERT, ruRoBERTa, and ruGPT3, for the fact-checking task.

The rest of the paper is structured as follows. First, we overview the papers that are related to the field of fact-checking. In section 3, we discuss how we define the internal fact-checking task and what the fact is. Section 4 is devoted to the data we use in our experiments and various approaches to its collection. The methods and models we used and the description of the experiments are presented in Section 5. Finally, section 6 presents the evaluation and discussion.

2 Related work

The general task of fact-checking can be divided into several sequential steps (Guo et al., 2022) — first, the search of the sources and the collection of evidence necessary for verification verdict. Secondly, selecting the most relevant evidence to be used for verification. And finally, issuing a verdict using the collected evidence.

Thus, fact-checking can be separated into internal and external depending on the evidence source type. External fact-checking is the process of checking the actual accuracy of the content generated by a language model using external sources of information and data. This approach aims to determine the consistency of the generated text by comparing it with verifiable facts from some databases or sources such as news articles, academic journals, government reports, and other reliable sources. For internal fact-checking, a reliable source of evidence is predetermined by the downstream task. For example, we are checking the actual consistency of the source text with the content generated by the summarization model. The factual consistency of the summarization task is one of the most frequent cases, discussed in works (Wang et al., 2020) (Fabbri et al., 2021) (Kryściński et al., 2019b). In this case, the model's input text is evidence and aims to preserve the facts in the generated text output. This paper will focus on internal fact-checking for the text-generative downstream tasks and the factual consistency of language models.

2.1 Fact-checking Datasets

The bottleneck for building a fact-checking model is the need for labeled data for various languages. Most of the datasets are presented in English only. The FEVER dataset (Thorne et al., 2018a) is one of the most well-known fact-checking datasets in English, which contains claims extracted from Wikipedia documents. Each claim is assigned one of three labels: *Supported*, *Refuted* or *NotEnoughInfo*.

[1] `https://huggingface.co/datasets/akozlova/RuFacts`

For the first two classes, the annotators recorded the sentences forming the necessary evidence for their judgment. The evidence is texts from Wikipedia, and annotators write claims for verification. Another dataset for fact-checking is the Vitamin C dataset (Schuster et al., 2021) based on texts from Wikipedia. The largest publicly available multilingual dataset is the X-FACT dataset (Gupta and Srikumar, 2021), which includes 31,189 short statements labeled for factual correctness and covers 25 typologically diverse languages, including statements in Russian. As part of the FactRuEval (Starostin et al., 2016) competition, a publicly available corpus was created to evaluate fact extraction systems. The corpus can be used to detect facts of specific types in the texts but is not intended to be used for the fact-checking task. The Russian Commitment Bank dataset that is a part of the Russian SuperGLUE (Shavrina et al., 2020) benchmark can be considered a close variant of the task definition as it also validates the contradiction/entailment of some source premise. However, Natural Language Inference (NLI) is a much broader task and can not be defined as fact-preservation due to the inability of concrete fact selection.

2.2 Fact-checking Methods

There are various approaches to the problem of fact-checking using evidence. Question-answer systems are often used for fact-checking, the main task of which is to check the consistency of named entities in texts. According to previous research, scores based on question-answer systems correlate highly with a human judgment of facts. The approach (Wang et al., 2020) and similar question-answer approaches are based on the intuitive assumption that if we ask the same questions to both the summarized text and its source, we will get similar answers, but only if the generated text matches the source. The authors have shown that this approach significantly outperforms other automatic scoring measures in terms of correlation with human judgments of factual consistency. However, such approaches do not consider the completeness of the presentation of the original information, checking only individual facts.

The most common formulation of the fact-checking problem is to build a binary classifier based on a pre-trained language model, such as BERT, labeled *Supported* or *Unsupported* (Glockner et al., 2022; Guo et al., 2022). The paper is also based on the hypothesis from the FactCC [2] paper (Kryściński et al., 2019b) that errors made by paraphrasing models are most often associated with the use of incorrectly named entities, as well as numbers and pronouns. The authors base their work on the approach for generating training data for fact-checking to reduce manual markup costs. The training data is generated by applying a series of rule-based transformations to the sentences of the source documents. Examples are created by sampling individual sentences, later called claims, from source documents. The claims then undergo a series of text transformations resulting in new sentences with positive and negative labels. The advantage of using a synthetic dataset is that it generates large amounts of data at minimal costs.

The author of the paper (Lee et al., 2021) used a perplexity score from the language model to check the consistency between a claim and evidence. The researchers suggest including evidence in the perplexity calculation, using it as a prefix for a claim since perplexity measures the likelihood of a given sentence regarding a previously encountered text. They assume that unsupported claims have higher perplexity compared to supported claims.

Some approaches (Cao et al., 2020) are devoted to correcting factual errors in generated texts through post-editing. Usually, such text correction models are trained on adversarial examples built using heuristics to introduce errors. However, generating such examples using heuristics often needs to generalize better to actual model errors. In this paper, the authors propose to generate representative non-factual adversarial examples using infilling language models. The authors use a beam search of lower-ranked candidates from the language model to source potentially incorrect facts, creating a set of plausible and probable but incorrect synthetic texts for a particular correct text.

3 Task definition

The task of internal fact-checking can be considered from different perspectives. For example, based on the Named Entity Recognition (NER)/facts span detection in two texts or the classical task of NLI,

[2] https://github.com/salesforce/factCC

determining whether a “hypothesis” is true (entailment), false (contradiction), or undetermined (neutral) given a “premise”.

We are to combine such approaches and formulate the fact-checking problem as follows: Given a pair of texts (c, g), where c is a source human-written text, and g is the generated text by some generative model, that needs to be checked for factual consistency with the conditional input c. The fact-checking model must predict one of two labels for the generated output: the facts are ‘consistent‘ or ‘inconsistent‘.

Based on the problem statement, the requirements for a fact-checker include 1) examining the factual inconsistency, looking for the presence of facts that are not contained in the source text, and 2) verifying the completeness of the presentation of the source information. It’s worth mentioning, for instance, not all facts should be presented for the summarization task in the generated abstract, but at the same time, corruption or new facts, in this case, are unacceptable.

A similar definition is used in works that proposed an assessment of factual consistency evaluation methods (Honovich et al., 2022). They require the text to be faithful to its source text, regardless of the "correctness" concerning the "real world". To assess faithfulness, criteria are based on the information presented in the input text, not external knowledge.

Investigating the common errors of factual inconsistency in the corresponding works (Kryściński et al., 2019b) (Tam et al., 2022) we highlight the cases that cover the most frequently encountered contradictions in facts in generated texts. We further use them for the data augmentation procedure. The classes are the following:

- **NER (names, numbers, localizations)**. *Examples: “Lermontov” instead of “Pushkin”; “125.000 roubles” instead of “125 roubles”*
- **relations** *Examples: “grandmother” instead of “grandfather”; “chef” instead of “subaltern”*
- **negotiation** *Examples: “Natasha did not see her boss yesterday” and “Natasha saw her boss yesterday”*
- **gender** *Examples: “Natasha did not see her boss yesterday” and “Natasha did not see his boss yesterday”*
- **states (actions, positions)** *Examples: “Masha has eaten the apple” and “Masha is eating the apple”*

To sum it up, the fact-checking system needs to be based on these typical error cases, and the following conditions need to be complied with: 1) the facts are correct and not corrupted in both texts (source and generated); 2) any additional facts in the generated texts are not included; 3) the generated text includes all the main facts from the source text.

4 Data

4.1 Data Collection

Various data sources and approaches for data generation were used to create the training and test datasets for the fact-checking task. Our approach involves analyzing data at both the sentence level and within smaller texts. The data exhibits an average text length of 198 symbols, with a minimum length of 10 symbols and a maximum length of 3,402 symbols. The final dataset was formed using three main approaches: 1) texts generated by a paraphrase model 2) translations of datasets for fact-checking 3) text augmentation.

Text Generation. The most frequent usage of the fact-checking verification system is some generated output based on the original text. Thus, we take the generation results of the paraphrase model and summarization data for the basis of the dataset. The paraphraser[3] was chosen as it’s a free model that is provided as an API. The model was trained on 7000 examples from different sources of various domains: 1) text level - texts from different domains filtered with Bertscore (Zhang et al., 2019) and Rouge-L) 2) sentence level - the Russian version of Tapaco corpus (Scherrer, 2020) and filtered ParaphraserPlus (Gudkov et al., 2020) corpus. Russian news dataset for summarization[4] was used as the source data for models generation. From each text, a fragment consisting of 1, 2 or 3 sentences were

[3] https://habr.com/ru/company/sberdevices/blog/667106/

[4] https://huggingface.co/datasets/IlyaGusev/gazeta

taken. The collected fragments were used as input for generating statements using the paraphrase model and the evidence for the generated statements. Since the generated data may be factually inconsistent with the source texts, we annotate them manually for future reference.

Datasets Translation. The dataset also included English-language data from the FEVER fact-checking dataset (Thorne et al., 2018a) that was translated into Russian. In the FEVER dataset, the claims are classified as *Supported*, *Refuted* or *NotEnoughInfo*. For the first two classes, the annotators recorded the sentences forming the necessary evidence for their judgment. We use claims labeled *Supported* and *Refuted* and collected evidence in our work. The two NLLB-200 models[5][6] are tested for translation. We sample using the temperature of 0.85, top_k of 100, top_p of 0.8, max_length of 200 as generation parameters. We then choose the best translation using Question-Answering based metrics (Scialom et al., 2019). For each translation to assess, questions are successively generated from a source text by masking each of the named entities in this text. The results are triplets (input, question, answer), where input denotes the claim, the question refers to the sentence containing the masked entity, and the answer refers to this masked entity to retrieve. For each triple, an $F1$ score is calculated. As QA system we use the pre-trained ruBERT-large[7] fine-tuned on the SberQuAD[8] dataset. The resulting dataset included examples with a $F1$ score greater than 0.25.

Text Augmentation. The rule-based transformations (Kryściński et al., 2019b) were proposed as an alternative approach to syntectic data generation. A paraphrase dataset[9] was used as the source data. The original pairs of texts were factually consistent. A series of rule-based transformations were applied to one of the pairs obtaining factually inconsistent pairs, with one paraphrase as evidence and the other as a statement that would go through the transformations. The rule-based transformations consisted of several stages, based on the task definition criteria:

1. a randomly selected named entity in the statement was replaced with a different randomly selected named entity from the evidence text;
2. randomly selected numbers in the statement were replaced with randomly generated numbers;
3. the negative particle не was removed from the statement to change the context.

In the current work, we used the SpaCy library[10] to recognize entities. To generate additional factually inconsistent examples, available Russian corpora[11] were used. We apply the entity swapping transformation for Persons-1000[12] and Collection5[13] datasets annotated with PER, LOC, and ORG tags. For the Persons-1000 dataset, we also apply the number-swapping transformation. We use the sentence negation for the RuADReCT dataset (Tutubalina et al., 2021). We additionally manually annotate the augmented data for the test set; augmented data without manual annotation is used for the training set.

4.2 Test data

The test set consists of examples from all three sources: 26% translations, 6% augmented data, and 68% generated paraphrases. A description of the sources is presented in Section 4.1.

The test data for fact-checking was manually labeled via the crowd-sources platform Yandex.Toloka[14] (Pavlichenko et al., 2021). First, we made a classification task and asked annotators to check whether the facts in the two texts were correct. However, we faced several problems: 1) cheating and 2) misunderstanding the fact definition. It's proved that determining the truthfulness of a fact regarding a general "real world" is subjective and depends on the knowledge, values, and beliefs of the subject (Heidegger, 2005). To decrease these effects, we claim the annotators not just check the fact's coincidence but also highlight exactly the facts span. Human annotation submissions are collected

[5] https://huggingface.co/facebook/nllb-200-distilled-600M
[6] https://huggingface.co/facebook/nllb-200-distilled-1.3B
[7] https://huggingface.co/sberbank-ai/ruBert-large
[8] https://huggingface.co/datasets/sberquad
[9] https://huggingface.co/datasets/merionum/ru_paraphraser
[10] https://spacy.io/
[11] https://github.com/natasha/corus
[12] http://ai-center.botik.ru/Airec/index.php/ru/collections/28-persons-1000
[13] http://www.labinform.ru/pub/named_entities/
[14] https://toloka.ai/tolokers

and stored anonymously via the design presented in Figure 1. Each annotator is warned about potentially sensitive topics in data (e.g., politics, religion, societal minorities, etc.). The annotation details are provided in Table 1.

IAA	**Total**	**Overlap**	N_T	N_{page}	N_C	N_U	**ART**
80.2%	42$	5	8	3	50	74	113

Table 1: Details on the data collection project for the test set. **IAA** (inter-annotator agreement) refers to the IAA confidence scores. **Total** is the total cost of the annotation project. **Overlap** is the number of votes per example. N_T is the number of training tasks. N_{page} denotes the number of examples per page. N_C is the number of control examples. N_U is the number of users who annotated the tasks. **ART** means the average response time in seconds.

IAA	**Total**	**Overlap**	N_T	N_{page}	N_C	N_U	**ART**
75.1%	801$	3-5	8	3	50	181	103

Table 2: Details on the data collection project for the train set. **IAA** (inter-annotator agreement) refers to the IAA confidence scores. **Total** is the total cost of the annotation project. **Overlap** is the number of votes per example. N_T is the number of training tasks. N_{page} denotes the number of examples per page. N_C is the number of control examples. N_U is the number of users who annotated the tasks. **ART** means the average response time in seconds.

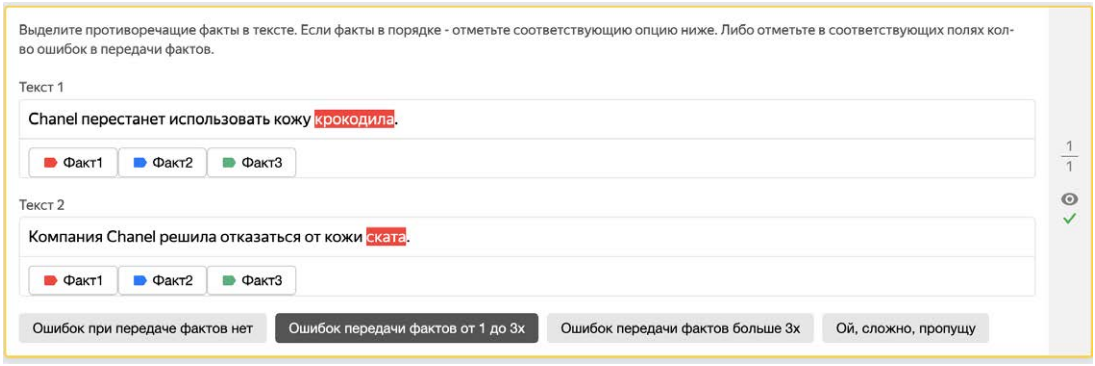


Figure 1: The example of Yandex.Toloka design setup. Two texts are provided, and annotators need to span the inconsistency of the facts. There is a required field with four options to set how many facts the texts contain.

We verify the annotator submissions' quality with control questions and exclude cheaters. The overlap is set to 5 to provide more reliable results and high confidence. We count IAA using majority votes, considering not just classification buttons but also the span overlap of annotators. Due to the complexity of the task, we exclude the examples in the set that contains less than three annotators' votes or has a low IAA. We balance the dataset to save the class distribution; dataset statistics are reported in Table 3.

4.3 Training data

Three training sets were prepared based on data from Section 4.1 to compare various approaches to creating training data for the fact-checking task.

- The first train set **Translated set** consists of translated English-language fact-checking dataset.
- The second train set **Augmented set** contains augmented data.
- The third train set **Labeled set** includes parts of the translations, augmented data and generated data. Translations and generated data were manually labeled via the crowd-sources platform Yan-

dex.Toloka. The annotation project was similar to the golden test set collection setting. The details of the train verification procedure are presented in Table 2.

Data Set	Consistent	Inconsistent	Total
Translated set	2150	2146	4296
Augmented set	1258	1434	2692
Labeled set	2994	3242	6236
Test set	250	250	500

Table 3: Statistics of data sets.

The final statistics of data sets are reported in Table 4. We split all sets into train and validation. For each dataset, we use 75% of the data as the training set and 25% as the validation set.

5 Experiments

Despite the span annotations in our data, in this paper, we define the task as a classification problem and conduct experiments for binary classification. We provide several baselines on the different train sets and fine-tune state-of-the-art models on this task.

5.1 Models

Baselines As baselines, we develop a classifier built on perplexity calculation and a classifier built on the cosine similarity calculation.

The perplexity-based approach (Lee et al., 2021) **ruGPT3-ppl** is based on including evidence in the perplexity calculation, using it as a prefix for a claim: $X = (x_{e_0}, ..., x_{e_E}, x_{c_0}, ..., x_{c_C})$, where E and C denote the number of evidence tokens and claim tokens, respectively. We obtain the perplexity of an input text as follows:

$$PPL(X) = \sqrt[C]{\prod_{i=0}^{C} \frac{1}{p_\theta(x_{c_i}|x_{e_0}, ..., x_{e_E}, ..., x_{c_{i-1}})}} \quad (1)$$

where X is an input text, C is the length of the claim. The ruGPT3-large model[15] is used to calculate perplexity. The ruGPT3 is a Russian adaptation of the autoregressive language model GPT3 (Brown et al., 2020).

The cosine similarity approach **LaBSE-sim** is based on calculating the cosine similarity between embeddings. We use the LaBSE model[16] (Feng et al., 2020) to obtain embeddings of the evidence e and claim c texts, then we calculate the cosine similarity between them:

$$\cos(\theta) = \frac{e \cdot c}{\|e\|\|c\|} \quad (2)$$

Optimal threshold values are determined for baseline models that effectively distinguish between factually consistent and inconsistent claims. The training set is utilized to identify the hyper-parameter value that yields the highest level of performance for the threshold parameter, denoted as th, without requiring any parameter updates to pre-existing language models.

Fine-tuned models We fine-tune pre-trained Transformer-based models on the collected training datasets to build baseline classifiers. Three state-of-the-art models of different size are considered:

- ruBERT-base[17] is a Russian BERT model (Devlin et al., 2019) trained 30 GB Russian filtered dataset (including domains: Wikipedia, news, part of the Taiga corpus, fiction, etc.),
- ruRoberta-large[18] is a RoBERTa model (Liu et al., 2019) trained on 250GB Russian dataset,

[15] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2
[16] https://huggingface.co/sentence-transformers/LaBSE
[17] https://huggingface.co/sberbank-ai/ruBert-base
[18] https://huggingface.co/sberbank-ai/ruRoberta-large

• ruGPT3-small[19] is a small version of ruGPT from the ruGPT-family [20].

We fine-tune ruBERT-base and ruRoberta-large models with a single-layer classifier on top. We concatenate the evidence e and the claim c, insert [SEP] token between them and add [CLS] to make the sequence. This sequence is fed as input to the model for binary classification.

For the ruGPT3-large, the input prompt sequence for the task is written as follows:

Доказательство: [e]

Утверждение: [c]

Доказательство подтверждает утверждение:

We fine-tuned ruGPT3-large to generate the target tokens Да (Yes) or Нет (No).

5.2 Experimental Setup

Evaluation metrics Since we consider the fact-checking task a binary classification problem for a balanced test set, we used accuracy as the primary metric to evaluate models. We also used precision, recall, and F1-score as additional metrics. For fine-tuned models, we report the average results across five runs with different random seeds (the standard deviation is presented in Table 4).

Training Details During our experiments, we set the maximum sequence length to 512 and used a batch size of 16. Models were trained for seven epochs using the Adam optimizer (Kingma and Ba, 2014). For ruGPT3, we used a learning rate of $5e{-}5$, while for ruBERT and ruRoberta we employed the Adam optimizer with a learning rate of $1e{-}5$ was used. The best model checkpoints were selected based on performance on the validation set.

6 Evaluation

Model	Training Set	Accuracy	F1	Precision	Recall
ruGPT3-ppl	Translated set	56.0	57.2	55.7	58.8
ruGPT3-ppl	Augmented set	56.2	63.8	54.4	77.2
ruGPT3-ppl	Labeled set	56.4	62.8	54.8	73.6
LaBSE-sim	Translated set	62.8	55.1	69.5	45.6
LaBSE-sim	Augmented set	51.6	67.3	50.8	99.6
LaBSE-sim	Labeled set	63.2	65.4	61.7	69.6
ruBERT-base	Translated set	57.4 (±0.52)	33.3 (±1.31)	76.8 (±2.59)	21.3 (±1.11)
ruBERT-base	Augmented set	52.4 (±1.07)	63.0 (±0.54)	51.5 (±0.70)	81.1 (±0.59)
ruBERT-base	Labeled set	63.4 (±0.59)	65.7 (±1.12)	61.9 (±0.43)	70.0 (±2.57)
ruRoBERTa-large	Translated set	60.2 (±0.66)	41.8 (±3.27)	**78.0** (±4.77)	28.8 (±3.65)
ruRoBERTa-large	Augmented set	55.3 (±0.83)	63.5 (±1.60)	53.7 (±0.58)	77.8 (±4.54)
ruRoBERTa-large	Labeled set	**66.0** (±1.49)	**68.0** (±1.03)	64.5 (±2.43)	72.2 (±3.73)
ruGPT3-small	Translated set	53.8 (±1.17)	49.3 (±2.38)	54.6 (±1.38)	45.1 (±3.75)
ruGPT3-small	Augmented set	42.2 (±0.58)	56.4 (±1.18)	45.3 (±0.48)	74.7 (±2.83)
ruGPT3-small	Labeled set	54.4 (±1.99)	57.1 (±0.98)	54.2 (±2.44)	60.9 (±4.76)

Table 4: Results of models fine-tuned on each training set and evaluated on the test set. We report the mean and standard deviation (in parentheses) across 5 runs with different random seeds for fine-tuned models.

Our experiments assess the impact of different training datasets on model performance. We report the results in Table 4, which displays the accuracy of the fine-tuned models on *Translated*, *Augmented*, and *Labeled* training sets, evaluated on our manually labeled test set. Based on our accuracy metrics, all models perform best when trained on the *Labeled* set. Specifically, the ruRoBERTa-large model trained on the *Labeled* set achieves the highest accuracy score of 66.0% accuracy and F1-score of 68.0%. These

[19] `https://huggingface.co/sberbank-ai/rugpt3small_based_on_gpt2`

[20] https://sbercloud.ru/ru/datahub/rugpt3family

results can be attributed (i) to the diversity of data sources included in the sample and (ii) to the manual annotation of the collected data, which enhances the quality of data labeling.

Experimental results reveal a decrease in performance metrics when using the *Translated* set for fine-tuning. This can be attributed to the fact that the *Translated* set is composed of automatically translated texts, which may contain mistranslations, especially in the case of named entities and language peculiarities. Therefore, using such translated data may result in poorer model performance compared to the *Labeled* set, which benefits from manual annotation, contains various data sources, and is more reliable.

In our experiments, we observed that using LaBSE-sim on the *Augmented* set resulted in a high F1-score comparable to the best-performing ruRoBERTa-large model, and low, almost random accuracy metrics. This can be attributed to the high recall but low precision of the LaBSE-sim approach. It appears that there is a possibility that the finding of the threshold on synthetic augmented sets can increase model recall in the cases of simple fact contradictions and replacements similar to the FactCC approach. However, this method may not be sufficient for catching more complex fact inconsistencies, as the test set contains more complex cases that cannot be identified solely based on factual inconsistency class replacements.

According to our results, the perplexity-based approach, ruGPT3-ppl, outperforms the ruGPT-small fine-tuned on the classification task. This coincides with the Russian SuperGLUE leaderboard[21], which shows that the ruGPT3-small is not performing well in classification tasks, particularly those based on NLI, perhaps due to its generative pre-training nature. In contrast, the ruGPT3-ppl approach demonstrates consistent results. We suggest that a larger model, such as the ruGPT3 XL, may exhibit more generalization abilities and improve the perplexity-based approach's overall performance.

The experimental results on the proposed datasets demonstrate an overall accuracy close to 70%. This performance level is comparable to that achieved by state-of-the-art models on analogous benchmarks for the English language, such as the FEVER leaderboards (Thorne et al., 2018a) (Thorne et al., 2018b). Moreover, the TRUE benchmark for English also reported comparable F1 scores for a similar task and highlighted that NLI-based models, for example, Adversarial NLI (Nie et al., 2020), outperformed other approaches (Honovich et al., 2022). This observation is not surprising given the complexity of the collected dataset, which requires models to exhibit robust reasoning capabilities. In fact, the nature of factual consistency in the text is more intricate than just simple sentence structures, necessitating more nuanced and sophisticated approaches to capture and classify factual information accurately.

7 Conclusion

This paper investigates the problem of internal fact-checking and the ability of large language models to preserve factual consistency. We introduce a new evidence-based fact-checking dataset and benchmark for the Russian language, which is publicly available[22]. To expand the training set, we utilize data augmentation techniques and compare classification methods on various augmented datasets. Based on our analysis of model performances, we find out that the pre-trained ruRoBERTa-large model fine-tuned on manually annotated data yields the best results. Furthermore, we have launched a competition[23] and present a public leaderboard for the proposed task. In future research, we plan to explore using factual inconsistency spans for model training and treating the task as a token classification problem. Additionally, we aim to address the challenges associated with evaluating factual consistency and explore the integration of NLI-based methods into our current approach.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. // *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, P 610–623.

[21] https://russiansuperglue.com/leaderboard/2

[22] `https://huggingface.co/datasets/akozlova/RuFacts`

[23] `https://www.kaggle.com/competitions/internal-fact-checking-for-the-russian-language`

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Meng Cao, Yue Dong, Jiapeng Wu, and Jackie Chi Kit Cheung. 2020. Factual error correction for abstractive summarization models. *arXiv preprint arXiv:2010.08712*.

Subhajit Chaudhury, Sarathkrishna Swaminathan, Chulaka Gunasekara, Maxwell Crouse, Srinivas Ravishankar, Daiki Kimura, Keerthiram Murugesan, Ramón Fernandez Astudillo, Tahira Naseem, Pavan Kapanipathi, et al. 2022. X-factor: A cross-metric evaluation of factual correctness in abstractive summarization. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 7100–7110.

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.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Max Glockner, Yufang Hou, and Iryna Gurevych. 2022. Missing counter-evidence renders nlp fact-checking unrealistic for misinformation. *arXiv preprint arXiv:2210.13865*.

Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh. 2020. Automatically ranked russian paraphrase corpus for text generation. *arXiv preprint arXiv:2006.09719*.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Ashim Gupta and Vivek Srikumar. 2021. X-fact: A new benchmark dataset for multilingual fact checking. *arXiv preprint arXiv:2106.09248*.

Martin Heidegger. 2005. On the essence of truth. *Truth: Engagements across philosophical traditions*, P 244–260.

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. True: Re-evaluating factual consistency evaluation. *arXiv preprint arXiv:2204.04991*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Wojciech Kryściński, Nitish Shirish Keskar, Bryan McCann, Caiming Xiong, and Richard Socher. 2019a. Neural text summarization: A critical evaluation. *arXiv preprint arXiv:1908.08960*.

Wojciech Kryściński, Bryan McCann, Caiming Xiong, and Richard Socher. 2019b. Evaluating the factual consistency of abstractive text summarization. *arXiv preprint arXiv:1910.12840*.

Nayeon Lee, Yejin Bang, Andrea Madotto, Madian Khabsa, and Pascale Fung. 2021. Towards few-shot fact-checking via perplexity. *arXiv preprint arXiv:2103.09535*.

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. *arXiv preprint arXiv:1907.11692*.

Mohsen Mesgar, Edwin Simpson, and Iryna Gurevych. 2020. Improving factual consistency between a response and persona facts. *arXiv preprint arXiv:2005.00036*.

Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. // J. Vanschoren and S. Yeung, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. // *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Tal Schuster, Adam Fisch, and Regina Barzilay. 2021. Get your vitamin c! robust fact verification with contrastive evidence. *arXiv preprint arXiv:2103.08541*.

Thomas Scialom, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2019. Answers unite! unsupervised metrics for reinforced summarization models. *arXiv preprint arXiv:1909.01610*.

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. Russiansuperglue: A russian language understanding evaluation benchmark. *arXiv preprint arXiv:2010.15925*.

Anatoly S Starostin, Victor V Bocharov, Svetlana V Alexeeva, Anastasiya A Bodrova, Alexander S Chuchunkov, SS Dzhumaev, Irina V Efimenko, Dmitry V Granovsky, Viktor F Khoroshevsky, Irina V Krylova, et al. 2016. Factrueval 2016: evaluation of named entity recognition and fact extraction systems for russian.

Derek Tam, Anisha Mascarenhas, Shiyue Zhang, Sarah Kwan, Mohit Bansal, and Colin Raffel. 2022. Evaluating the factual consistency of large language models through summarization. *arXiv preprint arXiv:2211.08412*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018a. Fever: a large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018b. The FEVER2.0 shared task. // *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*.

Elena Tutubalina, Ilseyar Alimova, Zulfat Miftahutdinov, Andrey Sakhovskiy, Valentin Malykh, and Sergey Nikolenko. 2021. The russian drug reaction corpus and neural models for drug reactions and effectiveness detection in user reviews. *Bioinformatics*, 37(2):243–249.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. *arXiv preprint arXiv:2004.04228*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Paco Guzman, Luke Zettlemoyer, and Marjan Ghazvininejad. 2020. Detecting hallucinated content in conditional neural sequence generation. *arXiv preprint arXiv:2011.02593*.

Text complexity as a non-discrete value: Russian L2 text complexity dataset annotation based on Elo rating system

Laposhina Antonina Nikolaevna
Pushkin State Russian Language Institute, Moscow, Russia
antonina.laposhina@gmail.com

Abstract

The task of assessing text complexity for L2 learners can be approached as either a classification or regression problem, depending on the chosen scale. The primary bottleneck in such research lies in the limited availability of appropriate data samples. This study presents a combined approach to create a dataset of Russian texts for L2 learners, placed on a continuous scale of complexity, involving expert pairwise comparisons and the Elo rating system. For this pilot dataset, 104 texts from Russian L2 textbooks, TORFL tests, and authentic sources were selected and annotated. The resulting data is useful for evaluation of the automated models for assessing text complexity.

Keywords: text complexity; Russian as a foreign language; Elo ratings; text complexity dataset; pairwise annotation
DOI: 10.28995/2075-7182-2023-22-278-286

Сложность текста как недискретная величина: экспертная разметка сложности текстов по РКИ на основе рейтингов Эло

Лапошина Антонина Николаевна
Государственный институт русского языка им. А. С. Пушкина, Москва
antonina.laposhina@gmail.com

Аннотация

В исследовании представлен подход к созданию коллекции текстов, аннотированных по сложности для изучающих русский язык как иностранный, на непрерывной шкале, базирующейся на уровнях CEFR. Подход основан на попарной экспертной оценке текстов и системе рейтингов Эло. Исследование выполнено на 104 текстах из специализированных пособий по РКИ и аутентичных источников. Полученные данные полезны для оценки предсказательных моделей уровня сложности текста для изучающих русский язык как иностранный.

Ключевые слова: сложность текста; русский язык как иностранный; рейтинги Эло; попарное сравнение

1 Introduction

The crucial initial step in text complexity studies is to establish a complexity scale and acquire a collection of text samples marked with this scale. The model is then developed and tested based on this data. Depending on the chosen scale, the task of text complexity evaluation can be resolved as a classification problem (resulting in the anticipated class, grades, levels) (Karpov et al. 2014; Francois, Fairon 2012; Reynolds 2016) or a regression problem (yielding any decimal number on a specified scale) (Kate et al. 2010; Seiffe et al. 2022). Hence, not only does the algorithm's design depend on the selection of the scale, but the researcher's fundamental perspective on the concept of text complexity as discrete levels or as a continuum of difficulty.

The primary bottleneck in such research lies in the limited availability of appropriate training data. Most existing datasets consist of discrete complexity levels, such as school materials annotated by grade (Solovyov et al. 2018), age or abstract units (Pitler, Nenkova 2008), or «easy-difficult» binary scale (Sharoff et al. 2008). Regarding the assessment of the text complexity for L2 learners, the Common European Framework of Reference (CEFR) is the preferred choice for the majority of researchers

(Reynolds 2016; Karpov et al. 2014; Schwarm and Ostendorf 2005; Laposhina et al. 2018; Corlatescu et al. 2022).

1.1. CEFR levels as a complexity scale

The Common European Framework of Reference for Language Proficiency (CEFR) establishes universal standards that are utilized worldwide to determine language proficiency levels and serve as a means to acknowledge qualifications obtained from diverse educational systems. In its current version, the 2018 descriptors, the CEFR scale comprises 7 levels ranging from pre-A1 to C2. Nevertheless, even the CEFR descriptor`s authors acknowledge the conventional nature of the proposed scale. «All categories in the humanities and liberal arts are in any case conventional, socially constructed concepts. Like the colors of the rainbow, language proficiency is actually a continuum. Yet, as with the rainbow, despite the fuzziness of the boundaries between colors, we tend to see some colors more than others. Yet to communicate, we simplify and focus on six main colors» (Common European Framework 2018: 34). Private initiatives, such as the sub-level system shown in Figure 1, based on the main CEFR scale material and used in the Polyskills institutes network, further support the practical necessity for a more detailed level scale.

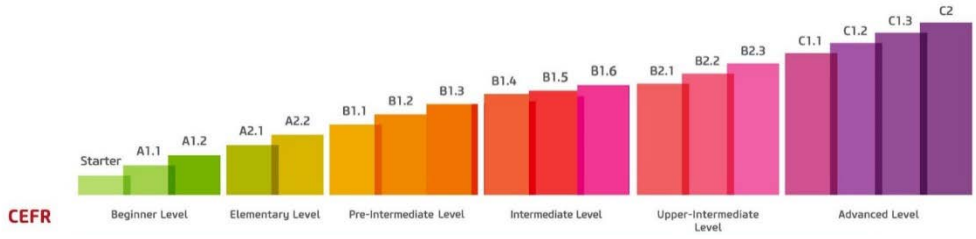


Figure 1: Detailed visualization of CEFR levels for ease of use in teaching practice

In studies on Russian L2 materials, it has been found that teachers commonly use unofficial terms to specify the placement of a particular text within a CEFR level, such as «beginning of B1» «end of B1» «B1+» etc (Laposhina, 2018). Consequently, the formal presentation of text complexity as a conventional scale of levels does not always suit the users' needs and requires more precise information about the place of the text on it.

1.2. Datasets for L2 text complexity assessment task

An automated approach to the complexity assessment of the Russian L2 texts has several examples, most of them are based on datasets with discrete levels, such as the corpus of textbooks annotated by publishers on the CEFR scale can be used (Reynolds 2016; Karpov et al. 2014; Batinic et al. 2016; Laposhina et al. 2018; Corlatescu et al. 2022). However, to create a non-discrete scale, expert annotation is necessary, which can be time-consuming and expensive. Besides, some studies report a low level of expert agreement in the direct task of assigning a text to one of the levels (Laposhina, 2018). To optimize this step, the problem of text complexity annotation can be modified to a pairwise comparison problem of the complexity of two texts (De Clercq et al., 2014; Chen et al.,2013).

Consider the scenario where the creation of a needed dataset is attempted not de novo, but instead utilizing an existing dataset with discrete levels and just refining them by pairwise comparison. The Elo rating system, originally designed to assess the relative strength of chess players (Elo, 1978), has been applied to rank various types of data, including the complexity of the educational content difficulty (Mangaroska et al. 2019), compiling a set of lexical and grammatical topics for Russian L2 learners and evaluating a student's proficiency level in these topics (Jue Hou et al. 2019).

In this article, we examine a combined approach to the ranking of Russian texts for L2 learners on a continuous scale of complexity, which involves expert pairwise comparisons and the Elo rating system. The resulting data is useful for the creation and evaluation of automated models for assessing text complexity for Russian L2 learners.

2 Materials and methods

2.1. Data

For this pilot study, 104 texts from Russian L2 textbooks, reading sections of TORFL tests, and different samples of authentic sources - news sites, blogs, and other media were selected. As an initial complexity level, we used the information about the CEFR level indicated by textbook editors; texts from authentic sources got initial levels C1 (blogs, news, and non-fiction notes) and C2 (academic and official text fragments). The length of the text samples for levels A2-C2 varies from 98 to 127 tokens to save a relatively complete idea of the text fragment. Texts for A1 level are usually shorter, so their length varies from 55 to 103 tokens. The composition of the text sample is shown in Table 1.

Text source\ Number of texts	**A1**	**A2**	**B1**	**B2**	**C1**	**C2**	**Total**
Russian L2 textbooks	13	15	15	13	3	0	**55**
TORFL test reading section	2	2	3	3	2	0	**12**
Authentic sources (news, blogs, online magazines)	0	0	0	0	20	13	**33**
Total	**15**	**17**	**18**	**16**	**25**	**13**	**104**

Table 1: Number of text samples per initial levels and text sources

2.2. Annotation process

To collect data, we have developed a special web interface for pairwise comparison of texts. First, the window with instructions is demonstrated, and a test comparison of two texts with obvious results (elementary and very complex text) so that we could make sure that the expert understood the task correctly. After successfully passing the instruction part, experts were asked to compare texts in pairs, having 3 options: «Left text is more difficult», «Right text is more difficult» and «It's difficult to answer». Pairs of text samples for the main annotation track were generated randomly. In order to save annotators' resources and not show too obvious text pairs the main annotation track, we set an additional rule that a pair of texts should have an equal initial text level or +- one level (e.g. B1 vs B1; B2 vs C1, etc.). An example of an interface is shown in Figure 2.

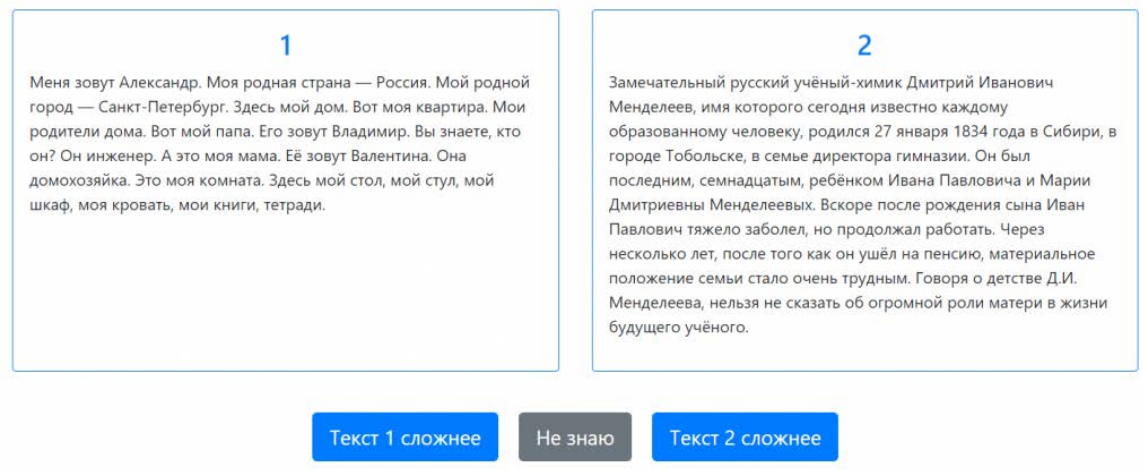


Figure 2. Interface for pairwise annotation of text complexity

The annotation process of this study can be called «expert crowdsourcing». On the other hand, the project was open to all potential annotators who successfully passed the validation stage (as described below). However, the promotion of the annotation project was limited to professional communities of Russian as foreign language teachers. The sample size consisted of 102 anonymous annotation sessions, each of which entailed 10 pairwise comparisons.

To ensure the adequacy and accuracy of the expert's work, we implemented an algorithm for primary verification. Specifically, pairs of texts with obvious right answers (text with initial level 1 or 2 VS 5 or 6) were shown twice during each annotation session. The annotation sessions were considered reliable if both verification tasks were completed accurately. Of the 102 annotation sessions, two had errors in both verification tasks, 17 made mistakes in one of the tasks, and 83 passed both verification tasks and were thus deemed acceptable for analysis.

2.3. Elo ratings

The concept behind the ratings introduced by the Hungarian mathematician Arpad Elo is to calculate the value of a player's victory based on the predictability of his victory. Using a chess analogy, the initial level of the text is «the rating of the player in the preliminary tournament table», which identifies who are the «grandmasters», «beginners», and «players of comparable skill». «Match» is a pairwise comparison of texts by an expert, where the «winner» is the text marked as more difficult.

The initial level of a text is a CEFR level declared by authors and editors, converted to a numeric format (A1 = 1, A2 = 2, B1 = 3, etc.); texts from authentic sources received starting levels 5 and 6 depending on the genre. As a result of each comparison session, texts received points: 1 - if the text annotated was more difficult; 0 - if the text annotated was simpler than the other; 0.5 - if the expert found it difficult to answer.

To adapt this idea to the text complexity task, we used formulas proposed in (Pelanek, 2016). According to them, the probability of «win» for text *i* in a «match» with text *j* is calculated with the following formula:

$$M_{ij} = \frac{1}{1+\exp(L_j - L_i)}$$

where L_i is the level of the text *i* and L_j is the level of the text *j* at the moment of comparison. New level of text *i* as a result of its comparison with text *j* was calculated with the formula:

$$L'_i = L_i + K(P_{ij} - M_{ij})$$

Where L'_i is the new assessment of the text, L_i is the level of the text at the time of comparison, P_{ij} is the score that gets *i* in a «match» (comparison) with text *j*, M_{ij} is the mathematical expectation that the *i*-th text will be more difficult than the *j*-th one. The factor K controls the maximum level adjustment that is possible at one round of comparison, so we set it to 0.25, following the (Ontaelio, 2016). The L values of both texts are updated after each comparison session.

The step-by-step example of comparing two texts is presented below: text 1 «Mailman» (example 1) is a fragment of an authentic text of an interview with a starting level of 5, text 2 «Burglary» (example 2) is a fragment of a journalistic text from the Russain L2 textbook with a level declared by authors of B2, i.e. with initial level 4.

(1) *До того как я сюда устроилась, думала, что почта — это уже прошлый век, вроде городского телефона: мало кто ей пользуется. Но, оказывается, на почту приходит множество людей! Конечно, загрузка у всех отделений разная, но наше, например, находится недалеко от метро, и народу здесь всегда хватает. У меня бывает больше 150 человек в день, всего работают три окошка, то есть получается порядка 500 человек ежедневно. Норма обслуживания на каждого клиента — восемь минут, и это очень мало, конечно. Такого, что скучно и не знаешь, чем себя занять, у нас не бывает. Всегда много клиентов, запросы у всех разные, только и успевай шевелиться.*

(2) *По статистике больше всего квартирных краж совершается в новых районах, так как новосёлы ещё плохо знакомы с соседями. Большая часть краж совершается с 9 до 12 часов (больше половины краж) и с 12 до 15 часов (четверть случаев). Воры предпочитают квартиры на первых и последних этажах: часто залезают в квартиры с крыш. Открытое окно или балкон – серьёзная ошибка. Неважно, на каком этаже вы живёте. Нередко воры залезают в понравившуюся им квартиру из квартиры этажом выше или ниже через балкон. Пытаясь узнать, дома хозяева или уехали, воры придумывают нехитрые манипуляции: периодически звонят в дверь (если хозяева откроют, то всегда можно представиться сотрудником компании, устанавливающей спутниковую связь или продавцом, предлагающем купить картошку, сахар и так далее).*

Annotator N felt that the text 2 about burglaries was more difficult. Therefore, according to the outcome of the comparison, text 1 gets 0 points, and text 2 gets 1 point. Based on the initial levels (5 and 4, respectively), the mathematical expectation of such an outcome will be:

$$M_{mailman} = \frac{1}{1+\exp(4-5)} = 0.73$$

$$M_{burglary} = \frac{1}{1+\exp(5-4)} = 0.27$$

In other words, based on the initial level of these two texts, with a probability of 73% text 1 should have «win» (be marked as more difficult). The annotator, on the contrary, chose text 2 as more difficult, although the probability of such an outcome was only 27%. The new levels for the two given texts will be equal to:

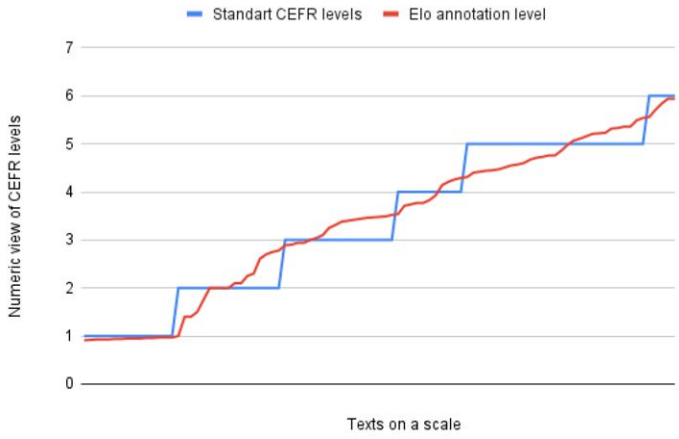
$$NewL_{mailman} = 5+0.25(0-0.73) = 4.82$$

$$NewL_{burglary} = 4+0.25(1-0.27) = 4.18$$

As a result of this comparison, the level of text 1 «Mailman» decreased, and the level of text 2 «Burglary» increased. Then the next comparison takes place, where the initial levels will be considered to be a new value. In total, text 1 participated in 24 comparisons with different texts, and as a result, its level decreased from 5 to 4.6.

3 Results

As a result of the pairwise annotation and calculations described above, we have obtained a collection of 104 texts smoothly distributed along the text complexity CEFR-based scale. Figure 3 illustrates a comparison of the distribution of texts on a scale and their initial CEFR levels.



Following the expert annotation process, the minimum and maximum values of the difficulty level were altered. Whereas the initial collection was marked on a scale of 1 (A1) to 6 (C2), the minimum level value decreased to 0.9, and the maximum increased to 6.8. Consequently, the study generated samples of texts that even native speakers find challenging. Interestingly, the most difficult text in the

collection turned out to be a fragment of an education bill. Figure 4 shows a detailed example of how texts with initial level 3 (B1) were distributed after assessments by expert annotators, from 2.5 to 3.7.

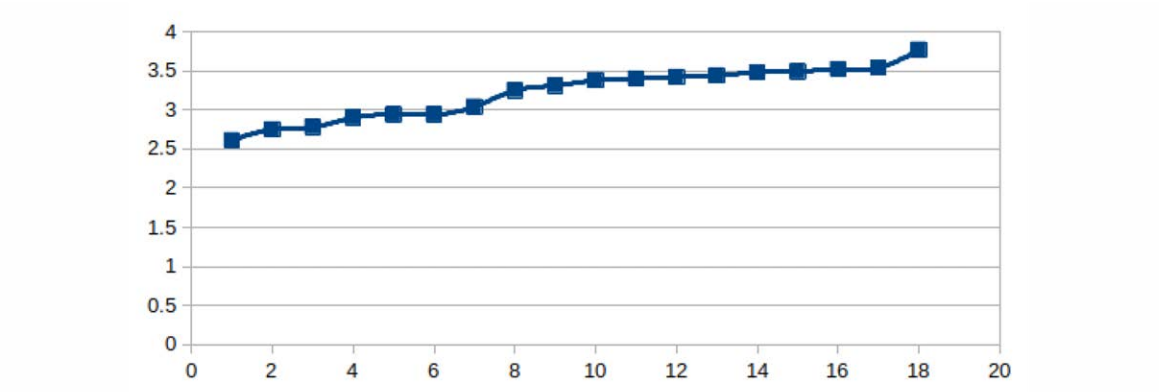


Figure 4: Texts with initial level 3 (B1) after annotation: x-axis: text number in the ranked list, y-axis: the new level value after the annotation, where 2 is equal to A2, 3 is equal to B1, etc.

As illustrated in the figure, the changes in text levels were not revolutionary; level 3 texts were smoothly distributed on a difficulty scale ranging from 2.5 to 3.7. However, some texts shifted to the end of the previous level 2, which is an important finding for evaluating the quality of the text complexity assessment model. Additionally, such a view of the level of text complexity aligns naturally with the idea of language acquisition as a gradual progression from simple to complex.

3.1. Assessment of the validity of expert answers

Annotation design using Elo rating system protects data to some extent from inconsistent markup: even if an expert N made an unexpected decision, next experts and next comparisons will be able to «shift» given text on the scale, thus creating an average expert opinion about the right place of this text on the complexity scale.

For additional verification, we inserted one specific pair of texts into each session, on the basis of which it became possible to calculate the agreement of the annotators. The percent agreement was found to be 79%, indicating an acceptable level of agreement.

3.2. The resulting data as a test set

One of the main purposes of this dataset was to be a test set for the algorithm of the text complexity assessment for Russian L2 learners. In the previous study we developed the ML system trained on 800 texts from Russian L2 textbooks and a set of linguistic features, including lexical, morphological, grammatical, and syntactic ones (Laposhina et al. 2018). The examples of linguistic features are shown in Table 2.

Group of features	**Examples of features**
Lexical	average word length; percentage of words longer than 4 syllables; lexical diversity (TTR); lexical diversity (MLTD TTR); lexical density; text coverage with a frequency list of 1000, 5000 and 10000 of the most common words from a frequency dictionary; text coverage with vocabulary lists for L2 learners; percentage of abstract words
Grammatical	percentage of each POS in text; percentage of words in the genitive case in text; percentage of verbs in finite forms in text; percentage of words with 1st person tag in text
Syntactic	average sentence length; number of adversarial conjunctions per text; number of coordinating conjunctions per text; average number of punctuators per sentence; text coverage with a list of the 500 most frequent POS trigrams

Table 2: Linguistic features for model training

We have experimented with two linear regression algorithms: ordinary least squares Linear Regression and Ridge Regression (linear least squares with l2 regularization, alpha=1.0) from the scikit-learn library. The best result was achieved by Ridge regression trained on 44 best correlation linguistic features. For the model evaluation, we implemented a twenty-fold cross-validation test that showed an accuracy of 0.82 (±0.05).

However, using standard metrics like mean absolute error and comparing the output of our regression model which is a fractional number to test data from textbooks that is an integer on a discrete scale may not be an efficient approach. For instance, text i from the test set was given from the end of the A2 textbook (so the expected level is 2). The prediction for text i is 3,18 (that may be interpreted as the beginning of B1 level). In terms of linguodidactics, it is not a big mistake (the end of A2 course vs the beginning of B1 course), but it is in terms of mean absolute error.

The present paper aims to fill this gap and provide a test set with texts smoothed on the non-discrete scale. Below are our results of comparing the metrics of the same regression model with two ways of a test set annotation: standard discrete CEFR levels and Elo-based non-discrete levels. Importantly, the texts from this dataset were not used in the model training process.

To evaluate the accuracy of the regression model, which involves comparing actual and predicted values, a widely used approach is to calculate the correlation between the two sets of data. In this study, the Elo-based level scale shows a higher correlation coefficient and a lower mean absolute error compared to the CEFR level scale. Both correlations are statistically significant with p-values less than 0.05 (see Table 3).

Type of complexity scale	**Pearson's correlation coefficient with predicted level**	**p-value**	**Mean absolute error**
CEFR levels	0.81	< 0.05	0.85
Elo-based levels	0.86	< 0.05	0.77

Table 3: Pearson correlation coefficient and mean absolute error values of the predicted and observed levels depending on the chosen scale

To gain more detailed understanding of the comparison results, we analyzed the extent of the discrepancy between the expert opinions and the mathematical model predictions. The severity of the error is dependent on the magnitude of the difference between the expert opinion and the model result. For instance, an error of 0.5 signifies that the model was incorrect by half of a level, which is an acceptable margin of error, as it falls within the range of variation among expert opinions. An error of 1 level or greater suggests more significant discrepancies that require our attention. To estimate the overall magnitude of the prediction error, we used the mean absolute error metric. For the dataset analyzed in this study, the mean absolute error value was 0.77, indicating that, on average, the model's predictions are off by one level. Interestingly, the model tended to overestimate complexity levels in 30% of cases, while underestimating them in 70% of cases. Table 4 displays the distribution of absolute errors between predicted values obtained from a standard CEFR-level-based dataset and an Elo rating system dataset.

Absolute error	**Percent of cases, Elo dataset**	**Percent of cases, CEFR levels**
0-0.5 (good prediction)	38 %	41%
0.51-1 (acceptable prediction)	32%	27%
1.01 - 2 (wrong prediction)	28%	25%
> 2 (dramatically wrong prediction)	2%	7%

Table 4: The proportion of values of the average absolute error of the regression model on the resulting dataset

We consider a difference of less than 0.5 between the predicted and actual values as a correct prediction, which constitutes the majority of cases (38%). The difference greater than 0.5, but within the same level, is an acceptable quality prediction, which represents 32% of the cases. Overall, the model provides correct predictions in 70% of the cases, while being off by more than one level in the remaining 30% of the cases.

The enhanced interpretability of the error report is noteworthy. Now the absolute error distance means the real distance of the text complexity value from the level marked by the experts. This is especially important at the boundaries between levels. For instance, if a text designed for a course ending at level B1 is incorrectly predicted as a text belonging to the subsequent level B2, it will be classified as an error not in an entire level, but rather in a few tenths. The dataset is in the public domain and can be used for scientific purposes.

4 Discussion and conclusion

Construction of suitable datasets is a crucial challenge in the practical implementation of machine learning models, including the L2 linguodidactics field. Given that every language constitutes a multi-faceted living system, any categorization and partitioning into discrete levels are somewhat arbitrary. In this research, we proposed the method of creation of a dataset of texts ranked along the continuous scale of complexity for L2 learners based on CEFR levels. To accomplish this, we relied on the pairwise evaluation of the text complexity by experts and processed the resultant annotations using Elo rating system. This approach provides a non-discrete scale of text complexity, which is more in line with the view of the text complexity as a continuum of difficulty.

Among the limitations of the method, we note the small size of the dataset, which makes it possible to consider it only as a test set, but not a training data collection. Secondly, the assessment of the annotator agreement posed certain challenges. Since the main idea of the method is to compare the text with as many other texts as possible, and pairs of texts for comparison are formed randomly, there are very few identical pairs of comparisons on the basis of which the annotators' agreement can be calculated, unless it is set algorithmically, as was done in this study.

Acknowledgements

The article was prepared in full within the state assignment of Ministry of Education and Science of the Russian Federation for 2020–2024 (No. FZNM-2020-0005).

References

[1] Batinic, D., Birzer, S. Developing an English Language Placement Test for Undergraduate Students: A Crowdsourcing Approach // Educational Technology Society, 18(4), P. 259–271.

[2] Chen, X., Bennett, P. N., Collins-Thompson, K., Horvitz, E. (2013). Pairwise ranking aggregation in a crowdsourced setting. In Proceedings of the sixth ACM international conference on Web search and data mining (pp. 193–202). ACM.

[3] Clercq, O. D., Hoste, V., Desmet, B., Van Oosten, P., De Cock, M., Macken, L. (2014). Using the Crowd for Readability Prediction. Natural Language Engineering, 20(3), 293–325.

[4] Corlatescu, Dragos, Ștefan Ruseti Mihai Dascalu. (2022). ReaderBench: Multilevel analysis of Russian text characteristics. Russian Journal of Linguistics, 26(2), 342–370. https://doi.org/10.22363/2687-0088-30145

[5] Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors / B. North, E. Piccardo, T. Goodier. – Strasbourg: Council of Europe Publishing, 2018. – 227 p.

[6] Elo, A. E. (1978). The rating of chessplayers, past and present. Arco Pub.

[7] Francois, T., Fairon, C. (2012). An 'AI readability' formula for French as a foreign language. In Proceedings of the 2012 Joint Conference on Empirical methods in natural language processing and computational natural language learning (pp. 466–477).

[8] Jue Hou, Maximilian W. Koppatz, Jose Mar´ıa Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, Roman Yangarber. (2015). Modeling language learning using specialized Elo ratings. International Journal of Artificial Intelligence in Education, 25(1), 1–19.

[9] Karpov, N., Baranova, J., Vitugin, F. (2014). Single-sentence readability prediction in Russian. In Proceedings of Analysis of Images, Social Networks, and Texts conference (AIST) (pp. 91–100).

[10] Kate, R. J., Luo, X., Patwardhan, S., Franz, M., Florian, R., Mooney, R. J., Roukos, S., Welty, C. (2010). Learning to Predict Readability using Diverse Linguistic Features. In Proceedings of the 23rd International Conference on Computational Linguistics (pp. 546–554). Association for Computational Linguistics.

[11] Laposhina, A. N. (2018). Insights from an experimental study on the text complexity for Russian as a foreign language. In Proceedings of the VI Congress of ROPRYAL (pp. 1544–1549). ROPRYAL.

[12] Laposhina, A. N., Veselovskaya, T. S., Lebedeva, M. U., Kupreshchenko, O. F. (2018). Automated Text Readability Assessment For Russian Second Language Learners. In Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference «Dialogue 2018» (Issue 17 (24), pp. 396–406). Moscow.

[13] Mangaroska, K., Vesin, B., Giannakos, M. (2019). Elo-Rating Method: Towards Adaptive Assessment in E-Learning. In Proceedings of the 19th IEEE International Conference on Advanced Learning Technologies (ICALT) (pp. 380–382). IEEE.

[14] Ontaelio, O. (2016, may 19). Count the invisible: how to reliably test the vocabulary [Soschitat' nezrimoe: dostoverno opredelyaem slovarnyj zapac]. Habr.ru. https://habr.com/ru/companies/skyeng/articles/301214/

[15] Pelanek, R. (2016). Applications of the Elo Rating System in Adaptive Educational Systems. Computers & Education, 98, pp. 169-179.

[16] Pitler, E., Nenkova, A. (2008). Revisiting Readability: A Unified Framework for Predicting Text Quality. In Proceedings of the Conference on Empirical Methods in Natural Language Processing - EMNLP ’08, page 186, Honolulu, Hawaii. Association for Computational Linguistics.

[17] Reynolds, R. (2016). Insights from Russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. In Proceedings of the 11th Workshop on the Innovative Use of NLP for Building Educational Applications

[18] Schwarm, S. E. Ostendorf, M. (2005). Reading level assessment using support vector machines and statistical language models. In Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics (ACL ’05). Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 523–530.

[19] Seiffe, L., Kallel, F., Naderi, B., Moller, S. Roller, R. (2022). Subjective Text Complexity Assessment for German. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 707–714 Marseille, 20-25 June 2022.

[20] Sharoff S., Kurella S., Hartley A. (2008). Seeking needles in the web’s haystack: Finding texts suitable for language learners. In Proceedings of the 8th Teaching and Language Corpora Conference, (TaLC-8), Lisbon, Portugal.

[21] Solovyev V., Ivanov V., Solnyshkina M. (2018). Assessment of Reading Difficulty Levels in Russian Academic Texts: Approaches and Metrics. 3049–3058.

Whose word? Problems of lexicographic representation of ideologically marked words (the lexicon of the Russian-Ukrainian conflict)

Levontina I. B.
Vinogradov Russian Language Institute of the Russian Academy of Sciences / Volkhonka 18/2, Moscow
irina.levontina@mail.ru

Shmeleva E. Ya.
Vinogradov Russian Language Institute of the Russian Academy of Sciences / Volkhonka 18/2, Moscow
eshkind@mail.ru

Abstract

The article deals with the problems of presenting ideologically marked words in the dictionary. It is based on the analysis of the words that appeared in the Russian language or received new meanings during the Russian-Ukrainian conflict. The difficulty of the lexicographic representation of such words is that their evaluative potential is mobile, for example, offensive nicknames can be assimilated by "offended" ones and become neutral words. Ideologically marked words can either exist in the lexicon for a long time or be quickly replaced by other lexical units. Therefore, in the interpretation of ideologically marked words, it is advisable to indicate the approximate time of their existence. In addition to temporary indicators, in the dictionary entry of such words, it is necessary to indicate whose word it is, that is, on whose behalf an assessment is given to a person or event. Since we believe that explanatory dictionaries should contain not only common names, but also proper names, the article also discusses geographical names.

Keywords: lexicographic representation, ideologically marked words, one's own and someone else's words, appropriation of someone else's word

DOI: 10.28995/2075-7182-2023-22-287-294

Чье слово? Проблемы лексикографического представления идеологически маркированных слов (лексика российско-украинского конфликта)

Левонтина И. Б.
Институт русского языка им. В. В. Виноградова РАН / Волхонка 18/2, Москва
irina.levontina@mail.ru

Шмелева Е. Я.
Институт русского языка им. В. В. Виноградова РАН / Волхонка 18/2, Москва
eshkind@mail.ru

Аннотация

В статье рассматриваются проблемы, возникающие при представлении в толковом словаре идеологически маркированных слов. Материалом статьи послужили слова, появившиеся в русском языке или получившие новые значения во время российско-украинского конфликта. Трудность лексикографического представления таких слов состоит в том, что их оценочный потенциал подвижен, например, обидные прозвища могут осваиваться «обиженными» и становиться нейтральными словам. Идеологически маркированные слова могут как существовать в языке длительное время, так и быстро заменяться другими – естественным путем или приказом «сверху». Тем самым в толковании идеологически маркированных слов желательно указывать приблизительное время их бытования. Помимо временных, в толкованиях таких слов нужно указывать «партийные» характеристики– чье это слово, от чьего имени дается оценка человеку или событию. Поскольку мы считаем, что в толковых словарях должны быть не только нарицательные, но и собственные имена, в статье рассматриваются также географические названия, которые ощущаются как свои или чужие конфликтующими сторонами.

Ключевые слова: лексикографическое представление, идеологически маркированные слова, слова свои и чужие, присвоение чужого слова

1 «Партийная принадлежность» как элемент лексикографического описания

В теоретической семантике и в лексикографии подробно изучается фигура наблюдателя (см., например, [Апресян 1995: 639–644], [Падучева 2018: 48–73]). Обсуждается положение наблюдателя в пространстве, его личные оценки и т. д. Однако оказывается, что для ряда слов важна и «партийная» принадлежность наблюдателя. Для лексикографии такие слова составляют серьезную проблему. Как писал И. А. Бодуэн де Куртенэ в Предисловии к новому, исправленному и дополненному изданию словаря Даля: «Если и в обыкновенное мирное время значение слов постоянно меняется и разнообразится, смотря по принадлежности индивидов не только к той или иной местности, но даже к тому или иному сословию, классу общества и даже “партии”, – то тем необходимее далеко идущие изменения значений слов в только что пережитое и еще до сих пор переживаемое <…> время. Раз словарь претендует на относительную полноту, он должен по возможности принимать в соображение и это разнообразие, не менее важное, чем, например, разнообразие по местным говорам» [Бодуэн де Куртенэ 1904: IX].

Как известно, тираж первого тома Толкового словаря русского языка под ред. Д. Н. Ушакова (далее Словарь Ушакова), вышедший в 1934 году, был фактически уничтожен за объективизм, «беззубость» и «политическую незаостренность». В цензурированном издании, первый том которого был опубликован в 1935 году, в толковании многих слов «политическая заостренность» уже вполне представлена. Так, слово *либерализм* описано с совершенно определенной точки зрения: «1. Система политических идей, взглядов и стремлений, свойственная идеологам промышленной буржуазии эпохи ее подъема, отстаивающая, трусливо и непоследовательно, политические свободы в интересах «свободы приобретения» и эксплуатации пролетариата. 3. Буржуазное щегольство терпимостью, свободолюбием (устар.). 4. Преступная снисходительность, попустительство (нов. неодобр.)» [выделено нами]. У слова *государство* на месте нейтрального определения «страна, управляемая своим собственным правительством» появилось два значения – «наше» и «их» государство. Это касается не только собственно политических терминов, ср., например, правку статьи слова *быдло*. В издании 1934 г. **БЫДЛО**, а, *ср.*, чаще *собир.* [польск. bydło – скот] (обл., бран.). О тупых, безвольных людях, покорных насилию. В «Дополнениях и поправках к I тому» она звучит уже так: «…В устах помещиков-крепостников – презрительное обозначение крестьянской массы как безвольного, бессловесного и покорного стада, опекаемого помещиком». Эта тенденция сохранялась и в последующих словарях. Например, хотя Словарь русского языка С. И. Ожегова (далее Словарь Ожегова) во многом сохранял преемственность со Словарем Ушакова, его толкования также отражают «линию партии»: в нем гораздо более нейтрально описаны многие политические термины, зато добавлена, например, отрицательная оценка в слове *космополитизм*. Словосочетание *Белая гвардия* в Словаре Ушакова толкуется как «(полит.) контрреволюционные войска»; в Словаре Ожегова как «общее название контрреволюционных войск в Гражданскую войну», а в Толковом словаре русского языка С. И. Ожегова, Н. Ю. Шведовой (издание 1992 года): «в годы гражданской войны: общее название русских военных формирований, боровшихся за восстановление законной власти в России». Интересно, что в «Большом толковом словаре русского языка» под ред. С. А. Кузнецова (далее БТС, издание 2014 г.) авторы выбирают нейтральное определение «общее название контрреволюционных войск в Гражданскую войну в Советской России в 1918–1920 гг.»

2 О слове *каратель*

Ощущение слова как «своего» или «вражеского» активизируется во время социальных и политических конфликтов. Показательна история слова *каратель* (об этом слове см. также [Левонтина 2021: 418–421]).

Сам глагол *карать* и его производные ведут себя по-разному. *Карать* указывает на *справедливое возмездие*, часто исходящее от *высшей силы*. Различия между глаголами *наказывать* и *карать* подробно описаны Ю. Д. Апресяном в «Новом объяснительном словаре синонимов русского языка». Те же идеи выражаются и словом *кара*. Если преступник застрелил заложника, который попытался бежать, — это не *кара*. В прилагательном *карательный* эти идеи представлены слабее: идея власти есть, а идея правоты частично или совсем стерлась — *карательная психиатрия,*

карательные органы. А вот существительное *каратель* полностью изменило оценку. Первоначально и оно подразумевало тот же круг представлений, что и *кара*, и *карать*:

(1) *Центром дома был папа. Он являлся для всех высшим авторитетом, для нас — высшим судьею и карателем* (В. В. Вересаев. «В юные годы», 1927).

В Гражданскую войну словом *каратель* стали называть военных, терроризирующих мирное население, причем использовалось оно с разных сторон: встречаются и *колчаковские каратели*, и *большевистские каратели*. В Великую Отечественную войну *каратели* – это фашисты, которые уничтожают целые деревни за помощь партизанам[1].

В последнее время оценочный потенциал слова *каратель* активно и разнообразно используется в пропагандистских целях. Журналистка Анна Наринская в связи с белорусскими (белорусскими) протестами 2020–2021 гг. написала на своей странице в Фейсбуке: «Совершенно уверена, что важным рычагом состоятельности белорусского протеста стало повсеместное переименование “силовиков” — омоновцев, нацгвардейцев, милиции и т. д. — в “*карателей*” (это, по-моему, запустил канал NEXTA, но в любом случае это прижилось). <…> Они — *каратели*. Это их название. Язык проясняет и определяет многое. И дает важнейшее для таких ситуаций разделение на “мы” и “они”. / Мы — это мы. А они — *каратели*».

Интересно, что на украинской почве судьба этого слова еще более сложная. Здесь ввести слово *каратели* в обиход попыталась как раз российская пропаганда, в одном ряду с *бандеровцами, хунтой* и др. (см. [Yavorska 2020]), а с другой стороны, это слово употребляется проукраинской стороной по отношению к российским военным.

Итак, в полном толковом словаре русского языка нужно отразить несколько режимов употребления слова *каратель*: устаревший с положительной оценкой и более поздние в нескольких исторических контекстах по отношению к противнику с резко негативной оценкой. Лексикографический формат представления этой информации может быть разным в зависимости от конкретного словаря – это могут быть пометы, исторические комментарии или примеры употребления с пояснениями.

Следует уточнить, что сказанное относится не только к словам, но и к устойчивым сочетаниям, см. ниже о выражениях *белое пальто* и *хороший русский*.

3 Российско-украинский конфликт: оценочные слова

В русском языке российско-украинского конфликта уже накопился целый словарь оценочных слов[2]. Негативные обозначения Украины и украинцев (*Хохляндия, Бандерштат, Укропия, хохлы, укры, укропы, салоеды, бандеры, бендеровцы, укронацисты, укрофашисты*) используются давно, а *майдауны* и *майданутые* даже уже не так частотны – Майдан у многих россиян изгладился из памяти. Как и саркастическое обозначение *свидомые* [Reuther 2016]. С другой стороны, негативное наименование россиян *ватники*, *колорады* (по расцветке георгиевской ленточки, особенно популярные в 2014) тоже теперь встретишь не так часто, зато про россиян говорят *орки, рашисты*.

Особая трудность лексикографического представления таких слов состоит в том, что их оценочный потенциал может быстро меняться. В частности, прозвища и обидные названия иногда присваиваются (это хорошо известно по словам *санкюлоты*, *декаденты* и др.), однако присваиваются не всегда. Например, наименование *укроп* легко присвоилось, и украинцы даже стали делать себе нашивки и майки с изображением этого растения. До некоторой степени присвоились и наименования *ватник*, *ватница*: люди говорят о себе *я ватник*. В связи с обсуждением этого слова был даже проведен модный показ моделей ватников. При этом слова *вата* и *колорады* не обнаруживают тенденции к присвоению. Слова сосуществуют, одни вытесняют другие, а иногда одни директивно заменяют другие, как было со словами *ополченцы* и *сторонники федерализации* (они же в языке другой стороны – *сепары*). Желательно при лексикографическом представлении таких слов учитывать все эти обстоятельства. Конечно, конкретные даты указать трудно, но следует отмечать хотя бы приблизительное время бытования слова.

[1] Впрочем, в интернете есть упоминания бронемашины «Каратель» (предположительно для ФСБ).

[2] В украинском языке, конечно, тоже есть аналогичный словарь, но мы его здесь не рассматриваем.

В этом отношении показателен случай с подачей слова *хунта* в БТС, который размещен на портале gramota.ru. Там сообщается, что словарь публикуется в авторской редакции 2014 года. Четвертое значение слова *хунта* на Грамоте выглядит так: «4. Разг. Группа лиц, действующих по какому-л. соглашению, сговору (обычно с неблаговидными целями). *В Институте действует х*». Между тем, многие люди помнят, что уже в 2014 году описание слова выглядело иначе и включало указание на «Киевскую хунту», пришедшую к власти в результате «переворота» 2014 года, что вызвало тогда бурную реакцию пользователей. Действительно, в издании БТС 2016 года (Китай: The Commercial Press) читаем: «*хунта* 4. Разг. Группа лиц, действующих по тайному соглашению, сговору с неблаговидными целями. *Киевская х.* (о людях, незаконно пришедших к власти на Украине в результате государственного переворота в конце февраля 2014 года)». Оказывается, такое описание было и в авторской редакции 2014 года, но этот фрагмент статьи был быстро с портала удален (с согласия автора), однако попал в следующие издания. Этот случай очень характерен. Безусловно, слово *хунта* имеет соответствующие употребления, однако очевидно, что употребить это слово таким образом могут лишь люди определенных политических воззрений[3]. При описании подобных идеологически маркированных слов следовало бы снабжать их указанием на то, с какой группой себя ассоциирует говорящий. Выражение «Киевская хунта» уместно в речи противников Евромайдана (ср. уже упоминавшуюся нами формулировку из Словаря Ушакова «в устах…»). При этом за прошедшие годы выражение «Киевская хунта» начинает устаревать и в речи тех же людей сменяется другими выражениями, например, «нацистский режим»:

(2) *Очередной расстрел российских военнопленных в значительной мере лежит на совести западных кураторов Киева, которые создали этот нацистский режим, вырастили поколение украинцев, одержимых идеей ненависти и национального превосходства* (https://www.vedomosti.ru/politics/news/2023/02/09/962398-mid-vozlozhil-vinu-na-kuratorov)[4].

4 Географические названия

Не только жители России и Украины, но и сами страны получают в конфликтном дискурсе оценочные наименования. Украина – это *Укропия* или *Бандерштат*, а Россия – *Рашка*, а также *Эрэфия, Орда, Мордор, Оркостан*, а «в России» – это *на болотах*, *на России* или *за поребриком*. *На России* стали массово говорить не так давно, это последствие дискуссии по поводу *в Украине / на Украине*. Выражение *за поребриком* идет от знаменитого донбасского видео 2014 года, где «ополченец» и «сторонник федерализации» кричит: «За поребрик отойдите!» (http://youtu.be/qmxBjsU2rig; ср. [Левонтина 2021: 317–319]), что там было воспринято как доказательство российской принадлежности *ихтамнетов.* В считанные дни у слова появилось и новое значение: *поребриками* стали называть тех, кого в Крыму звали «зелеными человечками» («*Поребрики* незаконно подключили российские телеканалы»). Позже Россию с украинской стороны стали называть *за поребриком*, а потом *Запоребрией*.

Надо заметить, что вообще наименование географических объектов – это важный фронт языковой войны. На этом построена фабула рассказа Александра Солженицына «Случай на станции Кочетовка», где молоденький лейтенант Зотов сначала симпатизирует отставшему от своего эшелона ополченцу Тверитинову, а затем мгновенно разочаровывается в нем и сдает Тверитинова НКВД, ср, диалог:

(3) *— Это считайте уже под Сталинградом.*
— Под Сталинградом, — кивнул Тверитинов. Но лоб его наморщился. Он сделал рассеянное усилие и переспросил: — Позвольте… Сталинград… А как он назывался раньше?
<…>
— Раньше он назывался Царицын. (Значит, не окруженец. Подослан! Агент! Наверно, белоэмигрант, потому и манеры такие.)

[3] https://odnarodyna.org/article/ukrainskaya-khunta-trebuet-ne-nazyvat-eyo-khuntoy
[4] В цитатах из Интернета сохраняется орфография и пунктуация источника.

— Ах, верно, верно, Царицын. Оборона Царицына.
(Да не офицер ли он переодетый? То-то карту спрашивал…)

Через переименование происходит символическое присвоение пространства. При территориальных конфликтах одни и те же земли называются разными сторонами (*временно*) *оккупированные территории, освобожденные территории, присоединенные территории, новые территории* (сейчас становится в России официальным наименованием) и даже просто *территории*.

Кстати, еще в 2014 году можно было определить позицию говорящего по тому, произносит ли он *Юго-восток Украины* или *восток Украины*. Обозначение *юго-восток* было географически неточным, но «пожелательным» для сторонников «Новороссии». Кроме того, хотя и юг, и юго- восток – это просто разные направления, но *юго-восток* звучит отчасти как «юг и восток», то есть несколько внушительнее.

В интернете постоянно идут дискуссии: *– Правильно Днепропетровск – Нет, правильно Днепр! – Нет, ПРАВИЛЬНО Днепропетровск*. В последние месяцы идет битва за донецкий город Бахмут, который в 1924 году был переименован в Артёмовск в честь советского государственного деятеля Сергеева, известного по псевдониму Артем, а в 2016 году снова переименован в Бахмут. Военное противостояние отражается и в языке военных сводок. Российские источники последовательно называют этот город Артемовск, а украинские – Бахмут. Когда речь идет о решениях украинских властей одни российские издания используют стратегию de dicto **Error! Reference source not found.**, то другие – стратегию de re (5)[5].

(4) *Объяснен смысл решения Зеленского «продолжать оборонять» Бахмут* (mk.ru›…resheniya-zelenskogo…oboronyat-bakhmut.html).

(5) *Зеленский: сражение за Артемовск стало одним из самых тяжелых для украинской армии* (https://news.rambler.ru/conflicts/50327619-zelenskiy-srazhenie-za-artemovsk-stalo-odnim-iz-samyh-tyazhelyh-dlya-ukrainskoy-armii/?ysclid=lfa9m5z4vy352569661).

Но в целом очевидно, что дать «правильное» название этому городу противоборствующим сторонам не менее важно, чем его захватить, ср.:

(6) *До городской администрации Бахмута осталось 670 метров. Потом он будет Артемовском* (https://vk.com/wall-40316705_47746800?ysclid=lfa9y1uv9493693711).

Вообще нужно заметить, что русская лексикография традиционно упускает из виду имена собственные, в частности географические названия, хотя в других лексикографических традициях имена собственные помещаются в словаре наряду с нарицательными. Наверно, будет правильно включать имена собственные и в русские толковые словари, как это сделано в поздних изданиях Грамматического словаря А. А. Зализняка. И в этих словарях при таких географических наименованиях, как *Бахмут* / *Артемовск*; *Санкт-Петербург* / *Петроград* / *Ленинград*, *Питер*; *Прибалтика* / *Страны Балтии* тоже понадобятся временные и «партийные» характеристики. Причем в толковом словаре важно учитывать не только энциклопедические сведения об официальных переименованиях, но и те оценки, которые люди вкладывают в выбор того или иного из существующих параллельно наименований.

5 «Свое» и «чужое» слово

Было бы неверно считать, что речь идет о русском языке только по разные стороны фронта или, как теперь говорят, *линии соприкосновения*. Внутри России тоже используются слова разной «партийной принадлежности». Например, *аналоговнетный* (об оружии) – это слово противника вооруженного конфликта (да и вообще современного российского государства), не верящего

[5] При стратегии de dicto говорящий использует номинации, которые счел бы адекватными и субъект передаваемого мнения; при номинации de re говорящий все переименовывает в соответствии со своими представлениями о реальности [Шмелев 1997: 472].

российской пропаганде. А вот *тик-ток войска* – выражение, которое могут употребить обе стороны, причем по отношению к одному и тому же объекту.

Выражение *за ленточкой* – на территории противника или на передовой чаще встречается в чатах, в которых родственники *мобиков* (мобилизованных) и *чмобиков* (мобилизованных во время так называемой «частичной мобилизации») обсуждают, где сейчас их мужья и сыновья, чем им можно помочь, ср.:

(7) *После начала частичной мобилизации среди призывников и их близких распространилось выражение «за ленточкой»* (barnaul.bezformata.com›listnews…za-lentochkoy…).

Если, например, *заминусовать* и *задвухсотить* – обозначения гибели врага, которые обе стороны могут использовать вполне симметрично, то *отправиться / попасть на концерт Кобзона* или *вернуться в (черном) пакете* в этом отношении сильно маркированы. Интересно, что обозначения *мясо* и *фарш*, при всей их чудовищности, вполне используются по отношению к российским солдатам и в российском z-дискурсе – например, в контексте критики стратегии «Бабы новых нарожают», а также по отношению к мобилизованным или *ЧВК со стороны кадровых военных* (ср. выражение из современного военного жаргона *мясная атака*).

5.1 *Завойнисты*, *СВОшники*, *зетники*, *Z-патриоты* vs *нетвойнисты*, *х**войнисты*, *нетвойняшки*

Особенно показательны в этом отношении обозначения сторонников и противников *СВО*, которые появились практически сразу же после ее начала. Они сильно различаются с точки зрения оценочного потенциала. Слова *СВОшники* и *Z-патриоты* вполне могут присваиваться:

(8) *Ничего хорошего господа скулящие «нетвойняшки» «украинаненападала» и прочие скулящие сочувствующие бандеритату, а также «всепропало», сдальщикихерсона и остальным паникёрам да оскорбляющие СВО, от нас СВОшников не ждите нам уже известны ваши домашние адреса …* (9111.ru›questions/7777777772053565/).

(9) *Уважаемые участники конкурса! Результаты Всероссийского конкурса патриотического рисунка «Z патриот» будут объявлены 31 октября 2022г.* (zpatrioticpictures.ru).

Так же, как другой стороной охотно присваивается обозначение *х**войнисты*. Замечательно, что хотя это слово очень грубое, негативная оценка в нем не так закреплена. Конечно, оно употребляется и в осуждение, но может использоваться и одобрительно, поскольку связывается и с известным лозунгом, и с названием антивоенного телеграм-канала.

Но есть слова, которые не присваиваются, как, скажем, слово *всёнетакоднозначники*. Таково и слово *нетвойняшка*, которое выражает презрительное отношение к противникам войны:

(10) *Нетвойняшки задрожали,*
С самокатиков упали.<>
Все давай чихать, хромать –
Не желают воевать!
newostrie.ru›Блог›Военкомище

(11) *В минувшие выходные «нетвойняшки» снова пытались нетвойнять, проявляя как обычно недюжинные умственные способности* (vk.com›wall7382371_6517)

См. также пример (8). Как видно из этих примеров, слово *нетвойняшка* отражает вполне определенную концепцию: против *СВО* выступают благополучные москвичи, либеральные интеллигенты, хипстеры и так называемые *креаклы* (креативный класс). Они изображаются достойными всяческого презрения инфантильными трусами. Кстати, такая же модель насаждалась во время «Снежной революции» 2010–2011 – «революция норок» (в том смысле, что на протесты выходят дамочки в норковых шубах) [Левонтина 2021: 559–561].

Идея, что «против» изнеженные мальчики, а «за» - настоящие мужики, которую, кстати, пытались проводить также в социальной рекламе (*Мальчики уехали, мужчины остались*), отразилась и в концепте «Кузьмича» так и не завоевавшем особой популярности[6]. Кузьмичами в пропагандистских текстах с симпатией именуют мужчин средних лет, которые не достигли особого социального успеха в мирной жизни, но нашли себя в условиях *СВО*.

В большом толковом словаре такие слова, как *креакл*, *нетвойняшка*, *Кузьмич* требуют не только традиционных стилистических помет типа *презр*., но и хронологических, и «партийных» характеристик. Кроме того, для слов *креакл* и *нетвойняшка* можно указать предположительное авторство (П. Пряников и Н. Осипова соответственно).

5.2 *Белое пальто* и *хороший русский*

Конечно, язык отражает и более частные смысловые противопоставления в каждом из лагерей. Так, в лагере *нетвойнистов* есть, например, интересные выражения *белое пальто* и *хороший русский*.

Выражение *белое пальто* само по себе не связано с темой *СВО*. Оно восходит к иронической фразе, приписываемой Валерии Новодворской: «А я в белом пальто стою красивая»; на значение повлиял и анекдот «А я в белом фраке». *Белое пальто* – выражение, которым сейчас описывают человека, считающего себя носителем идеального нравственного чувства и камертоном коллективной совести. В последний год это в первую очередь русские люди, которые постоянно указывают другим русским людям, что те недостаточно каются, неправильно страдают и ели на масленицу блины.

Интересна судьба выражения *хороший русский*. Оно пошло от неловкого высказывания Гарри Каспарова, который имел в виду, что среди русских людей есть «нетвойнисты», которые оказались в трудном положении и заслуживают поддержки. Это выражение очень быстро стало ироническим, причем его уже употребляют не только противники, но и сторонники *СВО*:

(12) а я такой хороший русский,
меня наказывать нельзя.
<…> мне Гарри Кимович Каспаров
дал отпущение грехов.
(Хор:) Смотри: подписано Каспаров,
Пономарев и Альфредкох.
Игорь Петров
Хороший русский романс

Отметим, что выражения *белое пальто* и *хороший русский* имеют разный статус в языке и должны по-разному подаваться в словаре. Коллокация *белое пальто* (при какой вокабуле она будет подаваться зависит от принципов словаря) уже полностью устоялась и употребляется достаточно широко. Но все же сведения о его бытовании в русском языке желательно приводить в словарной статье. Что же касается выражения *хороший русский*, то оно вообще непонятно вне конкретного исторического контекста.

6 Заключение

Как мы пытались показать, для многих слов адекватное лексикографическое представление невозможно без описания того, в контексте какого конфликта, какой стороной и в какое время это слово используется. И хотя Бодуэн де Куртенэ писал, что «…полное беспристрастие в этом отношении почти не мыслимо, ибо и учёные вообще, и языковеды в частности – прежде всего люди, и как люди, они даже в научных своих трудах не могут не отражать своих личных взглядов, считаемых ими правильными и справедливыми» [Бодуэн де Куртенэ 1904: IX]. хороший словарь должен сохранить речь всех сторон, при этом не обезличивая словоупотребление и не расставляя

[6] *«Незатейливый человечишка, обиженный судьбою». Почему история о «кузьмичах» не ложится на народное сознание* (https://www.fontanka.ru/2022/11/30/71858834/).

оценок. Как кажется, что арсенал современной лексикографии представляет для этого достаточно возможностей.

Благодарности

Мы благодарны Алексею Шмелеву, указавшему нам на высказывания Бодуэна де Куртенэ процитированные нами в данной статье, а также оказавшему помощь при подготовке окончательного текста статьи.

References[7]

[1] Apresyan Yu. D. (1995), Deixis in vocabulary and grammar and a naive model of the world [Deiksis v leksike i grammatike i naivnaya model' mira], Selected Works. Vol. 2. Integral description of the language and system lexicography, [Izbrannye trudy. Tom 2. Integral'noe opisanie yazyka i sistemnaya leksikografiya]. Moscow, pp. 629-650.
[2] Boduen de Kurtene I. A. (1904), Preface to the new, corrected and supplemented edition of Dahl's dictionary. [Predislovie k novomu, ispravlennomu i dopolnennomu izdaniyu slovarya Dalya.], St. Petersburg – Moscow.
[3] Levontina I B. (2021), Word of honor, [Chestnoe slovo], Moscow, 576 p.
[4] Paducheva E. V. (2018), Egocentric language units,.[Egotsentricheskie edinitsy yazyka], Moscow, 440 p.
[5] Reuther, Tilmann. (2016), Kolorady, maidauny, separy, ukry, dvukhsotyi, truba, «krokodil»: zur lexik des russischen im ukrainekonflikt, Wiener Slawistischer Almanach, Band 77, pp.301-328.
[6] Shmelev A. D. (1997), Techniques of linguistic demagogy. Appeal to reality as a demagogic technique, [Priemy yazykovoi demagogii. Apellyatsiya k real'nosti kak demagogicheskii priem], Bulygina T. V., Shmelev A. D. Linguistic conceptualization of the world (based on Russian grammar), [Yazykovaya kontseptualizatsiya mira (na materiale russkoi grammatiki)], Moscow, pp. 461-477.
[7] Yavorska, Galina (2020), "Karateli" v Ukraïni i Bilorusi. An example of linguistic resistance, [«Карателі» в Україні і Білорусі. Приклад лінгвістичного опору], Access mode: https://www.academia.edu/43986882/_Карателі_в_Україні_і_Білорусі_Приклад_лінгвістичного_опору

References[8]

[1] Апресян Ю. Д. Дейксис в лексике и грамматике и наивная модель мира // Избранные труды. Том 2. Интегральное описание языка и системная лексикография. М., 1995. С. 629–650.
[2] Бодуэн де Куртенэ И. А. Предисловие к новому, исправленному и дополненному изданию словаря Даля. С.-Петербург – Москва, 1904.
[3] Булыгина Т. В., Шмелев А. Д. Приемы языковой демагогии. Апелляция к реальности как демагогический прием // Языковая концептуализация мира (на материале русской грамматики). М., 1997. С. 461–477.
[4] Левонтина И Б. Честное слово. М. 2021. 576 с.
[5] Падучева Е. В. Эгоцентрические единицы языка. М, 2018. 440 с.
[6] Reuther, Tilmann. Колорады, майдауны, сепары, укры, двухсотый, труба, «крокодил»: zur lexik des russischen im ukrainekonflikt // Wiener Slawistischer Almanach, Band 77 (2016), 301-328.
[7] Yavorska, Galina. «Карателі» в Україні і Білорусі. Приклад лінгвістичного опору. https://www.academia.edu/43986882/_Карателі_в_Україні_і_Білорусі_Приклад_лінгвістичного_опору

[7] References, Scopus version
[8] References, РИНЦ version

Parameter-Efficient Tuning of Transformer Models for Anglicism Detection and Substitution in Russian

Daniil Lukichev
HSE University, Sber
Moscow, Russia
peroprosi@gmail.com

Darya Kryanina
HSE University
Moscow, Russia
daryd388@gmail.com

Anastasia Bystrova
HSE University
Moscow, Russia
eyer89@gmail.com

Alena Fenogenova
SberDevices
Moscow, Russia
alenush93@gmail.ru

Maria Tikhonova
HSE University, SberDevices
Moscow, Russia
m_tikhonova94@mail.ru

Abstract

This article is devoted to the problem of Anglicisms in texts in Russian: the tasks of detection and automatic rewriting of the text with the substitution of Anglicisms by their Russian-language equivalents. Within the framework of the study, we present a parallel corpus of Anglicisms and models that identify Anglicisms in the text and replace them with the Russian equivalent, preserving the stylistics of the original text.

Keywords: Anglicisms, paraphrase, natural language processing, machine learning, language models, style-transfer

DOI: 10.28995/2075-7182-2023-22-295-306

Эффективное по числу обучаемых параметров обучение трансформерных моделей для задач детекции и замены англицизмов в русском языке

Лукичев Даниил
НИУ ВШЭ, Sber
Москва, Россия
peroprosi@gmail.com

Крянина Дарья
НИУ ВШЭ
Москва, Россия
daryd388@gmail.com

Быстрова Анастасия
НИУ ВШЭ
Москва, Россия
eyer89@gmail.com

Феногенова Алена
SberDevices
Москва, Россия
alenush93@gmail.ru

Тихонова Мария
НИУ ВШЭ, SberDevices
Москва, Россия
m_tikhonova94@mail.ru

Аннотация

Данная статья посвящена проблеме англицизмов в текстах на русском языке: задачам детекции и автоматического переписывания текста с заменой англицизмов на их русскоязычные аналоги. В рамках исследования мы представляем параллельный корпус, а также модель, которая выявляет англицизмы в тексте и заменяет их на русский эквивалент, сохраняя стилистику исходного текста.

Ключевые слова: англицизмы, парафраз текста, обработка естественного языка, машинное обучение, языковые модели, стилевой трансфер

1 Introduction

Language reflects the society to which it belongs. Its lexis reflects undergoing changes in political, scientific, technological, and other spheres of life. As new scientific and technical inventions emerge

regularly, new words (neologisms) are coined to denote new concepts. The English language has a vast influence in the context of globalisation, exerted by global economic, social, and cultural processes over national ones. "The English language finds itself at the centre of the paradoxes which arise from globalisation. It provides the lingua franca essential to the deepening integration of global service-based economies. It facilitates transnational encounters and allows nations, institutions, and individuals worldwide to communicate their world view and identities" (Graddol, 2006).

English nowadays is an international language of communication, business, education, and innovation. English has affected most languages in the past 100 years (Görlach, 2002b). For this reason, Görlach (2002a) called the English language "the world's biggest lexical exporter" , as most of the newly-coined words are English. Moreover, statistics show that 14.7 English neologisms are created per day[1], making English a highly productive *Source Language* (or SL, in short).

A significant number of English words are integrated into different spheres of human activity (e.g., modern and youth culture, civil and political life, IT, science, education, sports, medicine) in the form of loanwords. English borrowings (or Anglicisms), thus, form a vast lexical stratum in many languages, including Russian. However, often the meaning of these loanwords is uncertain or domain-specific and incomprehensible to people outside a particular field or social strata. Therefore, Anglicisms may impede effective communication between representatives of different generations, professions, subcultures. Furthermore, Anglicisms are inappropriate in some official and scientific discourse unless they "refer to terminology or common vocabulary recorded by explanatory dictionaries of the Russian language" (Апетян, 2011). In this regard, we frequently have to adjust our writing and speaking styles to a particular audience, social context, or formality of the occasion. In addition, the Anglicisms detection and substitution task is relevant in *Natural Language Processing* (or NLP, in short). Anglicisms often pose challenges for this sphere (for example, machine translation, rewriting and summarization, text-to-speech) as many systems are often dependent on the lexicon.

This paper presents methods for automatic Anglicism detection and their elimination via paraphrasing the original text with these loanwords replaced by their native equivalents. These methods can contribute to many NLP systems enhancing the accuracy of large language models or machine translation systems. Moreover, they can make contribution to language correction and proofreading applications. By identifying potential loanwords, the Anglicism detector can assist writers and editors in to ensure grammatical and stylistic accuracy in written content. Altogether, our models can improve the text's overall readability by replacing Anglicisms with more natural and understandable phrases in the target language. Such tools can be particularly useful in business, education, science, and journalism, where clear and effective communication is crucial. In addition, we present a parallel corpus of Anglicisms in Russian[2] and the code is available on our GitHub repository[3].

Thus, the contribution of our paper is three-fold: (I) first, we present a parallel corpus for the Anglicisms detection and their substitution with the detailed Anglicism markup, (II) we train and evaluate several models for Anglicisms detection (III) we present, several generation models for Anglicisms substitution.

The rest of the paper is structured as follows: in section 2, we overview the papers related to this research. Next, in section 3, we formally define the task. Section 4 describes the Anglicism dataset, section 5 discusses the methods we used, section 6 describes the metrics we used and the experimental setup, and section 7 presents evaluation results. Finally, section 8 concludes the paper.

2 Related work

The task of Anglicisms detection is relevant in NLP research: these words often refer to out-of-vocabulary words, and as many systems are often dependent on the lexicon, it poses various problems for machine translation, text processing, speech recognition, Natural Language Understanding, and text-to-speech synthesis (Jawahar et al., 2021), (Weller et al., 2022) (Pritzen et al., 2021). And the global trend

[1] https://languagemonitor.com/

[2] https://huggingface.co/datasets/shershen/ru_anglicism

[3] https://github.com/dalukichev/anglicism_removing

is gaining momentum: code-switching (the mixing of languages within a single conversation or text), the predominance of Anglicisms over the *Receptor Language* (or RL, in short) equivalents, the emergence of hybrid languages (e.g., Frenglish, Denglisch, Runglish, or Spanglish).

There are multiple works related to Anglicisms detection in different languages, e.g. detecting Anglicisms in Spanish (Álvarez Mellado and Lignos, 2022). The article describes the creation of an annotated corpus of Spanish text containing examples of unassimilated borrowings, which can be used to train machine learning models to identify such borrowings in new texts. The corpus has 370,000 tokens. The authors also propose several approaches to modelling unassimilated borrowings, including machine learning algorithms such as decision trees, support vector machines and rule-based systems that rely on linguistic features such as phonetics, morphology, and syntax. CRF, BiLSTM-CRF, and Transformer-based models were used to assess their performance on a new annotated corpus of Spanish newswire full of unassimilated lexical borrowings. The results of this work demonstrate that a BiLSTM-CRF model beats results produced by a multilingual BERT-based model.

Another idea for borrowed word detection is presented in (Miller et al., 2020), where the authors focus on phonological and phonotactic aspects of words in a language for the detection in monolingual word lists using such methods as Markov Models, Bag of Sounds and Neural Networks. The authors presented the idea to train a lexical language model on a dataset of annotated borrowings and then use it in detection for previously unseen word loans. The model performed well when tested on artificially generated words, but the three methods proved ineffective on a sample of actual words taken from WOLD [4]. Failure analysis shows that to achieve a positive result in the detection task, many borrowed words from a given language and coherent and consistent word properties are required. For our task, this problem was also taken into account.

Detecting Anglicisms in the Russian language has some peculiar features due to their transliteration into the Cyrillic script (comparison: Youtube [en] - ютьюб/ютуб [ru]; big data [en] - биг дата [ru]), lexicalization and some internal processes in the language (loanwords constitute an effective mechanism for word formation). The authors (Fenogenova et al., 2016) proposed an automated method for Cyrillic-written Anglicism detection based on the idea that speakers tend to preserve phonetic and orthographic properties of the borrowed words. The proposed method involves a combination of two approaches: 1) a linguistic approach based on identifying patterns of English words in Russian text, and 2) a machine learning approach that utilises a feature-based classifier to predict whether a given word is an Anglicism. Using transliteration (ru-en), phonetic transcribing(en-ru) and morphological analysis methods and various filters, authors compose a list of "unknown Anglicism" pairs. They used the Levenshtein distance (Levenshtein and others, 1966) with thresholds (2-3) to measure the similarity between two words in a pair, and the possible candidates' shortlist was created. With the help of Skip-Gram and CBOW, the list of hypotheses was shortened: if words are semantically and phonetically similar and are close in the word2vec model, they can be considered borrowings.

The substitution of Anglicisms in a text can be viewed as a paraphrasing task. In research mentioned in (Egonmwan and Chali, 2019), the authors present a new method for text paraphrasing based on the seq2seq and Transformer-based (Vaswani et al., 2017) models. As a result, the authors proposed a new TRANSEQ framework that combines the efficiency of the transformer model and seq2seq and improves the current state-of-the-art (Gupta et al., 2017) of QUORA and MSCOCO paraphrase data.

In our work, we trained the models for Anglicisms detection and their substitution using different variations of prompt-tuning techniques. The prompt-tuning method was proposed in (Lester et al., 2021). The fundamental concept of this approach involves training soft prompts, which are incorporated into the input sequence passed to the model while all other parameters of the model are frozen.

This idea was further developed in (Liu et al., 2021), where the authors introduce the concept of deep prompt tuning, which involves adding prompts in different layers as prefix tokens. In (Konodyuk and Tikhonova, 2022), the authors studied the applicability of the prompt-tuning method for the Russian language: they showed that it could be a good alternative to model training techniques.

In addition, in our research, we experiment with low-rank adaptation methods (or LoRA) proposed in

[4]World Loanword Database: `https://wold.clld.org`

(Hu et al., 2021). This method compresses the original language model into a low-rank representation that captures the essential information for the target task. This compression is achieved through a low-rank matrix factorization, which decomposes the original weight matrices of the model into two low-rank matrices. Once the low-rank representation of the original language model is obtained, the compressed model is fine-tuned on the target task using a small amount of labelled data. The fine-tuning process updates the compressed parameters of the model to suit the target task better while preserving the most important information from the original model. The authors demonstrated the effectiveness of the LoRA method in several NLP tasks. In addition, they showed that the LoRA approach generates compressed models that exhibit significantly smaller sizes than the original models while still achieving comparable or better performance on the target tasks.

3 Task Definition

In this paper, we formulate the Anglicism substitution (or elimination) problem as the task of rewriting a sentence by replacing Anglicisms with their Russian equivalents.

In our work, we define an Anglicism based on the definition of Görlach(Görlach, 2002b): "a word or idiom that is recognizably English in its form (spelling, pronunciation, morphology, or at least one of the three), but is accepted as an item in the vocabulary of the receptor language".

According to Pulcini (2012), there are different types of lexical borrowings:

1. **phrasal borrowings**: usually multi-word units or whole phrases, i.e. collocations, idioms, proverbs. (*e.g.,* "она, конечно, *бест оф зе бест*" (best of the best), "*ху из ху*" (who is who)).
2. **lexical borrowings**: words or multi-word units.
   (a) *direct*: formal evidence of the SL is detectable.
      i. loanword – borrowed from SL; meaning in RL is close to meaning in SL (*e.g.*, *голкипер* - goalkeeper, *нон-стоп* - non-stop)
      ii. hybrid – a combination of SL and RL elements (*e.g.,* (OVER-) + adv./adj.: *овер-дофига* домашки, *овер-пресный* рассказ)
   (b) *indirect*: the SL model is reproduced in the RL through native elements.
      i. Calques – reproduce the etymon in the form and meaning or meaning only.
         A. loan translation – translation of SL item into RL (*e.g.,* небоскреб - skyscraper, утечка мозгов - brain drain, промывка мозгов - brainwashing);
         B. loan rendition – compound or multi-word unit, one part of which is translated from SL and the other is a loose equivalent of the SL part (e.g., *топовый* (TOP + овый: adj.affix) блогер, *оффлайновое* (OFFLINE + овое: adj.affix) издание, *фолловить* (FOLLOW + ить: verb.affix) звезду, *фаниться* (FUN + ить +ся: verb refl.affix));
         C. loan creation – RL freely renders the SL equivalent (*e.g.*, синий чулок - blue stocking).
      ii. Semantic loans - an already existing item in the RL takes a new meaning after a SL one. (*e.g.*, обои (на экране) - wallpaper, карта - bank card)

In addition, it is noteworthy to mention such a phenomenon as *Pseudo-Anglicisms*, which are either:

- lexical units borrowed from English into another language, which have a meaning differing from the SL, and which are used in contexts and situations in which they would never appear in English (смокинг(smoking) -> dinner jacket, автостоп (autostop) -> hitch-hiking, паркинг(parking) -> parking lot));
- Russian formations created by combining English morphemes or imitating the phonetic shape of English words (e.g., фейс контроль - "face control", рекордсмен - "recordsman" - (record holder) (Дьяков, 2012).

In this paper, both Anglicisms and pseudo-Anglicisms are the objects of our interest. Therefore, examples of pseudo-Anglicisms were included in the dataset along with Anglicisms (for simplicity, we refer to both types simply as Anglicisms).

Borrowed words, as was mentioned earlier, are altered to fit the phonetic and grammatical structure of

the language. As English and Russian employ different alphabetic systems, loanwords from English are transliterated into the native Cyrillic-based writing system, where Anglicisms usually adopt the structure of the English source word and typically have the set of endings presented in Table 1.

-ер [-er]	спикер [speaker], бартер [bartender], стриммер [streamer]
-инг [-ing]	консалтинг [consulting]
-мен [-man]	спортсмен [sportsman]
-мент [ment]	энтертеймент [entertainment], истеблишмент [establishment]
-ист [-ist]	активист [activist], лоббист [lobbyist]
-зер [-ser]	мерчендайзер [merchandiser], тизер [teaser]
-изм [-ism]	расизм [racism], нарциссизм [narcissism]
-енд(энд) [-end]	уикэнд [weekend], хэппиэнд [happy end], бэкенд [backend]
-аут [-out]	таймаут [time out], камингаут [coming out], чилаут [chill out]
-ент/ант [-ent]	оппонент [opponent], резидент [resident], фигурант [figurant]
-джер [-ger]	мессенджер [messenger], тинейджер [teenager]
-бэк [-back]	флешбэк [flashback], фидбэк [feedback], хэтчбэк [hatchback]

Table 1: Anglicism endings in Russian

In the Russian language, Anglicisms usually undergo a process known as **domestication**, which poses challenges to NLP systems due to the lack of standardization and inconsistency in the usage of domesticated and non-domesticated borrowings. Domestication refers to how a language adapts foreign words or expressions to fit into its linguistic system, making them sound more natural and familiar to native speakers. This process is usually accompanied by altering the word's spelling, pronunciation, or meaning to better fit into the RL's linguistic system. In addition, the borrowed word is altered to fit the phonetic and grammatical structure of the language. For example, *софт* (software); "грозятся закидать *дизами*" (dislikes); "нужно установить обнову на *винду*" (Windows).

4 Data

To create an Anglicisms dataset, we collected 1084 sentences which contained 472 unique words from different domains. This data was collected semi-automatically from several sources (the Russian National Corpus[5], dictionaries (e.g., A.I. Dyakov's[6], dictionary of Anglicisms in Russian language, Russian Wikidictionary[7]), several Internet resources such as Kartaslov[8], Habr[9], Pikabu[10], as well as blogs and social media sources.

To create a parallel corpus, we paraphrased each sentence replacing all Anglicisms with their Russian equivalents, which were taken from multilingual dictionaries[11],[12] and Wikipedia[13]. All sentences were validated and paraphrased manually by the linguists. It should also be noted that replacing an Anglicism with a single word was not always possible. In some cases, they were substituted with collocations or set expressions (фидбэк (feedback) - обратная связь, краудфандинг (crowdfunding) - коллективный сбор средств, фандрайзинг (fundraising) - сбор средств, оффер (job offer) - предложение по трудоустройству, приглашение на работу).

Thus, we obtained a novel corpus for Anglicisms detection and substitution in the Russian Language[14]. It consists of parallel text pairs: an original sentence with Anglicisms and a sentence in which their Rus-

[5] https://ruscorpora.ru
[6] http://Anglicismdictionary.ru
[7] https://ru.wiktionary.org/wiki/РЭРѳСЂхРуР«СЃРчСS:Р№РхР«РњР«РуРчРчРёСК/ru
[8] https://kartaslov.ru
[9] https://habr.com/
[10] https://pikabu.ru/
[11] Multitran: https://www.multitran.com/
[12] Cambridge dictionary: https://dictionary.cambridge.org/dictionary/english-russian/
[13] https://ru.wikipedia.org/wiki/
[14] https://huggingface.co/datasets/shershen/ru_anglicism

Word	Form	Sentence	Paraphrase without Anglicisms
агриться	сагрилась	Пойдем пока она не сагрилась на нас.	Пойдем пока она н не разозлилась на нас.
кринж	кринжового	Ничего более кринжового я в жизни не видел.	Ничего более постыдного я в жизни не видел.
трушный	трушным	Рядом с тобой даже Джонни Бой был трушным пацаном.	Рядом с тобой даже Джонни Бой был настоящим пацаном.
слот, позер	слоты, позеры	Во дворе эти позеры заняли все парковочные слоты.	Во дворе эти притворщики заняли все парковочные места.
эпикфейл	эпикфейла	Моему злорадству по поводу эпикфейла сего сайта нет предела.	Моему злорадству по поводу провала сего сайта нет предела.

Table 2: A snippet from the Anglicism dataset.

Sentence (English)
Let's go before she gets angry at us.
That's the most cringe-worthy thing I've ever seen in my life.
Next to you, even Johnny Boy was a real kid.
In the yard, these posers took up all the parking slots.
My gloating over the epic fail of this site has no limits.

Table 3: Anglicism dataset format. Translation of the sentences from Table 2. Due to the Anglicism specifics, both sentences (with and without Anglicisms) are translated into English the same way.

sian analogues replace them. A snippet from the dataset is presented in Table 2 (the English translation of the sentences is given in Table 3).

The resulting dataset consists of 1084 sentence pairs divided into train and test parts (999 for the train part and 85 for the test part). The test part includes 30 unique Anglicisms which are not encountered in the train part.

The modest size of the dataset can be partially explained by the fact that in our work, we decided to prioritize the data quality before its quantity. That coincides with the results of the recent research (Zhou et al., 2023), which shows that a relatively small amount of high-quality data can be more beneficial than large low-quality datasets. Thus, we put additional effort into collecting data and selecting good Anglicism examples, which took additional time and resources. Namely, to ensure the annotation quality and to avoid potential errors, we avoided using such annotation services as Yandex.Toloka[15] and paraphrased all sentences with the help of professional linguists, which was more expensive and time-consuming. As a result, we obtained a relatively modest but high-quality dataset. In addition, it should be noted that we took into account the current dataset size and selected suitable methods, such as prompt-tuning and LoRA (see section 5), which can be successfully applied to such amounts of data (Konodyuk and Tikhonova, 2022).

5 Method

Our approach consists of two parts: 1) a model for Anglicisms detection and 2) a paraphrasing model, which rewrites a sentence, replacing the Anglicisms with their Russian-language equivalents.

5.1 Prompt-tuning

Both parts of the algorithm use different variations of prompt-tuning(Lester et al., 2021). Prompting is a technique that provides additional information to the language model to condition during the generation of output Y. Typically, this is achieved by adding a series of tokens P to the input X, resulting in a new input $[P; X]$. The model's parameters remain fixed while it maximizes the possibility of generating the correct Y:

[15] `https://toloka.yandex.ru`

$$Y = \arg\max Prob_\theta(Y|[P;X]).$$

The generative model incorporates the prompt tokens P, into the model's embedding table, parameterized by frozen θ. Finding an optimal prompt involves selecting prompt tokens from a fixed vocabulary of embeddings, either through manual search or non-differentiable search methods. Prompt tuning, on the other hand, enables the prompt to have its own dedicated parameters, θ_p, that can be updated. Prompt tuning involves using a fixed prompt of special tokens, with only the embeddings of these prompt tokens being updatable. In essence, prompt tuning eliminates the requirement for the prompt P to be parameterized by θ, as in traditional prompting.

There are different types of initialization of added embeddings:

1. embeddings of random words from the dictionary
2. embeddings of class labels from the task
3. random initialization (does not work well)

We use this variant of prompt-tuning for the Anglicism substitution part, applied in combination with the paraphrase decoder-based models. In our approach, embeddings of random tokens from the first layer of the model are used.

As for Anglicisms detection, we, among other approaches, deployed advanced prompt-tuning. However, in the original prompt-tuning, only continuous prompts are incorporated into the input embedding sequence, which presents two major drawbacks. First, the sequence length limitations impose constraints on the number of trainable parameters. Secondly, the impact of the input embeddings on model predictions is relatively indirect. To overcome these obstacles, *P-Tuning v2* (Liu et al., 2021) introduces the concept of deep prompt tuning, which involves adding prompts in different layers as prefix tokens. This approach enables tuning more task-specific parameters (between 0.1 and 3 per cent), providing greater per-task capacity while remaining parameter-efficient. Additionally, prompts added to deeper layers have a more direct impact on the model's predictions.

5.2 Anglicism detection

We regard the Anglicism detection problem as a token classification task. Tokens that are Anglicisms are labelled as **1**, and the remaining are labelled as **0**. For this task, we evaluated three models:

- **ruBert-tiny**[16]: a small BERT-like model;
- **ruRoberta-large**[17]: a large Russian language RoBERTa model;
- **XLM-RoBERTa**[18]: a large multilingual RoBERTa model.

Since large models tend to overfit on a small amount of data, we used different approaches for training small and large models. Namely, for small models, we used a relatively low learning rate (see section 6.2 for the details). For the large models, we implemented the P-Tuning v2 technique. In addition, we have incorporated the trained tensors into each model layer, effectively decreasing the number of trainable parameters to prevent overfitting. All together, this enables to fine-tune large models for the downstream task, even with limited data.

5.3 Anglicism substitution

We used the prompt-tuning technique to train a paraphrasing model for Anglicism substitution. The important aspect of this approach is to specify the position of trained embeddings within the model's input. In our work, we used the following types of prompts formats:

- **only sent**: <prompt> sentence with Anglicisms <prompt> its paraphrase without Anglicisms
- **sent + angl**: <prompt> sentence with Anglicisms <prompt> Anglicism <prompt> its paraphrase without Anglicisms

In the first format, the embeddings that have been trained are positioned both at the beginning of the sample and between the sentence and its paraphrase, which does not contain Anglicisms. In the

[16] `https://huggingface.co/cointegrated/rubert-tiny`
[17] `https://huggingface.co/sberbank-ai/ruRoberta-large`
[18] `https://huggingface.co/xlm-roberta-base`

second prompt format, we also pass an Anglicism as a model input together with the original sentence and the sentence paraphrase. For this approach, we need the knowledge of Anglicisms to format our examples. We used an Anglicism detector trained at the Anglicism detection stage. Namely, we utilised ruRoBERTa-large detector, which showed the best results in our experiments on Anglicism detection (see section 6 for the details). Thus, the second approach incorporates two models. The detection model identifies Anglicisms in the sentence and then feeds them, along with trained embeddings, to the input of the paraphrasing model.

We utilise the large-scale Russian language model ruGPT3-Large[19] and a multilingual GPT-based model mGPT[20].

For the low-rank adaptation approach, we add the product of two matrices with dimensions $H \times K$ to all attention layers, where H is the dimension of the hidden state of the model, and K is a small value. In our experiments, we use $K = 4$, which was motivated by the research conducted in (Hu et al., 2021).

6 Experiments

6.1 Evaluation

Anglicism detection As long as we consider the Anglicism detection task as a binary token classification problem, we use binary classification metrics (F1, precision, and recall) for evaluation.

Anglicism substitution As for the Anglicism substitution, we evaluate this part using the following metrics, which are commonly used for generative tasks and the paraphrase tasks in particular:

1. CHRF++[21](Popović, 2015)
2. BLEU score(Papineni et al., 2002)
3. Rouge-L(Lin, 2004)
4. BERTScore(Zhang et al., 2019)
5. LaBSE(Feng et al., 2020)[22]

All metrics listed above are computed between gold paraphrases and model predictions and averaged over the test set.

6.2 Experimental setup

One of the essential hyperparameters of prompt tuning is the length of the prompt. In our research, we use the following prompt lengths:

- *detection*: in our methodology, we introduce prompts of length 100 to each attention layer and optimize them using the learning rate $1e-3$. Additionally, the linear head is optimized with a learning rate of $1e-5$, with a batch size of 8 and for a duration of 10 epochs.
- *sentence-paraphrase approach*: we add a prompt of length 50 before the sentence and a prompt of length 40 between the sentence and the paraphrase. We optimize prompts with a learning rate of $1e-3$ and linear head with a learning rate of $1e-5$ with a batch size of 8 and for 5 epochs.
- *sentence-anglicism-paraphrase approach*: we add a prompt of length 50 before the sentence, a prompt of length 20 between the sentence and the Anglicism and a prompt of length 40 between the Anglicism and the paraphrase. We optimize prompts with a learning rate of 1e-3 and linear head with a learning rate of $1e-5$ with a batch size of 8 and for 5 epochs.

In low-rank adaptation approaches, the models are trained with the learning rate $1e-5$, which is kept the same for both the model and linear head parameters, using a batch size of 8 and for a total of 15 epochs. The AdamW optimizer (Loshchilov and Hutter, 2017) and linear scheduler with warm-up are employed in all the experiments.

[19] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2
[20] https://huggingface.co/sberbank-ai/mGPT
[21] https://huggingface.co/spaces/evaluate-metric/chrf
[22] https://huggingface.co/sentence-transformers/LaBSE

7 Results

7.1 Anglicism detection

Analyzing the results of Anglicism detection (see Table 4), it can be observed that ruRoberta-large shows the best quality surpassing other models in all metrics. XLM-RoBERTa also produces competitive results, while ruBert tiny performs much worse. We hypothesize that such low performance can be explained by the fact that the model was fine-tuned without prompt tuning, and even though it contains a small number of parameters, it still began to overfit too quickly on the small dataset.

The obtained results coincide with the work of (Leidig et al., 2014), where the authors tried the combination of several features (G2P confidence, grapheme perplexity, Google hits count) to detect Anglicisms in German and achieved a 0.75 F1 score. The work (Mellado et al., 2021) devoted to the same task for Spanish, presented in IberLef 2021, reported F1 scores ranging from 0.37 to 0.85. In addition, another research for the Norwegian language (Andersen, 2005) is devoted to Anglicism extraction using a combination of methods (rule-based, lexicon-based, and chargram-based). In their work, such a combined approach yielded the most favourable outcome, achieving an overall 0.96 accuracy score for correctly annotated forms and a precision rate of 0.76, which is comparable with our results.

Model	**F1**	**Precision**	**Recall**
ruBERT-tiny (fine-tuning)	0.62	0.59	0.66
ruRoBERTa-large (prompt-tuning)	**0.72**	**0.69**	**0.80**
XLM-RoBERTa (prompt-tuning)	0.70	0.67	0.78

Table 4: Anglicism detection results. Detailed metrics descriptions are given in subsection 6.1.

Besides the general Anglicism detection evaluation, we also performed an additional study of Anglicism detection mistakes. For this, we analyzed the predictions of the best model, that is, the ruRoBERTa-large (prompt-tuning) model (see Table 5 for the most typical mistakes).

Sentence	**Model prediction (token level)**
В ЛДЦ “Кутузовский” в Москве вы можете пройти полное чек-ап обследование всего организма.	чек-
Если не знаешь как начать дейтиться, то этот коуч научит тебя.	дейт, коуч
Можешь рассчитывать даже на апельсиновый фреш в моём исполнении!	ап

Table 5: Typical Anglicism detection mistakes of the ruRoBERTa-large (prompt-tuning) model.

From the mistake analysis, several conclusions can be made:

1. The model demonstrates a restricted capability in accurately identifying Anglicisms that consist of multiple words connected by hyphens. Although the model can identify such Anglicisms, lowering the sensitivity threshold of the linear classification layer resolves this issue.
2. In the process of tokenization, some Anglicisms are tokenized as several tokens. As a result, the model sometimes marks only the English root as an Anglicism, omitting suffixes and inflections.
3. The model occasionally generates false positive errors by incorrectly marking tokens resembling English word parts as Anglicisms.

7.2 Anglicism substitution

As for the Anglicism substitution results (see Table 6), the two model variants can be highlighted here. Namely, ruGPT3 sent+angl outperforms other models by CHFR++ and BLEU, and ruGPT3 LoRA yields the best score by Rouge-L, BERTscore, and LaBSE. This result was obtained due to the fact that in the first approach, the model did not always replace Anglicism in the sentence. In contrast, in the second

approach, the model replaced Anglicism more often, but sometimes not with the same word as in our golden paraphrase. Nevertheless, the substitution the model proposed was semantically close to the golden one. Therefore, metrics measuring semantic proximity, BERTScore and LaBSE turned out to be higher in the second approach. The low-rank adaptation approach has demonstrated its efficiency as it maximizes the potential of large pre-trained models by optimizing all model layers, albeit in a specific manner. The hypothesis that multilingual models cope better with Anglicisms detection and substitution has not been confirmed.

It should also be noted that we solve the Anglicism substitution problem as the generative task and, therefore, employ generative metrics for their evaluation. Thus, due to the possible plurality of the correct answers and the variety of generated output and distinctiveness, these metrics are not expected to reach the theoretical maximum when assessing the effectiveness of generative models like the one in our approach.

Model	CHRF++	BLEU	Rouge-L	BERTScore	LaBSE
ruGPT3 only sent	0.79	0.58	0.74	0.89	0.91
ruGPT3 sent+angl	**0.81**	**0.72**	0.77	0.91	0.93
mGPT3 only sent	0.75	0.64	0.73	0.89	0.92
mGPT3 sent+angl	0.78	0.68	0.75	0.90	0.91
ruGPT3 LoRA	0.76	0.67	**0.8**	**0.92**	**0.94**
mGPT3 LoRA	0.71	0.62	0.78	0.90	0.91

Table 6: Anglicism substitution results. Detailed metrics descriptions are given in subsection 6.1.

Analyzing the predictions of ruGPT3 Lora, which yielded the best scores by most of the metrics, two main types of mistakes can be highlighted:

1. The model leaves the sentence unchanged. This usually happens with uncommon Anglicisms, which are, by being rare, tokenized into several tokens. For example, in the sentence "Футболист Лионель Месси является амбассадором Adidas." the Anglicism "амбассадором" is tokenized into four tokens, and the model fails to replace it.
2. The model replaces an Anglicism with a wrong word changing the meaning (e.g., "Она скринит наши переписки." paraphrased as "Она проверяет наши переписки."). This is most likely due to the fact that the model failed to learn the correct meaning of the Anglicism.

8 Conclusion

This article is devoted to Anglicism detection in Russian and their substitution with Russian equivalents to ensure effective communication across various social and professional strata. In this work, we presented a parallel corpus of Anglicism, several models for Anglicism detection and a set of generative models for Anglicism substitution. In addition, we compared a series of experiments and performed a comprehensive model evaluation. All the code and all the models are available in our repository[23] and the dataset can be downloaded[24] from HuggingFace project.

As a part of future work, we plan to augment the existing dataset with both new Anglicisms and new sentences with the current one. We hope that such data augmentation will improve the result.

8.1 Possible Misuse

We believe that our research should not be involved in creating content that affects the individual or communal well-being in any way, including

- legislative application or censorship;
- mis- and disinformation;
- infringement of the rights of access to information.

[23] https://github.com/dalukichev/anglicism_removing

[24] https://huggingface.co/datasets/shershen/ru_anglicism

8.2 Biases and data quality

The Anglicism corpus includes large segments representing the Internet domain, and therefore, it may possibly contain a variety of stereotypes and biases. Proper evaluation is still needed to explore possible model vulnerabilities in terms of generalizing on the new data and specific new data.

References

Gisle Andersen. 2005. Assessing algorithms for automatic extraction of anglicisms in norwegian texts. 01.

Elozino Egonmwan and Yllias Chali. 2019. Transformer and seq2seq model for paraphrase generation, November.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding.

Alena Fenogenova, Ilia Karpov, and Viktor Kazorin. 2016. A general method applicable to the search for anglicisms in russian social network texts. // *2016 IEEE Artificial Intelligence and Natural Language Conference (AINL)*, P 1–6. IEEE.

Manfred Görlach. 2002a. *An annotated bibliography of European anglicisms*. OUP Oxford.

Manfred Görlach. 2002b. *English in Europe*. OUP Oxford.

David Graddol. 2006. *English next*, volume 62. British council London.

Ankush Gupta, Arvind Agarwal, Prawaan Singh, and Piyush Rai. 2017. A deep generative framework for paraphrase generation.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. 2021. lora. *CoRR*, abs/2106.09685.

Ganesh Jawahar, El Moatez Billah Nagoudi, Muhammad Abdul-Mageed, and Laks Lakshmanan, V.S. 2021. Exploring text-to-text transformers for English to Hinglish machine translation with synthetic code-mixing. // *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, P 36–46, Online, June. Association for Computational Linguistics.

Nikita Konodyuk and Maria Tikhonova. 2022. Continuous prompt tuning for russian: how to learn prompts efficiently with rugpt3? // *Recent Trends in Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 30–40. Springer.

Sebastian Leidig, Tim Schlippe, and Tanja Schultz. 2014. Automatic detection of anglicisms for the pronunciation dictionary generation: A case study on our german it corpus. 05.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. // *Soviet physics doklady*, volume 10, P 707–710. Soviet Union.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. // *Text summarization branches out*, P 74–81.

Xiao Liu, Kaixuan Ji, Yicheng Fu, Zhengxiao Du, Zhilin Yang, and Jie Tang. 2021. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *CoRR*, abs/2110.07602.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam.

Elena Mellado, Luis Espinosa-Anke, Julio Arroyo, Constantine Lignos, and Jordi Porta. 2021. Overview of adobo 2021: Automatic detection of unassimilated borrowings in the spanish press, 10.

John E. Miller, Tiago Tresoldi, Roberto Zariquiey, César A. Beltrán Castañón, Natalia Morozova, and Johann-Mattis List. 2020. Using lexical language models to detect borrowings in monolingual wordlists, Dec.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. // *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, P 311–318.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. // *Proceedings of the Tenth Workshop on Statistical Machine Translation*, P 392–395, Lisbon, Portugal, September. Association for Computational Linguistics.

Julia Pritzen, Michael Gref, Christoph Schmidt, and Dietlind Zühlke. 2021. A comparative pronunciation mapping approach using g2p conversion for anglicisms in german speech recognition. // *Speech Communication; 14th ITG Conference*, P 1–5. VDE.

Virginia Pulcini, Cristiano Furiassi, and Félix Rodríguez González. 2012. The lexical influence of english on european languages. *The anglicization of European lexis*, 1.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Orion Weller, Matthias Sperber, Telmo Pires, Hendra Setiawan, Christian Gollan, Dominic Telaar, and Matthias Paulik. 2022. End-to-end speech translation for code switched speech. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 1435–1448, Dublin, Ireland, May. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

Elena Álvarez Mellado and Constantine Lignos. 2022. Detecting unassimilated borrowings in spanish: An annotated corpus and approaches to modeling.

Светлана Геннадьевна Апетян. 2011. Англицизмы в структуре масс-медийного и официально-делового дискурсов (лексико-семантический и когнитивно-прагматический аспекты).

Анатолий Иванович Дьяков. 2012. УРОВНИ ЗАИМСТВОВАНИЯ АНГЛИЦИЗМОВ В РУССКОМ ЯЗЫКЕ. *Известия Южного федерального университета. Филологические науки*, (2):113–124.

Disambiguation in context in the Russian National Corpus: 20 yeas later

Olga Lyashevskaya
HSE University
Vinogradov Russian Language Institute RAS
Moscow, Russia
olesar@yandex.ru

Ilia Afanasev
HSE University
MTS AI
Moscow, Russia
szrnamerg@gmail.com

Stefan Rebrikov
HSE University
Kurchatov Institute
Moscow, Russia
robstef85@gmail.com

Yana Shishkina
HSE University
Moscow Institute of Physics and Technology
Moscow, Russia
yanaalekseevna2000@mail.ru

Elena Suleymanova **Igor Trofimov** **Natalia Vlasova**
A. K. Ailamazyan Program Systems Institute of RAS
Pereslavl-Zalessky, Russia
yes2helen@gmail.com itrofimov@gmail.com nathalie.vlassova@gmail.com

Abstract

An updated annotation of the Main, Media, and some other corpora of the Russian National Corpus (RNC) features the part-of-speech and other morphological information, lemmas, dependency structures, and constituency types. Transformer-based architectures are used to resolve the homonymy in context according to a schema based on the manually disambiguated subcorpus of the Main corpus (morphology and lexicon) and UD-SynTagRus (syntax). The paper discusses the challenges in applying the models to texts of different registers, orthographies, and time periods, on the one hand, and making the new version convenient for users accustomed to the old search practices, on the other. The re-annotated corpus data form the basis for the enhancement of the RNC tools such as word and n-gram frequency lists, collocations, corpus comparison, and Word at a glance.

Keywords: morphological tagging; dependency parsing; lemmatization; disambiguation; NLP evaluation; Russian National Corpus; Russian

DOI: 10.28995/2075-7182-2023-22-307-318

Разрешение неоднозначности в контексте для Национального корпуса русского языка: 20 лет спустя

О. Н. Ляшевская[1,2], **И. А. Афанасьев**[1,3], **С. А. Ребриков**[1,4], **Я. А. Шишкина**[1,5], **Е. А. Сулейманова**[6], **И. В. Трофимов**[6], **Н. А. Власова**[6]

[1]Национальный исследовательский университет «Высшая школа экономики»
[2]Институт русского языка им. В. В. Виноградова РАН
[3]МТС ИИ
[4]НИЦ «Курчатовский институт»
[5]МФТИ
Москва, Россия
[6]Институт программных систем им. А. К. Айламазяна РАН
г. Переславль-Залесский, Ярославская обл., Россия
olesar@yandex.ru, {szrnamerg, robstef85}@gmail.com, yanaalekseevna2000@mail.ru,
{yes2helen, itrofimov, nathalie.vlassova}@gmail.com

Аннотация

Обновление разметки Основного, Газетного и ряда других корпусов Национального корпуса русского языка (НКРЯ) касается информации о части речи, других морфологических признаках, леммах (словарных формах слов), структурах зависимостей предложения и типах составляющих. Для разрешения лингвистической неоднозначности в контексте используются нейросетевые архитектуры на основе трансформеров. Разметка воспроизводит схему, применяемую в подкорпусе Основного корпуса со снятой вручную грамматической омонимией (морфология и леммы) и UD-SynTagRus (синтаксис). В статье рассматриваются проблемы применения моделей к текстам, написанным в различных функциональных стилях, орфографиях и в разные периоды времени. Поскольку в ряде случаев текстовому фрагменту в заданном контексте можно сопоставить более одного теоретически возможного лингвистического разбора, необходимо принимать во внимание поддержку множественных разборов. Кроме того, обсуждаются вопросы совместимости старой и новой разметки в плане адаптации пользователей к новому поисковому функционалу корпуса. Автоматически дизамбигуированные данные больших корпусов позволили улучшить существующие и разработать новые сервисы поисковой платформы НКРЯ, такие как частотные списки слов и n-грамм, коллокации, сравнение корпусов и портрет слова.

Ключевые слова: автоматическое разрешение лексико-грамматической неоднозначности, морфологическая разметка, синтаксическая разметка, русский язык, Национальный корпус русского языка

1 Introduction

For almost 20 years, the lexico-grammatical annotation of the Russian National Corpus (RNC) existed in three formats. (1) In the Syntactic corpus (SynTagRus, 1.4 MW), each word was provided with one and only one morphological and lemma analysis appropriate in context, and each sentence was analysed as one syntactic dependency tree. (2) In the the manually disambiguated subcorpus of the Main corpus ("Snyatnik", 6 MW) and in the Educational corpus (0,6 MW), only morphology and lemmas were analysed based on a somewhat different tagset and grammatical dictionary compared to SynTagRus. The majority of historical RNC corpora were annotated generally in the same way and oriented on their own markup schemas, tagsets, and dictionaries. (3) Finally, there were no disambiguation in the largest part of the modern Russian texts (more than 1 billion words) and Church Slavonic texts (5,3 MW): each word corresponded to as many analyses as the grammatical dictionary stores, regardless of the context. If the word form of a modern language is not attested in the dictionary, the MyStem hypothesis module assigns a few of the most probable annotations to it (Segalovich, 2003; Zobnin and Nosyrev, 2015).

One of the objectives of the Corpus 2.0 project (2020-2022) was to add syntactic annotations and resolve lexical and morphological ambiguity in modern Russian texts. Firstly, this allows users to constraint the search window by defining syntactic relations between elements or setting up a certain type of clause or phrase within which the elements should occur. Secondly, this makes it possible to significantly reduce the number of irrelevant examples in the search output. Thirdly, other search facilities such as lexical groups-based search, frequency lists, collocations, associated words, etc. definitely benefit from the less noisy annotation input. Fourthly, the use of syntactic n-grams based on dependency parses (Goldberg and Orwant, 2013) in addition to ordinary sequential n-gram opens the way to a new kind of high-quality tools for researchers. All these changes also involve technical improvements in the infrastructure of the corpus search engine such as reducing the size of the search indices and the time spent performing the calculations, extending the amount of annotated data and information conveyed to the user.

2 Related Work

The approaches to the three grammar tasks that form the basic NLP pipeline, namely, part-of-speech/morphological tagging, lemmatisation, and dependency parsing, rapidly developed for the last half a century (Hann, 1974) (Spyns, 1996) (Aduriz et al., 1996) (Branco and Silva, 2003)] (Qi et al., 2020) (Kumar et al., 2022). Currently pipeline models that combine part-of-speech/morphological tagging, lemmatisation , and parsing, dominate the landscape (Straka and Straková, 2017) (Kondratyuk, 2019) (Kanerva et al., 2021). However, despite this pursuit to develop the language-independent tagger for benchmark datasets (Toleu et al., 2022) that provide satisfying for all the included languages, yet moderate for each of them results, there is a growing concern that low-resourced language NLP, and

probably NLP in general, is going to suffer from the trend (Alonso-Alonso et al., 2022). Frw works clearly state the intention to make a universal tagger, which is based upon the multi-lingual training and switching parameters to fine-tune for a single language (Üstün et al., 2020). The models, trained for the particular task-language pair, still seem to deserve attention, as (Dyer, 2022) states for the case of Wolof language.

Automatic morphological tagging systems currently employ the pair of dominating approaches, the single-language rule-based one (Gambäck, 2012), and the machine learning-based one, which can assume both monolingual (Berdičevskis et al., 2016) (Qi et al., 2018) (Qi et al., 2020) (Scherrer, 2021) and multi-lingual (Straka and Straková, 2017) forms. Instead of targeting the multi-lingual level, now morphological tagging shifts into the multi-lect one to be able to deal with the very close (Obeid et al., 2022), yet significantly different lects, as is the cast with Arabic (Inoue et al., 2022) (Fashwan and Alansary, 2022). This also provokes a lot of discussion for morphological tagging of low-resouced languages (Blum, 2022) (Wiemerslage et al., 2022). The discussion about data quality takes place within the common morphology tagging discourse (Muradoglu and Hulden, 2022). New methods are being developed, for instance, graph-based part-of-speech tagging (ImaniGooghari et al., 2022), or using compressed FastText models (Nevěřilová, 2022). Specifically concerning Russian, joined morphological analysis and morpheme segmentation models were proposed recently (Bolshakova and Sapin, 2022).

Lemmatisation follows the same patterns that morphological tagging does. Currently, there is a division between the universal lemmatisation tools (Straka and Straková, 2017) (Bergmanis and Goldwater, 2018) (Kanerva et al., 2021), and language, or domain-specific (Fernández, 2020) The sequence-to-sequence architecture (Sutskever et al., 2014) (Cho et al., 2014) prevails now, and within it the encoder-decoder transformers dominate (Lewis et al., 2020) The ensemble models that enhance lemmatisation efficiency with external resources (Milintsevich and Sirts, 2021) are gaining popularity (de Graaf et al., 2022)

Dependency parsing is probably the most dynamically developing area of the three, as it still presents the highest challenge of the three for the automated corpus tools. New methods are constantly being implemented: the last three years witnessed a combination of the second-order graph-based and headed-span-based projective dependency parsing (Yang and Tu, 2022), the domain adaptation (Li et al., 2022) and the dependency parsing being treated as machine reading comprehension (MRC)-based span-span prediction (Gan et al., 2022) and using structure preserving embeddings for dependency parsing (Kádár et al., 2021) The state-of-the-art method, biaffine parsing, is modified (Xu et al., 2022). The previously under-utilised concepts, such as *nuclei* (semantically independent units consisting of a content word together with its grammatical markers, regardless of whether the latter are realised in dependent words or not (Basirat and Nivre, 2021)), are introduced to the frameworks. The data augmentation techniques are implemented to enhance the performance of the models (Goodwin et al., 2022). (Eggleston and O'Connor, 2022) and (Langedijk et al., 2022) introduce cross-lect dependency parsing, getting in line with papers that consider low-resourced languages (Tian et al., 2022) and zero-shot (de Lhoneux et al., 2022) (Shi et al., 2022) dependency parsing. The issues of the dataset construction that affect evaluation are discussed in (Krasner et al., 2022) Artificial performance inflation is a problem that should be addressed across the pipeline of morphological tagging, lemmatisation and part-of-speech tagging (Goldman et al., 2022).

3 Data for Training and Evaluation

We conducted experiments involving a diverse panel of text samples. A variety of genres, types, domains, time periods of creation, and orthographies were presented in the following datasets for modern Russian (1700-2020s):

- `SynTagRus UD 2.8` - 1,1 M tokens (contemporary fiction, popular science, newspaper and journal articles dated between 1960 and 2016, texts of online news etc.). This portion of the RNC Syntactic Corpus converted to the Universal Dependencies (UD) format was the main training dataset used in the GramEval-2020 shared task.
- `SynTagRus UD 2015` - 400k tokens. An addition to the RNC Syntactic Corpus annotated in 2015-

GramEval-2020 (Taiga)	dev	test	New RNC datasets	dev	test
fiction	1.0k	1.0k	prose-XX	10.4k	20.0 k
news	1.0k	1.0k	newspapers-XXI	7.8k	14.4k
poetry	1.0k	1.0k	prose-XIX	41.7k	80.7k
social	1.0k	1.0k	poetry-XIX	1.4k	1.4k
wiki	1.0k	1.0k	old-orthography	14.8k	14.8k
			old-orthography-XVIII	6.1k	6.1k
			Middle Russian: LEG	16.5k	39.0k
			bezobrazov		519.0k

Table 1: Size of the validation and test sets, tokens.

2020; converted and added to UD v.2.9. New genres: wikipedia.

- Taiga - 200 k tokens. Modern text samples extracted from Taiga Corpus, MorphoRuEval-2017 and GramEval-2020 shared tasks collections. Genres include electronic communication (VK, Twitter and other social media, YouTube comments, questions & answers from otvet.mail.ru, reviews from reviews.yandex.ru); poetry from stihi.ru (naïve poetry) and RNC Corpus of Russian poetry; fiction; news (lenta.ru etc.); wiki (Russian wikipedia). Taiga includes, among others, development and test data of the GramEval-2020 shared task (modern Russian), which was subdivided into the following subsets: fiction, news, poetry, social, wiki.
- newspapers-XXI - 34 k tokens. Samples extracted from the RNC National media and Regional and international media corpora.
- prose-XX - 423 k tokens. Texts of the 20th c. and the beginning of the 21th c. in modern orthography (RNC Main corpus). Fiction includes stories by V. M. Shukshin, I. V. Evdokimov, and M. K. Pervukhin, non-fiction - diaries and memories, journalism covers general news, finance, church news, recipes and tips.
- prose-XIX - 108 k tokens. Texts of the 19th c. in modern orthography (RNC Main corpus). The dataset includes drama by A. V. Sukhovo-Kobylin, A. Pisemsky, M. Gorky, etc., fiction by N. V. Gogol, S. T. Aksakov, E. A. Salias etc., non-fiction on history, hygiene, memories and essays.
- poetry-XIX - 50 k tokens. Samples from the RNC Russian Poetry Corpus written before 1917 and provided in modern orthography.
- old-orthography - 108 k tokens. Texts of the 19th - early 20th cc. in pre-revolutionary orthography (S. T. Aksakov, P. A. Kulish, M. Pogodin, A. Spaso-Kukotsky, N. I. Grech)
- old-orthography-XVIII - 6 k tokens. 18th century texts in old orthography (by Peter the Great, S. Pufendorf, P. I. Pogoretsky, F. A. Emin)

As for historical Russian data (1400-1700s), we used official legal and business writing texts, as the other RNC Middle Russian collections, like vernacular gramotki, were distinctly different in the occurrences of old grammatical forms and constructions, in phonetic features reflected in orthography, and in genre-specific lexical distributions. We split the taken texts into two datasets:

- LEG(acy) texts written in 15th – 17th cc. (ca. 1.1 M tokens), and
- Bezobrazov - recently added to the RNC texts of the latter half of the 17th c. from Bezobrazov's archive (500 k tokens).

Table 1 summarises the size of the development and test data used in experiments. In the experiments reported below, the models were trained on a joined modern Russian training dataset (1700-2020s) or historical Russian data (1400-1700s).

All data are presented in the CONLL-U format and annotated according to the Russian UD-Ext scheme (Lyashevskaya, 2019). This scheme assumes the use of a standard inventory of the UD-Russian dependency relations and common RNC and UD policy for lemmatisation. Enhanced dependency relations are not provided. To make morphological annotations of the RNC Main corpus and Russian UD compatible,

the following features are added to the GramEval2020 and SynTagRus data and used in all new datasets:

- parts of speech: PRED for predicatives (eg. *можно, холодно, жаль*), ADVPRO for pronominal adverbs (eg. *тут*), PREDPRO for pronominal predicatives (eg. *некого*), PARENTH for parentheticals (eg. *конечно*), ANUM for ordinal numerals (eg. *второй*).
- grammatical features: Transit={Tran,Intr} for transitivity, Case={Acc2,Loc2} for secondary cases, Degree=Cmp2 for comparatives with the prefix *по-*, Anom=Yes for anomalous forms.

PoS-tags that are absent from the UD format were added by automatic replacement with the use of wordlists. Some PoS-tags were added manually, e.g. ANUM for numerals written with numbers, PRED for ambiguous words. PoS-tag disambiguation (e.g. *холодно* - ADV vs. ADJ vs. PRED; мало NUM vs. ADVPRO vs. PRED) and corresponding correction of dependency relations were performed manually. Necessary grammatical features were corrected or added using the wordlists and lists of tokens with manual correction. The transitivity feature was manually checked in context with the dependency relations correction.

4 Rubic: a Model for Tagging and Parsing

The study is divided into the following parts. In the first one we examine the previous results of the GramEval-2020 shared task. From this data, we form our expectations for the next suitable model to achieve in morphological tagging, lemmatisation, and dependency parsing. The second stage of the research is the description of the new model, and its results on the GramEval data. In some tasks, the model is challenged by the other models, specifically trained for this task on the particular dataset, to explore the possible enhancements. The third part of the study is dedicated to the analysis of the key errata that the proposed model makes, and whether the other models struggle with the same issues.

The model that we are starting with, our baseline, is the one that has been previously used for the annotation of the RNC corpus data, qbic (Anastasyev, 2020), a winner of the GramEval-2020 shared task. Qbic is a RuBERT encoder accompanied by three classifier decoders performing the part-of-speech classification, lemmatisation, and dependency parsing, respectively. Lemmatisation is conducted in two stages, with the classifier assigning the particular rule to a token, after which the rules themselves are applied. Each lemmatisation rule specifies the number of characters to be cut and a combination of characters to be added, thus comprising a total of 1000 to 2000 rules, depending on the amount of training data (cf. also “less than 1,000 classes of rules in total” in (Michurina et al., 2021)). The rules form in the following manner:

- Training set yields sequences of transformations that are required to transform a token into its lemma (delete postfix/suffix of a certain length > add some sequence of characters to the end > capitalise/decapitalise)
- We take the sequences of transformations that are met more than 3 times (to exclude noise)
- The remaining sequences become rules

Table 2 shows the performance of qbic on the re-annotated GramEval-2020 datasets. A standard CONLL18 script was used to calculate accuracy scores for parts of speech (PoS), morphological features, lemmas, and labeled attachment score for syntactic dependencies (LAS, basic relation inventory, ie. `nummod` and `nummod:gov` are considered the same). The model performed in a satisfactory way in most of the aspects. However, its performance on dependency parsing was below expectations. Non-standard patterns in poetry, social media texts, and wiki presented an especially hard challenge for it. Additionally, qbic was not robust in full morphological tagging and lemmatisation in the case of social media, poetry, diaries, and encyclopedic texts, which contain abbreviations, non-standard punctuation, transcript notes, rare named entities, and especially in the case of the RNC subcorpus of older orthographies (ca. 13M tokens).

To meet this challenge, we present Rubic, a model that utilises the same architecture as qbic, with enhancements, see Figure 1. For an encoder, we use sberbank-ai/ruBert pretrained on 30 GB data. In our model, the lemmatisation module receives additional information from the part-of-speech tagging classifier. Rubic checks lemma candidates against a supplementary dictionary compiled manually. The dictionary is a pair of lemma and part of speech, split by tab, e.g. *автоматизм NOUN*. Besides that,

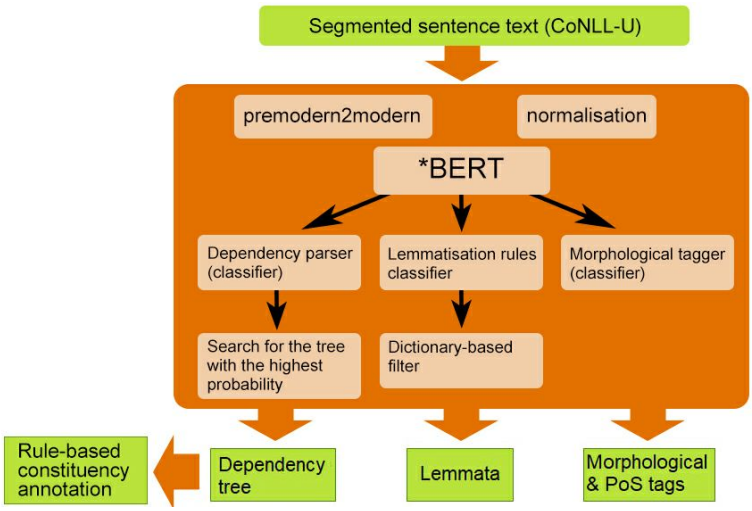


Figure 1: Key principles and architecture of Rubic.

Dataset	fiction	news	poetry	social	wiki
PoS	98.0	96.6	96.9	94.7	92.7
Morph.features	98.7	96.1	96.7	94.7	94.4
Lemmatisation	98.0	98.2	95.3	96.0	93.6
LAS	89.6	91.2	81.4	80.7	78.1

Table 2: Accuracy score of qbic on GramEval-2020 dataset, %

the symbol sequences unlikely to occur in Russian texts are preprocessed. We specifically set up Rubic to process data with non-standard orthography by implementing a graphic premodern2modern heuristic, and mapping the tokens in older orthography to tokens in modern orthography.

We perform data augmentation when training Rubic. We use the calculation of "the lexical usefulness weight" that prioritise the use of rare tokens for the further pipeline of data augmentation. If a sentence contains two, and exactly two quotation marks, we add another sentence to the dataset, that contains guillemets instead (we add 450 sentences via this heuristic). We use the heuristic of jo-fication, transforming e into $\ddot{e}$, in words, where it is possible (we add more than 800 sentences via this heuristic). We use the capitalisation heuristic, when the tokens are randomly capitalised for the purposes of better recognition (we acquire nearly 2000 additional sentences via this heuristic. %; we take only 20% of the sentences, generated by the previous heuristic).

With all these enhancements, the results of the model expectedly grow. We provide the difference between accuracy scores in Table 3. Rubic improves in parsing, and some improvements can be seen in tagging and lemmatisation. It underperforms on the fiction dataset, and wiki morphology presents it with some challenges. All this may also signal about overfitting, so we use the other datasets of the modern Russian language: CONLL18, and IWPT21. The results are presented in Table 4.

We also evaluated Rubic on the RNC test sets prepared specifically for the task of full corpus re-annotation. The results are shown in Table 5. In all datasets, Rubic performs well on major and most frequent part of speech categories such as verbs, nouns, proper nouns, prepositions, and coordinate conjunctions. Noun case accuracy is above 98% in all datasets except poetry and old orthography-XVIII. Mixing adjectives vs. participles, adjectives vs. adverbs is higher in the latter datasets and Taiga. Annotation of predicatives and corresponding syntactic structures is problematic in poetry, fiction and non-fiction written in the 20th c. and earlier, in which a wider variety of constructions and lexical fillers is available. Expectedly, parsing quality drops on longer sentences, and non-standard symbols, non-

Dataset	fiction	news	poetry	social	wiki
PoS	+0.1	+1.4	+1.7	+1.0	+0.5
Morph.features	-0.1	+0.3	+0.1	+0.6	-0.4
Lemmatisation	-0.3	+0.0	+0.2	+0.6	+0.5
LAS	+0.5	+0.8	+1.3	+0.3	+2.8

Table 3: Change in accuracy score for Rubic compared to qbic, %, GramEval-2020 datasets

Dataset	CONLL18	IWPT21
PoS	99.23	99.14
Morph.features	98.27	98.19
Lemmatisation	97.49	97.83
LAS	95.51	95.47

Table 4: Accuracy score of Rubic on standard modern Russian datasets, %

standard place of punctuation marks and other non-letters, and out-of-vocabulary abbreviations misleads the model.

5 Lemmatisation: Further Experiments

Rubic, thus, does not overfit for GramEval-2020 datasets. However, we wanted to see if there is a possibility to enhance its performance. To test this, we picked the lemmatisation task and trained two BART-large-based lemmatiser models (Lewis et al., 2020). This is a sequence-to-sequence state-of-the-art multilingual method that can help to reveal critical points in which Rubic needs enhancement.

The comparison is based on the following data: modern RNC datasets, historical LEG and Bezobrazov datasets. Both Rubic and BART-large were separately fine-tuned for modern and historical data. The results of comparison between BART-large and Rubic are in Table 6.

The news dataset witnesses a better performance of Rubic, by 0.1 per cent: the Rubic heuristics adapt the model for the specific language variety. However, it seems that the texts of the Middle Russian period require much more intricate heuristics, which leads to the striking 12 to 20, depending on data quality, per cent difference between BART-large and Rubic accuracy in favour of the former. Overall, BART-large beats Rubic by a significant margin of 0.4 to 3 per cent. The main challenges are non-standard orthography and syntactic structures of XIX century poetry, which encourage a more generalising approach of BART-large.

The Rubic model, despite implemented heuristics, is challenged by two main classes of words: non-productive verb models (*скорбатъ* instead of *скорбетъ* 'mourn'), and proper names (*Любовя* instead of *Любовъ* 'Lyubov'). The non-standard modern orthography also takes its toll: *наср@ла* is returned instead of *насратъ* 'do not give a damn about smth' likely due to the special symbol that was not normalised. Sometimes model generates empty lemmata, due to the rule-based nature of its lemmatiser module.

BART-large sequence-to-sequence architecture helps to deal with the aforementioned problems. It still overgeneralises, creating the syntagmae, similar to *-жо-* in verbs (*ожотъся* instead of *ожечься* 'get fired by'), or choosing the more general ending, completely confusing the word class, cf. *Стоцка* instead of *Стоцкая* 'Stotskaja'. Generalisation also leads into the model being unable to deal with orthography issues (odd *с* in *естесственный* 'natural'; odd *о* in *-пр-*, cf. *предупорежденне* instead of *предупреждение* 'warning'). Probably, the same factor leads to the appearance of hyphens in lemmas for the words that were transitioned from string to string somewhere in the data, sometimes with character replacing, for instance, in *пеп-льница* instead of *пепельница* 'ashpot'. Compound pronouns, such as *ни о чём* 'about nothing', often lose their negative particle (*ни*) part. The words that contain similar

Dataset	Taiga	newspa-pers-XXI	prose -XX	prose -XIX	poetry -XIX	old ortho-graphy	old ortho-graphy-XVIII
PoS	97.8	99.0	98.9	99.2	97.4	98.9	95.8
Morph.features	94.6	97.3	97.2	97.7	94.2	95.9	90.1
Lemmatisation	97.6	99.1	98.3	98.9	95.9	97.5	93.7
LAS	85.7	95.1	94.1	94.6	85.6	94.0	83.7

Table 5: The accuracy score of Rubic on RNC datasets, %

Dataset	Rubic, accuracy, %	BART-large, accuracy, %
Taiga	97.6	**98.0**
newspapers-XXI	**99.1**	99.0
prose-XX	98.3	**98.7**
prose-XIX	98.9	**99.3**
poetry-XIX	95.9	**98.9**
old orthography	97.4	**98.7**
old orthography-XVIII	93.7	**93.8**
LEG(al) test, 1400-1700	85.4	**98.0**
Bezobrazov	73.8	**85.0 (92.6 with normalisation)**

Table 6: Lemmatisation accuracy scores for Rubic and BART-large models on RNC datasets. The best results are highlighted in bold.

syllables, such as *царица* 'empress', are often reduced to a single syllable, in this case, *ца*: probably, the original BART-large dataset was trained to eliminate reduplication. The model clearly lacks knowledge of how the lemmas in particular language should look, which leads to generating adjective lemmas that after the adjectival affix *-ск-* have *-ив-* instead of *-ий-*. The model often does not pay attention to the morphology tagging (generated verb lemmas with *Aspect=Perf* tag often contain *-ывать*, which is a strong marker of continuous aspect in Russian verbs; prefix *по-* for *Degree=Cmp2* adjectives generated lemmas).

BART-large experiments show that sequence-to-sequence is not a necessarily ideal solution. It appears to be slow when annotating large amount of texts. However, this method reveals room for improvement of models like Rubic, particularly when it concerns the dataset construction, non-standard orthography, and low-productive paradigms, such as proper names and some verb classes. We are going to dedicate further research to these particular issues.

6 Corpus annotation and future development

At the moment, Main corpus, Regional Media, and Educational corpora are annotated by Rubic. In order to make it easier for users to switch from the old version to the new one, two lemma layers – annotations provided by Mystem and Rubic – are searchable. By default, the search is conducted on the layer automatically disambiguated by Rubic only.

We decided to apply three techniques to improve the Rubic outcome. Firstly, although the neural model is set up to produce only one analysis per token, in the case of theoretically plausible equivalent linguistic interpretations (eg. adjective vs. participle, see the practice of the manually disambiguated RNC subcorpus) additional morphological and lexical analyses were provided by rules. Secondly, lemmas that occur 30 times and more in the corpus and are not found in the Mystem dictionary, were checked and corrected manually. Thirdly, a number of heuristics were applied to the dependency annotations to provide search by constituency types and unlabeled tree configurations (eg. search within subordinate clauses; within participial phrases; search words that do not have dependents).

In the future, based on the results of the users' feedback, more disambiguated RNC corpora will be made available, with necessary adjustments in the annotation methods. RNC services such as frequency lists, graphs by year, lemma-based corpus portraits and comparison, collocation tools, Word at a glance sketch tool, and search by lexico-semantic features, depend critically on the quality of data lemmatisation. More work should be done in terms of finding new text classes on which the models underperform and adding relevant excepts to training; balancing the training collection by text types; balancing learning rate for different task. Decoding of abbreviated words is likely to be formulated as a separate since the distribution of such forms in large corpora cannot be modeled in the same way as lemmatisation rules.

The project's repository containing supplementary materials is available at: `https://github.com/olesar/RNC2.0`.

Acknowledgements

This work was carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793: "Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research".

References

Itziar Aduriz, Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Nerea Ezeiza, and Ruben Urizar. 1996. Euslem: A lemmatiser/tagger for basque. // Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström, and Catalina Röjder Papmehl, *Proceedings of the 7th EURALEX International Congress*, P 27–35, Göteborg, Sweden, aug. Novum Grafiska AB.

Iago Alonso-Alonso, David Vilares, and Carlos Gómez-Rodríguez. 2022. The fragility of multi-treebank parsing evaluation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5345–5359, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Daniil Anastasyev. 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.

Ali Basirat and Joakim Nivre. 2021. Syntactic nuclei in dependency parsing – a multilingual exploration. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1376–1387, Online, April. Association for Computational Linguistics.

Aleksandrs Berdičevskis, Hanna Eckhoff, and Tatiana Gavrilova. 2016. The beginning of a beautiful friendship: rule-based and statistical analysis of Middle Russian. // *Komp'yuternaya lingvistika i intellektual'nye tekhnologii. Trudy mezhdunarodnoj konferencii «Dialog»*, P 99–111, Moscow, Russia. RSSU.

Toms Bergmanis and Sharon Goldwater. 2018. Context sensitive neural lemmatization with Lematus. // *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, P 1391–1400, New Orleans, Louisiana, June. Association for Computational Linguistics.

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, P 1–9, Dublin, Ireland, May. Association for Computational Linguistics.

Elena I Bolshakova and Alexander S Sapin. 2022. Building a combined morphological model for Russian word forms. // *Analysis of Images, Social Networks and Texts: 10th International Conference, AIST 2021, Tbilisi, Georgia, December 16–18, 2021, Revised Selected Papers*, P 45–55. Springer.

Antònio Branco and Jõao Silva. 2003. Portuguese specific issues in the rapid development of state of the art taggers. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 7–9, Paris. European Language Resources Association.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder–decoder for statistical machine translation. // *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 1724–1734, Doha, Qatar, October. Association for Computational Linguistics.

Evelien de Graaf, Silvia Stopponi, Jasper K. Bos, Saskia Peels-Matthey, and Malvina Nissim. 2022. AGILe: The first lemmatizer for Ancient Greek inscriptions. // *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, P 5334–5344, Marseille, France, June. European Language Resources Association.

Miryam de Lhoneux, Sheng Zhang, and Anders Søgaard. 2022. Zero-shot dependency parsing with worst-case aware automated curriculum learning. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 578–587, Dublin, Ireland, May. Association for Computational Linguistics.

Bill Dyer. 2022. New syntactic insights for automated Wolof Universal Dependency parsing. // *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, P 5–12, Dublin, Ireland, May. Association for Computational Linguistics.

Chloe Eggleston and Brendan O'Connor. 2022. Cross-dialect social media dependency parsing for social scientific entity attribute analysis. // *Proceedings of the Eighth Workshop on Noisy User-generated Text (W-NUT 2022)*, P 38–50, Gyeongju, Republic of Korea, October. Association for Computational Linguistics.

Amany Fashwan and Sameh Alansary. 2022. Developing a tag-set and extracting the morphological lexicons to build a morphological analyzer for Egyptian Arabic. // *Proceedings of the The Seventh Arabic Natural Language Processing Workshop (WANLP)*, P 142–160, Abu Dhabi, United Arab Emirates (Hybrid), December. Association for Computational Linguistics.

Laura García Fernández. 2020. A contribution to old english lexicography. *NOWELE / North-Western European Language Evolution*, 73(2):236–251.

Björn Gambäck. 2012. Tagging and verifying an amharic news corpus. // *Proceedings of the Workshop on Language Technology for Normalisation of Less-Resourced Languages (SALTMIL8/AfLaT2012)*, P 79–84, Paris. European Language Resources Association.

Leilei Gan, Yuxian Meng, Kun Kuang, Xiaofei Sun, Chun Fan, Fei Wu, and Jiwei Li. 2022. Dependency parsing as MRC-based span-span prediction. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 2427–2437, Dublin, Ireland, May. Association for Computational Linguistics.

Yoav Goldberg and Jon Orwant. 2013. A dataset of syntactic-ngrams over time from a very large corpus of english books. // *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 1: Proceedings of the Main Conference and the Shared Task: Semantic Textual Similarity*, P 241–247.

Omer Goldman, David Guriel, and Reut Tsarfaty. 2022. (un)solving morphological inflection: Lemma overlap artificially inflates models' performance. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 864–870, Dublin, Ireland, May. Association for Computational Linguistics.

Emily Goodwin, Siva Reddy, Timothy O'Donnell, and Dzmitry Bahdanau. 2022. Compositional generalization in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6482–6493, Dublin, Ireland, May. Association for Computational Linguistics.

Michael Hann. 1974. Principles of automatic lemmatisation. *ITL Review of Applied Linguistics*, 23(1):3–22.

Ayyoob ImaniGooghari, Silvia Severini, Masoud Jalili Sabet, François Yvon, and Hinrich Schütze. 2022. Graph-based multilingual label propagation for low-resource part-of-speech tagging. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 1577–1589, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Go Inoue, Salam Khalifa, and Nizar Habash. 2022. Morphosyntactic tagging with pre-trained language models for arabic and its dialects. // *Proceedings of the Findings of the Association for Computational Linguistics: ACL2022*, Dublin, Ireland, May. Association for Computational Linguistics.

Ákos Kádár, Lan Xiao, Mete Kemertas, Federico Fancellu, Allan Jepson, and Afsaneh Fazly. 2021. Dependency parsing with structure preserving embeddings. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 1684–1697, Online, April. Association for Computational Linguistics.

Jenna Kanerva, Filip Ginter, and Tapio Salakoski. 2021. Universal lemmatizer: A sequence-to-sequence model for lemmatizing universal dependencies treebanks. *Natural Language Engineering*, 27(5):545–574.

Dan Kondratyuk. 2019. Cross-lingual lemmatization and morphology tagging with two-stage multilingual BERT fine-tuning. // *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, P 12–18, Florence, Italy, August. Association for Computational Linguistics.

Nathaniel Krasner, Miriam Wanner, and Antonios Anastasopoulos. 2022. Revisiting the effects of leakage on dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 2925–2934, Dublin, Ireland, May. Association for Computational Linguistics.

C S Ayush Kumar, Advaith Maharana, Srinath Murali, Premjith B, and Soman Kp. 2022. BERT-based sequence labelling approach for dependency parsing in Tamil. // *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, P 1–8, Dublin, Ireland, May. Association for Computational Linguistics.

Anna Langedijk, Verna Dankers, Phillip Lippe, Sander Bos, Bryan Cardenas Guevara, Helen Yannakoudakis, and Ekaterina Shutova. 2022. Meta-learning for fast cross-lingual adaptation in dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 8503–8520, Dublin, Ireland, May. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 7871–7880, Online, July. Association for Computational Linguistics.

Ying Li, Shuaike Li, and Min Zhang. 2022. Semi-supervised domain adaptation for dependency parsing with dynamic matching network. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 1035–1045, Dublin, Ireland, May. Association for Computational Linguistics.

Olga Lyashevskaya. 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.

Mariia Michurina, Alexandra Ivoylova, Nikolay Kopylov, and Daniil Selegey. 2021. Morphological annotation of social media corpora with reference to its reliability for linguistic research. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 492–504.

Kirill Milintsevich and Kairit Sirts. 2021. Enhancing sequence-to-sequence neural lemmatization with external resources. // *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, P 3112–3122, Online, April. Association for Computational Linguistics.

Saliha Muradoglu and Mans Hulden. 2022. Eeny, meeny, miny, moe. how to choose data for morphological inflection. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, P 7294–7303, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Zuzana Nevěřilová. 2022. Compressed FastText Models for Czech Tagger. // *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022*, P 79–87, Tribun EU. European Language Resources Association.

Ossama Obeid, Go Inoue, and Nizar Habash. 2022. Camelira: An Arabic multi-dialect morphological disambiguator. // *Proceedings of the The 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 319–326, Abu Dhabi, UAE, December. Association for Computational Linguistics.

Peng Qi, Timothy Dozat, Yuhao Zhang, and Christopher D. Manning. 2018. Universal Dependency parsing from scratch. // *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 160–170, Brussels, Belgium, October. Association for Computational Linguistics.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. 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*, P 101–108, Online, July. Association for Computational Linguistics.

Yves Scherrer, 2021. *Adaptation of Morphosyntactic Taggers*, P 138–166. Studies in Natural Language Processing. Cambridge University Press.

Ilya Segalovich. 2003. A fast morphological algorithm with unknown word guessing induced by a dictionary for a web search engine. // *MLMTA*, P 273–280, 01.

Freda Shi, Kevin Gimpel, and Karen Livescu. 2022. Substructure distribution projection for zero-shot cross-lingual dependency parsing. // *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, P 6547–6563, Dublin, Ireland, May. Association for Computational Linguistics.

Peter Spyns. 1996. A tagger/lemmatiser for Dutch medical language. // *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, P 1147–1150, USA. Association for Computational Linguistics.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. // *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, P 88–99, Vancouver, Canada, August. Association for Computational Linguistics.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. // *Advances in neural information processing systems*, P 3104–3112.

Yuanhe Tian, Yan Song, and Fei Xia. 2022. Enhancing structure-aware encoder with extremely limited data for graph-based dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5438–5449, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Alymzhan Toleu, Gulmira Tolegen, and Rustam Mussabayev. 2022. Language-independent approach for morphological disambiguation. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5288–5297, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Ahmet Üstün, Arianna Bisazza, Gosse Bouma, and Gertjan van Noord. 2020. UDapter: Language adaptation for truly Universal Dependency parsing. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 2302–2315, Online, November. Association for Computational Linguistics.

Adam Wiemerslage, Miikka Silfverberg, Changbing Yang, Arya D. McCarthy, Garrett Nicolai, Eliana Colunga, and Katharina Kann. 2022. Morphological processing of low-resource languages: Where we are and what's next. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 988–1007. Association for Computational Linguistics.

Ziyao Xu, Houfeng Wang, and Bingdong Wang. 2022. Multi-layer pseudo-Siamese biaffine model for dependency parsing. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 5476–5487, Gyeongju, Republic of Korea, October. International Committee on Computational Linguistics.

Songlin Yang and Kewei Tu. 2022. Combining (second-order) graph-based and headed-span-based projective dependency parsing. // *Findings of the Association for Computational Linguistics: ACL 2022*, P 1428–1434, Dublin, Ireland, May. Association for Computational Linguistics.

AI Zobnin and GV Nosyrev. 2015. Morfologicheskij analizator MyStem 3.0. *Trudy Instituta russkogo yazyka im. VV Vinogradova*, 6:300–310.

Multimodal Hedges for Companion Robots: A Politeness Strategy or an Emotional Expression?

Maria Malkina
MSLU
maria020602@mail.ru

Anna Zinina
Kurchatov Institute,
RSUH, MSLU
anna.zinina.22@gmail.com

Nikita Arinkin
Kurchatov Institute, RSUH
nikita.arinkin@gmail.com

Artemiy Kotov
Kurchatov Institute, RSUH
kotov@harpia.ru

Abstract

We examine the use of multimodal hedges (a politeness strategy, like saying *A kind of!*) by companion robots in two symmetric situations: (a) user makes a mistake and the robot affects user's social face by indicating this mistake, (b) robot makes a mistake, loses its social face and may compensate it with a hedge. Within our first hypothesis we test the politeness theory, applied to robots: the robot with hedges should be perceived as more polite, threat to its social face should be reduced. Within our second hypothesis we test the assumption that multimodal hedges, as the expression (or simulation) of internal confusion, may make the robot more emotional and attractive. In our first experiment two robots assisted users in language learning and indicated their mistakes by saying *Incorrect!* The first robot used hedges in speech and gestures, while the second robot used gestures, supporting the negation. In our second experiment two robots answered university exam questions and made minor mistakes. The first robot used hedges, while the second robot used addressive strategy in speech and gestures, e. g. moved its hand to the user and said *That's it!* We have discovered that the use of hedges as the politeness strategy in both situations makes the robot *comfortable to communicate with*. But robot with hedges looks more *polite* only in the experiment, where it affects user's social face, and not when the robot makes mistakes. However, the usage of hedges as an emotional cue works in both cases: the robot with hedges seems to be *cute* and *sympathy provoking* both when it attacks user's social face or loses its own social face. This spectrum of hedge usage can demonstrate its transition from an expressive cue of a negative emotion (nervousness) to a marker of speaker's friendliness and competence.

Keywords: multimodal communication; companion robots; emotional computing; face threatening acts; theory of politeness

DOI: 10.28995/2075-7182-2023-22-319-326

Мультимодальные хеджи для роботов-компаньонов: стратегия вежливости или эмоциональная экспрессия?

Малкина М. П.
МГЛУ
maria020602@mail.ru

Зинина А. А.
НИЦ Курчатовский институт,
РГГУ, МГЛУ
anna.zinina.22@gmail.com

Аринкин Н. А.
НИЦ Курчатовский институт,
РГГУ
nikita.arinkin@gmail.com

Котов А. А.
НИЦ Курчатовский институт,
РГГУ
kotov@harpia.ru

Аннотация

Мы исследуем использование мультимодальных хеджей (стратегия вежливости, например, во фразе *Типа того!*) роботами-компаньонами в двух симметричных ситуациях: (а) пользователь совершает ошибку, и робот угрожает социальному лицу пользователя, указывая на эту ошибку, (б) робот совершает ошибку, теряет своё социальное лицо и может компенсировать это хеджем. В рамках нашей первой гипотезы мы проверяем теорию вежливости в применении к роботам: робот с хеджами должен восприниматься как более вежливый, угроза его социальному лицу должна быть снижена. В рамках нашей второй гипотезы мы проверяем предположение о том, что мультимодальные хеджи, как выражение (или имитация) внутреннего замешательства, могут сделать робота более эмоциональным и привлекательным. В нашем первом эксперименте два робота помогали пользователям в изучении языка и указывали на их ошибки, говоря *«Неправильно!»* Первый робот использовал хеджи в речи и жестах, в то время как второй робот использовал жесты, поддерживающие отрицание. В нашем втором эксперименте два робота отвечали на вопросы университетского экзамена и допускали незначительные ошибки. Первый робот использовал хеджи, в то время как второй робот использовал стратегию апелляции в речи и жестах, например, махал рукой в сторону пользователя и говорил: «Вот так!» Мы обнаружили, что использование хеджей в качестве стратегии вежливости в обеих ситуациях делает общение с роботом более комфортным. При этом робот с хеджами выглядит более вежливым только в эксперименте, где он угрожает социальному лицу пользователя, но не когда сам робот совершает ошибки. Однако использование хеджей для выражения эмоций работает в обоих случаях: робот с хеджами кажется симпатичным и вызывает сочувствие, когда он угрожает социальному лицу пользователя или когда он теряет собственное социальное лицо. Этот спектр использования хеджей может продемонстрировать переход хеджа от средства выражения негативной эмоции (неуверенности) к средству обозначения дружелюбия и компетентности говорящего.

Ключевые слова: мультимодальная коммуникация; роботы-компаньоны; эмоциональные компьютерные системы; угроза социальному лицу; теория вежливости

1 Introduction

Robots may encounter different communicative tensions while failing to execute a user's instruction, and thus, failing a user's trust, or while correcting a user, and thus, deprecating his competence. The linguistic theory of politeness [1] describes these situations as a *threat to social face* – of the speaker or of the hearer – which can be compensated by the use of politeness strategies. These strategies may mitigate the face loss and make the communication more polite and pleasant, while still permitting to transfer the required message. Hedge is an expression of approximation: *You are quite right*. The theory of politeness describes hedges as a strategy of negative politeness [1: 145] and prefers these utterances to direct judgements, like *You are right!* At the same time, hedge can also serve as a discourse marker of (a) uncertainty and hesitation, when the speaker is not confident about the judgement and adds a hedge to make it less definite, (b) dialogue turn taking, when a speaker says *I guess* to gain people's attention [2]. In multimodal behavior hedges can be combined with nonverbal signs of hesitation or confusion. In our study we want to evaluate the perception of multimodal hedges in two different situations: where the speaker threatens the social face of the hearer or his own social face – see [1: 67]. We shall execute these studies in interactive communications with two companion robots, as a robot can precisely reproduce the required behavioral patterns in interactive situations. Although the experimental talks with robots may not exactly imitate natural human communication, robots may maintain interactive communication with people in exact and determined way that cannot be achieved in interactive human-to-human experiments.

In our study within human communication with companion robots, we put forward two hypotheses: (1) the expression of verbal and non-verbal hedges makes the speaker more *polite* and *comfortable to communicate with*, (2) multimodal hedges are the expression of emotions that can make a communication *friendlier*, and the speaker – more *sympathy provoking*. Our goal is to find the boundaries of the theory of politeness, applied to communication with robots, and study the conditions, where a hedge is perceived as (a) a mean of politeness, or (b) a marker of internal nervousness and hesitation. The application of politeness strategies to the robots giving advice may have very promising perspectives [3]. Robots communicating with people may naturally fail (be corrected by humans) or correct a human, thus, requiring some politeness strategies to support natural communication.

To test the hypotheses, we have executed two experiments, where (a) robot affects user's social face by correcting user's mistakes, and (b) robot loses its social face by making slight pre-programmed mistakes in its answers. In each experiment, one of the robots uses hedges, while the other does not. We evaluated human perception of the robots via surveys. As the two robots are identical in their behavior

(except for the hedges), we are able to justify the differences in evaluations by the usage of hedges by one the robots.

We have been concentrating on the situations of communication, where success or failure is linked to some oral production. For the first experiment we were looking for a setup, where the user makes real mistakes and the robot has to indicate these mistakes to the user. We have chosen a situation of word learning, where the human participant practices memorizing words of a foreign language, while the robot corrects its mistakes. For the symmetric experiment we were looking for a situation, where the robot fails in its oral production. We have selected an exam situation, where a participant asks the robot some exam questions, and the robot answers with slight mistakes in its statements. Each experiment was performed with two robots, where the first robot used hedges, while the second robot used gestures and speech, supporting its judgement: addressive strategy or negation.

2 Experiment 1: Robot affects user's social face by indicating his mistakes

To study the situation where a speaker affects the social face of the hearer, we have simulated a word learning environment, where the user (hearer) was learning Latin words with the companion robot (speaker). 38 participants took part in the experiment, mean age 19. Each participant started the experiment with one of the two robots, the order was randomized for each participant. The experiment with each robot was divided into two stages: word acquisition and word training (see Figure 1). During the acquisition phase each of the Latin words was introduced to the participant on a screen with a translation into Russian. Pre-recorded pronunciation of a Latin word by a professional Latin teacher was transmitted via the speakers. The robot then announced a keyword to help remember the Latin word. Keywords were phonetically similar Russian words, selected in a preliminary survey (n = 42, mean age 22, 28 females). The robot used Yandex speech API service for speech production.

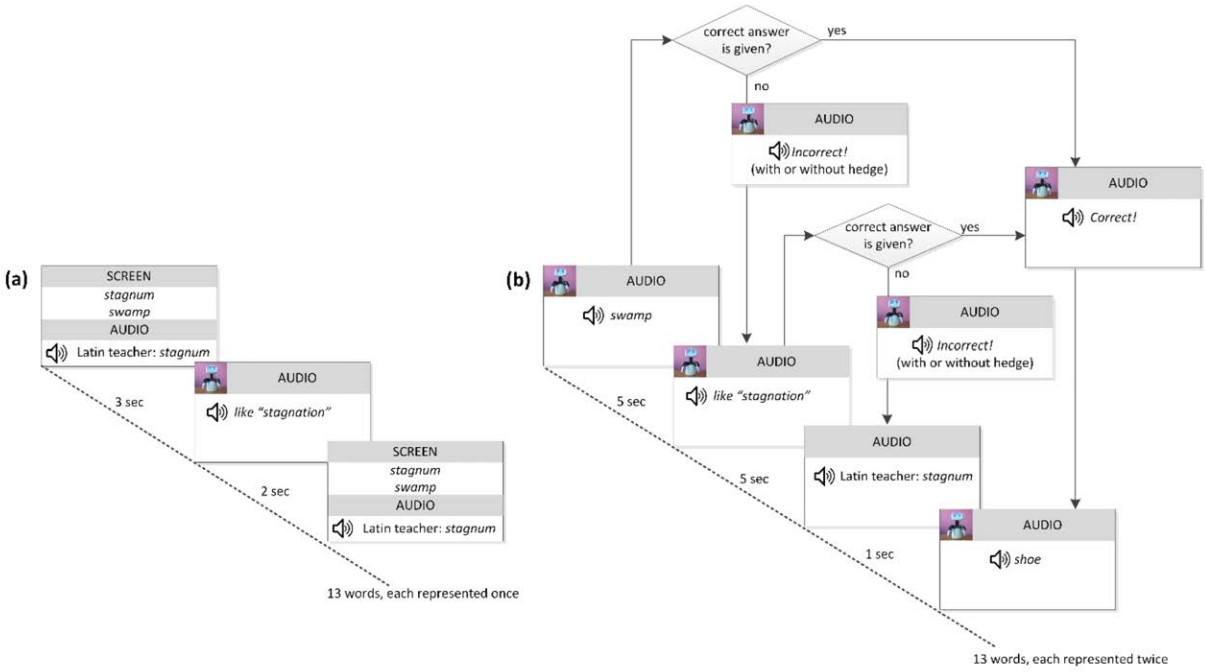


Figure 1: Scheme of Experiment 1 for a single condition. Words with keywords (hints) are introduced on the acquisition phase (a). On the training phase (b) the robot asks to translate each word and replies with a negative reaction (*Incorrect!*) with or without hedge – or with a positive reaction (*Correct!*).

During the training phase the robot announced a word in Russian and waited 5 seconds for the translation into Latin. If no correct answer was given within 5 seconds (silence was treated as *an incorrect answer*), the robot announced that the answer is not correct, offered the keyword and waited another 5 seconds. If no correct answer was given, the robot reacted as to an incorrect answer, the correct translation was announced by the Latin teacher (pre-recorded audio via the speakers) and the robot moved to

the next word in 1 second. Correct answers were marked by the experimenter from another room via Wizard-of-Oz scheme to start the "positive" reaction, while robot's reactions to errors/silence were automatic. The order of words was randomized; on the training phase each word was offered twice. Computer screens were not used on the learning phase, participants only communicated with robots: they saw the robots and heard robots' speech as well as the correct pronunciation of the words by the Latin teacher via the speakers.

Two robots differed in two experimental conditions: the first robot accompanied its reactions to incorrect answers by multimodal hedges, for example, by saying *No! A bit incorrect!* and manipulating its hands, while the other robot said only *Incorrect!* and used gestures, supporting the negation, like shaking its head or hand. The gestures were selected from the Russian Emotional Corpus [4, 5, 6] as typical multimodal behavioral patterns for the corresponding utterances; gestures were reproduced on the robot to be used in the experimental protocols. Behavioral protocols for the robots were designed in the Behavior Markup Language [7].

After word learning with one of the robots, participants filled out a questionary to evaluate the interaction and moved to the table with the other robot to study the next batch of Latin words. After the sessions with the two robots, participants were invited to another room to check the learned words and fill out the final questionary to compare the robots.

The experiment did not show any significant difference in the efficiency of word learning. However, the robot with hedges was preferred as a potential learning partner: 42% of the participants chose the robot with hedges, 21% with negations, and 37% evaluated robots equally. Not all the participants noticed the difference between the two robots, but many of them implicitly preferred the one with multimodal hedges. At the same time, several subjects explicitly noticed the differences, but have preferred the "strict" robot that clearly corrected the errors, as this type of control suited them and corresponded to the traditional role of a "strict teacher".

3 Experiment 2: Robot loses its social face by making mistakes

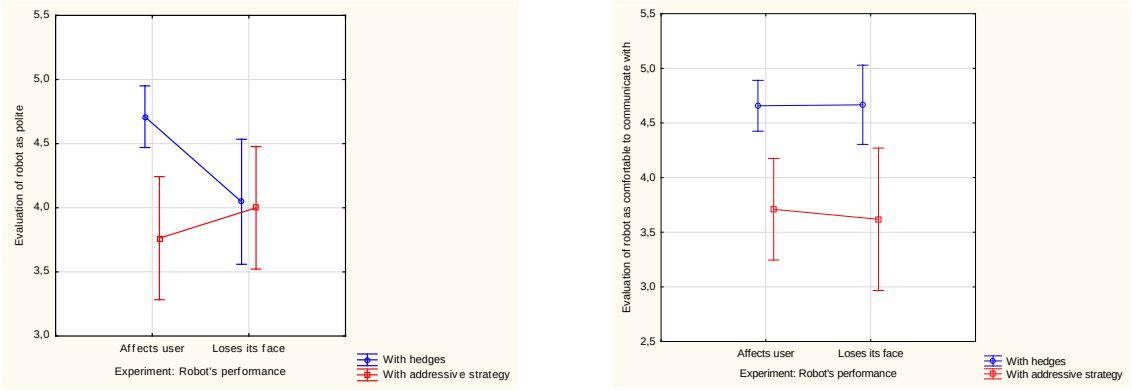
Within the second experiment the robot had to experience failures in its speech production and compensate it with a hedge. We have selected a situation, where students interviewed the robot on the questions of an actual university course "Introduction to Semiotics". 21 participants took part in the experiment with mean age 20. The list of 8 exam questions with the correct answers was reviewed by participants before the experiment and remained on the table during the experiment. Participants had to interview one of the robots, asking one question after another, and then – the other robot. The order of robots was randomized for each participant. After user's question, the corresponding answer to be given by the robot was selected by the experimenter via Wizard-of-Oz scheme. So, the robot could answer questions in randomized order, as it was, indeed, suggested by some participants. The questions were similar for the two experimental conditions. Each answer contained a slight pre-programmed inaccuracy: the robot indicated *century* (instead of the exact year), indicated only one option out of three, or made a mistake in the second name of a scientist. The mistakes were similar for the two experimental conditions. Users had the ability to control robot's mistakes as they had the correct answers on the table during the whole experiment. Robot's answer consisted of three parts, the robot (a) hesitated – looked aside or upward, joined its hands, (b) reported the answer with no gestures (eye movements were allowed), (c) for the 1st condition – demonstrated a hedge with speech and gestures, for example, said *I think so*, bit its lip and manipulated hands (see Figure 2a), and for the 2nd condition – demonstrated addressive strategy, for example, said *That's it* and waved its hand towards the human (see Figure 2b). Parts (a) and (b) of the reaction, including answers and mistakes, were identical for the two conditions. Between the answers robots demonstrated slight movements, typical for inactive behavior. After the interaction with each robot a participant had to fill out a questionary, reporting, if the robot *hesitated, was nervous, made a lot of mistakes, answered confidently, was comfortable to communicate with*, etc. Participants had also to evaluate the perceived psychological characteristics of the robot, by rating it as *friendly, competent, sympathy provoking, apathetic, emotional* (etc.) on 5 points scale from *very unlikely* to *very likely*.

(a) (b)

Figure 2: F-2 Robot with (a) hedge – biting the lip and manipulating hands, or (b) addressive gesture

4 Results

Regarding the usage of hedges within the politeness theory (the first hypothesis), in the first experiment, **the robot, attacking user's social face and using hedges,** was perceived as more *polite* ($p < 0,01$, Mann-Whitney U Test) (Fig. 3a), on the contrary, **robot without hedges** was evaluated as *more hostile* ($p < 0,01$), *indifferent* ($p < 0,01$) and *condemning* ($p < 0,01$); its corrections were *more confusing* to a user ($p < 0,01$). Robot with hedges was evaluated as *more trying to establish contact* ($p < 0,05$), as compared to the robot without hedges. In the second experiment, **the robot, making mistakes and using hedges,** did not appear to be *more polite* (no significant results). While the evaluation of the robot with hedges *as polite* was significant only for the first experiment, robots with hedges in both experiments were evaluated as more *comfortable to communicate with* (Mann-Whitney U Test, $p < 0,05$) (Fig. 3b).



(a) (b)

Figure 3: Hedges within the politeness theory. Robot seems *polite* when it uses hedges while affecting a user, not when it's losing its face (a). Robot seems *more comfortable to communicate with*, when it uses hedges in both conditions (b).

Regarding the usage of hedges to establish an emotional contact (the second hypothesis), in the first experiment, **the robot, attacking user's social face and using hedges,** is considered as *competent* ($p < 0,05$), *responsive* ($p < 0,01$), *caring* ($p < 0,05$). Also, this robot was evaluated *as calm* ($p < 0,01$), as compared to the robot without hedges.

In the second experiment: **robot, making mistakes and using hedges,** was evaluated as *hesitating* ($p < 0,01$) and *nervous* ($p < 0,05$), while **the robot with addressive strategy** was *answering clearly* ($p < 0,05$) and more *detached* ($p < 0,05$).

In both experiments, robots with hedges are perceived as *friendly* (Mann-Whitney U Test, $p < 0,01$) (Fig. 4), *sympathy provoking/cute* (Mann-Whitney U Test, $p < 0,01$) and *good-hearted* (Mann-Whitney U Test, $p < 0,01$).

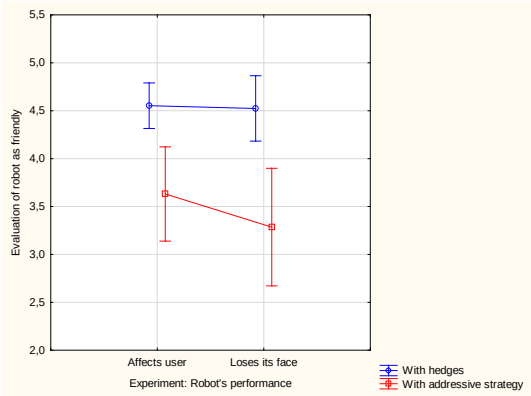


Figure 4: Hedges to establish an emotional contact: robot seems *friendly*, when it uses hedges in both conditions.

5 Discussion

Our verification of hedges as a strategy of politeness in its "strong" definition (*the usage of hedges makes speaker sound more polite*) applies only to the situation, where the speaker attacks the social face of the hearer: e. g., corrects hearer's mistakes. At the same time, the understanding of hedges as a politeness strategy in more "moderate" definition (*the usage of hedges makes conversation more comfortable*) applies to both situations: when the speaker attacks social face of the opponent or loses his face due to his own mistake.

The first observation may seem trivial: indeed, the robot using *politeness strategy* seems more *polite*. At the same time, this starting point testifies that the politeness strategy does apply to robots (at least, within the modelled setup), as some people (schoolchildren) prefer the robot without politeness strategies and consider it as more modern, close to the speech of school children [8]. As an ambiguous expressive pattern, a hedge may contribute not only to the expression of politeness, but also to the expression of emotional and cognitive states: *nervousness* and *hesitation*. Our verification of hedges as a mean to convey the internal emotional state gave quite compound results. Indeed, the robot, giving wrong answers, is perceived as *hesitating* and *nervous*, so hedges can indicate the internal emotional state. At the same time, hedges (as an indication of internal confusion) can provoke some complementary emotions of the hearer, like compassion and sympathy. Some experiment participants – students – did underline that they associated themselves with the robot, who makes mistakes in exam answers and hesitates. So, a negative emotional state (nervousness) can provoke a positive emotional state of the hearer and establish the emotional contact in general: robots with hedges we perceived as *friendly/cute/good-hearted* in both situations.

At the same time, in the situation where the speaker (the robot) controls the hearer by asking the lexical questions and indicating hearer's mistakes, the speaker's hedges make him *competent* and *responsive*. We suggest that its use of hedges naturally allows a human to assign to the speaker *locus of control* (teacher's role) and, thus, treat the speaker as more *competent* and *responsive*. This observation contributes both to the first and second hypotheses.

While the robot with hedges in the second experiment was *more nervous*, in the first experiment it was, on the contrary, considered as *calmer*. We suggest that while in the second experiment hedges played their primary expressive role (the expression of hesitation and nervousness), in a situation, where the speaker governs the hearer, hedges (as voluntary usage of a politeness strategy) indicate speaker's degree of self-control, thus, he is considered as *calmer*, as compared to the speaker without hedges.

In the experiment 2, we have compared hedges with addressive gestures. The robot with addressive gestures was considered as *answering clearly*, which can be treated as a contribution of addressive gestures (as compared to hedges). At the same time, addressive strategies in this condition cannot be considered as a form of positive politeness, as they made the robot look *detached* – not *empathetic*, as it could be suggested, if the addressive gestures contributed to positive politeness.

5 Conclusion

As can be demonstrated in the experiments with companion robots, multimodal hedges contribute to the politeness in different situations by making the speaker more *comfortable to communicate with*. At the same time, hedges make the speaker more *polite* only when he affects the social face of the hearer, e. g. corrects hearer's mistakes.

The compared results of the two experiments allow us to suggest the following spectrum of communicative functions for hedges. Hedges, as a language formula, prototypically express inexactness and tentativeness. They initially correspond to the emotional expression of *hesitation* and *nervousness* of the speaker. Indeed, a speaker, who makes mistakes and uses hedges is evaluated as *nervous* and *hesitating*. This emotional state can invoke the compassion of the hearer and make him perceive the speaker as a friendly interlocutor in a wide range of situations: where speaker loses his face or attacks the faces of others. This usage of hedges corresponds to a wider definition of politeness strategies, as a hedge makes communication more **comfortable** – both, when the speaker loses his social face or has to attack the social face of the hearer. The ability of the speaker to use hedges in a situation, where he governs and corrects the hearer, makes him sound *caring* and *responsive*: i. e. the hearer agrees with the transfer of control to the speaker, who uses hedges. Moreover, the hearer considers a speaker with hedges as *more competent*. And finally, hedges contribute to making the speaker more polite – the core function of hedges, as described by the theory of politeness. However, this applies only to the situations, when the speaker threatens the social face of the hearer. This corresponds to the narrow understanding of a hedge as a politeness strategy.

This spectrum demonstrates the transition of hedge from an expressive negative emotional reaction (*nervousness*, *hesitation*) to a marker of speaker's *care* and *competence* and finally – to a politeness strategy.

Acknowledgements

The project is in part supported by the Russian Science Foundation, project No 19-18-00547, https://rscf.ru/project/19-18-00547/

References

[1] Brown Penelope, Levinson Stephen C. Politeness: Some universals in language usage. – Cambridge: Cambridge university press, 1987 – Vol. 4.

[2] Paggio Patrizia, Navarretta Costanza (2017), The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations, Language Resources and Evaluation, Vol. 51(2), pp. 463-494.

[3] Strait Megan, Canning Cody, Scheutz Matthias. Let me tell you! Investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance // Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction. – 2014. – P. 479-486.

[4] Kotov A. A., Budyanskaya E. (2015), The Russian Emotional Corpus: Communication in Natural Emotional Situations, Computational Linguistics and Intellectual Technologies, M., Vol. 11(18), pp. 296–306.

[5] Zinina A., Arinkin N., Zaidelman L., Kotov A. (2018), Development of a communicative behavior model for F-2 robot based on the REC multimodal corpus, Computational linguistics and intelligent technologies, Vol. 17(24), pp. 831-844.

[6] Kotov A. A., Zinina A. A. (2015), Functional analysis of nonverbal communicative behavior [Funkcional'nyj analiz neverbal'nogogo kommunikativnogo povedeniya], Computational Linguistics and Intellectual Technologies [Komp'yuternaya Lingvistika i Intellektual'nye Tekhnologii], M., Vol. 14(21), pp. 308-320.

[7] Kopp Stefan, et al. Towards a common framework for multimodal generation: The behavior markup language // Proceedings of the Intelligent Virtual Agents: 6th International Conference, IVA 2006, Proceedings 6 – Springer Berlin Heidelberg, 2006. – P. 205-217.

[8] Zinina, A. A., Zaidelman, L. Y., Kotov, A. A., Arinkin, N. A. (2020), The perception of robot's emotional gestures and speech by children solving a spatial puzzle, Computational Linguistics and Intellectual Technologies, M., Vol. 19 (26), pp. 811-826.

References

[1] Brown Penelope, Levinson Stephen C. Politeness: Some universals in language usage. – Cambridge: Cambridge university press, 1987 – Vol. 4.
[2] Paggio Patrizia, Navarretta Costanza. The Danish NOMCO corpus: multimodal interaction in first acquaintance conversations // Language Resources and Evaluation – 2017. – Vol. 51(2) – P. 463-494.
[3] Strait Megan, Canning Cody, Scheutz Matthias. Let me tell you! investigating the effects of robot communication strategies in advice-giving situations based on robot appearance, interaction modality and distance // Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction. – 2014. – P. 479-486.
[4] Kotov A. A., Budyanskaya E. The Russian Emotional Corpus: Communication in Natural Emotional Situations, Computational Linguistics and Intellectual Technologies. – 2015. – Vol. 11(18). – P. 296–306.
[5] Zinina A., Arinkin N., Zaidelman L., Kotov A. Development of a communicative behavior model for F-2 robot based on the REC multimodal corpus // Computational linguistics and intelligent technologies. – 2018. – Vol. 17(24) – P. 831-844.
[6] Котов А.А., Зинина А.А. Функциональный анализ невербального коммуникативного поведения. // Компьютерная лингвистика и интеллектуальные технологии. – М.: РГГУ, 2015, Том 1(14). – С. 308-320.
[7] Kopp Stefan, et al. Towards a common framework for multimodal generation: The behavior markup language // Proceedings of the Intelligent Virtual Agents: 6th International Conference, IVA 2006, Proceedings 6 – Springer Berlin Heidelberg, 2006. – P. 205-217.
[8] Zinina, A. A., Zaidelman, L. Y., Kotov, A. A., Arinkin, N. A. The perception of robot's emotional gestures and speech by children solving a spatial puzzle // Computational Linguistics and Intellectual Technologies. – 2020. – Vol. 19 (26) – P. 811-826.

Augmentation methods for spelling corruptions

Nikita Martynov
SberDevices
Moscow
nikita.martynov.98@list.ru

Mark Baushenko
SberDevices
Moscow
MABaushenko@sberbank.ru

Alexander Abramov
SberDevices
Moscow
Abramov.A.Sergee@sberbank.ru

Alena Fenogenova
SberDevices
Moscow
alenush93@gmail.com

Abstract

The problem of automatic spelling correction is vital to applications such as search engines, chatbots, spell-checking in browsers and text editors. The investigation of spell-checking problems can be divided into several parts: error detection, emulation of the error distribution on the new data for model training, and automatic spelling correction. As the data augmentation technique, the adversarial training via error distribution emulation increases a model's generalization capabilities; it can address many other challenges: from overcoming a limited amount of training data to regularizing the training objectives of the models. In this work, we propose a novel multi-domain dataset for spelling correction. On this basis, we provide a comparative study of augmentation methods that can be used to emulate the automatic error distribution. We also compare the distribution of the single-domain dataset with the errors from the multi-domain and present a tool that can emulate human misspellings.

Keywords: spelling correction, augmentation strategies, adversarial attacks, error detection
DOI: 10.28995/2075-7182-2023-22-327-349

Методы аугментации для задачи автоматического исправления орфографии

Никита Мартынов
SberDevices
Moscow
nikita.martynov.98@list.ru

Марк Баушенко
SberDevices
Moscow
MABaushenko@sberbank.ru

Александр Абрамов
SberDevices
Moscow
Abramov.A.Sergee@sberbank.ru

Алена Феногенова
SberDevices
Moscow
alenush93@gmail.com

Аннотация

Автоматическая коррекция орфографии актуальна для многих приложений, таких как поисковые системы, чат-боты, текстовых редакторах и тд. Системы автоматического распознавания и исправления опечаток часто используют в кач-ве метода аугментации данных. Это повышает метрики оценки на низкоуровневых задачах, увеличивает обобщающую способность модели и её робастность. В этой работе мы впервые представляем новый многодоменный набор данных для исправления орфографии. На его основе мы предлагаем несколько подходов к аугментации данных и проводим сравнительную оценку методов увеличения данных с различными распределениями ошибок, которые можно в дальнейшем использовать для эмуляции автоматического распределения ошибок.

Ключевые слова: проверка орфографии, автоматическое определение ошибок, методы аугментации данных

1 Introduction

The task of automatic spelling correction (or spell-checking) is crucial for many applications in different areas, including correction of search queries, spell checking in browsers, text editors etc. There are plenty of methods for spelling detection and correction. In recent research, with new big language models, the generation of texts without spelling errors expands new horizons. There are various methods of automatic text corruption and augmentations for further model training on parallel texts. However, more reliable information on human error distribution in the text data needs to be found. How well existing approaches can approximate the natural error distribution is still an open question. The influence on the quality of the generative models trained on such data is also a new field for investigation.

In this paper, we deal with several of these research problems. Due to the lack of data for the Russian language of different domains with spelling errors, we present a new parallel dataset for spelling correction. We propose two methods for spelling correction. On this basis, we conduct a comparative study of these augmentation algorithms that can be used to emulate spelling error distribution. Our key contributions to the paper are the following:

- **We introduce a novel multi-domain dataset for spelling correction**. We compare the public single-domain dataset from the Shared Task SpellRuEval-2016 (Sorokin et al.,) with the obtained golden multi-domain set and prove that the domain distributions differ in various domains.
- **We propose two approaches to generate spelling error distribution.**
  - We introduce the augmentation method that emulates human spelling errors based on statistical data and heuristics from keyboard usage. Such a method can produce corrupted text without any labelled data. The obtained spelling error distribution from texts corrupted with this method is compared with the golden test sets spelling error distribution.
  - We provide the augmentation tool based on the method that gathers the error distribution from the parallel corpus and can replicate the obtained source distribution on a new text based on classic Levenshtein operations (Lhoussain et al., 2015) (deletion, insertion, substitutions). We clone the error distribution from the golden set and compare the emulated with the original distribution.

The remainder is structured as follows. First, we overview the approaches to spell correction 2, the available datasets and methods for error augmentations. Section 3 describes the data sources and the annotation procedure for creating the Russian multi-domain corpus. In section 4, we observe the augmentation methods and models used and provide the description of the comparable experiments. The statistical evaluation is presented in Section 5.

2 Related Work

The problem of spelling correction has a long history of research. It attracted intensive attention in the early era of modern NLP. The most significant early works are the edit distance model, introduced by Levenshtein (Levenshtein and others, 1966) and further by Damerau (Damerau, 1964). Weighted variants of error distances were considered in (Kemighan et al., 1990) and Brill and Moore (Brill and Moore, 2000). The latest also proposed the noisy channel error correction model based on n-grams. Toutanova and Moore (Toutanova and Moore, 2002) added a pronunciation model for spelling correction. A broad historical overview of the problem is presented in the paper (Shavrina, 2017), where the author discusses the history of methods of automatic spelling correction and the requirements faced by the systems implementing such methods at different historical stages.

The interest in this field for the Russian language appears after SpellRuEval-2016 (Sorokin et al.,) competition. The authors created the single domain dataset for social media texts and provided the first benchmark and standard for spelling correction problems. Among other public popular solutions for Russian language are Yandex.Speller [1], DeepPavlov [2] method based on Damerau Levenshtein and

[1] https://yandex.ru/dev/speller/

[2] https://docs.deeppavlov.ai/en/master/features/models/spelling_correction.html

KenLM, Hunspell [3], Jamspell [4]. It's necessary to mention a multilingual source of parallel spell data – GitHub Typo Corpus (Hagiwara and Mita, 2019). It is a large-scale, multilingual dataset of misspellings and grammatical errors along with their corrections harvested from GitHub. For state-of-the-art spelling systems, the generative models [5] are applied. For its training, the parallel corpus needs to be built from scratch; emulating spelling errors or augmentation of the existing datasets is required.

The approaches for error augmentations are common and applied in further research. For example, they are incorporated in the GEM benchmark (Dhole et al., 2021), and its augmentation NL-Augmenter library [6]. The (Benes and Burget, 2020) examines the effect of data augmentation for training language models for speech recognition and investigates the behaviour of perplexity estimated on augmented data. For the Russian Language frameworks RuTransform (Taktasheva et al., 2022) [7] adds noise to data via spelling corruption. It contains the ButterFingers method, employed at the word level, as well as the sentence-level techniques of word swapping (EDA_{SWAP}) and token deletion (EDA_{DELETE}). The ButterFingers method, derived from the NL-Augmenter, constitutes a typo-based perturbation approach that adds noise into textual data and Case methos introduces noise to data through case alteration. This is accomplished by simulating spelling errors made by humans through character swaps, taking into account the keyboard distance between the characters. Notably, these methods are applicable to both the Russian and English languages.

3 Data

We acquire text data from publicly available sources out of five domains to create a multi-domain corpus. Due to human and time constraints, all the texts are automatically checked for the presence of spelling mistakes. For the sentences with potential misspellings, we set up a two-stage human annotation procedure. As a result, we select 1711 parallel sentences based on the agreement between annotators. You can see the full breakdown in Table 1.

3.1 Data sources

The choice of the domains of our primary interest lays upon the following criteria:

- The texts from a particular domain must be misspellings-prone.
- The representation of a domain should reflect the frequency of misspellings present within it. By assuming that texts belonging to a particular domain inherently contain spelling errors, it follows that a larger corpus of texts would naturally yield a greater number of sentences, thus expanding the dataset.
- Finally, the resulting domains must be diverse in terms of vocabulary, grammatical structures, slang, jargon etc. It ensures we capture different types, positions and co-occurrences of misspellings.

These conditions lead to the following choice of domains and corresponding datasets.

Aranea web-corpus (Benko, 2014) is a family of multilanguage gigaword web-corpora collected from Internet resources. The texts in the corpora are evenly distributed across periods, writing styles and topics they cover. We randomly picked the sentences from Araneum Russicum [8], which is harvested from the Russian part of the web.

Literature is a collection of Russian poems and prose of different classical literary works. We randomly picked sentences from the source dataset[9] that were gathered from Ilibrary, LitLib, and Wikisource.

News, as the name suggests, covers news articles on various topics such as sports, politics, environment, economy etc. The passages are randomly picked from the summarization dataset Gazeta.ru. (Gusev, 2020)

[3] https://github.com/pyhunspell/pyhunspell
[4] https://github.com/bakwc/JamSpell
[5] https://huggingface.co/UrukHan/t5-russian-spell
[6] https://github.com/GEM-benchmark/NL-Augmenter
[7] https://github.com/RussianNLP/rutransform
[8] http://ucts.uniba.sk/aranea_about/_russicum.html
[9] https://www.kaggle.com/datasets/d0rj3228/russian-literature

Social media is the text domain from social media platforms marked with specific hashtags. These texts are typically short, written in an informal style and may contain slang, emojis and obscene lexis.
Strategic Documents is part of the dataset the Ministry of Economic Development of the Russian Federation collected. Texts are written in a bureaucratic manner, rich in embedded entities, and have complex syntactic and discourse structures. The full version of the dataset has been previously used in the RuREBus shared task (Ivanin et al., 2020).

Datasets	Raw texts	Yandex.Speller	Filtered texts	First stage	Second stage
Aranea web-corpus	45512	3761	985	859	756
Literature	24635	1808	494	262	260
News	2001	245	245	245	245
Social media	25883	3000	281	208	200
Strategic Documents	44458	2000	284	250	250
TOTAL	**142489**	**10814**	**2289**	**1824**	**1711**

Table 1: The number of sentences on all stages of the dataset creation among all domains. *Raw texts* is several texts in the source; *Yandex.Speller* is a number of texts marked by Yandex.Speller that can contain misspellings. *Filtered texts* reflects texts sent to manual labeling; *First stage* corresponds to the texts passed to second stage of labeling; *Second stage* is a number of resulting sentences.

3.2 Candidate selection

First, we automatically detect mistakes with Yandex.Speller[10]. We find out that Yandex.Speller is often triggered by proper names, slang, abbreviations, obsolete and rare word forms (see Table 2 for illustrative examples) that do not contain any spelling errors.

Second, in this paper we do not consider specific vocabulary, e.g. slang, jargonisms, colloquialisms etc., as an error, as we see them as style markers that reflect distinctive domain features. For example, the word *"ето"* in a sentence *"тут ето, коты синхронизировались"* *("here it is, the cats are synchronized")* from Social media domain is not correct in terms of a standard language. Still, this word is presumably used to endow a sentence with a particular emotional expression. Nevertheless, we do not allow all the misspellings in specific vocabulary - we only keep those written deliberately. For example, in a sentence *"Когда типо болеешь и не пошел в универ: "* *("When you are supposedly sick and did not go to university:")* we have word "типо" which is just incorrect form of "типа" and does not carry any emotional or stylistic pallet. The preservation of lexicon of this kind is crucial considering practical value associated with systems trained on such data. In this work, we agreed to let annotators, who are native Russian speakers and passed the language exam, decide whether spelling errors in particular cases need to be corrected given the general instructions (see Section 3.3 for details).

Due to these two observations, we had to manually revise all the candidates that Yandex.Speller suggested.

3.3 Annotation

Next, we set up two-stage annotation project via a crowd-sourcing platform Toloka[11] (Pavlichenko et al., 2021):

1. **Data gathering stage**: we provide the texts with possible mistakes to annotators and ask them to write the sentence correctly;
2. **Validation stage**: we provide annotators with the pair of sentences (source and its corresponding correction from the previous stage) and ask them to check if the correction is right.

The designs of both projects are presented in Figures 4(see Appendix A 7).

We prepared instructions for annotators for each task. The instructions ask annotators to correct misspellings if it does not alter the original style of the text. Instructions do not provide rigorous criteria

[10]https://yandex.ru/dev/speller/
[11]https://toloka.ai/tolokers

Datasets	Sentence	Type
Aranea web-corpus	**Паррикар** говорит: пусть русские приезжают в Индию, веселятся, тратят деньги. ***Parrikar** says: let the Russians come to India, have fun, spend money.*	Proper name
Literature	**Лгание** Муция Сцеволы до сих пор не обличено ***The lies** of Mucius Scaevola have not yet been exposed*	Obsolete word
News	Лидером **антитопа** стал Мэттью Макконахи, звезда «Настоящего детектива». *The leader of the **antitope** was Matthew McConaughey, the star of True Detective.*	Rare word
Social media	Студент отправил файл с **домашкой** и удалил. спрашиваю: где файл? *The student sent a file with **homework** and deleted it. I ask: where is the file?*	Slang
Strategic Documents	**Кмо** - число объектов культурного наследия, по которым проведен мониторинг ***CMO** - number of objects of cultural heritage, for which monitoring was carried out*	Abbreviations

Table 2: False triggered examples of Yandex.Speller across all **Datasets** with **Type** of misleading trigger attached. All sentences are from corresponding datasets. The boldface indicates words that Yandex.Speller considers misspellings.

on the matter of distinguishing the nature of an error in terms of its origin - whether it came from an urge to endow a sentence with particular stylistic features or from unintentional spelling violation since it is time-consuming and laborious to describe every possible case of employing slang, dialect, colloquialisms, etc. instead of proper language. Instructions also do not distinguish errors that come from the geographical or social background of the source. Instead, we rely on annotators' knowledge and understanding of a language since, in this work, the important factor is to preserve the original style of the text.

To ensure we receive qualified expertise, we set up test iteration on a small subset of the data for both stages. We manually validated the test results and selected annotators, who processed at least six samples (2% of the total test iteration) and did not make a single error. After test iteration, we cut 85% and 86% of labellers for gathering and validation stages.

We especially urge annotators to correct mistakes associated with the substitution of the letters "ё" "й" and "щ" for corresponding "е" "и" and "ш" and not to explain abbreviations and correct punctuation errors. Each annotator is also warned about potentially sensitive topics in data (e.g., politics, societal minorities, and religion).

The annotation details are provided in Table 4, and statistics of confidence levels across all datasets on both stages are provided in Table 3.

Datasets	N_{FI}	C_{FI}	N_{FO}	C_{FO}	C_{SI}	N_{SO}	C_{SO}
Aranea web-corpus	985	78.95	859	85.77	96.38	756	97.95
Literature	494	72.56	262	80.32	99.94	260	99.95
News	245	99	245	99	245	99.94	99.95
Social media	281	67.81	208	79.67	99.93	200	99.934
Strategic Documents	284	79.77	86.14	250	99.94	250	99.95

Table 3: Details on the confidence levels on both stages across all datasets. N_{FI} is a number of samples labelled in the first stage. We proceed with samples with confidence above 67% after the first stage and 90% after the second stage. N_{FO} and N_{SO} are the number of texts selected after the first and second stages, respectively. C_{FI}, C_{FO}, C_{SI} and C_{SO} refer to confidence levels calculated on the corresponding stage and subset in %. C_{FI} and C_{FO} are calculated as the expected value of annotators' support of the most popular correction. C_{SI} and C_{SO} are calculated based on aggregation of annotators' skills.[12]

[12]https://toloka.ai/docs/guide/result-aggregation/#aggr-by-skill

Task	IAA	Total	Overlap	N_T	N_{page}	N_C	N_U	ART
Part 1. Test Iteration	77.98	14$	3	7	4	50	96	132
Part 2. Test iteration	89.09	7.9$	3	8	5	46	74	77
Part 1. Gathering	79.10	112$	3	-	4	-	14	165
Part 2. Validation	99.23	92$	3	-	5	-	10	111

Table 4: Details on the data collection projects for the Golden test set. **IAA** refers to the IAA confindence scores, %. IAA of Part 1 is calculated as the expected value of annotators' support of the most popular correction over all labelled texts. IAA of Part 2 is calculated as an average value of confidence scores (see C_{SI} and C_{SO} in Table 3) over all labelled texts. **Total** is the total cost of the annotation project. **Overlap** is the number of votes per example. N_T is the number of training tasks. N_{page} denotes the number of examples per page. N_C is the number of control examples. N_U is the number of users who annotated the tasks. **ART** means the average response time in seconds.

4 Method

To prove the uniqueness and utility of a dataset, we compare distributions of its spelling errors with those of SpellRuEval-2016 (Sorokin et al.,) and synthetically generated misspellings. To generate errors, we employ two approaches. The first is based on the most common spelling errors, statistics and heuristics and can produce corrupted text without any labelled data. The second approach, on the contrary, needs annotated parallel samples to scan source misspellings and try to emulate the spanning errors process to replicate source distributions. We dedicate the following two subsections to describing both methods in more detail.

4.1 Augmentex

For the first time, we present a tool for augmenting text data and conducting black-box attacks on machine-learning models. Augmentex is based on statistical data and heuristics based on human behaviour when using the keyboard and supports two separate augmentation modes:

1. at the character level;
2. at the word level (each of which has 7 and 5 methods, respectively).

You can control the number of augmentations using three primary parameters: min_aug, max_aug and aug_rate. The last is responsible for the number of augmentation applications by specifying the number of percentages of words or characters of the source string to which methods should be applied. The first two arguments set the lower and upper limits of the number of methods applications. These parameters are necessary, as the source string must only remain completely with augmentations or, on the contrary, is not significantly distorted by them. For convenience, batch data processing was done, in which one can specify how many percentages of the source corpus of texts one needs to apply augmentations. So far, the methods have support for the Russian language, but the variety of languages will expand in the future.

Below, in sub-paragraphs 4.1.1 and 4.1.2, we will describe the operation of all methods in detail.

4.1.1 Methods at the character level

The application scenario is the same for all the methods below: based on the parameters described in paragraph 4.1, the integer N is determined. After that, N characters are randomly selected, and one method is applied to each.

Shift method. This method is based on the heuristic that when printing text, a computer user can sometimes press the $shift\ key$ on a keyboard; in this case, a completely different character will be printed. To do this, we have created a dictionary in which the keys are numbers from 0 to 9, and all letters are in uppercase and lowercase. As values for each key, we put the corresponding keys when the $shift\ key$ is pressed. As a result, we got a dictionary power equal to 76: 33 letters in both registers and 10 digits.

Spelling error method. This method is based on statistical error data collected by researchers from the project KartaSlov[13]. The data contains frequent words of the Russian language and variants of their incorrect spelling (both spelling errors and typos). All erroneous spellings are equipped with weights that can estimate the relative frequency of occurrence of specific errors. The obtained error matrix is a matrix of relative frequencies when instead of correctly using the letter X, the letter Y will be mistakenly used. The reason for the error can be either spelling or a typo. There are correct uses along the lines and erroneous ones along the columns. Each row is individually assigned to one by the maximum value. Thus, the most frequent error in each row will weigh 1.0. The heat map can be viewed in Figure 5 in Appendix B 8.

For ease of use, each line was normalized and written into a dictionary, where the key is a letter of the Russian alphabet. The value is a float list of length 33 with the probability of making a mistake in writing a letter from the key. While applying the method to a particular character, we get a list of probabilities and randomly select a new character according to the probabilities in this list.

Typo method. This method is based on heuristics when a computer user misses a key and accidentally touches an adjacent key. We have created a dictionary where the key is 1 of 33 characters of the Russian alphabet or 1 of 10 digits. By default, each character on the keyboard has six neighbours if you do not consider the extreme characters. For example, the character "п" will have 6 neighbors: "е", "н", "р", "и", "м" and "а". Therefore, we put a list containing neighbouring characters as values. When applying the method, adjacent characters are selected equally likely and replaced by the original character.

Method of deleting a character. When calling this method, an empty string is returned instead of the original character.

Random character insertion method. When calling this method, a place for insertion is randomly selected and a random character from the dictionary is inserted (for the Russian language, this is 33 letters).

Character repetition method. The method is based on the heuristic of the key sticking during typing and as a result of the repetition of consecutive characters in the text. It has an additional parameter $mult\_num$, which is responsible for the upper limit of the number of repetitions of the original character. During the application, the number of repetitions is randomly selected from the range of integers from 1 to $mult\_num$, and the original character repeated as many times is returned.

Character permutation method. This method is based on the heuristic that when typing text quickly, the user often confuses the order of pressing the keys, resulting in consecutive characters having the reverse order of writing. We replace the original character with the following places to model human behaviour.

4.1.2 Methods at the word level

The logic of applying the methods will be similar to that described in paragraph 4.1.1, but the word level is used instead of the character. These methods primarily aim to introduce various language errors (lexical errors, agreement, etc.). Some can be used to add spelling errors and typos at the character level (in this paper, we consider only the spelling errors). We present here the description of all the library features, as it's potentially valuable for future research to investigate the imitation of more complex types of errors than orthographic.

Word replacement method. This method is very similar to its character counterpart – The Spelling error method. Only now, the dictionary acts as an error matrix, where the keys are words without errors, and the value is a list of pairs of the form (a word with an error, the probability of writing this word). It has 22187 keys and 4.1 pairs on average. Researchers collected these statistics from KartaSlov, as we mentioned earlier.

Word deletion method. During the application of the method, an empty string is returned instead of the original word. The logic of the work is similar to the method of removing the character.

[13]https://github.com/dkulagin/kartaslov/tree/master/dataset/orfo_and_typos

Word permutation method. This method rearranges two adjacent words in places. It simulates a syntax error when the word order in a sentence is broken.

The method of adding parasite words. A corpus of the most common parasite words from various open sources was collected to implement this method. The cardinality of the set is equal to 70 words. When applied, one of the parasite words is equally likely inserted into a random place in the sentence, which models the illiterate use of words in speech. They clog up the text's meaning, making it indistinct and difficult to understand.

Capital Letter method. This method changes the case of the first letter in the word. It models the incorrect spelling of proper names in the Russian language.

4.2 Statistic-based spelling corruption

The goal of statistic-based corruption is to mimic misspellings distributions scanned from source texts. The algorithm consists of two consecutive parts: analysis of errors in given sentences, which results in corresponding distributions and applying these distributions to correct texts.

This method needs a parallel dataset, where pair of samples consists of a source sentence, which potentially has spelling errors, and a corresponding correct sentence. Datasets for a spellchecking task often come without any annotation on where the error is located in the source sentence. To analyze spelling errors, we have to know their exact positions. It can be achieved either by manual annotation or automatically. In this work, we implement an algorithm that detects the position of misspelling and its category following predefined types of string edits. The idea behind the algorithm is to calculate Levenshtein distances (Levenshtein, 1966) between all the prefixes of the source sentence and correction and traverse it back.

4.2.1 Error analysis

To analyse the errors, we first have to define the notion of spelling error, types of spelling errors and types of distributions that we model. First, in this paper, we accept only one option of proper spelling. All datasets described in the current work are parallel and have corrected sentences for each corresponding sentence with errors. We consider these corrections *proper spelling*. This arrangement is necessary to suggest the following precedents, which result in errors in correct spelling:

- **Insertion**: insertion of a character;
- **Deletion**: deletion of a character;
- **Substitution**: substitution of a character for another non-identical character;
- **Transposition**: switching places of two contiguous characters;
- **Extra separator**: insertion of a gap;
- **Missing separator**: deletion of a gap;

Characters are represented only by letters of the Russian alphabet. We do not include punctuation signs and letters from other languages. We define a spelling error as an event that can be described by one and only one of the listed precedents. We add uniqueness property to the definition of spelling error to avoid interpretations of a particular event as a composition of multiple precedents. For example, the transposition of two contiguous letters gives the same result as substitution of these letters on one another.

Since we defined the notion of spelling error, we can now describe it with corresponding types. We set the type of error as a random variable T, which can take one of the six possible categories. Each category is a precedent. This assumption is correct because we restricted spelling errors to be described by only one of the precedents. Because T takes one of the six possible outcomes, we assume T follows multinoulli distribution D_T. To describe D_T, we have to estimate the probabilities of each outcome. In this paper, we calculate the number of appearances of each error type, normalize them by the total number of misspellings and use these estimates as parametrization for T.

Another important attribute of an error, that should be studied, is its position in a sentence. We calculate the relative position of a misspelling by dividing its absolute position by the number of characters in a sentence. We treat the relative position as a random variable P distributed according to unknown continuous distribution D_P and take values from the interval [0, 1]. For simplicity, we split this interval

into ten equal non-intersecting semi-open subintervals and model the probability that P will fall in one of them. Since the particular value of P can only take one subinterval, we can say that random variable $\hat{P}$, which describes the categorization of P, follows multinoulli distribution $D_{\hat{P}}$. Analogously, we model it by counting encounters of different subintervals and normalizing it to valid discrete distribution. To analyze different types of errors more thoroughly, we consider P and corresponding $\hat{P}$ to be unique for each misspelling type.

The last characteristic we want to keep track of is the number of spelling errors per sentence. The random variable N, which takes integer numbers starting from zero, can describe this characteristic. For simplicity again, we suggest that N follows multinoulli distribution D_N, with the number of possible outcomes equal to the maximum number of errors in a sentence. We use the same procedure to estimate parameters of D_N.

4.2.2 Text corruption

Since we know how to estimate parameters of D_T, $\hat{P}$ for each type of misspelling and D_N, we can use these distributions to corrupt the correct text and expect corresponding distributions of corrupted texts to be similar to those of source texts. We sample the number of misspellings from D_N for each sentence in a corpus of presumably correct sentences. Then for each misspelling, we sample its type from D_T and its subinterval from $\hat{P}$, corresponding to the selected type. To calculate the exact position of an error in a sentence, we scale back the boundaries of subinterval according to the number of characters in a sentence and sample random positions within these boundaries. We check if sampled position satisfies predefined conditions for the particular type of error. For example, we do not allow the deletion of punctuation signs or the insertion of a double gap. If conditions do not hold, we sample position again or skip this misspelling. If position is found, we apply a selected type of error and proceed to the next misspelling or following sentence. The pseudocode for this procedure can be seen in listing 8 in Appendix B 8.

5 Evaluation

The evaluation process is separated into two parts. First, we evaluate our multi-domain dataset and compare misspellings distributions, described in Section 4.2, with corresponding distributions of SpellRuEval-2016 (Sorokin et al.,) to ensure we bring novelty in the field of automatic spelling correction explained by multi-domain nature of the corpus. Second, we want to evaluate methods of spelling corruption proposed in Section 4. These tools aim to mimic human spelling errors to some degree of accuracy. We generate synthetic misspellings with both methods on the correct sentences of our dataset. We then compare synthetic and natural error distributions analogously to the first part of the evaluation.

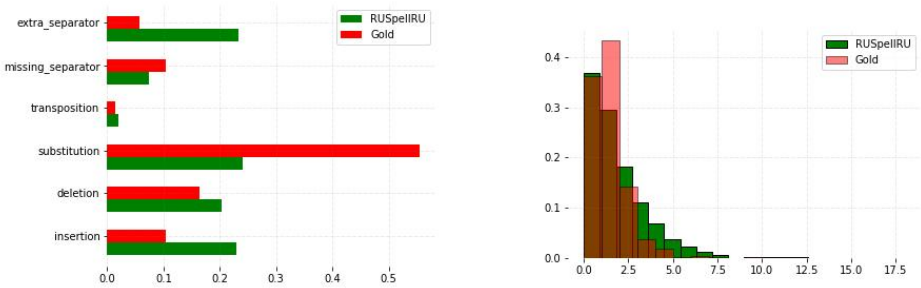
This study primarily focuses on the description and evaluation of proposed methods, rather than conducting a comparative analysis with existing analogues. Specifically, the ButterFingers method is applied to lowercase letters, without considering other symbols or characters. The Case method lacks specific thresholds for incorporating misspellings, resulting in a scenario where the text remains unaltered without any substitutions. Furthermore, it should be noted that the Augmentex tool offers a broader range of perturbation techniques, making it challenging to establish a comprehensive comparison with mentioned tools.

To compare distributions, we employ two approaches: visualization analysis and numeric metrics. The visualization part is represented by histograms that depict distributions of realizations of T, P and N.

We also employ a two-sample variation of Kolmogorov–Smirnov test (Dimitrova et al., 2020) as a numeric metric. Kolmogorov–Smirnov test (Dimitrova et al., 2020) is designed to suit continuous distributions. It does not require normality and can be used with arbitrary distributions and subsets of arbitrary sizes. Thus, in this work, we prefer Kolmogorov–Smirnov test (Dimitrova et al., 2020) over correlation metrics and other tests. It produces scores representing the supremum distance between two empirical distribution functions corresponding to each sample. Then, based on these scores, p-values are calculated under the null hypothesis, which says that two observed sets of values come from the same unknown distribution. We use these p-values in all the tables starting from Table 5 alongside with a significance level of 0.05, which in particular means that if the p-value is less than 0.05, then two given subsets of values do not come from the same underlying distribution.

We apply Kolmogorov–Smirnov test (Dimitrova et al., 2020) for D_P because in Section 4.2, we state that P follows the continuous distribution. N, on the contrary, follows discrete distribution and does not fit in Kolmogorov–Smirnov test (Dimitrova et al., 2020) continuous setup. For N, we use the discrete case of two-sample Kolmogorov–Smirnov test (Dimitrova et al., 2020), and for T, we do not use either of Kolmogorov–Smirnov test (Dimitrova et al., 2020) variations, because some categories are too scarce and estimates may have been incorrect.

5.1 Dataset evaluation



(a) Types of errors

(b) Number of errors per sentence.

Figure 1: The distributions of the errors by type and number in SpellRuEval-2016 and Gold testsets.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	0.0	0.0	0.0	0.0	0.003	**0**
Literature	0.0	1.0	0.257	0.736	0.0	0.003	**0**
News	0.0	0.239	1.0	0.262	0.0	0.0	**0**
Strategic Documents	0.001	0.743	0.253	1.0	0.0	0.08	**0**
Social media	0.0	0.0	0.0	0.0	1.0	0.0	**0**
Gold	0.001	0.002	0.001	0.08	0.0	1.0	**0**
SpellRuEval	0.0	0.0	0.0	0.0	0.001	0.0	1

Table 5: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for the number of errors per sentence. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **Aranea** refers to Aranea web-corpus, **SpellRuEval** refers to SpellRuEval-2016 (Sorokin et al.,) and **Gold** refers to our dataset. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020).

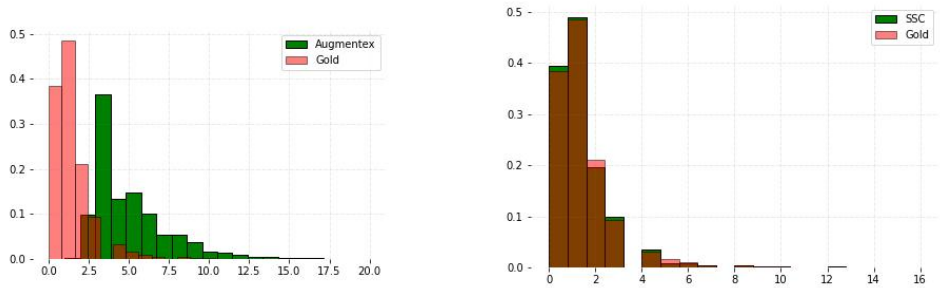
Detailed graphics and tables are in Appendix C 9. Several observations follow an analysis of graphs and tables. First, SpellRuEval-2016 (Sorokin et al.,) and multi-domain dataset seem to deviate in the distribution of types of spelling errors. While the latter has the dominant type of error - *substitution*, - SpellRuEval-2016 (Sorokin et al.,) almost evenly shares misspellings among its four most representative categories. A closer look at the remaining distributions of positions and corresponding tables suggests non-negligible difference in parts of the sentence, where spelling errors occur in the Gold dataset and SpellRuEval-2016 (Sorokin et al.,).

Second, p-values in Tables 5, 6, 8, 11, 12 suggest that Gold dataset differs from its constituents, at least according to corresponding distributions. This observation may be explained by the diverse nature of the source datasets and substantial deviations in properties of errors, which are brought by different domains. This leads to statistics yielded from the Gold dataset, which is a composition of source datasets, to be differ from those gathered from constituents.

Summing up the first part of the experiments, the multi-domain dataset and SpellRuEval-2016 (Sorokin et al.,) are different in proportions of misspellings, their positions in a sentence and domains that are included in corresponding corpora.

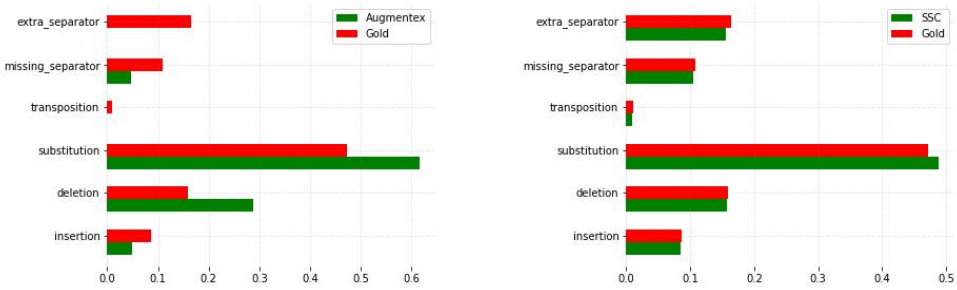
5.2 Spelling corruption methods evaluation

This subsection describes the results of evaluating the proposed spelling corruption methods. We generate synthetic spelling errors with the suggested algorithms on correct sentences of the multi-domain gold dataset. Then we do the same procedure in Section 5.1.



(a) Augmentex (b) SSC

Figure 2: The distributions of number (per sentence) of synthetically generated errors by the proposed methods for spelling corruption compared to the dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



(a) Augmentex (b) SSC

Figure 3: The distributions of types of synthetically generated errors by the proposed methods for spelling corruption compared to the dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	0.0	0.0	0.001	0.0	0.002	0.001	0.0
Literature	0.0	1.0	0.227	0.736	0.0	0.001	0.004	0.0
News	0.0	0.231	1.0	0.266	0.0	0.0	0.0	0.0
Strategic Documents	0.0	0.724	0.262	1.0	0.0	0.076	0.12	0.0
Social media	0.0	0.0	0.0	0.0	1.0	0.0	0.0	0.0
Gold	0.001	0.004	0.0	0.079	0.0	1.0	0.85	0.0
SSC	0.001	0.006	0.0	0.122	0.0	0.842	1.0	0.0
Augmentex	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

Table 6: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for the number of errors per sentence. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **Aranea** refers to Aranea web-corpus, **SpellRuEval** refers to SpellRuEval-2016 (Sorokin et al.,), **Gold** refers to Gold dataset, **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020).

The detailed graphics and tables are in Appendix D 10. We witness from a visualization point of view that the Statistic-based spelling corruption method fits well for distributions of the gold test set's number and types of spelling errors (see Figure 2 and Figure 3).

However, it should be noticed that we compare two methods on the complete range of sentence lengths. Research on the correlation between sequence length and the number of errors and probable degradation or enhancement of performance of two approaches is yet to be done as a promising aspect of our future work.

Both methods provide mostly high p-values produced by Kolmogorov–Smirnov test (Dimitrova et al., 2020) (see Tables 18, 17, 16, 15, 14, 13) between sets of relative positions of synthetically generated errors and corresponding misspellings from the gold set. Thus, both methods can approximate distributions of human spelling errors.

6 Conclusion

In this paper, we dealt with the spelling errors augmentation problem. We present the multi-domain parallel corpus for the Russian language for the first time. It represents the golden spelling error distribution we compare with the artificial ones. To generate artificial mistakes, we employ two approaches. The first is based on statistics and heuristics and can produce corrupted text without labelled data. The second approach, on the contrary, needs annotated parallel samples to examine source misspellings and replicate the spanning error distributions. The dataset is publicly available in the repository [14]. As part of our future research, we intend to enrich the existing dataset by incorporating data from new domains. Furthermore, an intriguing aspect to explore would be the examination of text distributions pertaining to input sources such as computer keyboards and mobile devices. We propose the inclusion of relevant metadata associated with these sources within the dataset, thereby enhancing its comprehensiveness and contextual relevance.

References

Karel Benes and Lukás Burget. 2020. Text augmentation for language models in high error recognition scenario. *CoRR*, abs/2011.06056.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) web corpora. *// Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17*, P 247–256. Springer.

Eric Brill and Robert C Moore. 2000. An improved error model for noisy channel spelling correction. *// Proceedings of the 38th annual meeting of the association for computational linguistics*, P 286–293.

Fred J Damerau. 1964. A technique for computer detection and correction of spelling errors. *Communications of the ACM*, 7(3):171–176.

Kaustubh D. Dhole, Varun Gangal, Sebastian Gehrmann, Aadesh Gupta, Zhenhao Li, Saad Mahamood, Abinaya Mahendiran, Simon Mille, Ashish Srivastava, Samson Tan, Tongshuang Wu, Jascha Sohl-Dickstein, Jinho D. Choi, Eduard Hovy, Ondrej Dusek, Sebastian Ruder, Sajant Anand, Nagender Aneja, Rabin Banjade, Lisa Barthe, Hanna Behnke, Ian Berlot-Attwell, Connor Boyle, Caroline Brun, Marco Antonio Sobrevilla Cabezudo, Samuel Cahyawijaya, Emile Chapuis, Wanxiang Che, Mukund Choudhary, Christian Clauss, Pierre Colombo, Filip Cornell, Gautier Dagan, Mayukh Das, Tanay Dixit, Thomas Dopierre, Paul-Alexis Dray, Suchitra Dubey, Tatiana Ekeinhor, Marco Di Giovanni, Rishabh Gupta, Rishabh Gupta, Louanes Hamla, Sang Han, Fabrice Harel-Canada, Antoine Honore, Ishan Jindal, Przemyslaw K. Joniak, Denis Kleyko, Venelin Kovatchev, Kalpesh Krishna, Ashutosh Kumar, Stefan Langer, Seungjae Ryan Lee, Corey James Levinson, Hualou Liang, Kaizhao Liang, Zhexiong Liu, Andrey Lukyanenko, Vukosi Marivate, Gerard de Melo, Simon Meoni, Maxime Meyer, Afnan Mir, Nafise Sadat Moosavi, Niklas Muennighoff, Timothy Sum Hon Mun, Kenton Murray, Marcin Namysl, Maria Obedkova, Priti Oli, Nivranshu Pasricha, Jan Pfister, Richard Plant, Vinay Prabhu, Vasile Pais, Libo Qin, Shahab Raji, Pawan Kumar Rajpoot, Vikas Raunak, Roy Rinberg, Nicolas Roberts, Juan Diego Rodriguez, Claude Roux, Vasconcellos P. H. S., Ananya B. Sai, Robin M. Schmidt, Thomas Scialom, Tshephisho Sefara, Saqib N. Shamsi, Xudong Shen, Haoyue Shi, Yiwen Shi, Anna Shvets, Nick Siegel, Damien Sileo, Jamie Simon, Chandan Singh, Roman Sitelew, Priyank Soni, Taylor Sorensen, William Soto, Aman Srivastava, KV Aditya Srivatsa, Tony Sun, Mukund Varma T, A Tabassum, Fiona Anting Tan, Ryan Teehan, Mo Tiwari,

[14] https://huggingface.co/datasets/RussianNLP/russian_multidomain_spellcheck

Marie Tolkiehn, Athena Wang, Zijian Wang, Gloria Wang, Zijie J. Wang, Fuxuan Wei, Bryan Wilie, Genta Indra Winata, Xinyi Wu, Witold Wydmański, Tianbao Xie, Usama Yaseen, M. Yee, Jing Zhang, and Yue Zhang. 2021. Nl-augmenter: A framework for task-sensitive natural language augmentation.

Dimitrina S. Dimitrova, Vladimir K. Kaishev, and Senren Tan. 2020. Computing the kolmogorov-smirnov distribution when the underlying cdf is purely discrete, mixed, or continuous. *Journal of Statistical Software*, 95(10):1–42.

Ilya Gusev. 2020. Dataset for automatic summarization of russian news. // *Artificial Intelligence and Natural Language: 9th Conference, AINL 2020, Helsinki, Finland, October 7–9, 2020, Proceedings 9*, P 122–134. Springer.

Masato Hagiwara and Masato Mita. 2019. Github typo corpus: A large-scale multilingual dataset of misspellings and grammatical errors. *CoRR*, abs/1911.12893.

Vitaly Ivanin, Ekaterina Artemova, Tatiana Batura, Vladimir Ivanov, Veronika Sarkisyan, Elena Tutubalina, and Ivan Smurov. 2020. Rurebus-2020 shared task: Russian relation extraction for business. // *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog" [Komp'iuternaia Lingvistika i Intellektual'nye Tehnologii: Trudy Mezhdunarodnoj Konferentsii "Dialog"]*, Moscow, Russia.

Mark D Kemighan, Kenneth Church, and William A Gale. 1990. A spelling correction program based on a noisy channel model. // *COLING 1990 Volume 2: Papers presented to the 13th International Conference on Computational Linguistics*.

Vladimir I Levenshtein et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. // *Soviet physics doklady*, volume 10, P 707–710. Soviet Union.

V. I. Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707, February.

Aouragh Si Lhoussain, Gueddah Hicham, and YOUSFI Abdellah. 2015. Adaptating the levenshtein distance to contextual spelling correction. *International Journal of Computer Science and Applications*, 12(1):127–133.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and vox diy: Benchmark dataset for crowdsourced audio transcription. // J. Vanschoren and S. Yeung, *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1.

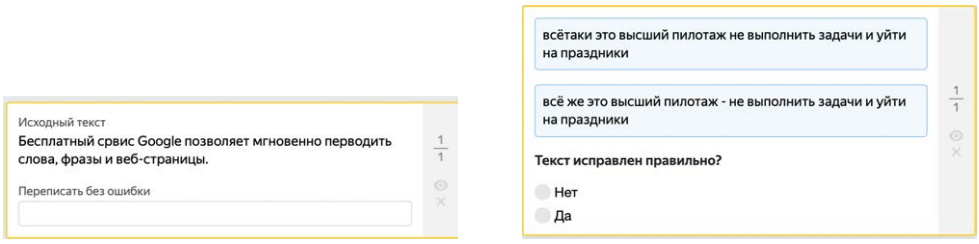
Tatiana Shavrina. 2017. Methods of misspelling detection and correction: A historical overview. *Voprosy Jazykoznanija*, (4):115–134.

AA Sorokin, AV Baytin, IE Galinskaya, and TO Shavrina. Spellrueval: the first competition on automatic spelling correction for russian.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, et al. 2022. Tape: Assessing few-shot russian language understanding. *arXiv preprint arXiv:2210.12813*.

Kristina Toutanova and Robert C Moore. 2002. Pronunciation modeling for improved spelling correction. // *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, P 144–151.

7 Appendix A



(a) Gathering. (b) Validation.

Figure 4: The example of the Yandex.Toloka design settings for the error gathering and validation steps.

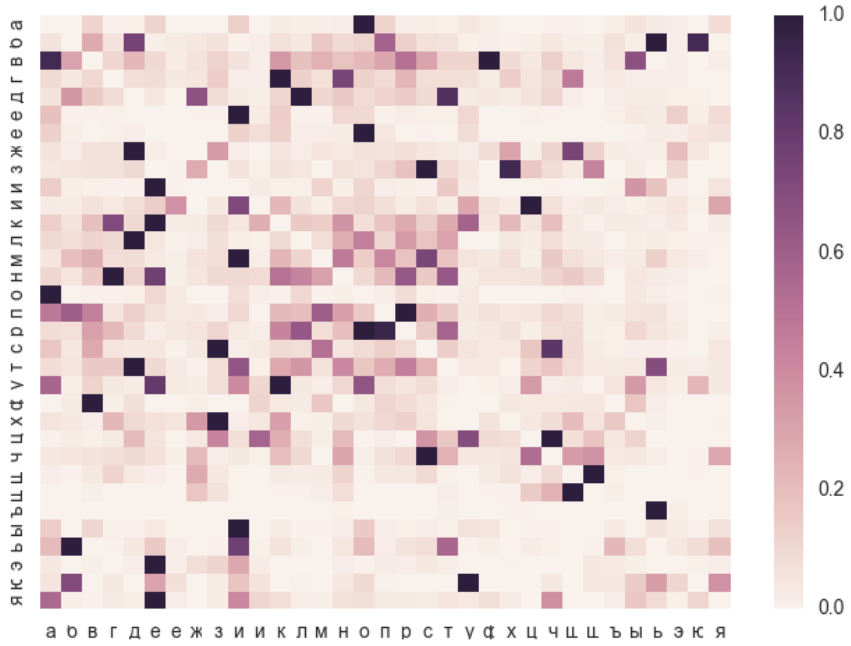


Figure 5: The heat map is read line by line. For example, for the letter "а", the most likely error is "о". All other errors are significantly less likely.

8 Appendix B

```
num_errors = D_N.sample() # sample number of errors
for error in num_errors:
    type = D_T.sample() # sample type of error
    subinterval = D_Ps[type].sample() # sample relative boundaries
    pos_left = len(sentence) * subinterval[0] # rescale boundaries back
    pos_right = len(sentence) * subinterval[1]

    counter = 0
    pos = choice(pos_left, pos_right) # sample position
    while not satisfy(type, pos): # check if conditions hold
        pos = choice(pos_left, pos_right)
        counter += 1
        if counter > max_tries: # if we tried every position in subinterval
            skip = True
            break
    if not skip:
        sentence = apply(sentence, pos, type) # insert the error
    skip = False
```

9 Appendix C

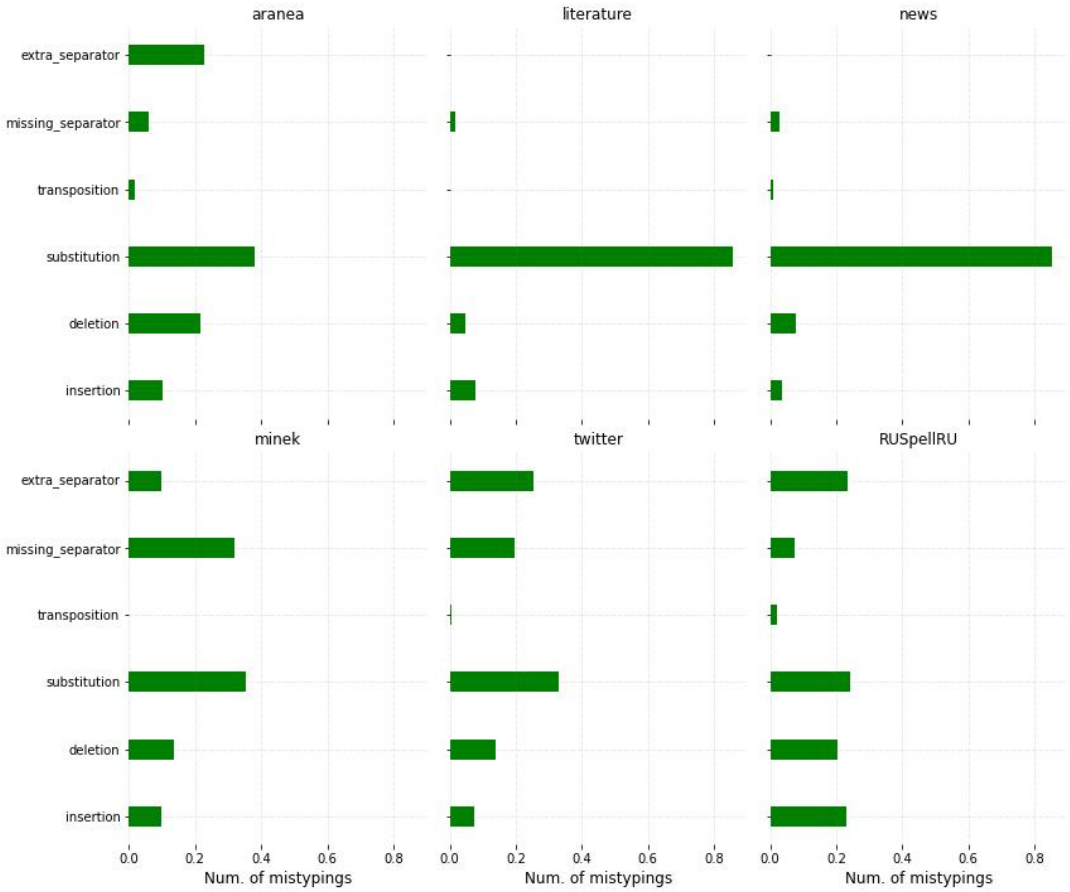


Figure 6: The frequencies of various types of errors encountered in different domains. *aranea, literature, news, minek, twitter* refer to domains of the porposed dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,). It is normalized counters of corresponding error types on the y-axis, which makes them estimates of probabilities of outcomes for T.

	Aranea	**Literature**	**News**	**Strategic Documents**	**Social media**	**Gold**	**SpellRuEval**
Aranea	1.0	0.122	0.536	0.249	0.722	0.983	0.001
Literature	0.122	1.0	0.562	0.389	0.449	0.275	0.522
News	0.536	0.562	1.0	0.842	0.842	0.674	0.227
Strategic Documents	0.249	0.389	0.842	1.0	0.773	0.519	0.009
Social media	0.722	0.449	0.842	0.773	1.0	0.927	0.108
Gold	0.983	0.275	0.674	0.519	0.927	1.0	**0.0**
SpellRuEval	0.001	0.522	0.227	0.009	0.108	**0.0**	1.0

Table 7: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *insertion-type errors*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution.

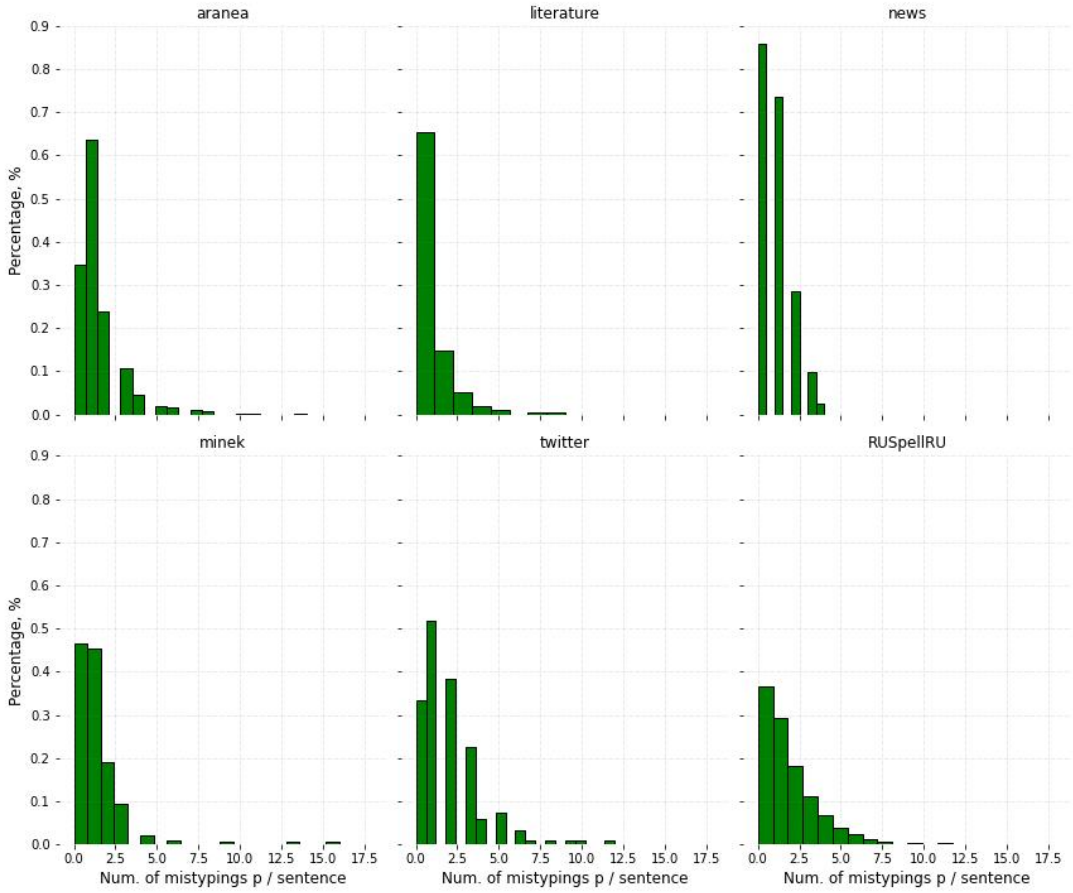
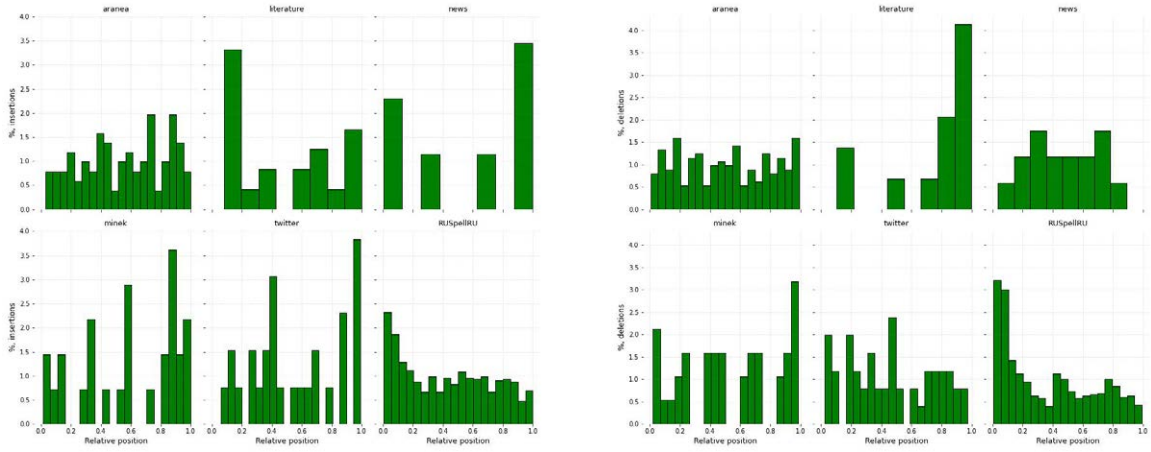
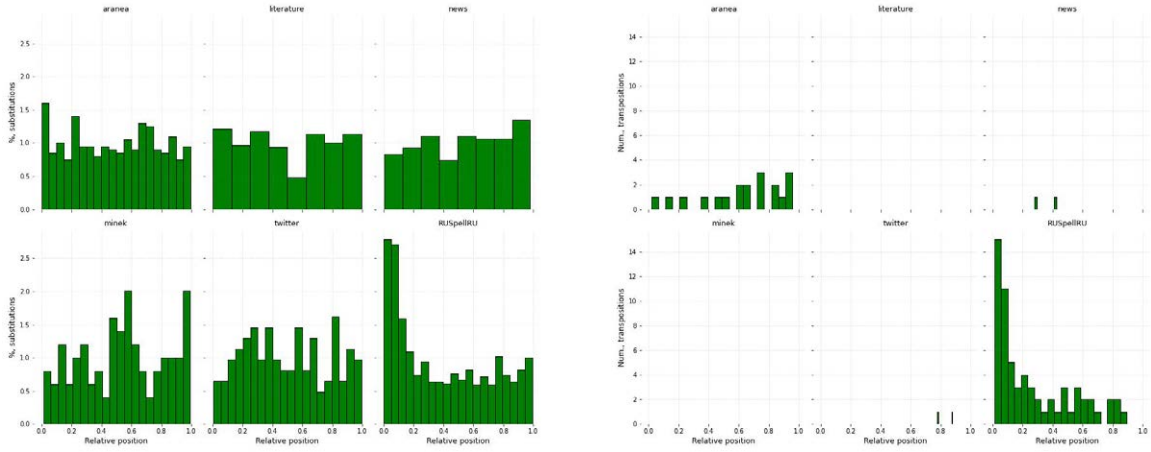


Figure 7: The number of spelling errors across domains in the proposed dataset compared to SpellRuEval-2016 (Sorokin et al.,). *aranea, literature, news, minek, twitter* refer to domains of our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).



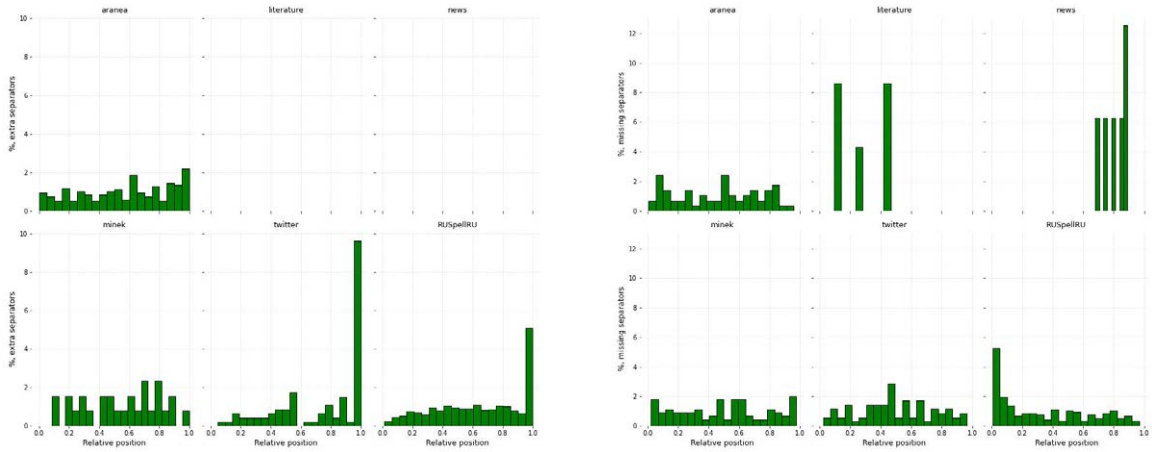
(a) Insertions (b) Deletions

Figure 8: Distributions of relative positions of corresponding types of errors across domains in the proposed dataset. *aranea, literature, news, minek, twitter* refer to domains of our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).



(a) Substitutions (b) Transpositions

Figure 9: Distributions of relative positions of corresponding types of errors across domains in the proposed dataset. *aranea, literature, news, minek, twitter* refer to domains of our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).

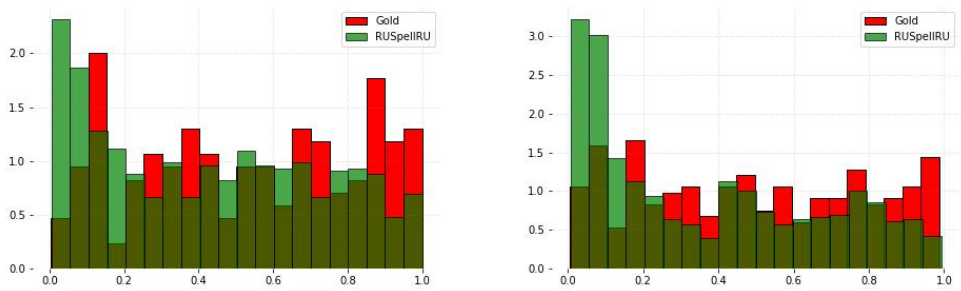


(a) Extra separators (b) Missing separators

Figure 10: Distributions of relative positions of corresponding types of errors across domains in the proposed dataset. *aranea, literature, news, minek, twitter* refer to domains of our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).

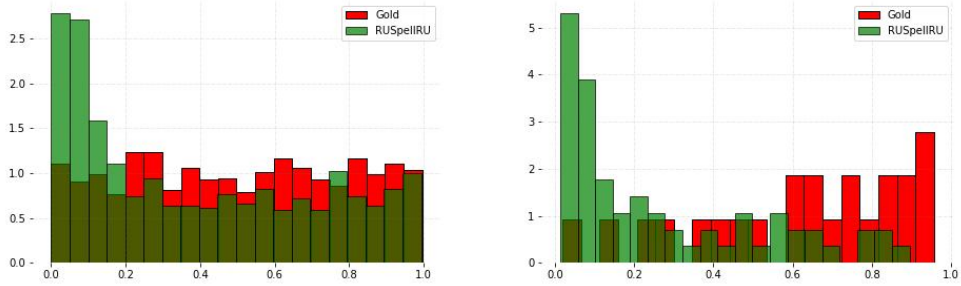
	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	0.002	0.547	0.687	0.479	1.0	**0.0**
Literature	0.002	1.0	0.003	0.015	0.002	0.002	**0.0**
News	0.547	0.003	1.0	0.498	0.838	0.501	0.057
Strategic Documents	0.687	0.015	0.498	1.0	0.318	0.835	0.006
Social media	0.479	0.002	0.838	0.318	1.0	0.48	0.005
Gold	1.0	0.002	0.501	0.835	0.48	1.0	**0.0**
SpellRuEval	**0.0**	**0.0**	0.057	0.006	0.005	**0.0**	1.0

Table 8: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *deletion-type errors*. Table entries are two-tailed p-vals given the null hypothesis that two subsets of samples come from the same distribution.



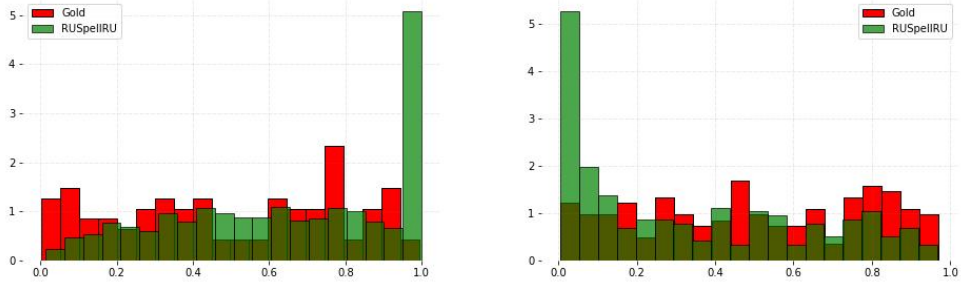
(a) Insertions (b) Deletions

Figure 11: Distributions of relative positions of corresponding types of errors between the multi-domain dataset and SpellRuEval-2016 (Sorokin et al.,). *Gold* refer to our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).



(a) Substitutions (b) Transpositions

Figure 12: Distributions of relative positions of corresponding types of errors between the multi-domain dataset and SpellRuEval-2016 (Sorokin et al.,). *Gold* refer to our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).



(a) Extra separators (b) Missing separators

Figure 13: Distributions of relative positions of corresponding types of errors between the multi-domain dataset and SpellRuEval-2016 (Sorokin et al.,). *Gold* refer to our dataset and *RUSpellRU* refer to SpellRuEval-2016 (Sorokin et al.,).

10 Appendix D

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	0.911	0.485	0.207	0.421	0.906	**0.0**
Literature	0.911	1.0	0.342	0.086	0.535	0.848	**0.0**
News	0.485	0.342	1.0	0.809	0.592	0.67	**0.0**
Strategic Documents	0.207	0.086	0.809	1.0	0.348	0.241	**0.0**
Social media	0.421	0.535	0.592	0.348	1.0	0.792	**0.0**
Gold	0.906	0.848	0.67	0.241	0.792	1.0	**0.0**
SpellRuEval	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	**0.0**	1.0

Table 9: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *substitution-type errors*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	-	0.143	-	0.267	1.0	**0.0**
Literature	-	-	-	-	-	-	-
News	0.143	-	1.0	-	0.333	0.187	0.28
Strategic Documents	-	-	-	-	-	-	-
Social media	0.267	-	0.333	-	1.0	0.3	0.009
Gold	1.0	-	0.187	-	0.3	1.0	**0.0**
SpellRuEval	**0.0**	-	0.28	-	0.009	**0.0**	1.0

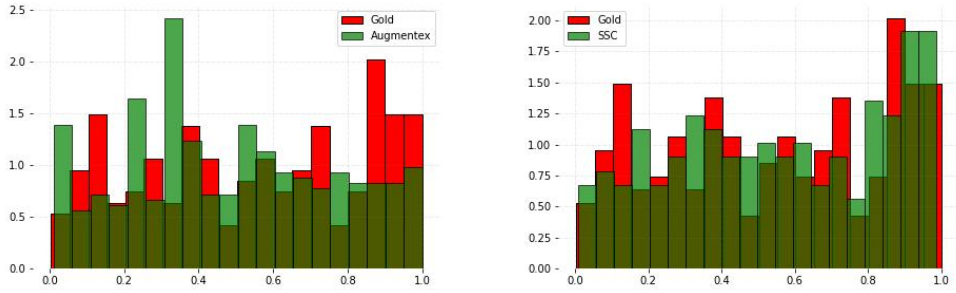
Table 10: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *transposition-type errors*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	-	-	0.585	0.0	0.066	**0.0**
Literature	-	-	-	-	-	-	-
News	-	-	-	-	-	-	-
Strategic Documents	0.585	-	-	1.0	0.0	0.15	0.046
Social media	0.0	-	-	0.0	1.0	0.0	**0.0**
Gold	0.066	-	-	0.15	0.0	1.0	0.003
SpellRuEval	**0.0**	-	-	0.046	**0.0**	0.003	1.0

Table 11: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *extra separators*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. Gaps indicate that samples from this domain are absent.

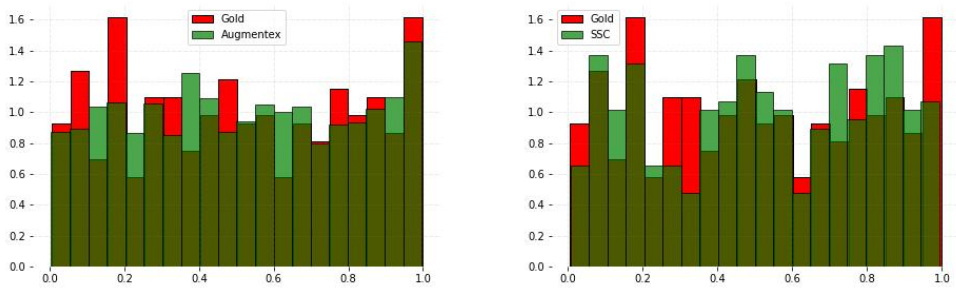
	Aranea	Literature	News	Strategic Documents	Social media	Gold	SpellRuEval
Aranea	1.0	0.071	0.002	0.63	0.459	0.917	0.008
Literature	0.071	1.0	0.004	0.056	0.074	0.057	0.502
News	0.002	0.004	1.0	0.002	0.001	0.001	**0.0**
Strategic Documents	0.63	0.056	0.002	1.0	0.658	0.983	**0.0**
Social media	0.459	0.074	0.001	0.658	1.0	0.808	**0.0**
Gold	0.917	0.057	0.001	0.983	0.808	1.0	**0.0**
SpellRuEval	0.008	0.502	**0.0**	**0.0**	**0.0**	**0.0**	1.0

Table 12: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *missing separators*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. Gaps indicate that samples from this domain are absent.



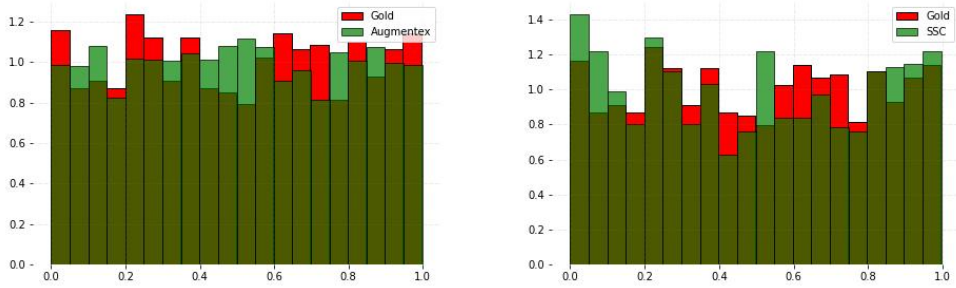
(a) Augmentex (b) SSC

Figure 14: Distributions of relative positions of synthetically generated insertions by the proposed methods for spelling corruption compared to *insertions* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



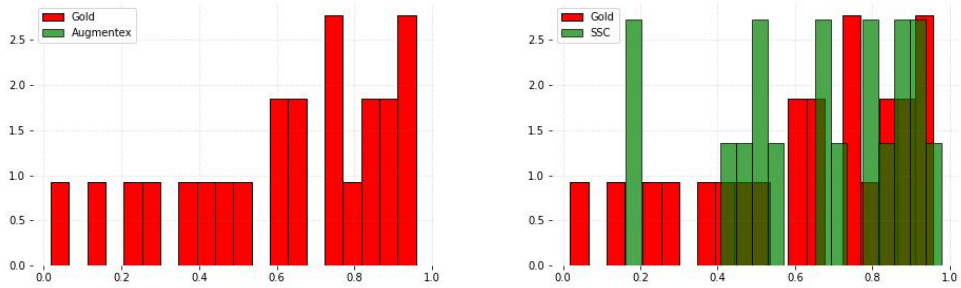
(a) Augmentex (b) SSC

Figure 15: Distributions of relative positions of synthetically generated deletions by the proposed methods for spelling corruption compared to *deletions* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



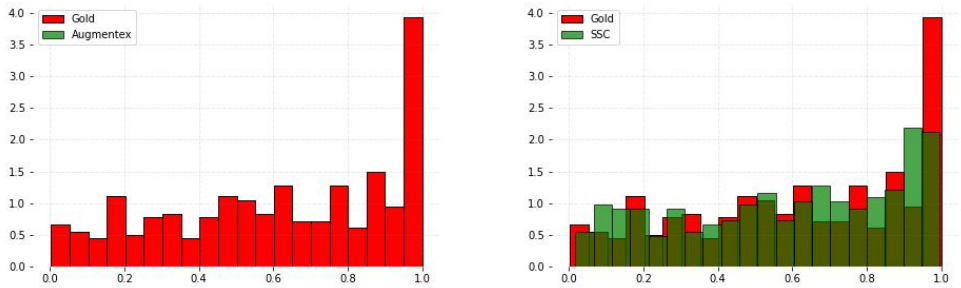
(a) Augmentex (b) SSC

Figure 16: Distributions of relative positions of synthetically generated substitutions by the proposed methods for spelling corruption compared to *substitutions* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



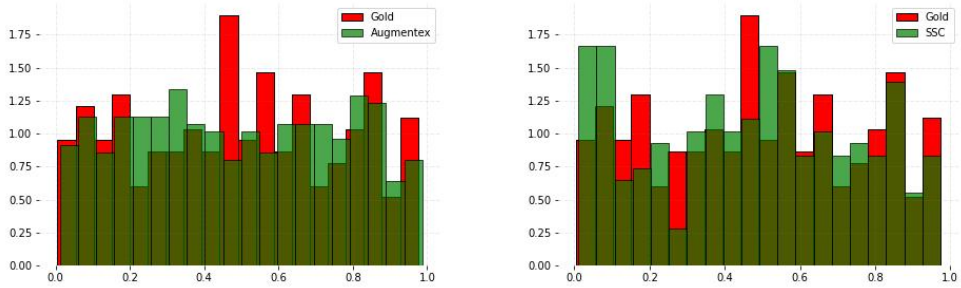
(a) Augmentex (b) SSC

Figure 17: Distributions of relative positions of synthetically generated transposition by both of the proposed methods for spelling corruption compared to *transposition* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



(a) Augmentex (b) SSC

Figure 18: Distributions of relative positions of synthetically generated extra separators by the proposed methods for spelling corruption compared to *extra separators* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.



(a) Augmentex (b) SSC

Figure 19: Distributions of relative positions of synthetically generated missing separators by the proposed methods for spelling corruption compared to *missing separators* in dataset. *Augmentex* and *SSC* refer to the methods described in Section 4.1 and Section 4.2 respectively and *Gold* refers to multi-domain dataset.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	0.122	0.536	0.249	0.722	0.983	0.738	0.077
Literature	0.122	1.0	0.562	0.389	0.449	0.275	0.146	0.16
News	0.536	0.562	1.0	0.842	0.842	0.674	0.801	0.316
Strategic Documents	0.249	0.389	0.842	1.0	0.773	0.519	0.479	0.023
Social media	0.722	0.449	0.842	0.773	1.0	0.927	0.903	0.51
Gold	0.983	0.275	0.674	0.519	0.927	1.0	0.924	0.017
SSC	0.738	0.146	0.801	0.479	0.903	0.924	1.0	0.021
Augmentex	0.077	0.16	0.316	0.023	0.51	0.017	0.021	1.0

Table 13: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *insertions*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020).

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	0.002	0.547	0.687	0.479	1.0	0.49	0.79
Literature	0.002	1.0	0.003	0.015	0.002	0.002	0.003	0.001
News	0.547	0.003	1.0	0.498	0.838	0.501	0.41	0.569
Strategic Documents	0.687	0.015	0.498	1.0	0.318	0.835	0.686	0.789
Social media	0.479	0.002	0.838	0.318	1.0	0.48	0.204	0.294
Gold	1.0	0.002	0.501	0.835	0.48	1.0	0.574	0.796
SSC	0.49	0.003	0.41	0.686	0.204	0.574	1.0	0.51
Augmentex	0.79	0.001	0.569	0.789	0.294	0.796	0.51	1.0

Table 14: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *deletions*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020).

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	0.911	0.485	0.207	0.421	0.906	0.406	0.562
Literature	0.911	1.0	0.342	0.086	0.535	0.848	0.742	0.583
News	0.485	0.342	1.0	0.809	0.592	0.67	0.233	0.179
Strategic Documents	0.207	0.086	0.809	1.0	0.348	0.241	0.135	0.231
Social media	0.421	0.535	0.592	0.348	1.0	0.792	0.273	0.72
Gold	0.906	0.848	0.67	0.241	0.792	1.0	0.139	1.0
SSC	0.406	0.742	0.233	0.135	0.273	0.139	1.0	1.0
Augmentex	0.562	0.583	0.179	0.231	0.72	1.0	1.0	1.0

Table 15: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *substitutions*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020).

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	-	0.143	-	0.267	1.0	0.976	-
Literature	-	-	-	-	-	-	-	-
News	0.143	-	1.0	-	0.333	0.187	0.063	-
Strategic Documents	-	-	-	-	-	-	-	-
Social media	0.267	-	0.333	-	1.0	0.3	0.474	-
Gold	1.0	-	0.187	-	0.3	1.0	0.941	-
SSC	0.976	-	0.063	-	0.474	0.941	1.0	-
Augmentex	-	-	-	-	-	-	-	-

Table 16: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *transpositions*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020). Gaps indicate that samples from this domain are absent.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	-	-	0.585	0.0	0.066	0.572	-
Literature	-	-	-	-	-	-	-	-
News	-	-	-	-	-	-	-	-
Strategic Documents	0.585	-	-	1.0	0.0	0.15	0.259	-
Social media	0.0	-	-	0.0	1.0	0.0	0.0	-
Gold	0.066	-	-	0.15	0.0	1.0	0.0	-
SSC	0.572	-	-	0.259	0.0	0.0	1.0	-
Augmentex	-	-	-	-	-	-	-	-

Table 17: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of *extra separators*. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020). Gaps indicate that samples from this domain are absent.

	Aranea	Literature	News	Strategic Documents	Social media	Gold	SSC	Augmentex
Aranea	1.0	0.071	0.002	0.63	0.459	0.917	0.976	0.833
Literature	0.071	1.0	0.004	0.056	0.074	0.057	0.093	0.092
News	0.002	0.004	1.0	0.002	0.001	0.001	0.001	0.003
Strategic Documents	0.63	0.056	0.002	1.0	0.658	0.983	0.707	0.87
Social media	0.459	0.074	0.001	0.658	1.0	0.808	0.454	0.559
Gold	0.917	0.057	0.001	0.983	0.808	1.0	0.477	0.701
SSC	0.976	0.093	0.001	0.707	0.454	0.477	1.0	0.298
Augmentex	0.833	0.092	0.003	0.87	0.559	0.701	0.298	1.0

Table 18: Kolmogorov–Smirnov test (Dimitrova et al., 2020) p-values for relative positions of missing separators. Table entries are two-tailed p-values given the null hypothesis that two subsets of samples come from the same distribution. **SSC** and **Augmentex** are methods described in Section 4.1 and Section 4.2 respectively. Reported values are averaged over 5 runs of the Kolmogorov–Smirnov test (Dimitrova et al., 2020). Gaps indicate that samples from this domain are absent.

Autocorrelations Decay in Texts and Applicability Limits of Language Models

Nikolay Mikhaylovskiy
Higher IT School, Tomsk State University, Tomsk, Russia, 634050
NTR Labs, Moscow, Russia, 129594
nickm@ntr.ai

Ilya Churilov
NTR Labs, Moscow, Russia, 129594
ichurilov@ntr.ai

Abstract

We show that the laws of autocorrelations decay in texts are closely related to applicability limits of language models. Using distributional semantics we empirically demonstrate that autocorrelations of words in texts decay according to a power law. We show that distributional semantics provides coherent autocorrelations decay exponents for texts translated to multiple languages. The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. We conclude that language models exhibiting Markovian behavior, including large autoregressive language models, may have limitations when applied to long texts, whether analysis or generation.

Keywords: autocorrelations decay laws, language models
DOI: 10.28995/2075-7182-2023-22-350-360

Убывание автокорреляций в текстах и границы применимости языковых моделей

Николай Михайловский
Высшая ИТ-Школа Томского Государственного Университета, Томск, Россия, 634050
ООО «НТР», Москва, Россия, 129594
nickm@ntr.ai

Илья Чурилов
ООО «НТР», Москва, Россия, 129594
ichurilov@ntr.ai

Аннотация

Показано, что законы затухания автокорреляций в текстах тесно связаны с пределами применимости языковых моделей. С использованием дистрибуционной семантики продемонстрировано, что автокорреляции слов в литературных текстах затухают по степенному закону. Показано, что дистрибуционная семантика обеспечивает когерентные показатели затухания автокорреляций для текстов, переведенных на несколько языков. Затухание автокорреляций в сгенерированных текстах количественно и часто качественно отличается от художественных текстов. Таким образом, языковые модели, демонстрирующие марковское поведение, включая большие авторегрессионные языковые модели, могут иметь ограниченную применимость к длинным текстам, будь то анализ или генерация.

Ключевые слова: большие языковые модели, законы убывания автокорреляции

1 Introduction

In this work, we endeavor into outlining statistically the limits of applicability of popular contemporary language models. To avoid any terminological doubt, when we write “models of the language”, we refer to any models that explain some linguistic phenomena, while “language models” refer to probabilistic

Grammar type (low → high)	Automaton	Memory
Regular (R)	Finite-state automaton (FSA)	Automaton state
Context-free (CF)	Push-down automaton (PDA)	+ infinite stack (only top entry accessible)
Context-sensitive (CS)	Linear bounded automaton (LBA)	+ bounded tape (all entries accessible)
Recursively enumerable (RE)	Turing machine (TM)	+ infinite tape (all entries accessible)

Table 1: Chomsky hierarchy of formal grammars (from [10])

language models as defined in Subsection 2.3 Probabilistic Language Models. While not long ago probabilistic language models were just models that assign probabilities to sequences of words [4], now they are the cornerstone of any task in computational linguistics through few-shot learning [6], prompt engineering [38] or fine-tuning [13]. On the other hand, current language models fail to catch long-range dependencies in the text consistently. For example, text generation with maximum likelihood target leads to rapid text degeneration, and consistent text generation requires probabilistic sampling and other tricks [22]. Large language models such as GPT-3 [6] push the boundary of “short text” rather far (specifically, to 2048 tokens), but do not remove the problem.

Our contributions in this work are the following:

- We explain how the laws of autocorrelations decay in texts are related to applicability of language models to long texts;
- We pioneer the use of pretrained word vectors for autocorrelation computations that allows us to study a widest range of autocorrelation distances;
- We show that the autocorrelations in literary texts decay according to power laws for all these distances;
- We show that distributional semantics typically provides coherent autocorrelations decay exponents for texts translated to multiple languages, unlike earlier flawed approaches;
- We show that the behavior of autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts.

2 Models of the Language

In this section, we briefly introduce models of the language that are important for the further considerations.

2.1 Formal Grammars

Formal grammars describe how to form strings from a language's alphabet that are valid according to the language's syntax. They were introduced by Chomsky in 1950s [7][8]. A formal grammar consists of a finite set of production rules in the form

$$left - hand\ side\ \rightarrow\ right - hand\ side, \tag{1}$$

where each side consists of a finite sequence of the following symbols:

- a finite set of nonterminal symbols (indicating that some production rule can yet be applied)
- a finite set of terminal symbols (indicating that no production rule can be applied)
- a start symbol (a distinguished nonterminal symbol)

Chomsky grammars constitute a hierarchy, see Table 1. While the original hierarchy implies strict inclusion of lower class grammars to higher ones, now there are several types of grammars known to fall between or partially overlap with the original classes (see, for example, [10]).

2.2 Distributional Semantics and Models

Distributional hypothesis assumes that linguistic items with similar distributions have similar meanings or function and was likely first introduced by Harris [20] in 1954 and was popularized in the form "a word is characterized by the company it keeps" by Firth [17]. The basic idea is to collect distributional information in, say, high-dimensional vectors, and then to define similarity in terms of some metric, say Euclidean distance or the angle between the vectors.

Early distributional approaches from 60s relied on hand-crafted features of the words [35], while more recent – on statistics of varied sorts. The first generation of statistical distributional semantics approaches included Latent Semantic Analysis (LSA) [11][12], Hyperspace Analogue to Language (HAL) [24][25], and many others, see [15] for a review. The second generation primarily consists of word2vec [31][32] and GloVe [37] models, the first, implicitly, and the second, explicitly adding the word analogy task into the training objective, so that similar relationships between words should be described by similar difference vectors between embeddings. The third generation of statistical distributional semantics models was started by emergence of BERT contextual word embeddings [13]. BERT have combined the word and its current context into a single vector embedding and used Masked Language Modelling training objective. A lot of recent work sprouted from BERT.

2.3 Probabilistic Language Models

Probabilistic language models consider sequences

$$t_{1:m} = \{t_1, t_2, \dots, t_m\} \tag{2}$$

of tokens from the lexicon $\mathcal{L}$. An autoregressive language model estimates the probability of such a sequence

$$P(t_{1:m}) = P(t_1)P(t_2|t_1)P(t_3|t_{1:2}) \dots P(t_m|t_{1:m-1}) = \prod_{k=1}^{m} P(t_k|t_{1:k-1}) \tag{3}$$

using the chain rule. Most models introduce the Markov [30] assumption that the probability of a token depends on the previous $n-1$ tokens only, thus approximating (3) with a truncated version

$$P(t_{1:m}) \approx \prod_{k=1}^{m} P(t_k|t_{k-n+1:k-1}) \tag{4}$$

While the language models based on recurrent [33], and specifically, LSTM [41] neural networks do not introduce the Markov assumption explicitly, we will shortly see that in practice they do exhibit Markovian behavior. On the other hand, it is long known that Markov models describe stochastic regular grammars [42].

3 Why Autocorrelations Decay Laws Matter?

In this section we explain why autocorrelation decay laws matter a lot to computational linguistics' near-future.

3.1 Computing Autocorrelations Using Distributional Semantics

Suppose we have a sequence of N vectors $V_i \in R^d, i \in [1, N]$. Autocorrelation function $C(\tau)$ is the average similarity between the vectors as a function of the lag $\tau = i - j$ between them. The simplest metric of vector similarity is the cosine distance $d(V_i, V_j) = \cos\angle(V_i, V_j) = \frac{V_i \cdot V_j}{\|V_i\|\|V_j\|}$, where $\cdot$ is a dot product between two vectors and $\|\ \ \|$ is an Euclidean norm of a vector. Thus,

$$C(\tau) = \frac{1}{N-\tau} \sum_{i=1}^{N-\tau} \frac{V_i \cdot V_{i+\tau}}{\|V_i\|\|V_{i+\tau}\|}. \tag{5}$$

$C(\tau)$ ranges from -1 for perfectly anticorrelated sequence (for $\tau = 1$ and $d = 1$ that would be $1, -1, 1, -1$ etc.) to 1 for a perfectly correlated one (for $\tau = 1$ and $d = 1$ that would be $1, 1, 1, 1$ etc.).

A distributional semantic assigns a vector to each word or context in a text. Thus, a text is transformed into a sequence of vectors, and we can calculate an autocorrelation function for the text.

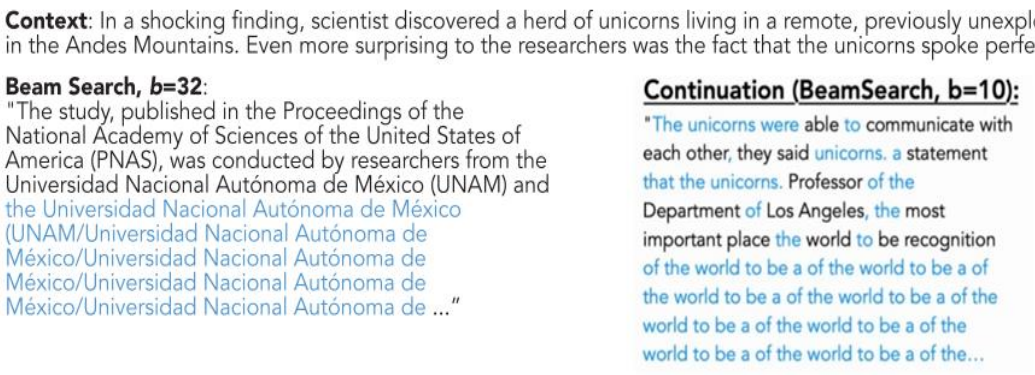


Figure 1: Beam search produces degenerate text (from [22])

Level	Architecture	Description
R-	Transformer	The encoder with stacked multi-head attention layers and dense layers.
R	RNN	A classical RNN with ReLU activations.
R+	LSTM	A classical LSTM.
DCF+	Stack-RNN	An RNN with an external stack, with PUSH, POP, and NO-OP actions.
NDCF	NDStack-RNN	An RNN with a nondeterministic stack, simulated using dynamic programming
CS	Tape-RNN	An RNN with a finite tape, as in a Turing machine (similar to Baby-NTM).

Table 2: Alignment of neural network architectures with Chomsky hierarchy (from [10])

3.2 Transformer Language Models Exhibit Markovian Behavior

In this paper, by Markovian behavior, we mean that large language models actually use only a limited context, often significantly less than the maximum context possible. Thus they implicitly or explicitly use the Markov assumption. Two separate phenomena classes that prove that transformer language models exhibit Markovian behavior are known, and in Section 5.5 we introduce the third one.

One such phenomenon is the rapid text degeneration when a transformer language model is used to generate text with maximum likelihood target [21][28]. Maximization-based decoding methods such as beam search lead to output text that is bland, incoherent, or gets stuck in repetitive loops [22] that are extremely reminiscent of positively recurrent Markov chains (see Figure 1).

The other phenomenon is studied in detail in [10]. The authors have found that the networks roughly match the computational models associated with the Chomsky hierarchy: RNNs can solve tasks up to the regular level, Stack-RNNs up to the DCF level, and Tape-RNNs up to the CS level. Finally, they observed that Transformers and LSTMs are less aligned with the Chomsky hierarchy: Transformers fail on regular tasks, while LSTMs can solve tasks more difficult than regular. The results of [10] are summarized in Table 2. As transformers can at most generalize to regular languages and Markov models describe stochastic regular grammars [42], we can safely say that transformers exhibit behavior no richer than regular.

3.3 Markovian Implies Exponential Correlations Decay, Probabilistic Context-Free Grammars Can Generate Power Laws

Assume that the sequence (2) is an output of a random source that takes values in $\mathcal{L}$. If the source is Markovian, it can be shown [23] that the autocorrelations (or, equivalently, mutual information between chunks of the text) decay exponentially. Namely, the following theorem holds:

Theorem 1 ([23]). Let M be a Markov matrix that generates a Markov process. If M is irreducible and aperiodic, then the asymptotic behavior of the mutual information $I(t_1, t_2)$ is exponential decay toward zero for $|t_2 - t_1| \gg 1$ with decay timescale $\log \frac{1}{|\lambda_2|}$, where λ_2 is the second largest eigenvalue of M. If M is reducible or periodic, I can instead decay to a constant; no Markov process whatsoever can produce power law decay.

One the other hand, the following theorem holds:

Theorem 3 ([23]). There exist a probabilistic context-free grammar such that the mutual information $I(A,B)$ between two symbols A and B in the terminal strings of the language decay like d^{-k}, where d is the number of symbols between A and B.

3.4 If the Natural Language Exhibits Power Law Correlations Decay, We Can Do Better Than Autoregressive Language Models

Summarizing the above, if texts in the natural languages exhibit exponential autocorrelations decay, autoregressive language models are good to analyze or generate texts of any length. On the other hand, if texts in the natural languages exhibit power law autocorrelations decay, building language models that exhibit at least hierarchical, context-free-grammar-ish, slow-correlation-decay behavior may be beneficial for a variety of downstream tasks. This may be not enough to model long texts successfully because natural languages cannot be completely described by context-free grammars (see, for example, [40]), but may be a meaningful step.

4 Studying Autocorrelations Decay Laws in Texts

4.1 Prior Research

Schenkel, Zhang, and Zhang [39] were likely the first to empirically find the power law autocorrelations decay in texts using a random walk model with an arbitrary mapping of characters to fixed length, 5 bit sequences. They studied 10 texts in English. The obvious drawback of their approach is dependency on encoding. Amit et al. [3] explored this problem in various translations of the Bible and have shown that the power law exponent depends on both the language and the codification. Testing multiple random mappings would provide a more reliable estimate of power law exponents, but such a research is a matter of future. Random walk models have later been used to find the power law in text by several researchers, including Ebeling and Neiman [14], Kokol et al. [26] (who, by the way, in our opinion have not found power-law autocorrelations in literary writing on distances studied, but found power-law autocorrelations in computer programs, in a perfect agreement with the fact that computer programs are described by context-free grammars), Pavlov et al. [36], who find multifractal structures in the text, and Manin [29], who attribute long-range correlations to slow variations in lexical composition within the text.

Alvarez-Lacalle et al. [2] used a version of first-generation distributional semantic model to study autocorrelations in 12 literary texts in English to find power law autocorrelations decay. Altmann, Cristadoro, and Degli [1] analyze 41 binary functions on words separately on ten English versions of international novels. They separate the effects of burstiness and long-range correlations in the power spectrum and find a power law correlations decay. Lin and Tegmark [23] in a short empirical part of their study use three text corpora: 100 MB from Wikipedia, the first 114 MB of a French corpus and 27 MB of English articles from slate.com. They observe the power law decay of mutual information, but note that the large portion of the long-range mutual information appears to be dominated by poems in the French sample and by the html-like syntax in the Wikipedia sample. They have also shown similar power decay laws for autocorrelations in natural music and exponential laws in generated music, the result reproduced by different means by Yamshchikov and Tikhonov [43]. Corral et al. [10] study intervals between consecutive appearances of specific words in literary texts in 4 languages, including Finnish (a rare study of highly agglutinative language) to find that most words have a universal dimensionless probability density function described by gamma distribution. Gillet and Ausloos [18] and Montemurro and Pury [34] study sequences built from word frequencies and word lengths to find the power law autocorrelations decay.

4.2 Research Questions

Given the prior art, many research question remain unanswered. The ones we address in this work are:

Q1. How accurately can we say that autocorrelations in texts decay according to a power law?
Q2. Can we reject the hypothesis of exponential decay of correlations?
Q3. Does the law of decay depend on the language of the text?
Q4. Over what range of distances does the decay in autocorrelations follow a power law?

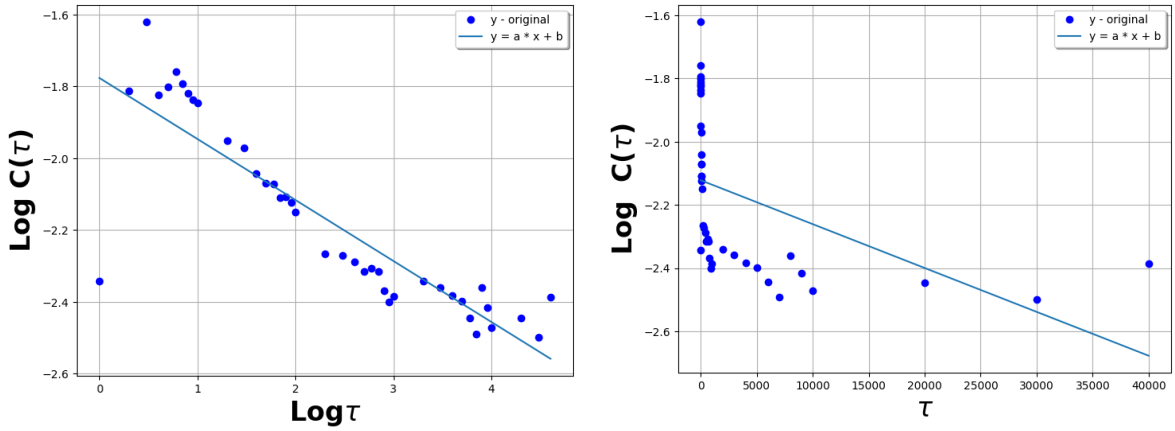


Figure 2: Autocorrelations in Don Quixote (English) computed using GloVe, a = 1, d = 300, $\tau \in [1, 40000]$ Left: log-log coordinates. Right: log-linear coordinates.

Q5. Are autocorrelations in LM-generated texts any different from literary texts?

4.3 Methods

In this work we use two distributional semantic models to estimate autocorrelations in long texts. One is a bag-of-words (BOW) embedding model of Alvarez-Lacalle et al. [2]. The other distributional semantic model we use is GloVe [37]. We use pretrained multilingual GloVe vector embeddings from [16]. We filter out both frequent and rare words filtering similarly to [2] when using BOW.

BOW assigns a vector of dimension d to each word first, and then averages these vectors over a window of the size a. This averaged vector is then assigned to a word in the center of averaging window. The exact procedure for BOW is described in detail in [2]. GloVe naturally maps each word to a vector; we then center the vector system by subtracting the average of vectors over the whole text, and, similarly, average over a window of the size a when we need direct comparison to BOW. After that in both cases we can compute the autocorrelation function following Section 3.1.

5 Experiments

5.1 The Dataset

For our studies we have collected a dataset of long literary and philosophical works in English, Spanish, French, German and Russian[i] each: Critique of Pure Reason, Don Quixote de la Mancha, Moby-Dick or, The Whale, The Adventures of Tom Sawyer, The Iliad, The Republic and War and Peace. The only translation absent is Moby-Dick in German, which happened to be substantially abridged. The texts have been obtained from Project Gutenberg, Wikisource, Royallib and lib.ru and preprocessed so as to fit our research purposes:

- copyright texts were removed from the files;
- author and translator notes were removed;
- table of contents and any indices were removed, except for the table of contents from Don Quixote;
- any links to illustrations have been removed;
- in the Russian version of War and Peace any non-Russian text have been replaced with Russian translations;
- etymology section was removed from Moby-Dick or, The Whale, where encountered, as some languages missed it.

5.2 Choosing Between Hypotheses of Power Law and Exponential Decay of Correlations

To address **Q1.** “How accurately can we say that autocorrelations in texts decay according to a power law?” and **Q2.** “Can we reject the hypothesis of exponential decay of correlations?” for each text, we

	Power Law					**Exponential Law**				
	BOW en	fr	es	ru	en	BOW en	fr	es	ru	en
The Adventures of Tom Sawyer	**0,16**	**0,11**	**0,16**	**0,14**	**0,21**	0,52	0,32	0,33	0,33	0,55
The Republic	**0,21**	**0,15**	**0,09**	**0,10**	**0,13**	0,58	0,28	0,25	0,31	0,38
Don Quixote	**0,20**	**0,11**	**0,12**	**0,09**	**0,20**	0,66	0,24	0,22	0,23	0,44
War and Peace	**0,20**	**0,13**	**0,11**	**0,08**	**0,09**	0,54	0,24	0,24	0,28	0,42
Critique of Pure Reason	**0,09**	**0,07**	**0,15**	**0,10**	**0,14**	0,27	0,17	0,20	0,21	0,25
The Iliad	**0,24**	2,37	**0,16**	**0,10**	**0,19**	0,63	**2,33**	0,17	0,19	0,54
Moby-Dick or, The Whale	**0,14**	**0,12**	**0,11**	**0,09**	**0,15**	0,40	0,22	0,22	0,22	0,47

Table 3: Goodness of fit of autocorrelation by power and exponential laws in terms of MAPE. BOW: a=200, d=100, $\tau \in [250, 4200]$ Glove: a = 1, d = 300, $\tau \in [\varepsilon, 40000]$

	BOW			GloVe		
	α	β	MAPE	α	β	MAPE
en	-0.7718	0.9545	0.1054	-0.7246	1.1582	0.1044
fr	-0.8836	1.1407	0.2154	-0.7749	1.1051	0.2150
es	-0.7601	0.9332	0.1057	-0.7083	0.9947	0.1271
ru	-0.7412	0.7874	0.0787	-0.6431	0.9173	0.0548
de	-0.8072	0.9542	0.1411	-0.8326	1.3478	0.1657

Table 4: Dependence of the autocorrelations power decay law in Don Quixote on the language and embedding. τ ranges from 200 to 4000 words, d=300, a = 200

have computed autocorrelations for a series of distances $\tau = n * 10^k$, $n \in [1, 9]$, and approximated the points produced by a straight line in both log-log and log-linear coordinates using the least squares regression. We have evaluated the goodness of fit of each model by MAPE (Mean Absolute Percentage Error). The range of τ for Glove was chosen from the first non-negative autocorrelation value ε (autocorrelations on small distances $\tau = [1, 2]$ happened to be sometimes negative).

The results for the English translation of Don Quixote are presented in the Figure 2. It can be seen that in log-log coordinates the regressed straight line approximates data well enough, unlike log-linear coordinates.

Table 3 lists the MAPE metrics of goodness of fit of autocorrelation by power and exponential laws (the smaller the better). It can be easily seen that for all the texts but one the hypothesis of exponential decay of correlations can be rejected. The peculiarity of the French translation of The Iliad is that the autocorrelation with $\tau = 1$ is very small but still positive, thus both producing significantly larger MAPE and ruining the approximation. Additional graphs are presented in the Appendix A.

5.3 Determining the Dependency of the Autocorrelations Decay Law on the Language of the Text

To study the dependency of the autocorrelations decay law on the language of the text, we have measured $C(\tau)$ for the same multilingual dataset as in Section 5.1 and fitted with power law $C(\tau) = \beta \cdot \tau^{\alpha}$. Table 4 presents results for Don Quixote. It can be easily seen that the parameters of power law, as well as the accuracy of the approximation are extremely consistent among languages for both embeddings, with standard deviation of exponent being 7% for BOW and 10% for GloVe. Moreover, the exponents for BOW and GloVe are also consistent within 15%, which we consider a very good agreement. This is in a stark contrast with the results from [3] that critically depend on the codification and language.

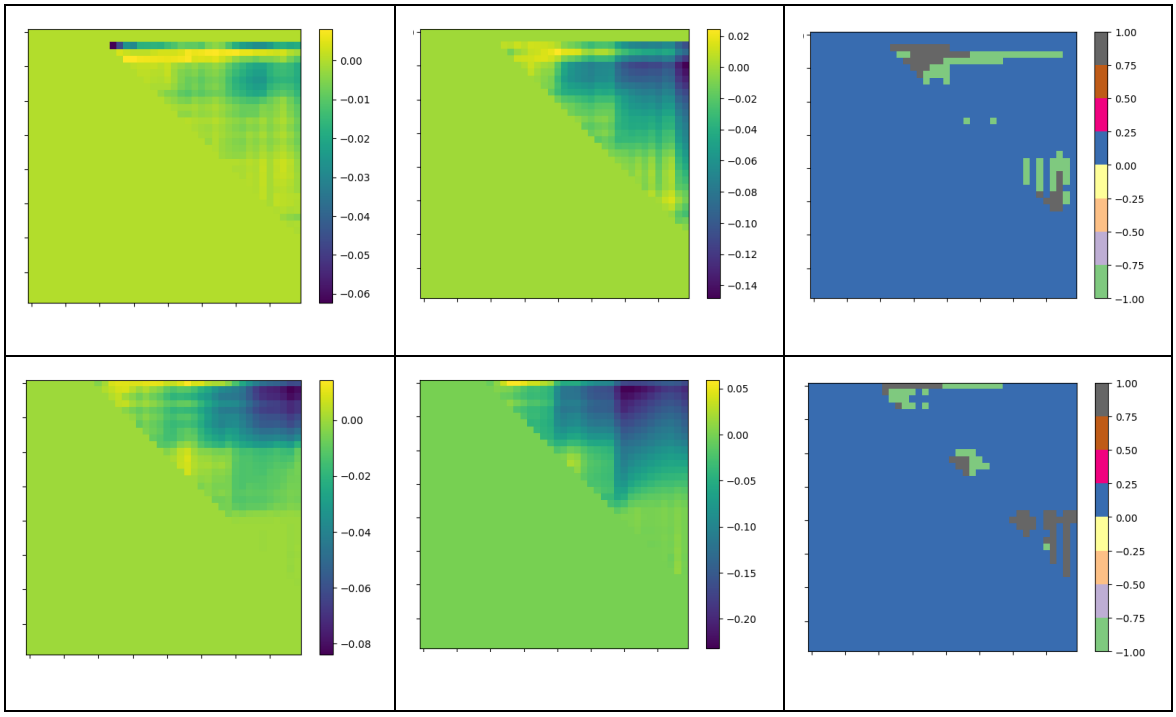


Figure 3: Autocorrelations in Critique of Pure Reason in English (top) and The Adventures of Tom Sawyer in Spanish (bottom) computed using GloVe, $a = 1, d = 300$. Vertical axis: start of τ range. Horizontal axis: end of τ range. Left: difference between power and log approximation MAPE. Middle: difference between power and exp approximation MAPE. Right: ranges where power (blue), exp (gray), and log (green) approximations are the best.

5.4 Determining the Range of Distances Where the Decay in Autocorrelations Can Be Considered Subject to a Power Law

As the BOW approach requires a sufficiently large window size a, we have studied the dependence of autocorrelations on distance ranges using GloVe embeddings with a window size $a = 1$. For each text we explored all the ranges of τ spanning at least a decimal order of magnitude, and fitted the autocorrelations with the best fitting log, power and exponential functions. We then mapped the differences between MAPE of power and other approximations, as well as the ranges where each function fits the data the best. The results for the Critique of Pure Reason in English and The Adventures of Tom Sawyer in Spanish are presented on Figure 3. Each small square on these images corresponds to a range of τ determined by its vertical (start) and horizontal (end) coordinates, for example, the full range of $\tau \in [1, 40000]$ corresponds to the top right corner. Additional graphs are presented in Appendix B.

It can be seen that for the shorter spans of τ the best approximations are sometimes logarithmic or exponential but their advantage is not significant, while for the longer ranges the best approximations are always power law. Additionally, the location of such ranges is hectic. We conclude that the cases where exponential or logarithmic approximation is better than the power law approximation represent natural short-range variability and cannot be considered a regularity.

5.5 Autocorrelations in Generated Texts

The behavior of autocorrelations is qualitatively different when the text is generated. The simplest way to generate an incoherent text is to shuffle words in a text. Figure 5 demonstrates that there is no specific autocorrelations decay law for an incoherent text.

To study autocorrelations in texts generated by large language models, we have used GPT-2 base [6] with the default generation parameters, and Structured State Space model S4 base [19] with the default generation parameters, and generated some 10K word continuous text with each model. The generated texts are listed in Appendix C and Appendix D, respectively. We then performed the same procedure as

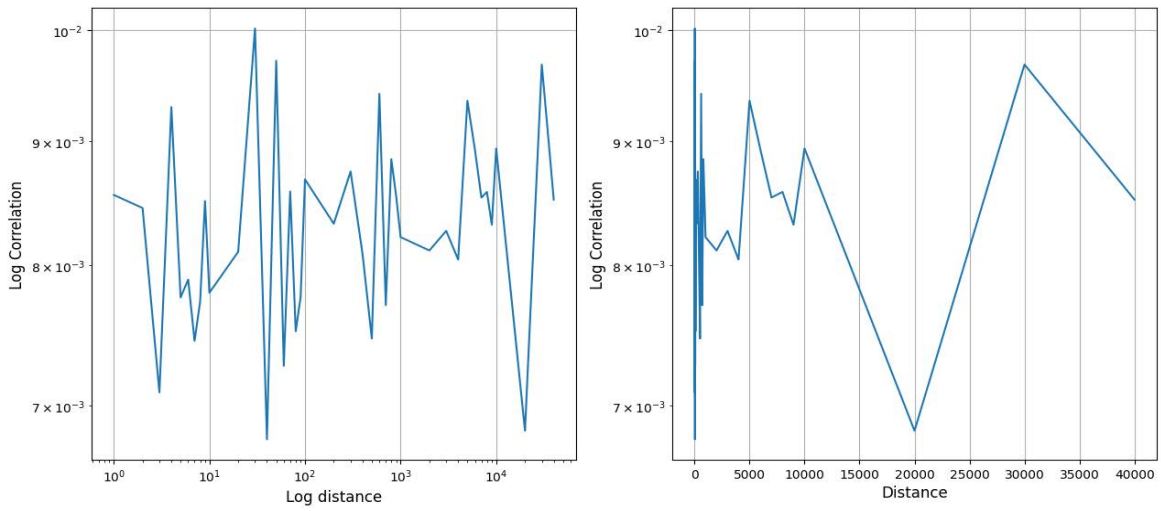


Figure 5: Autocorrelations in a randomly shuffled The Adventures of Tom Sawyer in Spanish computed using GloVe, a=1, d=300. Left: log-log, to right: log-linear coordinates

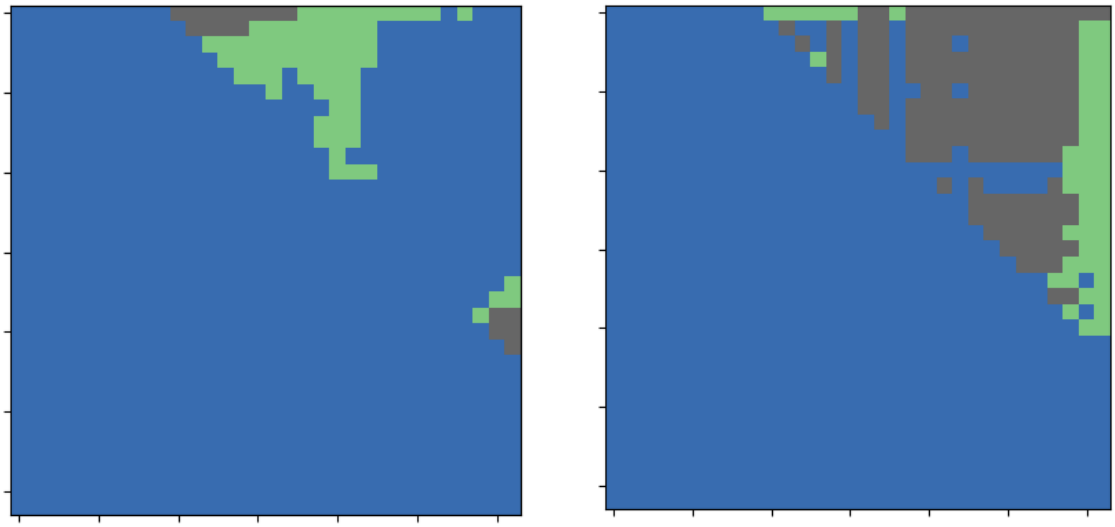


Figure 4: Autocorrelations in texts generated by GPT-2 (left) and S4 (right) models computed using GloVe, $a = 1, d = 300$, ranges where power (blue), exp (gray), and log (green) approximations are the best depicted. Vertical axis: start of τ range. Horizontal axis: end of τ range.

in Section 5.4, mapping ranges where each decay law provides the best approximation. The results are presented on Figure 4.

The autocorrelations decay in an exponential manner in the text generated by S4 model, while according to a power law on long distances and to log law – on short distances in the text generated by GPT-2. The autocorrelations in generated texts are significantly larger and decay much slower than the ones in the natural texts. In our S4 and GPT-2 generated examples, the power law coefficients are $a = -0.045, b = -0.71$ and $a = -0.027, b = -0.77$, respectively. At the same time we have not seen the coefficient a less than 0.1 for any natural text in English we have studied, and the average is closer to 0.2, indicating almost 10-fold gap between the power law decay rates in natural and generated texts. Typical values of coefficient b for natural texts are between -1.5 and -2, indicating at least 2-fold gap between natural and generated texts.

Thus we can say that the autocorrelations decay in generated texts are quantitatively and often qualitatively different from the literary texts. The conditions that influence the autocorrelations decay laws in generated texts may include sampling approach, temperature and other hyperparameters. This is a matter of future research.

6 Conclusions

We have shown empirically that autocorrelations in literary texts are decaying following the power law with the only upper limit being the length of the work itself and the hypothesis of exponential decay can be rejected for these distances. We have also shown empirically that the laws of autocorrelation decay, if measured using distributional semantics models are typically the same for the literary work translated to different languages. This contrasts previous findings that used flawed technique based on encoding-dependent random walks. Thus, we believe that distributional semantics models are a robust enough tool to measure autocorrelations in long texts.

The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. Based on the above, we can conclude that for long text processing one may need architectures different from the autoregressive ones, and many questions remain unanswered.

Acknowledgements

The authors are grateful to their colleagues at NTR Labs ML division for the discussions and support. Early versions of this work were discussed with Anton Kolonin, Dmitry Manin and Alexey Tikhonov. These discussions have improved our approach and research design, for which we are very grateful. We are also extremely grateful to Tatiana Sherstinova who discussed early versions of this work, suggested numerous improvements and provided a webinar venue at HSE to discuss this work publicly.

References

[1] Altmann E. G., Cristadoro G., Degli M. On the origin of long-range correlations in texts // PNAS. 2012. № 29 (109). C. 11582–11587.
[2] Alvarez-Lacalle E. et al. Hierarchical structures induce long-range dynamical correlations in written texts // PNAS. 2006. № 21 (103). C. 7956–7961.
[3] Amit M., Shmerler Y., Eisenberg Eli, Abraham M., Shnerb N.. Language and codification dependence of long-range correlations in texts // Fractals. 2012, №. 01 (02), C. 7-13.
[4] Bahl L.R., Jelinek F., Mercer R.L. A Maximum Likelihood Approach to Continuous Speech Recognition // IEEE Trans. Pattern Anal. Mach. Intell. 1983. Vol. PAMI-5, № 2. P. 179–190.
[5] Beltagy Iz, Peters M. E., Cohan A. Longformer: The Long-Document Transformer // arXiv:2004.05150
[6] Brown T.B. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. 2020. Vol. 2020, P. 1877–1901.
[7] Chomsky N. Three models for the description of language // IRE Transactions on Information Theory, 1956. № 2 (3): P. 113–124.
[8] Chomsky, N. Syntactic Structures. The Hague: Mouton, 1957.
[9] Corral A. et al. Universal Complex Structures in Written Language // Vol. arXiv:0901.2924v1, Access mode: https://arxiv.org/abs/0901.2924v1
[10] Delétang G. et al. Neural Networks and the Chomsky Hierarchy // International Conference on Learning Representations, 2023
[11] Deerwester S. C. et al. Improving information retrieval using latent semantic indexing. // Proceedings of the 51st Annual Meeting of the American Society for Information Science 1988, №25, P. 36–40.
[12] Deerwester S. C. et al. Indexing by latent semantic analysis // Journal of the American Society for Information Science. №41 (6), P. 391–407.
[13] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding // NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference. P. 4171–4186.
[14] Ebeling W., Neiman A. Long-range correlations between letters and sentences in texts // Physica A: Statistical Mechanics and its Applications, 1995, № 215 (3), P. 233-241
[15] Erk K. Vector space models of word meaning and phrase meaning: A survey // Language and Linguistics Compass, 2012, № 6(10), P. 635–653
[16] Ferreira D.C., Martins A.F.T., Almeida M.S.C. Jointly learning to embed and predict with multiple languages // 54th Annu. Meet. Assoc. Comput. Linguist. ACL 2016 - Long Pap. 2016. Vol. 4. P. 2019–2028.
[17] Firth, J.R. A synopsis of linguistic theory 1930-1955 // Studies in Linguistic Analysis, 1957, P. 1-32. Oxford: Philological Society.
[18] Gillet J., Ausloos M. A Comparison of natural (English) and artificial (Esperanto) languages. A Multifractal method based analysis // Vol. arXiv:0801.2510, Access mode: http://arxiv.org/abs/0801.2510

[19] Gu A., Goel K., Re C. Efficiently Modeling Long Sequences with Structured State Spaces // International Conference on Learning Representations. 2021, P. 1–32.
[20] Harris, Z. Distributional structure // *Word*, 1954, №10(23), P. 146-162.
[21] Holtzman A. et al. Learning to write with cooperative discriminators // ACL 2018 - 56th Ann. Meet. Assoc. Comput. Linguist. Proc. Conf. (Long Pap.) 2018. Vol. 1. P. 1638–1649.
[22] Holtzman A. et al. The curious case of neural text degeneration // Proceedings of the 2020 International Conference on Learning Representations. 2020. P. 2540.
[23] Lin H.W., Tegmark M. Critical behavior in physics and probabilistic formal languages // Entropy. 2017. Vol. 19, № 7. P. 1–25.
[24] Lund, K., Burgess, C., Atchley, R. A. Semantic and associative priming in a high-dimensional semantic space // Cognitive Science Proceedings (LEA), 1995, P. 660-665.
[25] Lund K., Burgess C. Producing high-dimensional semantic spaces from lexical co-occurrence // Behav. Res. Methods, Instruments, Comput. 1996. Vol. 28, № 2. P. 203–208.
[26] Kokol P. et al. Computer and natural language texts – a comparison based on long range correlations // J. Am. Soc. Inf. Sci. 1999. Vol. 50, № 14. P. 1295–1301.
[27] Kolchinsky A., Wolpert D.H. Semantic information, autonomous agency and non-equilibrium statistical physics // Interface Focus. 2018. Vol. 8, № 6
[28] Kulikov I. et al. Importance of search and evaluation strategies in neural dialogue modeling // INLG 2019 - 12th Int. Conf. Nat. Lang. Gener. Proc. Conf. 2019. P. 76–87
[29] Manin D.Y. On the nature of long-range letter correlations in texts // Vol. arXiv:0809.0103. Access mode: http://arxiv.org/abs/0809.0103. 2008. № 1. 1–14 p.
[30] Марковъ А.А. Примѣръ статистическаго изслѣдованія надъ текстомъ “Евгенія Онѣгина”, иллюстрирующій связь испытаній въ цѣпь // Извѣстія Императорской Академіи Наукъ. VI серія. 1913. Vol. 7, № 3. P. 153–162. In Russian. (English translation: Andrei Markov. 2006, An Example of Statistical Investigation of the Text Eugene Onegin Concerning the Connection of Samples in Chains. *Science in Context. 2006. Vol. 19, no. 4.* pages 591–600. DOI 10.1017/S0269889706001074.)
[31] Mikolov T. et al. Efficient estimation of word representations in vector space // 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings. International Conference on Learning Representations, ICLR, 2013.
[32] Mikolov T. et al. Distributed Representations of Words and Phrases and their Compositionality. // Proceedings of NIPS, 2013.
[33] Mikolov T. et al., Recurrent neural network based language model // Proc. of Interspeech 2010, pp. 1045–1048
[34] Montemurro M.A., Pury P.A. Long-range fractal correlations in literary corpora // Fractals. 2002. Vol. 10, № 4. P. 451–461.
[35] Osgood C., Suci G., Tannenbaum P. The measurement of meaning. — University of Illinois Press, 1957
[36] Pavlov A.N. et al. Scaling features of texts, images and time series // Phys. A Stat. Mech. its Appl. 2001. Vol. 300, № 1–2. P. 310–324.
[37] Pennington J., Socher R., and Manning C. GloVe: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, P. 1532–1543.
[38] Sanh V. et al. Multitask Prompted Training Enables Zero-Shot Task Generalization // ICLR. 2022.
[39] Schenkel A., Zhang J., Zhang Y.-C. Long Range Correlation In Human Writings. // Fractals. 1993. Vol. 4, № 3. P. 229–241.
[40] Shieber S.M. Evidence against the context-freeness of natural language // Linguist. Philos. 1985. Vol. 8, № 3. P. 333–343.
[41] Sundermeyer M., Schlüter R., Ney H. LSTM neural networks for language modeling // 13th Annu. Conf. Int. Speech Commun. Assoc. 2012, INTERSPEECH 2012. 2012. Vol. 1. P. 194–197.
[42] Thompson R., Booth T., Applying Probability Measures to Abstract Languages // IEEE Transactions on Computers, 1973, vol. 22, no. 05, pp. 442-450.
[43] Yamshchikov, I.P., Tikhonov, A. Music generation with variational recurrent autoencoder supported by history // SN Appl. Sci. 2, 1937, 2020. https://doi.org/10.1007/s42452-020-03715-w

[i] The dataset is available at https://github.com/nickm197/Longtexts

Named Entity-Oriented Sentiment Analysis with text2text Generation Approach

Ivan Moloshnikov Moloshnikov_IA@nrcki.ru
Maxim Skorokhodov Skorokhodov_MV@nrcki.ru
Aleksandr Naumov Naumov-AV@nrcki.ru
NRC "Kurchatov Institute"
Moscow, Russia

Roman Rybka Rybka_RB@nrcki.ru
NRC "Kurchatov Institute"
Russian Technological University "MIREA"
Moscow, Russia

Alexander Sboev Sboev_AG@nrcki.ru
National Research Nuclear University "MEPhI"
NRC "Kurchatov Institute"
Moscow, Russia

Abstract

This paper describes methods for sentiment analysis targeted toward named entities in Russian news texts. These methods are proposed as a solution for the Dialogue Evaluation 2023 competition in the RuSentNE shared task. This article presents two types of neural network models for multi-class classification. The first model is a recurrent neural network model with an attention mechanism and word vector representation extracted from language models. The second model is a neural network model for text2text generation. High accuracy is demonstrated by the generative model fine-tuned on the competition dataset and CABSAR open dataset. The proposed solution achieves 59.33 over two sentiment classes and 68.71 for three-class classification by f1-macro.

Keywords: multi-class classification, sentiment analysis, text2text generation, neural networks
DOI: 10.28995/2075-7182-2023-22-361-370

Анализ тональности по отношению к именованным сущностям с использованием подхода text2text generation

Иван Молошников Moloshnikov_IA@nrcki.ru
Максим Скороходов Skorokhodov_MV@nrcki.ru

Роман Рыбка Rybka_RB@nrcki.ru
РТУ "МИРЭА"

Александр Наумов Naumov-AV@nrcki.ru

Александр Сбоев Sboev_AG@nrcki.ru
НИЯУ "МИФИ"

НИЦ "Курчатовский институт"
Москва, Россия

Аннотация

В данной статье представлено описание решения задачи анализа тональности по отношению к заданным именованным сущностям в новостных текстах, выполненного в рамках на Dialogue Evaluation в 2023 году (RuSentNE). В статье исследуется две типа нейросетевых моделей для решения задачи мультиклассовой классификации: рекурентная нейросетевая модель с вниманием и векторным представлением слов, полученных из языковых моделей, а также нейросетевая модель для генерации текста в заданном формате. Лучшие результаты показала генеративная модель с подобранными гиперпараметрами и дополнительной настройкой на данных соревнования и доступного открытого корпуса CABSAR. Предложенное решение достигает точности по метрике F1-macro: 59.33 для двух классов тональности и 68.71 для трех классов.

Ключевые слова: мультиклассовая классификация, тональный анализ, text2text генерация, нейронные сети

1 Introduction

Sentiment analysis in relation to an entity in a news text is an important direction in the field of opinion mining and Natural Language Processing (NLP). The demand for effective approaches to targeted sentiment analysis grows with the increasing amount of news data (Brauwers and Frasincar, 2022; Zhang et al., 2022).

In recent years, solutions of this problem have transitioned from traditional machine learning methods (such as support vector machine or decision trees) to modern neural network models based on the Transformer architecture (Vaswani et al., 2017), in particular, large language models (LLM) like BERT (Devlin et al., 2018) or GPT (Radford et al., 2018).

Several methods for targeted sentiment analysis task were proposed based on these approaches. For example (Sun et al., 2019), by constructing an auxiliary sentence from the target, this task can be converted to a sentence-pair classification task. The authors of that paper (Sun et al., 2019) used a pre-trained BERT model fine-tuned on Sentihood and SemEval2014 Task 4 datasets. This method achieved the accuracy of 0.933. Another work (Ma et al., 2017) uses the Interactive Attention Network (IAN) with attention mechanism between a target (words that belong to the named entity) and its context. Put together with a recurrent neural network based on Long Short Term Memory (LSTM) layers, that network improved the accuracy by 5.6% compared to the ordinary LSTM on SemEval 2014 Task 4 dataset (Laptop part) (Kirange et al., 2014). An approach (Zhang and Lu, 2019) that used a pre-trained BERT model with point-wise feed-forward networks (PFFN) and Multi-Head Attention (MHA) increased the accuracy further by 4.25%, up to 76.35%, on the same dataset.

Generative models based on text generation (text2text) like GPT, BART (Lewis et al., 2019), or T5 (Raffel et al., 2020) can be used for the targeted sentiment analysis as well. A paper (Mishev et al., 2020) presented the BART language model with a dense layer for classification. This model was fine-tuned on SemEval 2017 Task 5 dataset (Cortis et al., 2017), achieving the f1-score of 0.95. Another work (Zhang et al., 2021) proposes an adaptation of a pre-trained T5 model. The authors induce the T5 model to generate text with sentiment elements for named entities. This approach demonstrates the f1-score of 69.42 on SemEval 2016 data (restaurant part) (Pontiki et al., 2016). These works show the efficiency of text2text models for the targeted sentiment analysis task and highlight the potential of using pre-trained generative models.

For the Russian language, solving entity-oriented sentiment analysis task is complicated by the limited amount of available datasets. Previous SentiRuEval competitions in 2015 and 2016 (Loukachevitch et al., 2015; Loukachevitch and Rubtsova, 2016) provided several open datasets. Labeled sentiment entities in the common case are for which sentiment was labeled were defined as words and expressions that denote some important characteristic of an entity (like ‘kitchen’ or ‘interior’ in SentiRuEval2015-reviews) or predefined company names (for tweets in SentiRuEval2015-tweets and SentiRuEval2016-tweets). Besides the competition datasets, an open corpus CABSAR has recently been introduced (Naumov et al., 2020). This corpus contains Russian-language sentences for three different domains: news, tweets, and posts from social networks. Each sentence includes labeling for named entities (Person and Location) and sentiment, labeled for each entity by three classes (positive, negative, and neutral). Sentiment labeling was performed by crowdsourcing.

The RuSentNE-2023 dataset (Golubev et al., 2023) significantly expands the available sets of labeled examples in the Russian language for the entity-oriented sentiment analysis task. Therefore, the purpose of this work is an investigation of two neural network methods for this task using the RuSentNE-2023 dataset:

1. the first method is based on a multi-class classification task. Here we have chosen a well-known neural network architecture based on a recurrent neural network model, which has demonstrated high efficiency in similar tasks. Word vector representations are obtained from large language models known to be efficient in various classification tasks (see section 3.1);
2. the second method is based on the text generation (text2text) approach. The T5 model for the Russian language is used. In this case, several variants of labeling data for output text sequences are tried (see section 3.2).

Dataset	Num. of samples	Avg. length (in chars)	NE sentiment class		
			Positive	Neutral	Negative
RuSentNE (train part)	6637	151.2	856(12.9%)	4774(71.9%)	1007(15.2%)
CABSAR	6705	129.5	2289(34.1%)	3068(45.8%)	1348(20.1%)

Table 1: Number of examples for each sentiment class for the datasets used.

Named entity tag name	RuSentNE-2023						CABSAR		
	Train-subpart			Valid-subpart					
	Pos.	Neg.	All	Pos.	Neg.	All	Pos.	Neg.	All
PERSON	339	290	1546	82	73	388	1962	1078	5070
ORGANIZATION	146	210	1168	40	51	319	327	270	1635
COUNTRY	109	168	1022	33	44	252	-	-	-
PROFESSION	68	108	1352	11	23	314	-	-	-
NATIONALITY	23	29	221	5	11	55	-	-	-

Table 2: Number of examples for each sentiment class by NER tags for the datasets used.

Our main contributions are:

1. two neural network methods are compared for the entity-oriented sentiment analysis task in Russian news texts;
2. the efficiency of merging several open-source datasets is evaluated for each method;
3. the dependence of the accuracy on applying methods for reducing computations during network fine-tuning is studied.

This paper is organized as follows: Section 2 describes the task and characteristics of the datasets used. Section 3 presents methods used for the entity-oriented sentiment analysis task, including neural network architectures, word vector representations, and pre-trained models. Section 4 demonstrates metrics for model validation, experiment results, and hyper-parameters of the final models.

2 Task and Datasets

The RuSentNE-2023 task (Golubev et al., 2023) is sentiment analysis in relation to named entities in a news text in the Russian language. Named entities of the following types are predetermined in the text: PERSON, ORGANIZATION, PROFESSION, COUNTRY, and NATIONALITY. The purpose of this task is to classify each of the given named entities into three sentiment classes: positive, negative, or neutral. The RuSentNE dataset contains train, development, and final-test parts. Each part includes sentences with labeled entities and their types. The train part has sentiment labels for named entities. The development part allows one to check the performance metric on the interface of the competition website [1].

Table 1 shows some statistics on the train part. Analysis of this table shows the following:

1. there is an imbalance of examples for different sentiment classes: entities of neutral class are predominant;
2. entities with different sentiment classes can be in the same sentence.

Since labels of the development part are only available online, to optimize the hyperparameters of the models, the train part of data is divided into 80% and 20% while preserving the representativeness of examples of sentiment classes. The first subpart (train-subpart, 5309 examples) is used to train models, and the second subpart (valid-subpart, 1325 examples) is used to estimate the efficiency of the models' hyperparameters. Table 2 shows the total number of examples for each sentiment class by NER tags for

[1] RuSentNE on CodaLab: `https://codalab.lisn.upsaclay.fr/competitions/9538`

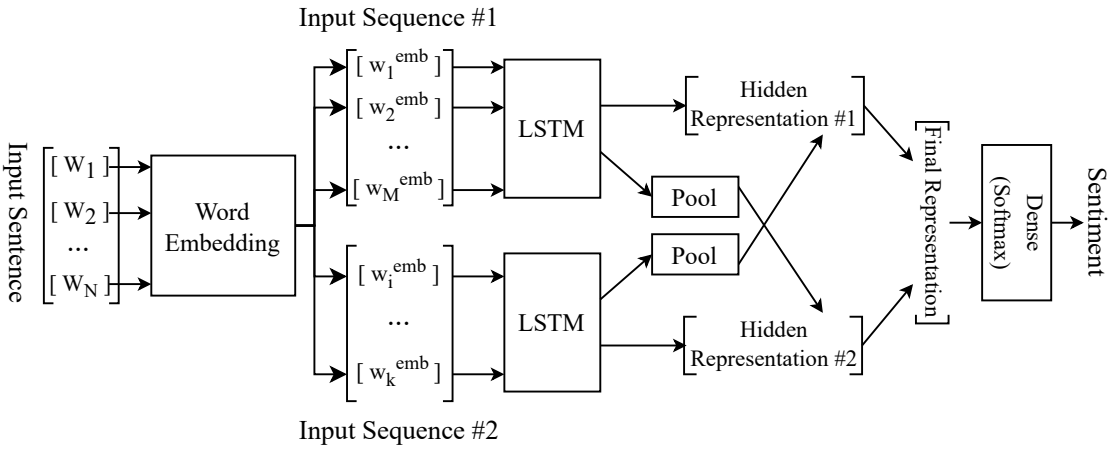


Figure 1: IAN model architecture.

train- and valid- subparts.

CABSAR (Naumov et al., 2020) is the closest corpus available in the Russian language to the dataset presented in the RuSentNE-2023 competition(Golubev et al., 2023). Therefore, this corpus was used to increase the number of train subpart examples. This corpus contains 6705 sentences in the Russian language from several sources: 2105 from LiveJournal blogs, 2603 from Lenta.ru news, and 1997 from Twitter. Named entity sentiment is labeled in these sentences by crowdsourcing. Table 1 and Table 2 show the number of samples and named entities for each sentiment class.

3 Methods

3.1 Multiclass Classification

This approach is based on a deep neural network with attention (Interactive Attention Network - IAN) (Ma et al., 2017). The authors of CABSAR (Naumov et al., 2020) used it to obtain baseline accuracy for the entity-oriented sentiment analysis task. Therefore, it was chosen to evaluate the accuracy of the RuSentNE-2023 task as a multiclass classification approach.

This method analyzes the input text sentence and splits it into two input sequences: for context (Input Sequence #1 in Figure 1) and target (Input Sequence #2 in Figure 1). The first input sequence is all the words of the sentence that contain a named entity, and the second input sequence is the words that belong to the same named entity for which sentiment is to be predicted. Word vectors obtained from these sequences are fed to a recurrent neural network based on LSTM layers with attention mechanism (see Figure 1).

The original IAN model used the GloVe (Pennington et al., 2014) as a word embedding model. The authors (Naumov et al., 2020) obtained a 0.7 f1-macro score on CABSAR using the ELMo language model (Peters et al., 2018) as a word embedding.

The following language models for word embedding are studied:

- ELMo (Peters et al., 2018) – word vector representations are formed based on Bidirectional LSTM layers. For the Russian language, a model trained on the Wikipedia text corpus is used from the DeepPavlov library [2].
- RuBERT (Kuratov and Arkhipov, 2019) is a model based on the Transformer architecture, obtained from Multilingual BERT pre-trained on 104 languages (Devlin et al., 2018). Then, that Multilingual BERT was trained on Wikipedia text corpus in the Russian language. The RuBERT used in this

[2]ELMo on Russian Wikipedia: http://docs.deeppavlov.ai/en/master/features/pretrained_vectors.html#elmo

Type of Input	Text of Sentence
A.	В роли **Тони Сопрано** Гандольфини удалось впервые создать образ гангстера с человеческим лицом. In the role of **Tony Soprano**, Gandolfini managed to create the image of a gangster with a human face for the first time.
B.	В роли **[Тони Сопрано]** Гандольфини удалось впервые создать образ гангстера с человеческим лицом In the role of **[Tony Soprano]**, Gandolfini managed to create the image of a gangster with a human face for the first time.
C.	**Тони Сопрано**. В роли **[PERSON]** Гандольфини удалось впервые создать образ гангстера с человеческим лицом **Tony Soprano**. In the role of **[PERSON]**, Gandolfini managed to create the image of a gangster with a human face for the first time.

Table 3: Options of data input for the text generation model.

paper is the large version of RuBERT is taken from the Huggingface[3] library.

- XLM-Roberta (Conneau et al., 2019) is a model based on the Transformer architecture, trained on 2.5 TB of data from CommonCrawl. The CommonCrawl data contains text in 100 languages, of which the Russian language is one of the most representative.

3.2 Text2text Generation

This approach is based on a generative neural network model with the Transformer architecture – T5 (Raffel et al., 2020). This model generates a new text from an input text. It consists of Encoder and Decoder blocks. The Encoder block accepts input text sequences as their word vector representation. The Decoder block generates new output text sequences.

Different options of input information for the text generation model were tested (see examples in Table 3):

A: source sentence text without any changes.

B: named entity in source sentence text is highlighted by square brackets, e.g. 'Tony Soprano' is replaced by '[Tony Soprano]'.

C: named entity type is replaced in source sentence text, e.g. 'Tony Soprano' is replaced by '[PERSON]'. The named entity text is added at the start of the source text.

Output text is one of the possible sentiments for the analyzed named entity: negative, neutral or positive.

Two T5-based models are considered within this approach: ruT5-base [4] and ruT5-large [5]. These models are an adaptation of T5-base and T5-large models for the Russian language. Wikipedia, books, news, and CommonCrawl texts were used to train them. The model dictionary size is 32101 tokens. The number of parameters is 220 million for the "base" model and 737 million for the "large" model.

4 Experiments

4.1 Metrics

As mentioned in the evaluation criteria of the RuSentNE-2023 competition, the main performance metric is the macro F1_pn-score, and the macro F1-score will be considered auxiliary. For macro-averaging,

[3]RuBERT-large: https://huggingface.co/ai-forever/RuBERT-large

[4]RuT5-base: `https://huggingface.co/ai-forever/ruT5-base`

[5]RuT5-large: `https://huggingface.co/ai-forever/ruT5-large`

№	Embbedding name	Add. data	Hyper. optim.	MLM tune	Valid-subpart		Final-test-part	
					F1-macro	F1_pn-score	F1-macro	F1_pn-score
1	ELMo	-	-	-	62.57	53.14	54.53	42.36
2	XLM-R-large	-	-	-	54.47	45.86	50.47	38.52
3	RuBERT-large	-	-	-	61.38	52.00	55.37	43.76
4	ELMo	+	-	-	61.10	50.32	54.96	44.14
5	XLM-R-large	+	-	-	50.49	42.38	51.12	41.97
6	RuBERT-large	+	-	-	60.68	51.53	55.94	46.20
7	XLM-R-large	-	+	-	58.07	47.36	54.66	42.16
8	RuBERT-large	-	+	-	65.82	56.29	57.09	44.67
9	XLM-R-large	+	+	-	64.63	54.44	57.45	45.36
10	RuBERT-large	+	+	-	66.79	56.80	59.46	48.17
11	XLM-R-large	-	+	+	64.04	54.22	56.78	44.24
12	RuBERT-large	-	+	+	67.76	58.68	56.17	43.85
13	XLM-R-large	+	+	+	64.69	54.99	56.05	46.16
14	RuBERT-large	+	+	+	65.88	56.21	54.80	44.77
-	RuSentNE-2023	-	-	-	-	-	56.71	40.92

Table 4: Results of the IAN model.

the F1-score calculation is averaged for each class separately. F1_pn-score is calculated by averaging the F1-score of two sentiment classes: negative and positive, excluding the neutral class.

4.2 Interaction Attention Network

The following experiments was performed with the IAN model:

- comparison of language models as word embeddings as part of the IAN model. In this case, hyperparameters were used from (Naumov et al., 2020). Only competition data are used for training;
- analysis of the impact of expanding the training samples by using additional data (CABSAR corpus);
- running hyperparameters optimization experiments with the RayTune library (Liaw et al., 2018) and selecting the more effective combination. The OpTuna framework (Akiba et al., 2019) was used as a search algorithm. The following hyperparameters were optimized: the size of the LSTM layer (hidden_dim), learning rate, batch size, etc.;
- pre-training of the language model used in IAN on the Masked-Language Modeling (MLM) task with 5000 steps and batch_size=64 on the train-subpart of the RuSentNE-2023 dataset. The model checkpoint was saved every 1000 steps, and the best one on the valid-subpart was selected.

The results of these experiments are shown in Table 4. Analysis of the results shows that the IAN model with word embeddings from the RuBERT-large model, using additional data, and with the hyperparameters optimization (exp. №10) has the best results among other methods: 48.17% and 59.46% by F1_pn-score and F1-score respectively on the final-test part of the data. It is better than the RuSentNE-2023 baseline by 7.25% and 2.75% respectively. In addition, there is an increase in scores in all experiments on the final-test part with using additional CABSAR data. Note that after hyperparameter optimization, IAN model with embeddings from RuBERT-large showed better results than with embeddings from XLM-R-large, although it has less parameters.

The best results of these models were obtained with the hyperparameters presented in Table 6.

4.3 ruT5 Model

Experiments with this model included: selecting the more effective option for input data representation, and evaluating the accuracy when using additional samples (CABSAR corpus) in the training part of the

Model Name	Extra Data	F1-macro	F1_pn-score
ruT5-base (type A.)	-	47.47	40.6
ruT5-base (type B.)	-	66.11	56.94
ruT5-base (type C.)	-	64.03	54.77
ruT5-large (type C.)	-	67.27	58.9
ruT5-base (type B.)	CABSAR	67.57	57.96
ruT5-base (type C.)	CABSAR	67.78	58.48
ruT5-large (type C.)	**CABSAR**	**68.71**	**59.33**
RuSentNE-2023	-	56.71	40.92

Table 5: Results of the text generation approach based on ruT5 model on the final-test part.

	ruT5-large	IAN-elmo	IAN-RuBERT-large	IAN-XLM-R-large
input text length	164	-	-	-
output text length	4	-	-	-
learning rate	10^{-5}	10^{-2}	$3.7 * 10^{-4}$	$2.2 * 10^{-5}$
batch size	64	4	64	128
LSTM hidden_dim	-	32	256	256
dropout	-	0.3		
train epochs	50	300 with early stopping		
optimizer	Adam			

Table 6: Hyperparameters for the best models.

data. Table 5 shows the results of experiments on the final-test part with ruT5 models.

As a result, the best model is ruT5-large with type "C" representation of input data, trained on the extended train part. Adding the CABSAR corpus, F1-score increases by 1.5%.

Text sequence generation is performed by Beam search with the number of beams equal to 2. For the final model, the hyperparameters were presented in Table 6.

The input text length (number of tokens) is set based on the maximum source sentence length in the competition dataset. The output text length is set based on the maximum number of tokens among the words "негативная" ("negative"), "нейтральная" ("neutral"), "позитивная" ("positive").

Calculations were conducted on the following equipment:

- ruT5-base model: Intel Xeon E5-2650v2 (2.6 GHz), 128 GB RAM, Nvidia Tesla K80;
- ruT5-large model: Intel Xeon E5-2630v4 (2.2 GHz), 64 GB RAM, Nvidia Tesla V100.

Additionally, a comparison of accuracy was performed for ruT5-large models (type C.) trained with and without Parameter-Efficient Fine-Tuning (PEFT)(Sourab Mangrulkar, 2022). In this case, the possibility of saving model accuracy was checked when training on small computing resources:

- Intel Xeon E5-2650v2 (2.6 GHz), 128 GB RAM, Nvidia Tesla K80

LORA(Hu et al., 2021) was used as the PEFT method. This method performs low-rank adaptation. It fixes weights of the pre-trained model and introduces trainable rank decomposition matrices into each level of the Transformer architecture. As a result, accuracy declined by 4% and 2% by F1_pn-score and F1-score respectively. However, it achieved a significant reduction in computing power requirements.

5 Discussion

A comparison of the best model score on the valid-subpart demonstrates a superior performance of the ruT5-large model for 4 of the 5 Named Entity (NE) tags (see Table 7). The accuracy of the sentiment classification for the NATIONALITY NE tag is similar for both models. The best accuracy (F1_pn-score

Named entity tag name	ruT5-large (type C)			IAN-RuBERT-large (exp-№10)		
	F1-micro	F1-macro	F1_pn-score	F1-micro	F1-macro	F1_pn-score
PERSON	73.71	69.98	65.19	70.1	64.57	57.75
ORGANIZATION	78.68	69.76	61.53	75.55	63.16	52.39
COUNTRY	85.32	81.05	76.28	80.56	71.23	62.59
PROFESSION	90.45	65.35	50.61	88.85	61.24	44.97
NATIONALITY	85.45	77.38	70.83	85.45	77.84	71.08

Table 7: Results of the best models by NER tags.

> 70) is achieved for the COUNTRY and NATIONALITY NE tags, and the worst (F1_pn-score 50) for PROFESSION. There are several factors involved in this, the most important of which is the balance of classes in the dataset used. For example, the proportion of the positive and negative sentiment classes to the total number of samples is 28% for COUNTRY and 25% for NATIONALITY NE tags. In contrast, the same proportion for the PROFESSION NE tag is 13%.

In this regard, improved accuracy can be achieved by:

- increasing the number of training data examples for the target task. This is confirmed by experiments with the addition of CABSAR data to the train-part of the RuSentNE-2023 dataset;
- applying more complex generative neural network models and training on larger datasets (e.g. GPT(Radford et al., 2018), T5-XXL(Raffel et al., 2020)).

Both datasets used in this paper extend the number of labeled examples for the joint task of named entities recognition and entity-oriented sentiment analysis for the Russian-language texts. However, they contain labels of mostly simple named entity samples, with a continuous word sequence and non-overlapping entities. The proportion of such complex samples for the RuSentNE-2023 and CABSAR datasets is 62 of 6637 (<1%) and 110 of 6705 (1.6%), respectively. Therefore, developing and researching named entity-oriented sentiment analysis methods for complex named entities is a very promising task.

6 Conclusion

This research shows the advantage of using a text generation approach for the entity-oriented sentiment analysis task. According to the results, the best accuracy was shown by the ruT5-large model with training on an extended dataset and a special input text representation. It was uploaded to the competition leaderboard as our final submission and showed a result of 59.33, which is 19% better than the baseline method in terms of the RuSentNE-2023 competition (Golubev et al., 2023). This result took the 5th place in the final rating leaderboard.

Our experiments with the multi-class classification model show that this method can be used for the target task. When using additional training data, a large language model for extracting word embeddings, and a hyperparameter optimization method, results were obtained that exceeded the baseline by 8%.

Further research will be focused on the improvement of input and output text data representation methods in generative neural network models, including for targeted sentiment analysis task.

References

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. // *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Gianni Brauwers and Flavius Frasincar. 2022. A survey on aspect-based sentiment classification. *ACM Computing Surveys*, 55(4):1–37.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. // *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, P 519–535.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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"*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

DK Kirange, Ratnadeep R Deshmukh, and MDK Kirange. 2014. Aspect based sentiment analysis semeval-2014 task 4. *Asian Journal of Computer Science and Information Technology (AJCSIT) Vol*, 4.

Yuri Kuratov and Mikhail Arkhipov. 2019. Adaptation of deep bidirectional multilingual transformers for russian language. *arXiv preprint arXiv:1905.07213*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

NV Loukachevitch and Yu V Rubtsova. 2016. Sentirueval-2016: overcoming time gap and data sparsity in tweet sentiment analysis. // *Computational Linguistics and Intellectual Technologies*, P 416–426.

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.

Dehong Ma, Sujian Li, Xiaodong Zhang, and Houfeng Wang. 2017. Interactive attention networks for aspect-level sentiment classification. *arXiv preprint arXiv:1709.00893*.

Kostadin Mishev, Ana Gjorgjevikj, Irena Vodenska, Lubomir T Chitkushev, and Dimitar Trajanov. 2020. Evaluation of sentiment analysis in finance: from lexicons to transformers. *IEEE access*, 8:131662–131682.

Aleksandr Naumov, R Rybka, A Sboev, A Selivanov, and A Gryaznov. 2020. Neural-network method for determining text author's sentiment to an aspect specified by the named entity. // *CEUR Workshop Proceedings*, P 134–143.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. // *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, P 1532–1543.

ME Peters, M Neumann, M Iyyer, M Gardner, C Clark, K Lee, and L Zettlemoyer. 2018. Deep contextualized word representations. arxiv 2018. *arXiv preprint arXiv:1802.05365*, 12.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Ion Androutsopoulos, Suresh Manandhar, Mohammed AL-Smadi, Mahmoud Al-Ayyoub, Yanyan Zhao, Bing Qin, Orphée De Clercq, et al. 2016. Semeval-2016 task 5: Aspect based sentiment analysis. // *ProWorkshop on Semantic Evaluation (SemEval-2016)*, P 19–30. Association for Computational Linguistics.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. 2018. Improving language understanding by generative pre-training.

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.

Lysandre Debut Younes Belkada Sayak Paul Sourab Mangrulkar, Sylvain Gugger. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

Chi Sun, Luyao Huang, and Xipeng Qiu. 2019. Utilizing bert for aspect-based sentiment analysis via constructing auxiliary sentence. *arXiv preprint arXiv:1903.09588*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Qiuyue Zhang and Ran Lu. 2019. A multi-attention network for aspect-level sentiment analysis. *Future Internet*, 11(7):157.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. *// 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.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*.

“Pears are big green”: gestures with concrete objects

Nikolaeva Y. V.
Lomonosov Moscow State University, Interdisciplinary Scientific and Educational School “Preservation of the World Cultural and Historical Heritage”, Moscow, Russia
julianikk@gmail.com

Abstract

The paper examines hand gestures when referring to inanimate referents. The aim of the study was to explore which factors determine the features of a gesture within the framework of modes of representation. Four main types of modes of representation were considered: drawing or shaping the form of the referent, acting, pointing, and presentation (PUOH); in addition, a new category of beat gestures was added.

As a result, it was shown that communicative dynamism or other referent characteristics such as control of the object or its inferability from the previous context do not fully determine the use of gestures with the referent. As an alternative hipothesis, we propose a notion of gesture information hierarchy, where discursive factors, such as previous mentions of the referent and the introduction or change of the protagonist along with the way an object is used determines the form of the gesture.

Keywords: gesticulation, reference, monologue, narration
DOI: 10.28995/2075-7182-2023-22-371-377

«Груши большие, зеленые»: жесты с конкретными референтами

Николаева Ю. В.
Междисциплинарная научно-образовательная школа
МГУ имени М. В. Ломоносова
«Сохранение мирового культурно-исторического наследия», Москва
julianikk@gmail.com

Аннотация

В статье рассматриваются жесты рук при описании одушевленных и неодушевленных референтов. Целью исследования было изучить, какие факторы определяют особенности жеста, с учетом модусов репрезентации, предложенных К. Мюллер. Были рассмотрены четыре модуса репрезентации: обозначение контура или формы референта, действие и репрезентация; кроме того, была добавлена новая категория (жестовые ударения или биты).

В результате было показано, что коммуникативный динамизм или другие характеристики референта, такие как контроль над объектом или выводимость из предыдущего контекста, не вполне объясняют использование жестов с этим референтом. В качестве альтернативной гипотезы мы предлагаем идею информационной иерархии жестов, где дискурсивные факторы, такие как предыдущие упоминания референта, введение или изменение протагониста, а также способ использования объекта определяют форму жеста.

Ключевые слова: жестикуляция, референция, монолог, рассказ

1 Credits

Annotations of manual gestures in the RUPEX corpus were made by A. Litvinenko and Y. Nikolaeva; vocal annotations were made by V. Podlesskaya and N. Korotaev.

2 Introduction

Co-speech Gesticulation is closely related to the content of speech and its context in general. Gestures represent the same communicative intention as words, so in gestures we can see realizations of the grammatical, semantic, and pragmatic meanings of corresponding words.

We used the RUPEX corpus [1] to study how animated and unanimated referents are illustrated with gestures, taking into account the referent's activation and how it is maintained throughout the narration. Six monologues (a total of 30 minutes) were analyzed, considering if it was the first or subsequent mention, the gesture type (if there was any), and episode boundaries. The aim of the study was to test the claim that first mentions of referents are more often accompanied by gestures, and these are C-VPT gestures.

3 Gesticulation, modes of representation and characteristics of a referent

According to McNeill [2, 3], the presence of gestures with a specific referent is determined by communicative dynamism, which is defined as the degree to which the information "pushes the communication forward" [4]. McNeill connects communicative dynamism to information status: the less accessible the information, the more probable a gesture with it is. Higher communicative dynamism also makes C-VPT (character viewpoint) gestures more probable. In contrast, there is other data showing that gestures tend to accompany referents that are reintroduced [5] or inferable from the preceding context [6] rather than first mentioned. Additionally, different gestures types can appear with different types of referents [7]: gestures with redundant information occurred with new referents, while non-redundant gestures occurred with already mentioned entities.

Also, McNeill [8] predicts, that "an absence of gesture is expected if there is a memory failure or its opposite, a complete predictability of the next step in discourse."

Considering these different results, it seems reasonable to consider different types of gestures. In [9], C. Müller discusses four modes of representation: drawing, molding, acting, and representing. They are combined into two groups: acting and representing. The first group of gestures depicts the actions of a protagonist, with the object being illustrated in gestures; the second group describes the form of the object. The first group is related to the character point of view (C-VPT), i.e., the gesturer acts as a character in his story; the second group is related to the observer point of view (O-VPT), i.e., the gesturer acts as an observer in relation to the events being described. Some authors suggest that the choice of point of view depends on the degree of activation of the referent in the narration. Thus, depending on the context, speakers can be expected to favor the first or second group of gesture modes.

Ortega and Özyürek [10] mention another aspect of gestural iconicity related to pantomime or personification. They noticed that this modus is used for animate objects; for controlled inanimate objects, the speakers chose action gestures, and for uncontrolled ones, drawing gestures. So, for the same clause, we can expect three different types of gestures depending on the speaker's mental representation and profiling of one of the few referents mentioned in the clause.

In sum, there are a few contradicting claims connecting the first or subsequent mentions of a referent (with the first mention, gestures in general are more probable) and the point of view or modus of repesentation (C-VPT gestures that relate to acting gestures in Müller's classification are more expected with the first mentioned referents, if we accept McNeill's perspective, or with controllable objects following Ortega and Özyürek).

Chu and Kita [11] found that speakers were more likely to produce speech-related gestures when the objects they saw triggered the action than when they did not. It is similar to the notion of control in [10], but makes the idea of how the object can be held more prominent. Another object feature can be its familiarity: if the speaker assumes that the addressee is unfamiliar with the object (for example, a toy with a ball tied to it), they may more often use a gesture with it [12].

4 Method

In this study based on "The Pear Story" retellings we used six monologues from the RUPEX corpus (recordings #04, #22, #23) to study gestures with all animate and inanimate referents (pears, bicycle, hat, baskets, apron, tie, pants, the girl's braids). Other inanimate referents were not added to the list

because they were rare and less often accompanied by gestures. We chose only the EDUs[1] that belonged to the main line of the story; if it was accompanied by a gesture, we marked the gesture type accordingly to the following procedure.

The RUPEX corpus divides gesture functions into four types: depictive, pointing, pragmatic, and beats, based on their formal characteristics, connection to speech and semantic features [13]. Beats are gestures with a simple form (usually realized through short up and down movement); they are supposed to highlight the corresponding fragment of speech, similar to the phrasal accent [14]. Pointing gestures have a typical form and convey information about the location of an object in the gesture space around the speaker. They are used to activate and reactivate the referent in the narrative. Depictive gestures are the most complex in form; they convey visual-spatial meanings and characterize the shape, movement, and mutual arrangement of objects. Depictive (representational) gestures are especially interesting for linguists since the complexity of their formal features makes it possible to study the relationships with different characteristics of speech, such as aspectuality, plurality, referent activation, etc. The fourth type is pragmatic gestures. They are characterized by a recognizable (partly emblematized) form and are associated not with the content of the story, but with the speaker's stance in relation to the events described, interaction between the interlocutors, discourse structure, etc.

For this study we reanalysed the classification and divided depictive gestures into two types: describing the form of the referent (O-VPT; molding or drawing gesture in [15]) or the character's actions (C-VPT; acting gesture gesture in [15]). As for pragmatics, most of them are PUOH in Müller's approach [16] or conduit metaphors in [2]; they were regarded as representing gestures in this study. There were few types of other pragmatics in the corpus and they were classified as beats based on their form (short downward movement coordinated with prosody). Thus, there were the following types of gestures:

1. beats (not mentioned in [15]),
2. representing (or pragmatic in [13]),
3. pointing and
4. iconic-OVPT (tracing in [15]),
5. iconic-CVPT (acting in [15]).

In the second part of our study, we organized gestures into three general types: gestures of presentation (1 and 2), gestures of form and position (3 and 4), and gestures of action (5).

for referents in the stories we noted for each EDU:

A. for animate referents:
1. the first and then
2. the second mention of a character,
3. the change of the protagonist (character reactivation),
4. subsequent references to a character.

B. for inanimate referents:
1. the first and then
2. the second mentioning of an inanimate referent from the list above,
3. subsequent references to the referent,
4. the absence of inanimate referents in the clause.

This study used a verbal transcription that assumes a very detailed division into EDUs, in which there were many cases of ellipsis and parcellation (splitting a syntactic clause into two or more EDUs), so we noted the animate referent even in those EDUs where it was not named explicitly, while for the episode the protagonist was retained. Additionally, we marked the cases where two animate referents were mentioned in one EDU.

5 Results

5.1 Communicative dynamism for referents and their gesture illustrations

In total, there were 763 EDUs related to the main line of the story

First, we tested the claim that new referents in an episode attract gestures in general or C-VPT gestures, Table 1 and Figures 1 and 2 present the results.

[1] Elementary discourse unit, defined primarily on the basis of prosodic criteriaю

EDU contained	No gesture	Beat	Representing	Pointing	Iconic-OVPT	Iconic-CVPT	Total
First menton of a character	22	0	2	1	1	2	28
Reintroduction of a character	57	1	3	3	4	9	77
Second mention of a character	18	0	3	1	3	3	28
Other mentions of a character	222	11	26	21	67	171	518
Two characters	38	0	0	1	1	11	51
Last mention of a character in the episode	43	3	2	2	1	9	60
First mention of an object	29	0	4	2	24	20	79
Second mention of an object	12	2	4	1	13	7	39
Other mentions of an object	86	9	20	23	30	159	327
No object	273	4	8	3	10	19	317

Table 1: Animate and inanimate referents with gestures

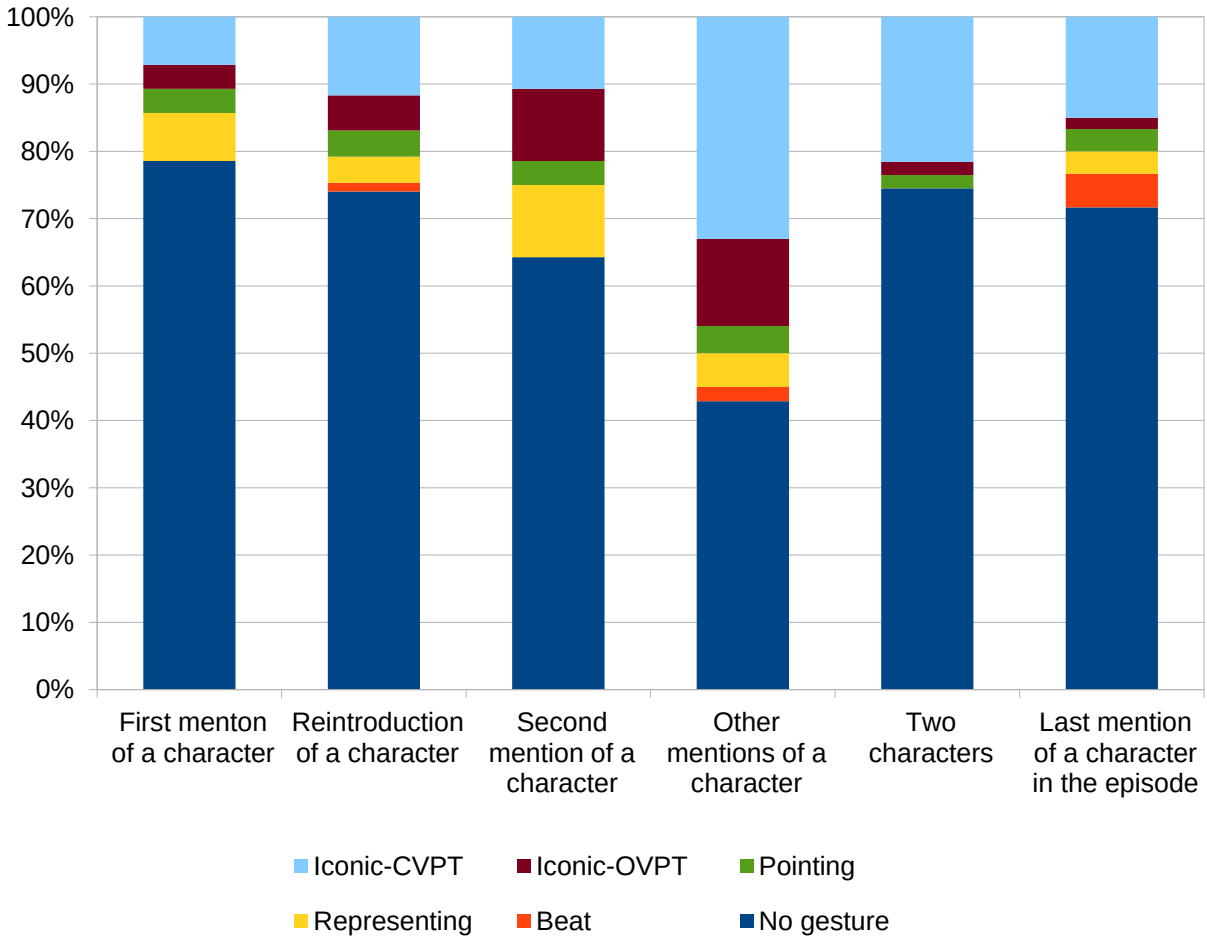


Figure 1: Animate referents with gestures

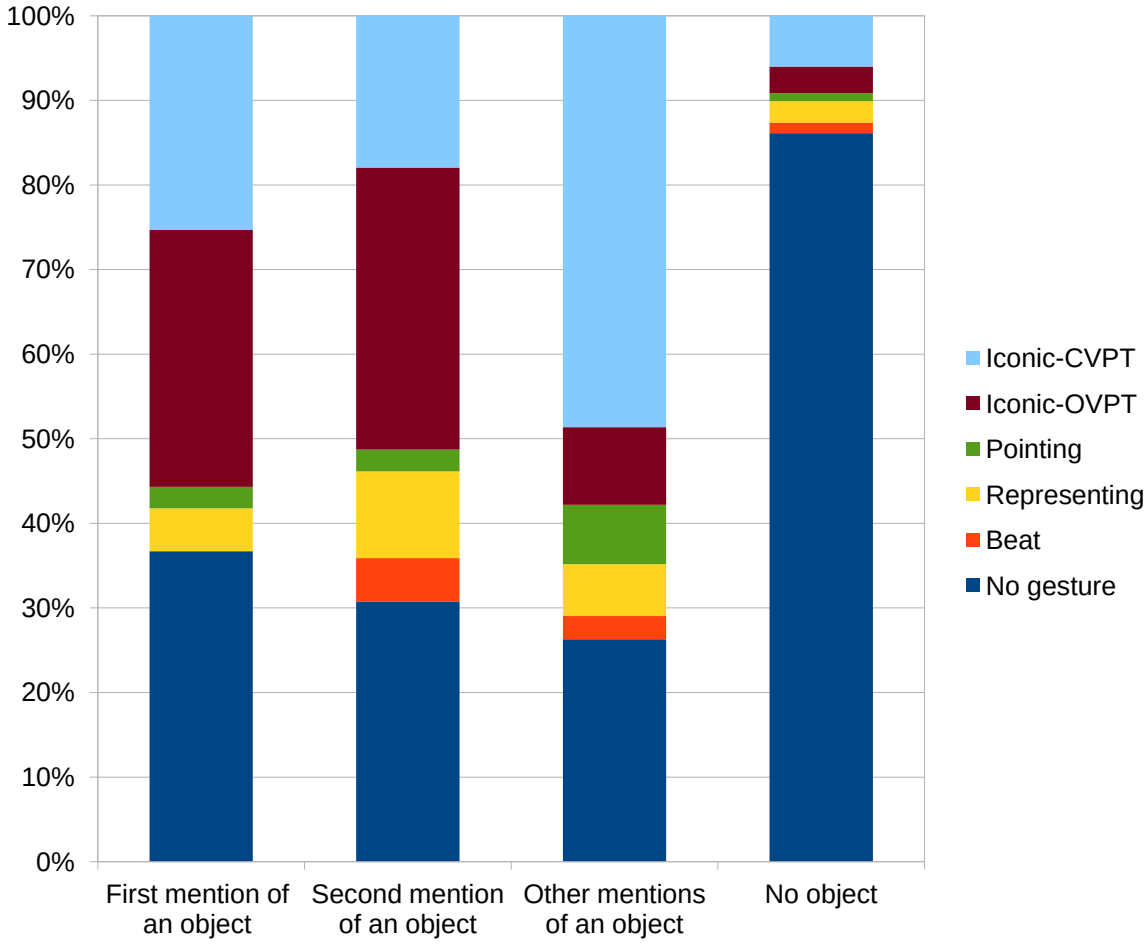


Figure 2: Inanimate referents with gestures

As shown in Table 1 and Figures 1 and 2, neither first nor second mention of a referent is more often accompanied with gestures. It contradicts the assumptions in [2, 3]. For animate referents, it is much more probable to see it with a gesture in the middle of the episode. Additionally, EDUs with inanimate referents are more often illustrated with gestures, than those without them, but there is still a tendency for central EDUs in the episode to be combined with gestures.

5.2 Hierarchy of gesture information within an episode

We supposed that position within the episode (but not the order of mentions) determines the type of accompanying gesture.

We propose a hierarchy of gestural activation according to the growth point hypothesis [17]. We suppose that verbal and gestural descriptions of a referent unfold in parallel, and we can expect gestural means of tracing a referent to be elaborated and differentiated along with verbal ones, but not necessarily in the same way. In this part of the research, we tested the hypothesis of gestural activation:

1. For the first gestural illustration of the protagonist, the speakers use gestures of presentation. They signal the importance or novelty of the simultaneously mentioned referent.

2. Gestures of form and position continue the gestural description of already introduced referents and add information about their appearance and/or position relative to other referents.

3. Action gestures are used to describe the protagonist's actions related to the main line.

4. Beats can mark episode boundaries too, but at the end of the episode (unlike gestures of presentation, which appear at its beginning). In our corpus, one of the speakers (04N) used 3 of the 13 beats at the last EDU of the episode.

There is a claim behind this approach, supposing that the gestural track has been maintained since the character is first named in the story.

All gestures can appear not with the first EDU but later, and at every of the three first stages of the referent description there can be more than one gesture of a particular type.

We tested the hypothesis with first mentions opposed to subsequent ones for animated referents. We checked which general gesture type was the first in two types of episodes: when the character is first mentioned in the narration and when he or she is reintroduced after an episode boundary (see Table 2).

	Presentation	Form and position	Action
First gestural illustration of a character	6	10	7
Gestural illustration after reintroduction of a character	7	12	21

Table 2: Gesture general type and their appearance within the episode

In fact, the choice of a gesture to start the story was more complex: in 4 of the 6 monologues examined, the very first gesture was of action. This might be due to the stimulus material for the story: the film begins with a close-up showing the hands of the gardener, who picks up pears and puts them in his apron. This movement was the first gesture in four of six subjects.

Figure 1 shows that with the first and even more often with the second mention of a referent the probability of a gesture of representation slightly grows, that can be interpreted as an indirect support for our claim.

6 Discussion

The choice of a particular form of depictive gesture when describing an inanimate referent is largely determined by its appearance and the way it is used. However, a preliminary analysis has already shown that other factors may influence the proposed scheme. The speaker uses gestures to consistently inform the addressee of the appearance of a new referent, its external features and the character's actions with the presented object.

At the same time, gestures reflect the history of previous references and sometimes signal an upcoming boundary of the episode.

Our findings do not support the idea that first mentions, inferability of a referent or other referent’s features directly influence the use of gestures, but we suppose that communicative dynamism can be a reason for gestures to be used, although the most dynamic clauses seem to be in the middle of an episode rather than in its beginning,

We distinguish three main modi of gestures as referential means in the narrative: the presentation of an object (announcing its existence, presentation or PUOH gestures); the description of the appearance of a character or object (iconic O-VPT); the position of the referent relative to the characters already mentioned (deictics); and iconic C-VPT which are related to actions of characters in the story. We partly support the idea that the modus of the gesture is determined by the ability to use the object, but we believe that along with this there is a procedure of 'introducing' the object, where first there will be either an indication of its existence or location, or a description of its properties and/or form, and this description can be quite extensive (more than one EDU and more than one gesture).

These observations, of course, only apply to a particular genre: the narration, an extended, coherent story about events in the real world. Other genres and types of discourse suggest different gestures use.

References

1 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, 2018. — Vol.15(2). — P. 191–200.
2 McNeill David. Hand and mind. — Chicago, IL: University of Chicago Press, 1992.
3 McNeill David, Levy Elena, Duncan Susan. Gesture in Discourse // The Handbook of Discourse Analysis, 2015. — P. 262-289.
4 Firbas Jan. On the concept of communicative dynamism in the theory of functional sentence perspective // Brno Studies in English, 1971. — Vol. 7. — P. 12–47.
5 Debreslioska Sandra, Gullberg Marianna. Information Status Predicts the Incidence of Gesture in Discourse: An Experimental Study // Discourse Processes, 2022. — Vol.59(10). — P. 791–827.
6 Debreslioska Sandra, Gullberg Marianna. What's New? Gestures Accompany Inferable Rather Than Brand-New Referents in Discourse // Frontiers in Psychology, 2020. — Vol. 11.
7 Foraker Stephany. Gesture and discourse // Integrating Gestures, 2011. — P. 279-292.
8 McNeill David. Gesture and Thought. — Chicago, IL: University of Chicago Press, 2005.
9 Müller Cornelia. Gestural modes of representation as techniques of depiction // C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & J. Bressem (Eds.), Body–Language–Communication: an international handbook on multimodality in human interaction, Volume 2, Berlin: De Gruyter Mouton, 2014. — P. 1687–1702.
10 Ortega Gerardo, Özyürek Asli. Systematic mappings between semantic categories and types of iconic representations in the manual modality: A normed database of silent gesture // Behavior Research Methods. — 2020. — Vol. 52. — P. 51–67.
11 Chu Mingyuan, Kita Sotaro. Co-thought and co-speech gestures are generated by the same action generation process // Journal of Experimental Psychology: Learning, Memory, and Cognition, 2016. — Vol. 42(2). — P. 257–270.
12 Campisi Emanuela, Özyürek Asli. Iconicity as a communicative strategy: Recipient design in multimodal demonstrations for adults and children // journal of Pragmatics, 2013. — Vol. 47(1). — P. 14-27.
13 Litvinenko Alla O., Kibrik Andrej A., Fedorova Olga V., Nikolaeva Julia V. Annotating hand movements in multichannel discourse: Gestures, adaptors and manual postures // Russian Journal of Cognitive Science, 2018. — Vol.5(2). — P. 4–17.
14 Prieto Pilar, Cravotta Alice, Kushch Olga, Rohrer Patrick, Vilà-Giménez Ingrid. Deconstructing beat gestures: a labelling proposal // Proceedings of the 9th international conference on speech prosody, 2018. — P. 201-205.
15 Müller Cornelia. Gestural modes of representation as techniques of depiction // C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, & J. Bressem (Eds.), Body–Language–Communication: an international handbook on multimodality in human interaction, Berlin: De Gruyter Mouton, 2014. — P. 1687–1702.
16 Müller Cornelia. Forms and uses of the Palm Up Open Hand: A case of a gesture family? // The semantics and pragmatics of everyday gestures, 2004. — P. 233-256.
17 McNeill David. The growth point hypothesis of language and gesture as a dynamic and integrated system // Cornelia Müller, Alan Cienki, Ellen Fricke, Silva Ladewig, David McNeill and Sedinha Tessendorf (Eds.), Body–Language–Communication: an international handbook on multimodality in human interaction, Volume 1, Berlin: De Gruyter Mouton, 2013. — P. 135–155.

Russian Constructicon 2.0: New Features and New Perspectives of the Biggest Constructicon Ever Built

Alexander Orlov
HSE University
alexander.orlov98@gmail.com

Zoia Butenko
HSE University
UiT The Arctic University of Norway
zoiab@uio.no

Daria Demidova
UiT The Arctic University of Norway
dashademidova1998@gmail.com

Vladimir Starchenko
HSE University
vsstarchenko@hse.ru

Ekaterina Rakhilina
HSE University
Vinogradov Institute for Russian language (Russian Academy of Sciences)
rakhilina@gmail.com

Olga Lyashevskaya
HSE University
Vinogradov Institute for Russian language (Russian Academy of Sciences)
olesar@yandex.ru

Abstract

Russian constructicon is an open-access linguistic database containing detailed descriptions of over 3,800 Russian grammatical constructions. In this paper we present a new, enlarged and updated version of Russian Constructicon (RusCxn) as well as new trajectories of development which were opened for the resource after the update. Since its first release, RusCxn, has undergone many significant changes. Our team has expanded the number of constructions present in the database 1,5 times, introduced new meta-information features such as glosses, significantly reworked the architecture and the design of Russian Constructicon's website, and improved the search facilities. The above-mentioned changes not only make RusCxn more attractive and convenient-to-use, but they can also greatly facilitate typological research in the field of Construction Grammar and improve the mapping between constructicography-orinented resources for different languages.

Keywords: Constrction Grammar; construction; constructicon.
DOI: 10.28995/2075-7182-2023-22-378-385

Русский Конструктикон 2.0: Новые особенности и новые перспективы развития самого большого в мире конструктикона

Александр Викторович Орлов
Национальный исследовательский университет «Высшая школа экономики»
alexander.orlov98@gmail.com

Зоя Алексеевна Бутенко
Национальный исследовательский университет «Высшая школа экономики»
Университет Тромсё — Норвежский арктический университет
zoiab@uio.no

Дарья Александровна Демидова
Университет Тромсё — Норвежский арктический университет
dashademidova1998@gmail.com

Владимир Миронович Старченко
Национальный исследовательский университет «Высшая школа экономики»
vsstarchenko@hse.ru

Екатерина Владимировна Рахилина
Национальный исследовательский университет «Высшая школа экономики»
Институт русского языка им. В. В. Виноградова РАН
rakhilina@gmail.com

Ольга Николаевна Ляшевская
Национальный исследовательский университет «Высшая школа экономики»
Институт русского языка им. В. В. Виноградова РАН
olesar@yandex.ru

Аннотация

Русский конструктикон — это бесплатная электронная лингвистическая база данных, содержащая подробные описания более 3800 русских грамматических конструкций. В этой статье мы хотим представить новую, расширенную и обновленную версию Русского Конструктикона, а также рассказать о новых перспективах развития ресурса, открывшихся после масштабного обновления. С момента своего первого выпуска Русский Конструктикон претерпел множество значительных изменений. Наша команда в 1,5 раза увеличила количество представленных на ресурсе конструкций, разработала новые типы мета-информации для описываемых конструкций, в частности глоссы, значительно переработала архитектуру и дизайн сайта Русского Конструктикона, а также улучшила механизм поиска. Эти изменения не только делают Русский Конструктикон более привлекательным и удобным в использовании, но также могут значительно облегчить типологические исследования в области грамматики конструкций и улучшить связь между конструктографическими ресурсами для разных языков.

Ключевые слова: грамматика конструкций; конструкция; конструктикон.

1 Introduction

In this paper we introduce a new, enlarged and upgraded version of “The Russian Constructicon (RusCxn)” as well as discuss the new prospects which became available for the resource after the upgrade.

1.1 Basic terms

The term *Constructicon* denotes both a system of constructions of a particular language, and a detailed description of this system, normally presented in a form of a searchable database.

Construction is a key term employed by Construction Grammar (CxG), which assumes that constructions are fundamental building blocks of a human language [1], [2], [3], [4]. Under this theory, any linguistic pattern or model can be recognized as a construction “as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist” [1: 5] – in other words, constructions are understood as fixed linguistic patterns lacking compositionality[1]. Thus, grammatical structures as small in scope as prefixation and as large in scope as passive voice can be viewed as constructions. While constructicons for some languages (i.e., English, Brazilian Portuguese) store all linguistic patterns compatible with the definition above, many prefer to narrow the object of their research, focusing only on *grammatical constructions* – predominantly multiword linguistic patterns, which (1) lie on the border of lexis and grammar and (2) which are partially schematic [6], [7]. [2]

[1] See more on compositionality in [5].

[2] Yet it is worth noting that such type of constructions better be called *quasi-grammatical* since, as stated in their definition, they express meanings which combine grammatical elements with the lexical ones. Albeit the term *grammatical construction* has been used in CxG works to denote such quasi-grammatical constructions since the release of [6], and we are not going to dispute it in this paper.

In the Russian Constructicon we also mainly focus on grammatical constructions. The examples of such constructions include *NP-Nom Cop что надо*[3] 'NP-Nom Cop what needed' – a construction signifying superior quality of NP, or *без пяти минут NP* 'without five minutes NP' – a construction meaning that NP will in the nearest future experience a change in its (social) status. Please note that both constructions express meanings which are neither fully lexical nor fully grammatical, and that neither meaning is clearly entailed from the component parts of a construction.

For the sake of simplicity, hereby we will use the term *construction* to specifically denote grammatical constructions as defined in [6].

1.2 Russian Constructicon 1.0

First presented to public in 2020 [7], the Russian Constructicon is a joint project of HSE University and The Arctic University of Norway. Upon its release, it became the largest constructicon of any language, featuring over 2,000 constructions. It was also among the most functional. In RusCxn, each construction was accompanied by a substantial amount of meta-information, incl. definition, examples, semantic type of a construction, syntactic structure of a construction, etc. (see more in [8]). RusCxn also became one of the two constructicons (the other being Swedish) which were specifically targeted not only towards the academics but also towards the L2 students of a particular language, providing CEFR levels and easy-to-read definitions of the constructions stored.

Clearly, with such a swift start, RusCxn had potential for becoming an exemplary constructicon. Yet, obviously, RusCxn came not without its own flaws. Some of these flaws, e.g., lack of glossing system, inability to form unique URLs for individual constructions, etc., seriously hindered both current usability and future perspectives of the resource. Therefore, after delivering the first version of the database to the public, RusCxn team continued efforts on the project.

Our team, which has expanded over time, implemented a number of fundamental changes in the Russian Constructicon since its first release: we have greatly increased the number of constructions present in the database, filled the gaps in RusCxn's instrumentarium and addressed some significant faults in the design of RusCxn's website. Hereby, we will refer to the new and upgraded version of RusCxn as 'The Russian Constructicon 2.0' (as opposed to RusCxn 1.0 version delivered in 2020).

This article aims to describe the biggest changes in the second iteration of the project. Section 2 addresses improvements on the 'theoretical side' of the Russian Constructicon, with 2.1 reporting on the increase in the number of constructions described in RusCxn and 2.2 - on the introduction of glossing system. Section 3 highlights the improvements on the 'computational side of thing', namely the launch of a new website for the resource. Section 4 describes further perspectives of the Russian Constructicon which became available after the big update.

2 RusCxn 2.0: contents update

2.1 Expanding the Number of Constructions Described

In the Russian Construction 2.0 the number of featured constructions increased from 2,200 to 3,800.

The new constructions originated from the following sources: (1) a list of phrases to depict manner, retrieved from ruscorpora.ru (~2,500 entries) [9], (2) Thesaurus dictionary of the Russian idioms [(>8,000 entries) [10], (3) a list of constructions collected manually from the Russian fiction books (~600 entries).

All the entries were manually examined by several annotators for compliance with our criteria. To begin with, it was necessary to make sure that the constructions under examination were not already present in our database (i.e., did not match the constructions from RusCxn 1.0). Thus, a significant number of entries (~ 800) were eliminated on the first stage because of repetition.

On the next stage, the units under consideration had to be checked for compliance with the definition of the construction adopted in RusCxn. A substantial number of entries from sources 1 and 2 (~ 6,000) were eliminated at this stage. For instance, [10] contains a large number of proverbs and sayings that do

[3] In the Russian Constructicon we developed a special system of notation for construction formulae. NP(-Nom) = noun phrase (with a noun in the nominative case), Cop = copula. See more in [7]. Hereby all examples of constructions are presented in accordance with how they appear on RusCxn's website, i.e., without transliteration.

not form constructions according to our definition, as they lack a free slot. In addition, lexical constructions, the semantics of which go well beyond the framework of quasi-grammatical meanings explored by RusCxn, were also excluded (cf. *VP в чем мать родила* 'lit. VP in which the mother gave birth' signifying *naked* – a fully lexical meaning). The process of annotation of phraseological units from [10] is described in more detail in [11].

Finally, we controlled for the frequency of use and stylistic coloring of the remaining constructions. Thus, we did not include into the final update constructions which have fallen out of use through time or were rather rare (as demonstrated by the data from Russian National Corpus). Because of the formal constraints imposed on us by the pedagogical nature of the resource, we were also unable to include any constructions that might appear rude or explicit to our users[4].

Subsequently, more than 1,600 constructions were added to the new version of the Russian Constructicon, increasing its volume by more than 50 per cent. These constructions are annotated in accordance with RusCxn rules and should be available to the public via a new website by the date of publication.

2.2 Introducing the Glossing System[5]

With definitions of constructions in multiple languages and semantic equivalents of Russian constructions in English and Norwegian, the Russian Constructicon was meant to be an internationally oriented resource since its release. Nevertheless, the usability of RusCxn 1.0 for non-Russian speaking linguists was significantly hindered by the lack of glosses or any other device non-Russophones could use to understand the inner structure of constructions. Therefore, in RusCxn 2.0, we developed a glossing system.

2.2.1 Glossing Format

Our glossing system is based in the Leipzig glossing format [12]. These are some examples of glossed constructions from the Russian Constructicon:

(1)	ко-му	как-ое	дел-о	Cop	до	NP-Gen
	who-DAT.SG	which-NOM.SG.N	deal-NOM.SG	Cop	to	NP-Gen
(2)	пош-л-и/пойд-ём	VP-Pfv.Fut/VP-Ipfv.Inf				
	go-PST-PL/go-FUT.1PL	VP-Pfv.Fut/VP-Ipfv.Inf				
(3)	больно	Adv/Adj/Pred				
	too<painfully	ADV/ADJ/PRED				

We describe the glossing rules in detail in [13]. In this article, we address only some key features of our system.

To begin with, in our glossing system, we utilize the symbol < for translating roots with multiple meanings, provided the literal/original meaning of a root is distinct from the contextual meaning. Consider word *больно* in (3): even though it originally means *painfully*, in this particular construction it is used as an intensifier, better translated as *too*. Under our glossing rules literal or original meaning of a root should appear to the right of <, whilst contextual meaning – to the left.

In addition to this, symbol < can be used to convey the origins of some function words. Cf. second *что* in (4) – a complementizer originating from a word meaning *what,* or *ишь* in (5), – a particle that comes from PRS.2SG of a verb *видеть* 'see'[14].

(4)	ну	и	что,	что	XP	
	PTCL	and	what	COMP<what	XP	
(5)	ишь,				как-ой	Adj-Nom Cop!
	PTCL<see{OBSOLETE}.PRS.2SG				which-NOM	Adj-Nom Cop

In the Russian Constructicon 2.0 we do not provide translation equivalents for complementizers, interjections, or particles, as it is rarely possible to find an exact equivalent for such words; yet we believe that preserving their source-meanings, where possible, may prove useful for some researchers.

Unlike many other glossing systems, one adopted at RusCxn 2.0 aspires to retain the original stylistics of roots and words glossed through the use of special stylistic labels. We currently have two labels:

[4] Given that the constructions excluded on stages 2 and 3 might still be of interest to some researchers, we plan to build a separate resource for hosting such entries.

[5] All the additions described in this section are to appear on the new version of the site by the date of publication.

OBSOLETE, marking usages which sound old-fashioned or which are no longer used in speech on their own, such as *сего* in (6), and SUBSTAND, marking colloquial and/or inappropriate in a written language usages such as *в лом* in (6).

(6) ни с того ни с се-го Cl
NEG from that.GEN.SG NEG from this{OBSOLETE}-GEN.SG Cl

(7) (NP-Dat) Cop в лом VP-Inf
(NP-DAT) COP in unwillingness{SUBSTAND}<break VP-INF

Glosses will occupy a separate section in the description of a construction. Since glosses are of interest to a limited number of users only, they will be displayed in an advanced meta-information section.

2.2.2 The Great Slash Problem – Resolved

In RusCxn slash (‘/’) is utilized to denote that some elements of a construction are in a free variation with each other, e.g., in *NP-Gen.Pl Cop выше крыш-и/голов-ы* ‘lit. NP-Gen.Pl Cop higher than the roof/head’ speaker can say either *roof* or *head* without any significant difference in meaning. Such notation, convenient at a first glance, becomes rather problematic when variative parts of a given construction consist of more than one word. Consider (8) where slash suddenly denotes free variation not between neighboring words *ущерб* and *во*, but rather between phrases *в ущерб* and *во вред*.

(8) VP в ущерб/во вред NP-Dat
VP in damage.ACC.SG/in harm.ACC.SG NP-DAT

The extreme example of this is construction (9), where multi-layered variation is found: first, *Pred* interleaves *Adj.Short*, and then the string of words *не так (уж и) Pred/Adj.Short* interleaves *не так уж и Adv VP*.

(9) NP-Nom Cop не так (уж и) Pred/Adj.Short/не так уж и Adv VP
NP-Nom Cop NEG so (FOC.PTCL and) Pred/Adj.Short/NEG so (FOC.PTCL and) Adv VP

Even though notations as (8) and (9) appear comprehensible to native speakers, they might be difficult to understand for L2 learners of the language because of the lack of clearly marked borders for variative parts, not to mention that such inconsistent notation is obviously unfit for automatized computational analysis. At the same time, marking borders for variative parts with square brackets or alike devices could pose a problem of its own, since, with too many notation symbols, construction formula risks to become unreadable.

After the introduction of a special section for glossing, this problem was partially resolved. Square brackets to mark the borders of variative parts were introduced in all construction formulae with slashes. Thus, (8) and (9), under new notation will look like this in the database:

(10) VP [в ущерб]/[во вред] NP-Dat
VP [in damage.ACC.SG]/[in harm.ACC.SG] NP-DAT

(11) [NP-Nom Cop не так (уж и) [Pred]/[Adj.Short]]/[не так (уж и) Adv VP]
[NP-Nom Cop NEG so (FOC.PTCL and) [Pred]/[Adj.Short]]/[NEG so (FOC.PTCL and) Adv VP]

To avoid overloading the construction entry with special symbols, the square brackets will not be visible in the main line introducing construction formula on the website, yet they will be visible in the gloss section.

2.2.3 Gloss Search

The Russian Constructicon 2.0 will feature a search facility for glosses, which will be part of Advanced search. This facility will allow to search constructions containing an individual gloss or a combination of glosses located at a particular distance from each other. A gloss can be represented by a translation (typed manually), a grammatical category (chosen from a pre-existing list), or a combination of a translation and a grammatical category. Translation part also allows for the use of regular expressions (e.g., one can find all constructions containing words glossed with ‘put’ and ‘lay’ with one query ‘put|lay’). The gloss search facility will be available via *Advanced Search* tab of our website and can be combined with other search facilities present in advanced search (semantic type of construction, syntactic structure of anchor, CEFR level, etc.).

3 RusCxn 2.0: Website Overhaul

The data from RusCxn 1.0 was available to users through a website built upon Github. The resource worked rather slow since the site had to cache all the data from a pre-made Google Spreadsheet with constructions at every opening. In addition to that, since the searching process was carried out at the expense of the front-end, the issuance of results was performed on the same webpage with a single unchanged URL. The descriptions of individual constructions also did not have individual URLs since this information was not pre-stored in any kind of a database. Thus, it was not technically possible to directly link RusCxn's data with the data from its other satellite resources, e.g. Constructesize![15], containing exercises on the constructions, or Pragmaticon[16], containing related discourse formulas [17]. It was also unfeasible to provide a direct link to a particular construction in the description of a different construction, for example, to signal their similarity or synonymity, or to formally depict construction families [18] in the database. To sum up, the RusCxn 1.0's website design was inconvenient for both ordinary users and academics, seriously limiting research possibilities and general perspectives of the resource.

In Russian Constructicon 2.0 we resolved the above-mentioned inconveniences by developing a totally new web-platform for the project. The new site is based on an SQL database. The resource is currently hosted ruscorpora.ru and is available through https://constructicon.ruscorpora.ru/. New design allows for assignment of unique URLs to each page with the description of a construction and to each search query; it also allows for a more swift and efficient processing of the data.

In addition to the creation of a new back-end for our resource, we introduced significant changes to the front-end. To accommodate new users, we added two sliders (in Russian and English) containing explanatory information about the resource. Each slider answers five basic questions about RusCxn in a simple and vivid language with several examples. The questions are *What is a construction?*, *What is constructicon?*, *What is the purpose of the Russian Constructicon?*, *What can you find here?*, and *Who built this resource and how?*. Now we also display sample queries in a search bar (cf. *не говоря о* in Fig.1) to better familiarize new users with the format of queries and the content of the resourse.

We additionally enhanced the appearance and the general usability of the website by changing a color scheme, text font, and a configuration of the plain text and widgets throughout the resource. For instance, the output window on the main page is now located under the search bar and is reduced in size to give way for the slider (yet the results are more readable than before due to the darkened color of the text and the font which prevents amalgamation).

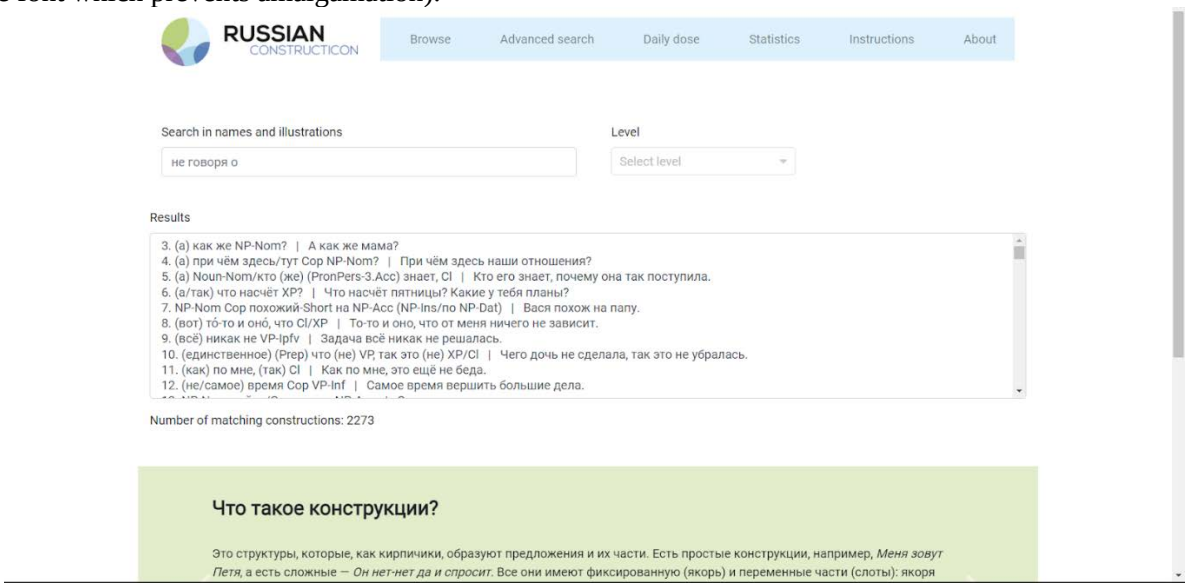


Figure 1: The main page of Russian Constructicon after the update

On top of that, we fixed several bugs in the searching mechanism, completely redesigned the Advanced search, and significantly changed the display of meta-information for a construction. Additionally, an option to choose language was introduced at the top of relevant pages (previously, the site would

feature duplicative sections like *Instructions Russian* and *Instructions English* in its header). The contents of the text pages were also rewritten to improve their comprehensibility.

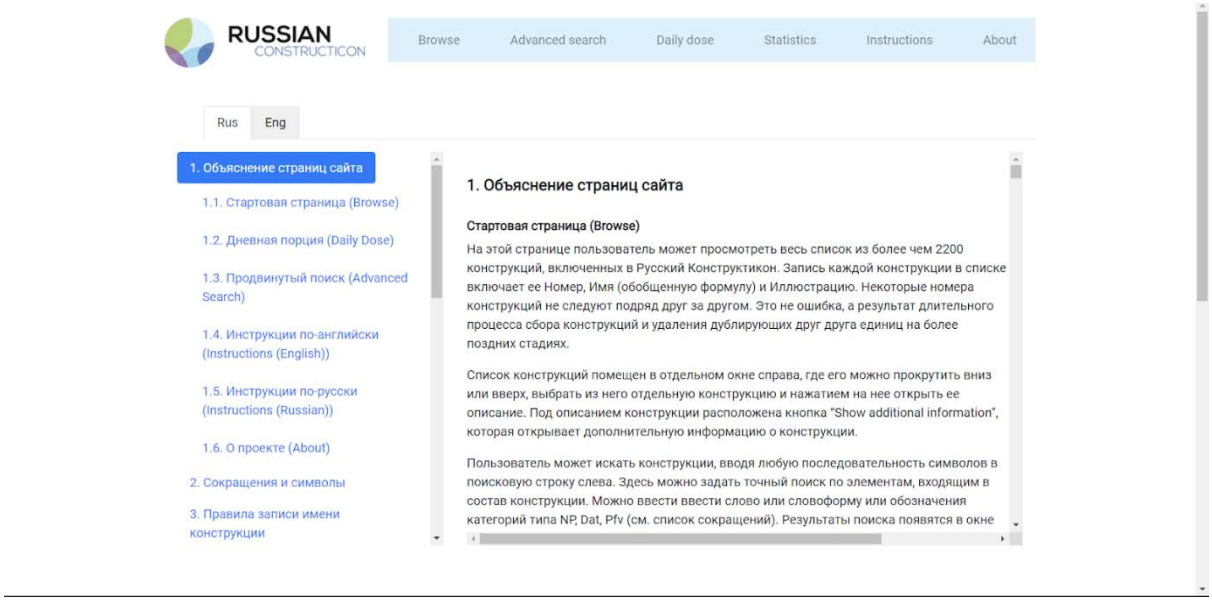


Figure 2: The instructions page of Russian Constructicon after the update

4 Instead of conclusion: New Perspectives for the Russian Constructicon

The changes implemented in the second version of RusCxn not only made it more attractive and convenient to use for both researchers and L2 learners, but also opened new prospects for improved integration with the constructicons for other languages and with the satellite resources of the Russian Constructicon.

First, we shall discuss how improvements in RusCxn 2.0 may facilitate typological research in Construction Grammar and improve connectivity between constructicons for different languages. The talks about somehow 'aligning' different databases with constructions to foster typological research in the field have been around since the first major conference on constructicography [19], as few cross-linguistic studies of constructions that existed at that time proved to be rather useful for both theoretical and applied linguistics [20]. Yet, up to date, there still exists no device or platform that could facilitate typological study of constructions from different languages. We reckon that such a platform should be based on a universal system of glossing, and we are happy to be the pioneers in this field. Even though currently RusCxn remains the only fully glossed resource of its kind, our team is actively working on Hill-Mari, Persian and Ukrainian constructicons, which all have the same architecture and, thus, will be easily mappable to each other, provided they also have glosses. We hope that researcher teams that work on constructicons for other languages will also join our endeavor, so that we can create a big typologically oriented platform for conducting constructicography studies at a fundamentally new level.

Besides that, we shall talk about improved cross-connectivity between RusCxn and other resources targeted at Russian constructicography, such as Constructesize![15], Pragmaticon[16], and Diachronicon (in development). These platforms include much data directly connectable to the constructions from RusCxn: exercises on constructions for L2 learners, diachronically related discourse formulas, and history and origins of some Russian constructions respectively. Nevertheless, in RusCxn 1.0 we were unable to easily map these data because of the lack of unique URLS for our constructions. Now, with a new website architecture, we can conduct studies involving these platforms more easily.

All in all, the Russian Constructicon has been in development for over eight years. In this article we present a second iteration of the resources, enlarged and updated. In the future we shall continue working on the project to remain on the cutting edge of constructicography with the largest and (possibly) the greatest constructicon ever made.

This work was partially carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15- 2020-793: “Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research”.

Funding

This work was partially carried out within the framework of the grant from the Ministry of Science and Higher Education of the Russian Federation within Agreement No. 075-15-2020-793: “Next-generation computational linguistics platform for the Russian language digital recording: infrastructure, resources, research”.

References

[1] Goldberg, A. E. (2006), Constructions at Work: The Nature of Generalizations in Language, Oxford University Press, Oxford.
[2] Fillmore, Ch. J., Kay, P., O’Connor, M. C. (1988), Regularity and idiomaticity in grammatical constructions: The case of let alone, Language, Vol. 64(3), pp. 501–538.
[3] Croft, W. (2001), Radical Construction Grammar, Oxford University Press, Oxford.
[4] Rakhilina, E. V. (ed.) (2010), Linguistics of constructions [Lingvistika konstrukcij], Azbukovnik, Moscow.
[5] Szabó, Z. G. Compositionality // The Stanford Encyclopedia of Philosophy (Fall 2022 Edition).
[6] Ehrlemark, A., Johansson, R., Lyngfelt, B. (2016), Retrieving Occurrences of Grammatical Constructions, Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, Osaka, Japan, pp. 815–824.
[7] Endresen A. A., Zhukova V. A., Mordashova D. D., Rakhilina E. V., Lyashevskaya O. N. (2020), The Russian Constructicon: A New Linguistic Resource, Its Design and Key Characteristics [Russkij Konstruktikon: novyj lingvisticheskij resurs, ego ustrojstvo i specifika] // Proceedings of the International Conference “Dialog 2020” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2020”, Moscow. Pp. 241-255.
[8] Janda, L. A., Lyashevskaya, O., Nesset, T., Rakhilina, E., Tyers, F. M. (2018), A constructicon for Russian: Filling in the Gaps, Lyngfelt, B., Borin, L., Ohara, K., Torrent, T. T. (eds.), Constructicography: Constructicon development across languages, John Benjamins, Amsterdam, pp. 165–181
[9] The Russian National Corpus (ruscorpora.ru). 2003—2023.
[10] Baranov, A. N., Dobrovol'skij, D. O., Kiseleva, K. L., Kozerenko, A. D., Voznesenskaya, M. M., & Korobova, M. M. (2007), Thesaurus dictionary of the Russian idioms [Slovar'-tezaurus sovremennoj russkoj idiomatiki], Avanta+, Moscow.
[11] Rakhilina E. V., Zhukova V. A., Demidova D. A., Kudryavceva P. S., Rozovskaya G. P., Endresen A. A., Janda L. A. (2022), Phraseology in the Perspective of Russian Constructicon [Frazeologiya v rakurse «Russkogo konstruktikona»] Vinogradov Institute for Russian language open series, Vol. 2 (32). Pp. 13-44.
[12] Leipzig Glossing Rules (2008), Conventions for interlinear morpheme-by-morpheme glosses. URL: https://www.eva.mpg.de/lingua/resources/glossing-rules.
[13] Demidova D., Makarova P. (forth.) Glossing Russian Constructicon. Arctic University of Norway open series, in prep.
[14] Vasmer, M. (1953), Russisches etymologisches Wörterbuch/1 A-K.
[15] Endresen, A., Zhukova V., Lonshakov G., Demidova D., Kalanova N., Bjørgve E., Lavén D. H., Janda L. A., Butenko Z., Perevoshchikova T. (2022), Con-struxercise! Hands-on learning of Russian constructions. A digital educational resource. https://constructicon.github.io/construxercise-rus/
[16] Buzanov, A., Bychkova, P., Molchanova, A., Postnikova, A., & Ryzhova, D. (2022). Multilingual Pragmaticon: Database of Discourse Formulae. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3331-3336).
[17] Bychkova, P., & Rakhilina, E. (forth.). Towards pragmatic construction typology: The case of discourse formulae // A. Barotto & S. Mattiola (Eds.), Discourse phenomena in typological perspective. John Benjamins Publishing Company, in press.
[18] Endresen, A., & Janda, L. A. (2020), Taking construction grammar one step further: Families, clusters, and networks of evaluative constructions in Russian. Frontiers in Psychology, 11.
[19] Lyngfelt, B., Torrent, T. T., Laviola, A., Bäckström, L., Hannesdóttir, A. H., & da Silva Matos, E. E. (2018), Aligning constructicons across languages: A trilingual comparison between English, Swedish, and Brazilian Portuguese. // Constructicography. John Benjamins, New York. pp. 255-302.
[20] Boas, H. C. (2010). Contrastive studies in construction grammar, John Benjamins, New York.

Linguistic Annotation Generation with ChatGPT: a Synthetic Dataset of Speech Functions for Discourse Annotation of Casual Conversations

Lidiia Ostyakova[♡,◇] ostyakova.ln@gmail.com
Kseniia Petukhova[◇] petukhova.ka@mipt.ru
Veronika Smilga [◇] smilgaveronika@gmail.com
Dilyara Zharikova[◇] dilyara.rimovna@gmail.com

[♡]HSE University
[◇]Moscow Institute of Physics and Technology

Abstract

This paper is devoted to examining the hierarchical and multilayered taxonomy of Speech Functions, encompassing pragmatics, turn-taking, feedback, and topic switching in open-domain conversations. To evaluate the distinctiveness of closely related pragmatic classes, we conducted comparative analyses involving both expert annotators and crowdsourcing workers. We then carried out classification experiments on a manually annotated dataset and a synthetic dataset generated using ChatGPT. We looked into the viability of using ChatGPT to produce data for such complex topics as discourse. Our findings contribute to the field of prompt engineering techniques for linguistic annotation in large language models, offering valuable insights for the development of more sophisticated dialogue systems.

Keywords: speech functions, ChatGPT, dialogue systems, discourse analysis, open-domain conversations
DOI: 10.28995/2075-7182-2023-22-386-403

Генерация лингвистических данных с помощью ChatGPT: создание синтетического корпуса речевых функций для разметки дискурса в диалогах на повседневные темы

Лидия Остякова[♡,◇] ostyakova.ln@gmail.com
Ксения Петухова[◇] petukhova.ka@mipt.ru
Вероника Смилга[◇] smilgaveronika@gmail.com
Диляра Жарикова[◇] dilyara.rimovna@gmail.com

[♡]Национальный исследовательский университет «Высшая школа экономики»
[◇]Московский физико-технический институт

Аннотация

Эта статья посвящена изучению иерархической и многоуровневой таксономии речевых функций. Чтобы оценить специфику близких прагматических классов, мы провели сравнительный анализ с участием как экспертов-аннотаторов, так и разметчиков краудсорсинга. Затем мы провели эксперименты по классификации аннотированного вручную набора данных и синтетического набора данных, сгенерированного с помощью ChatGPT. Мы рассмотрели возможность использования ChatGPT для получения данных для такой сложной сферы лингвистики, как дискурс. Данная работа вносит вклад в область лингвистической разметки данных.

Ключевые слова: речевые функции, ChatGPT, диалоговые системы, дискурс, общетематические диалоги

1 Introduction

The development of large language models (LLMs) such as ChatGPT, InstructGPT (Ouyang et al., 2022), DialoGPT (Zhang et al., 2019), GPT-3 (Brown et al., 2020), and others has contributed to the rapid expansion of Conversational AI. LLMs are often implemented in dialogue systems to generate replies to the user's utterances by using various prompt engineering techniques to elicit the required behaviour of the model. Incorporating LLMs makes conversational agents more adaptable, versatile, and simple to build. However, generative models need to be controlled within conversations with real users since they usually lack consistency, reliability, and common sense. Therefore, developers of conversational agents face a new challenge in light of the limitations of LLMs: the development of efficient methods to manage a dialogue flow.

Automatic discourse analysis is one of the most prominent ways of managing the dialogue flow in such systems because we can analyse and predict the structure of interconnected linguistic features: a topic, a speaker change, semantics, and pragmatics. For example, (Gu et al., 2021) present DialogBERT shifting the focus from utterance- to discourse-level in response generation. There are several fundamental theories for discourse analysis, such as Dialogue Act (DA) theory (Jurafsky et al., 1998), Segmented Discourse Representation Theory (SDRT) (Lascarides and Asher, 2007) , and Rhetorical Structure Theory (RST) (Mann and Thompson, 1987). Despite numerous applications to real-world problems, there is no standard approach to analysing discourse structures, particularly within open-domain dialogue systems. Despite the fact that discourse analysis is mostly oriented on pragmatics, tagsets usually reflect not pragmatic but grammar features of utterances (e.g., yes/no question, statement).

In this research paper, we focus on an alternative tagset developed by S.Eggins and D.Slade (Eggins and Slade, 2004) and explore its potential for use in dialogue systems. The taxonomy of speech functions is hierarchical and multilayered, including not only pragmatics but also turn-taking, feedback, and topic switching. Because the scheme includes classes with close pragmatics, we conducted additional research to determine whether it is possible to differentiate them for experts and crowdsourcing workers. Furthermore, we performed classification experiments on a manually annotated dataset as well as a synthetic dataset generated using ChatGPT. As a result, this paper contributes to the study of LLMs' prompt engineering techniques for linguistic annotation.

2 Discourse Analysis with Speech Function Theory

To get an idea of the structure of the dialogue and better manage the flow of the conversation, researchers often use an analysis of discourse structures. Such an analysis is used to represent dialogues at different linguistic levels, with a focus on pragmatics, i.e. functions of utterances or intentions of speakers. There are two common approaches to the research of discourse structures in the dialogues: Dialogue Act Theory (DA) (Jurafsky et al., 1998) and Segmented Discourse Representation theory (SDRT) (Lascarides and Asher, 2007). Within DA theory, each elementary discourse unit (EDU) is given a pragmatic characteristic, whereas SDRT, which is based on Rhetorical Structure Theory (Mann and Thompson, 1987), asserts a certain pragmatic class to a relation between two EDUs. The theory of dialogue acts is easier to apply to real-world problems since the task is carried out in one stage, unlike the SDRT approach, in which first the connections between statements must be determined and then only the connections are classified as discourse relations. For instance, a tagset of MIDAS, one of interpretations of DA theory, was used to select suitable replies in the Gunrock 2.0 chatbot, one of the participants in the Amazon Alexa Prize competition (Liang et al., 2020).

A number of tagsets were developed within DA theory and have gained prominence: DAMSL or Dialogue Act Markup in Several Layers (Core and Allen, 1997), Switchboard - DAMSL (Jurafsky, 1997), Meeting Recorder (Shriberg et al., 2004), and MIDAS (Yu and Yu, 2019). Interpretations differ in terms of discourse units, dialogue domains, and a number of described levels that results in inconsistent data (Table 1) although they usually have the same tags for general categories of utterances: statement, yes/no question, positive answer, negative answer.

Following SDTR, researches use one tagset in different task that inherits features of Rhetorical Structure Theory applied for text analysis. There are 16 labels for describing connections between utterances:

Clarification question, Comment, Question-answer pair etc (Li et al., 2020). However, existing datasets with such an annotation are task-oriented so they can not be used for analysis of casual conversations (see Table 1).

Theory	Dataset	Number of Utterances	Number of Labels	Domain
DA theory	SWITCHBOARD	205 000	60	open
	MRDA	180 000	54	open
SDRT theory	MOLWENI	88 303	16	technologies
	STAC	2 500	16	games

Table 1: Comparing of the most popular datasets with discourse annotation

Due to the lack of consistent conversational data with annotations that are good for open-domain dialogue systems, we decided to look into the potential of another taxonomy with classes similar to dialogue acts but with more functional dimensions for discourse analysis. It is important to mention that the theory of speech functions not only includes more complicated pragmatic categories than other taxonomies but also other layers of linguistic annotation that compound complicated discourse patterns united by a particular topic.

2.1 Speech Function Theory

(Eggins and Slade, 2004) developed a taxonomy of speech functions for discourse analysis of casual conversations extending M.K. Halliday's ideas about defining speakers' goals in dialogues. Speech functions combines features of DA theory and RST that reflects in connecting various layers of annotation in the system of dialogue turns and cross-dialogue discourse structure patterns (see Figure 1). Tagset developed by S.Eggins and D.Slade consists of speech functions representing different dimensions: Turn Management, Discourse Structure, Topic Organisation, Feedback (see Figure 1), Communicative Act, or Pragmatic Purpose.

Mostly, EDUs are defined by the functionality of dialogue acts within a particular theory used for discourse analysis. (Bunt et al., 2017) highlights the importance of defining EDUs by DA functions and even names units as functional segments. The speech function taxonomy differs from other approaches in terms of dialogue segmentation on EDUs as classes have more than one function. The taxonomy is divided into two levels of segmentation. The level of topics defines discourse patterns within conversations, while all speech functions are assigned at the sentence level. However, not all utterances are divided just into sentences; some of them are combined based on their common function or divided into several segments in other cases.

There are three high-level types of **discourse moves** in the taxonomy:

- Opening moves
- Sustaining moves
- Moves of Reaction

The purpose of **Opening moves** is to introduce new topics or start a conversation. According to S.Eggins and D.Slade, each Opening move indicates not only a new topic or the beginning of interaction between interlocutors within a conversation but also a **discourse pattern** (Eggins and Slade, 2004). **Sustaining moves** do not contribute to topic development but provide additional details and clarifications about the current topic given by the same speaker. They enhance the information discussed within it, while the speaker's role remains unchanged. **Moves of Reaction** are turns in dialogue where a speaker changes or responds to the previous utterance of the interlocutor that have more layers than the others. They are divided into two groups of speech functions representing different approaches to topic development. The React.Respond speech functions finish the conversation by not adding new challenges (e.g., questions changing conversational flow). React.Rejoinder, however, promotes discussion (see Appendix A).

Such a multilayered structure appears to be difficult to comprehend, especially given the uneven distribution of dimensions in tags. However, such complex dialogue modelling allows for the description of a conversational structure at various levels while taking into account topic shifts, discourse patterns, and

abstract intentions. The speech function annotation scheme, in contrast to other DA, SRDT taxonomies, has grammatical criteria for tag identification but does not include them in the tags. Besides that, speech functions feature a more subtle division into pragmatic classes comparing to other theories. For instance, most existing schemes for discourse analysis use the tag 'positive answer' for all cases when a speaker provides a yes-answer, while speech function theory distinguishes whether a speaker agrees with the provided information, acknowledges it, or affirms something.

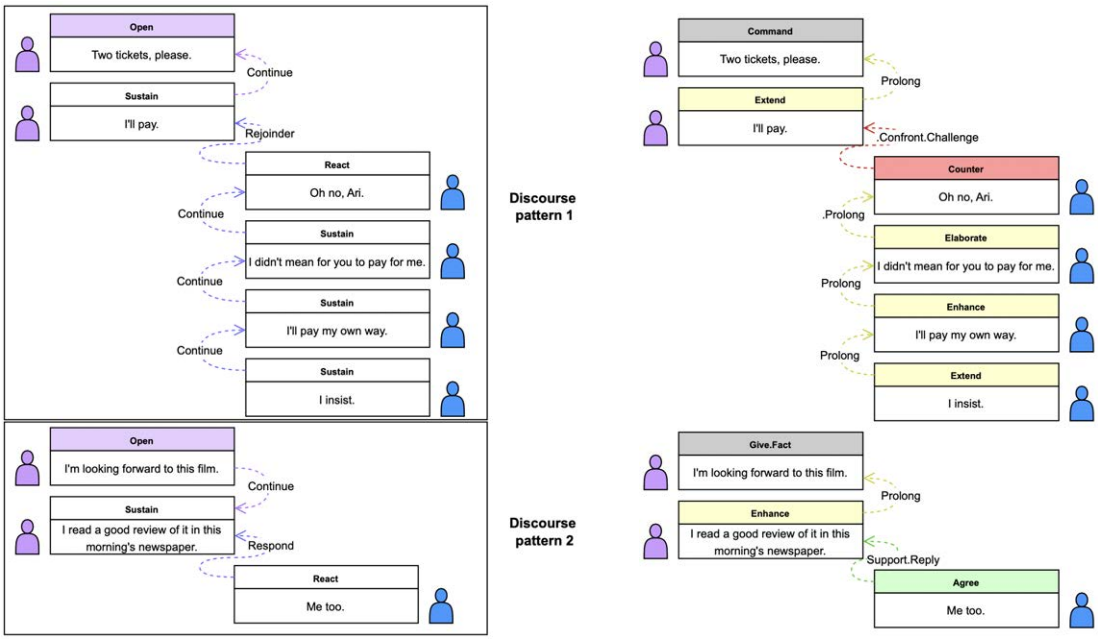


Figure 1: Discourse Patterns (left) and Feedback in Speech Functions (right)

3 Speech Function Dataset

Based on the classification developed in the framework of Speech Function Theory, we aim to obtain a dataset of open-domain dialogues with complex discourse annotation. The multidimensionality of the annotation scheme will allow to use the results in a variety of NLP tasks, especially those related to automatic discourse analysis.

As the basis for our speech function annotated dataset, we select DailyDialog, a dataset of human-written multi-turn dialogues on a variety of topics, widely used in evaluating open-domain dialogue systems. We preprocess DailyDialog data, removing duplicate dialogues and segmenting the remaining ones to split each utterance into several discourse units. To do so, we use a model for sentence segmentation that splits long and complex utterances into sentences and recovers punctuation.

As the first step of annotation, we employed three expert linguists to gather a small gold standard corpus with professionally annotated utterances. The resulting corpus consists of 75 dialogues (1264 utterances) annotated by three experts. We implemented an approach of double annotation with adjudication on our data, as it is commonly used for labelling discourse structures (Prasad et al., 2008; Webber et al., 2016; Zhou and Xue, 2015). We divided the dialogues into three equal parts, each annotated by two annotators independently. In cases of disagreement, the third expert not involved in annotating a particular part was responsible for adjudication and decided on final labels. The next step of the annotation process is crowd-sourcing annotation with the use of Toloka [1] crowdsourcing platform (Pavlichenko et al., 2021).

[1] https://toloka.ai/tolokers/

3.1 Inter-annotator Agreement: Experts vs. Crowdsourcing

(Mattar and Wachsmuth, 2012) implemented speech function annotation in a task-oriented dialogue system to aid in controlling a dialogue flow that demonstrated the possible potential of using the taxonomy for analyzing discourse structures. However, to work on automatic analysis using speech functions in open-domain dialogue systems, it was necessary to prove that the chosen taxonomy is reliable enough. So, we conducted several experiments on the annotation of casual conversations in English.

Annotation of discourse structures or dialogue acts is not trivial because it requires linguistic knowledge or trained workers (Yung et al., 2019). Besides that, perception of speakers' intentions in utterances differs across individuals, making the task even more difficult. We compared two results of annotation with speech functions completed by experts with professional backgrounds in linguistics and crowdsourced workers. We used Fleiss' Kappa (Fleiss and Cohen, 1973) for measuring inter-annotation agreement as it is considered to be the most common way to evaluate taxonomy reliability in tasks related to discourse analysis. However, this evaluation method has the limitation of not considering the common mistakes of annotators. That is why we measured not only inter-annotator agreement but also accuracy, weighted recall, and precision, as well as macro and micro F1 (Ghamrawi and McCallum, 2005), by comparing workers' annotations to results by experts.

Crowdsourcing is not the best option for labelling data with discourse structures since it is not possible to obtain high-quality annotations with linguistic labels from untrained workers (Kawahara et al., 2014). Nevertheless, it is important to test to what extent classes can be defined by non-professionals. For obtaining better results by crowdsourcing workers, we developed hierarchical guidelines consisting of easy questions about a topic and speaker change, the type of a sentence, the pragmatics of the utterance, and examples that allow better orientation in the scheme for untrained annotators (see Appendix B). In addition, extra methods for controlling the quality of annotation were devised to help us identify unreliable annotators, and some hints were included for crowdsourcing workers.

As a result of crowdsourcing, 675 utterances were cross-annotated by three non-professional workers each. It is important to note that crowdsourcing workers were different in each case that could also cause inconsistency. We evaluated the results for 16 high-level cut labels and the complete taxonomy to determine the weak points of the established hierarchical guidelines. Cut labels group the classes that are really close to each other in terms of pragmatics into one class (see Appendix B). When measuring the quality of crowdsourced annotation, we also examined cases of voting where not all annotators but the majority agree on a tag (see Table 2). As for cut labels, they were labeled with pretty good accuracy by crowdsourcing workers. Annotation of full tags is more challenging for non-experts, which is proven by all metrics. Macro F1 value shows that we have to pay attention to improving quality of annotating low-level classes (see Table 2). Measuring inter-annotator agreement using Fleiss' Kappa proves that the tags with close pragmatics are difficult for differentiating not only for untrained workers, but for experts as well. Still, in case of experts' annotation, Fleiss' kappa is more than 0.6, meaning that the chosen taxonomy is quite reliable (see Figure 5).

To sum up, crowdsourcing is a very consuming process in terms of time and resources, especially for such complicated annotation tasks related to linguistic data augmentation. Furthermore, this method of enlarging labeled data is not so effective as values of accuracy metrics and Fleiss' kappa have shown. The data labeled by crowdsourcing workers needs to be corrected by experts, which slows down and complicates the annotation process. That is why our next experiments on data augmentation were conducted using large language models.

3.2 Generating a Synthetic Speech Functions Dataset with ChatGPT

Data augmentation is a technique widely used in machine learning to increase the size of the training data. It can be especially useful when dealing with limited or imbalanced data, improving generalization and preventing overfitting. (Wei and Zou, 2019) describes a set of simple data augmentation methods that significantly improve the performance of models such as recurrent neural networks (RNNs) and convolutional neural networks (CNNs) on text classification tasks. In (Kobayashi, 2018), the authors

	Accuracy	Weighted Recall	Weighted Precision	Macro F1	Micro F1
Full tags	0.52	0.52	0.62	0.37	0.55
Full tags + voting	0.54	0.54	0.62	0.37	0.54
Cut labels	0.83	0.83	0.83	0.53	0.83
Cut labels + voting	0.87	0.87	0.85	0.53	0.87

Table 2: Evaluation of annotation by crowdsourcing workers

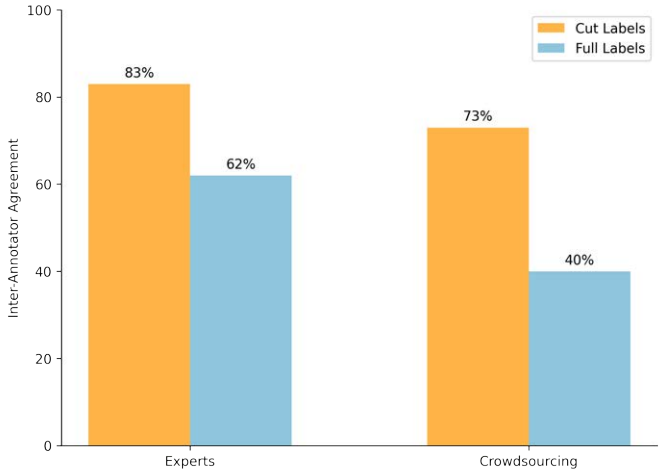


Figure 2: Inter-annotator Agreement

pretrain an LSTM on Wikipedia articles and fine-tune it on several labelled datasets to generate more sentences from training data by using the fine-tuned model to replace some words. Again, the proposed method improved the RNNs and CNNs performance on text classification tasks. In (Xie et al., 2020), the authors explore various advanced methods of data augmentation for language and vision tasks. On IMDb text classification dataset, their model trained on only 20 labelled examples mixed with augmented data outperforms the original state-of-the-art model trained on approximately 25000 labelled examples. Finally, (Kumar et al., 2020) describes how pre-trained text generation models like BART, BERT and GPT-2 can be used to generate augmented text data.

As we are now in the beginning of the process of building a speech functions dataset and lack annotated data, we decided to test whether we could effectively use data augmentation methods to build a decently performing classification model. In addition to that, any speech function dataset is by its nature imbalanced, as some speech functions are seen many times more rarely in conversations than the others, which would also make data augmentation methods effective. ChatGPT is a pretrained generative text model which was fine-tuned using reinforcement learning with human-feedback data. As reported in (OpenAI, 2022) and (Ouyang et al., 2022), InstructGPT and its sibling model ChatGPT perform particularly well when given instructions in natural language. Following (Kumar et al., 2020) and (Kobayashi, 2018) who use language models for textual data augmentation, we decided to use ChatGPT to generate synthetic data for our speech functions dataset.

The model was accessed via OpenAI API [2] and provided with hand-crafted instructions for each speech function class. We tried to implement different strategies in order to get more suitable, natural and various conversational data for particular classes:

- to make the model follow instructions developed for crowdsourcing and label the whole dialogue;

[2] `https://platform.openai.com/overview`

- to give instructions only with description of classes;
- to give just examples of classes;
- to give examples of one speech function;
- to give examples of several similar classes.

We had a lot of challenges putting the above-mentioned data generation strategies into action because of ChatGPT's limitations. The model overuses certain phrases that interfere with generating various conversational data. Even mentioning a change of topic and word collocations in prompts does not always lead to the variety of results needed. The instability of generative models does not allow to generate similar data with the same instructions. So, working on data augmentation, we had to control such cases of unstable generation and remove them from the data. As we were working with linguistic annotation, the model interpreted some labels differently than they were given in the instruction.

Considering all experiments, the final instruction included 1) the speech function name; 2) the speech function definition; 3) examples from the expert-annotated Gold Standard dataset; 4) guidelines for the model, i.e. "Generate 20 datapoints from these examples" (see Appendix C for a prompt example). We generated from 500 to 1000 datapoints for each class, approximately 25000 speech function examples in total (see AppendixD). We also generated examples to train a separate classification model to distinguish between declarative, interrogative, and miscellaneous (that includes emotional exclamations, greetings, goodbyes, etc.) classes.

4 Classification

We developed a multi-level annotation pipeline (see Figure 3) to annotate dialogues with Speech Functions. Firstly, a Topic Shift Classifier is applied to determine if an utterance initiates a new topic. Subsequently, an Upper Level Classifier annotates all utterances by identifying the type of the utterance. If the utterance is interrogative, the question classifier is then used to obtain the final label. If the utterance is declarative or miscellaneous, the Declarative Classifier or Miscellaneous Classifier is used, respectively. For utterances that were defined as commands, the final label is also 'COMMAND'. Definitions and examples of all final labels can be found in Table 5 of the Appendix.

The DeepPavlov library (Burtsev et al., 2018) was used to train classifiers for our project. For the Topic Shift Classifier, we trained double sequence binary classifier model based on `roberta-large-mnli`, where the input was a sequence of two consecutive utterances. The true label denotes the topic shift in the utterances. The model was trained with the following hyper-parameters: learning rate – 2e-5, optimizer – AdamW, input max length – 128. We applied the early-stopping to successfully train the model. Using pre-trained model allowed the classifier to transfer knowledge gained while pre-training on mnli to related task of shift identification (Konovalov et al., 2020; Gulyaev et al., 2020).

Similarly, for our remaining classifiers, we utilized double sequence classification based on `bert-base-cased` multi-class classification.

Table 3 shows the evaluation results on the test set of ChatGPT data. Table 4 displays the evaluation results for real data, i.e., dialogues that were manually annotated by the experts.

Classifier	**Accuracy**
Topic Shift	0.86
Upper Level	0.99
Questions	0.97
Declarative	0.94
Miscellaneous	0.99

Table 3: Evaluation results on ChatGPT data

Overall, it is evident that the accuracy of all classifiers, except the Topic Shift Classifier, is significantly lower on real data. The low level of classification quality for declarative and interrogative utterances can be explained by two main reasons. Firstly, distinguishing between Speech Functions within interrogative and declarative classes is challenging, even for humans, as shown in Table 2. Secondly, the data samples

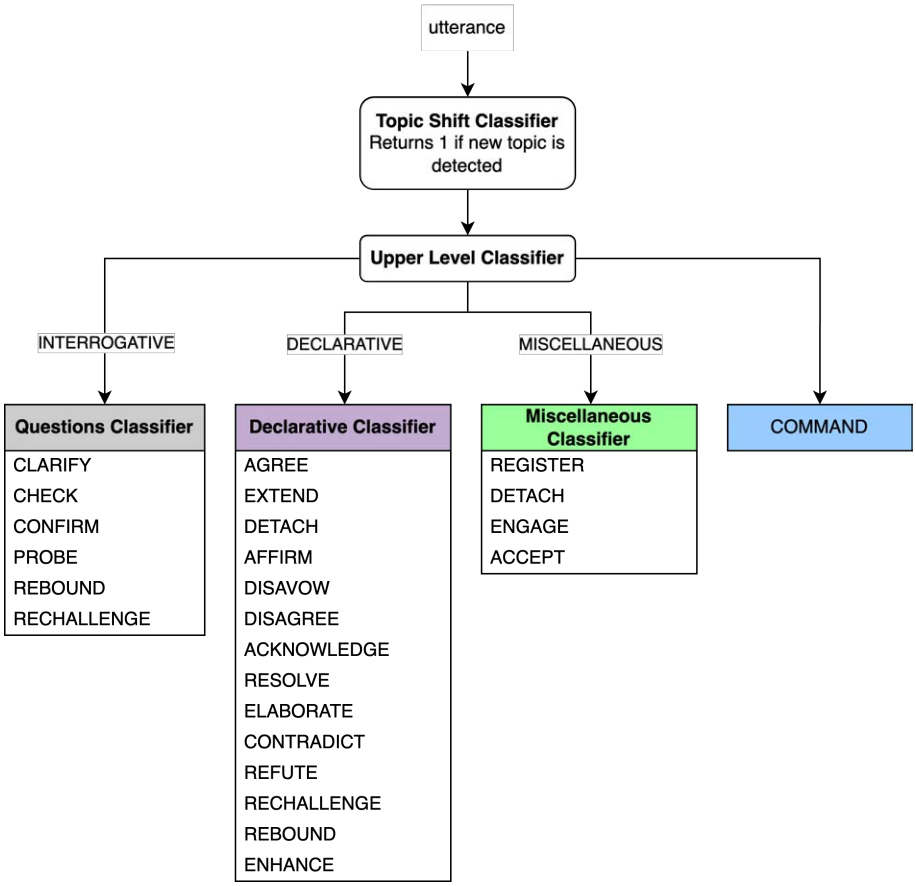


Figure 3: Annotation pipeline

Classifier	Accuracy	Weighted Recall	Weighted Precision	Weighted F1
Topic Shift	0.91	0.91	0.96	0.93
Upper Level	0.60	0.60	0.87	0.71
Questions	0.34	0.34	0.83	0.43
Declarative	0.20	0.20	0.31	0.24
Miscellaneous	0.81	0.81	0.87	0.84
Random Topic Shift	0.53	0.53	0.57	0.55
Random Upper Level	0.25	0.25	0.55	0.34
Random Questions	0.20	0.20	0.61	0.27
Random Declarative	0.09	0.09	0.20	0.15
Random Miscellaneous	0.23	0.23	0.55	0.33

Table 4: Evaluation results on real dialogues

generated with ChatGPT are very similar within classes. Although different prompts and examples were used during the generation process, samples are syntactically and semantically alike. Consequently, the model learned to differentiate between highly specific and similar samples of Speech Functions, while real conversations are much more unpredictable and varied, making it harder for the model to accurately classify them. Thus, for prompt illustrated in Figure 4, ChatGPT generated several similar examples on cuisine topic.

Here are some of them:

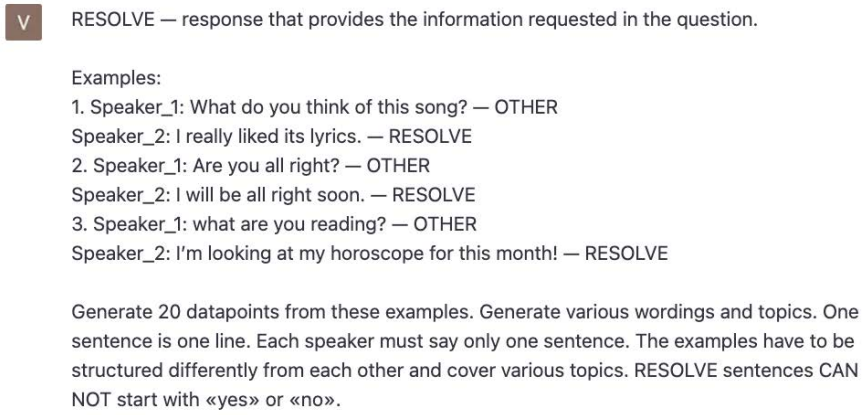


Figure 4: Prompt for generation of RESOLVE samples

- Speaker_1: What's your favorite type of cuisine? — OTHER
  Speaker_2: I love Mexican food, especially tacos! — RESOLVE
- Speaker_1: What's your favorite food? — OTHER
  Speaker_2: I love sushi and could eat it every day! — RESOLVE
- Speaker_1: What's your favorite type of cuisine? — OTHER
  Speaker_2: I love Japanese food, especially sushi and ramen. — RESOLVE

5 Conclusion and Future Work

This paper gives a thorough look at research done on a new way to analyze discourse in open-domain dialogue systems. Speech function theory sees the discourse structure of dialogues as a complex hierarchical system that connects linguistic levels and functional dimensions like taking turns, changing topics, pragmatics, and the interlocutor's feedback. Of particular research interest was the fact that low-levels of speech functions all reflect pragmatics, not semantics, as in many popular taxonomies. We checked the reliability of the taxonomy and did experiments on labelling dialogues on casual topics from the DailyDialog dataset, comparing inter-annotator agreement between experts with backgrounds in linguistics and untrained crowdsourcing workers. Considering the results of experts' annotation, it was proven that the scheme for annotation is reliable enough but still difficult because of close classes in terms of pragmatics.

In our study, we employed ChatGPT to generate synthetic data for our speech functions dataset as the human-labelled dataset is imbalanced which makes training a classifier more difficult. While exploring ChatGPT's capabilities, we found several strategies to create suitable conversational data for each speech function class. We encountered several challenges due to the nature of language models, such as overuse of certain phrases and instability in generation. However, by refining our instructions and incorporating expert-annotated examples from the Gold Standard dataset, we managed to generate 27,000 datapoints. Based on the generated data, we trained a custom multi-level annotation pipeline. The pipeline includes a Topic Shift Classifier, an Upper Level Classifier, a Question Classifier, a Declarative Classifier, and a Miscellaneous Classifier. The results show that the accuracy of the classifiers is significantly lower on real data, which can be attributed to the challenges of distinguishing between Speech Functions within interrogative and declarative classes and the limited variability of the data generated by ChatGPT.

Our next steps will involve running experiments on classification with ChatGPT because we could not achieve satisfactory results for speech function classification using data generation as an augmentation method. As LLMs pre-trained on instructions are becoming more popular instruments for data augmentation, implementing other models for labelling or generation may be beneficial to our research. In order to improve metrics for this classification task, we also intend to try training or fine-tuning other Transformer models on the annotated dialogues.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Harry Bunt, Volha Petukhova, and Alex Chengyu Fang. 2017. Revisiting the iso standard for dialogue act annotation. // *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)*.

Mikhail Burtsev, Alexander Seliverstov, Rafael Airapetyan, Mikhail Arkhipov, Dilyara Baymurzina, Nickolay Bushkov, Olga Gureenkova, Taras Khakhulin, Yuri Kuratov, Denis Kuznetsov, and Vasily Konovalov. 2018. Deeppavlov: Open-source library for dialogue systems. // *NIPS*.

Mark G Core and James Allen. 1997. Coding dialogs with the damsl annotation scheme. // *AAAI fall symposium on communicative action in humans and machines*, volume 56, P 28–35. Boston, MA.

Suzanne Eggins and Diana Slade. 2004. *Analysing casual conversation*. Equinox Publishing Ltd.

Joseph L Fleiss and Jacob Cohen. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement*, 33(3):613–619.

Nadia Ghamrawi and Andrew McCallum. 2005. Collective multi-label classification. // *Proceedings of the 14th ACM international conference on Information and knowledge management*, P 195–200.

Xiaodong Gu, Kang Min Yoo, and Jung-Woo Ha. 2021. Dialogbert: Discourse-aware response generation via learning to recover and rank utterances. // *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, P 12911–12919.

Pavel Gulyaev, Eugenia Elistratova, Vasily Konovalov, Yuri Kuratov, Leonid Pugachev, and Mikhail Burtsev. 2020. Goal-oriented multi-task bert-based dialogue state tracker.

Dan Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. // *Discourse Relations and Discourse Markers*.

Dan Jurafsky. 1997. Switchboard swbd-damsl shallow-discourse-function annotation coders manual. *www. dcs. shef. ac. uk/nlp/amities/files/bib/ics-tr-97-02. pdf*.

Daisuke Kawahara, Yuichiro Machida, Tomohide Shibata, Sadao Kurohashi, Hayato Kobayashi, and Manabu Sassano. 2014. Rapid development of a corpus with discourse annotations using two-stage crowdsourcing. // *Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*, P 269–278.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Vasily Konovalov, Pavel Gulyaev, Alexey Sorokin, Yury Kuratov, and Mikhail Burtsev. 2020. Exploring the bert cross-lingual transfer for reading comprehension. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologiithis*, P 445–453.

Varun Kumar, Ashutosh Choudhary, and Eunah Cho. 2020. Data augmentation using pre-trained transformer models. *arXiv preprint arXiv:2003.02245*.

Alex Lascarides and Nicholas Asher. 2007. Segmented discourse representation theory: Dynamic semantics with discourse structure. *Computing meaning*, P 87–124.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Kaihui Liang, Austin Chau, Yu Li, Xueyuan Lu, Dian Yu, Mingyang Zhou, Ishan Jain, Sam Davidson, Josh Arnold, Minh Nguyen, et al. 2020. Gunrock 2.0: A user adaptive social conversational system. *arXiv preprint arXiv:2011.08906*.

William C Mann and Sandra A Thompson. 1987. *Rhetorical structure theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Nikita Mattar and Ipke Wachsmuth. 2012. Small talk is more than chit-chat: Exploiting structures of casual conversations for a virtual agent. *// KI 2012: Advances in Artificial Intelligence: 35th Annual German Conference on AI, Saarbrücken, Germany, September 24-27, 2012. Proceedings 35*, P 119–130. Springer.

OpenAI. 2022. Introducing chatgpt. *OpenAI Blog*. Accessed: 2023-03-17.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Nikita Pavlichenko, Ivan Stelmakh, and Dmitry Ustalov. 2021. Crowdspeech and voxdiy: Benchmark datasets for crowdsourced audio transcription. *arXiv preprint arXiv:2107.01091*.

Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K Joshi, and Bonnie L Webber. 2008. The penn discourse treebank 2.0. *// LREC*.

Elizabeth Shriberg, Raj Dhillon, Sonali Bhagat, Jeremy Ang, and Hannah Carvey. 2004. The icsi meeting recorder dialog act (mrda) corpus. Technical report, INTERNATIONAL COMPUTER SCIENCE INST BERKELEY CA.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2016. A discourse-annotated corpus of conjoined vps. *// Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, P 22–31.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Dian Yu and Zhou Yu. 2019. Midas: A dialog act annotation scheme for open domain human machine spoken conversations. *arXiv preprint arXiv:1908.10023*.

Frances Yung, Vera Demberg, and Merel Scholman. 2019. Crowdsourcing discourse relation annotations by a two-step connective insertion task. *// Proceedings of the 13th Linguistic Annotation Workshop*, P 16–25.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Yuping Zhou and Nianwen Xue. 2015. The chinese discourse treebank: A chinese corpus annotated with discourse relations. *Language Resources and Evaluation*, 49:397–431.

A Speech Functions list

Speech Function	Definition
Open.Attend	These are usually greetings. NB: Used in the beginning of a conversation. Example: Hi!
Open.Demand	Demanding information. NB: Used in the beginning of a conversation. Example: What's Allenby doing these days?
Open.Give	Providing information. NB: Used in the beginning of a conversation. Example: I met his sister.
Open.Command	Making a request, an invitation or command to start a dialogue or discussion of a new topic. Example: Let's go for a walk!
Sustain.Continue.Prolong.**Extend**	Adding supplementary or contradictory information to the previous statement. A declarative sentence or phrase (may include and, but, except, on the other hand). Example: Just making sure you don't miss the boat. **I put it out on Monday mornings. I hear them. I hate trucks.**
Sustain.Continue.Prolong.**Elaborate**	Clarifying / rephrasing the previous statement or giving examples to it. A declarative sentence or phrase (may include for example, I mean, like). Example: Yeah but I don't like people... um... **I don't want to be INVOLVED with people.**
Sustain.Continue.Prolong.**Enhance**	Adding details to the previous statement, adding information about time, place, reason, etc. A declarative sentence or phrase (may include then, so, because). Example: Nor for much longer. **We're too messy for him.**
Sustain.Continue.**Monitor**	Checking the involvement of the listener or trying to pass on the role of speaker to them. Example: You met his sister that night we were doing the cutting and pasting up. **Do you remember?**
React.Rejoinder.Confront.Response.**Re-challenge**	Offering an alternative position, often an interrogative sentence. Example: David: Messi is the best. Nick: **Maybe Pele is the best one?**
React.Rejoinder.Support.Challenge.**Rebound**	Questioning the relevance, reliability of the previous statement, most often an interrogative sentence. Example: David: This conversation needs Allenby. Fay: **Oh he's in London.** So what can we do?
React.Rejoinder.Support.Response.**Resolve**	The response provides the information requested in the question. Example: Lina: What do you think of this song? Fay: **I really like its lyrics.**
React.Rejoinder.Support.Track.**Check**	Getting the previous speaker to repeat an element or the entire statement that the speaker has not heard or understood. Example: Straight into the what?
React.Rejoinder.Support.Track.**Clarify**	Asking a question to get additional information on the current topic of the conversation. Requesting to clarify the information already mentioned in the dialogue. Example: What, before bridge?

React.Rejoinder.Support.Track.**Confirm**	Asking for a confirmation of the information received. Example: David: Well, he rang Roman, he rang Roman a week ago. Nick: **Did he?**
React.Rejoinder.Support.Track.**Probe**	Requesting a confirmation of the information necessary to make clear the previous speaker's statement. The speaker themselves speculates about the information that they want to be confirmed. Example: Because Roman lives in Denning Road also?
React.Respond.Confront.Reply.**Contradict**	Refuting previous information. No, sentence with opposite polarity. If the previous sentence is negative, then this sentence is positive, and vice versa. NB! The speaker contradicts the information that he already knew before. Example: Fay: Suppose he gives you a hard time, Nick? Nick: **Oh I like David a lot.**
React.Respond.Confront.Reply.**Disagree**	Negative answer to a question or denial of a statement. No, negative sentence. Example: Fay: David always makes a mess in our room. May: **No, he's not so bad.**
React.Respond.Confront.Reply.**Disavow**	Denial of knowledge or understanding of information. Example: I don't know.
React.Respond.Support.Develop.**Elaborate**	Clarifying / rephrasing the previous statement or giving examples to it. A declarative sentence or phrase (may include for example, I mean, like). Example: Nick: Cause all you'd get is him bloody raving on. Fay: **He's a bridge player, a naughty bridge player.**
React.Respond.Support.Develop.**Enhance**	Adding details to the previous statement, adding information about time, place, reason, etc. A declarative sentence or phrase (may include then, so, because). Example: Fay: He kept telling me that. Nick: **The trouble with Roman though is that — you know he does still like cleaning up.**
React.Respond.Support.Develop.**Extend**	Adding supplementary or contradictory information to the previous statement. A declarative sentence or phrase (may include and, but, except, on the other hand). Extend: David: That's what the cleaner — your cleaner lady cleaned my place thought. Nick: **She won't come back to our place.**
React.Respond.Support.**Engage**	Drawing attention or a response to a greeting. Example: Hey, David.
React.Respond.Support.**Register**	A manifestation of emotions or a display of attention to the interlocutor. Example: Yeah.
React.Respond.Support.Reply.**Acknowledge**	Indicating knowledge or understanding of the information provided. Example: I know.
React.Respond.Support.Reply.**Affirm**	A positive answer to a question or confirmation of the information provided. Yes/its synonyms or affirmation. NB! The speaker confirms the information that he already knew before. Example: Nick: He went to London. Fay: **He did.**

React.Respond.Support.Reply.**Accept**	Expressing gratitude. Example: Thank you!
React.Respond.Support.Reply.**Agree**	Agreement with the information provided. In most cases, the information that the speaker agrees with is new to him. Yes/its synonyms or affirmation. Example: Steve: We're gonna make it. Mike: **Yeah, right.**

Table 5: Speech functions and their communicative roles in the dialogue

B Cut and full Speech Function labels

Cut labels	Full labels
Open.Demand	Open.Demand
Open.Give	Open.Give
Open.Command	Open.Command
Open.Attend	Open.Attend
React.Rejoinder.Confront.Response	React.Rejoinder.Confront.Response.Re-challenge
React.Rejoinder.Support.Track	React.Rejoinder.Support.Track.Probe
	React.Rejoinder.Support.Track.Check
	React.Rejoinder.Support.Track.Clarify
	React.Rejoinder.Support.Track.Confirm
Sustain.Continue. Prolong	Sustain.Continue. Prolong.Extend
	Sustain.Continue. Prolong.Enhance
	Sustain.Continue. Prolong.Elaborate
React.Rejoinder.Support.Challenge.Rebound	React.Rejoinder.Support.Challenge.Rebound
React.Respond.Support.Reply	React.Respond.Support.Reply.Affirm
	React.Respond.Support.Reply.Acknowledge
	React.Respond.Support.Reply.Agree
React.Respond.Support.Develop	React.Respond.Support.Develop.Extend
	React.Respond.Support.Develop.Enhance
	React.Respond.Support.Develop.Elaborate
React.Respond.Confront.Reply	React.Respond.Confront.Reply.Disagree
	React.Respond.Confront.Reply.Contradict
	React.Respond.Confront.Reply.Disavow
Sustain.Continue.Monitor	Sustain.Continue.Monitor
React.Respond.Support.Register	React.Respond.Support.Register
React.Respond.Support.Engage	React.Respond.Support.Engage
React.Respond.Support.Accept	React.Respond.Support.Accept
React.Rejoinder.Support.Response.Resolve	React.Rejoinder.Support.Response.Resolve

C Annotation interface and prompt example

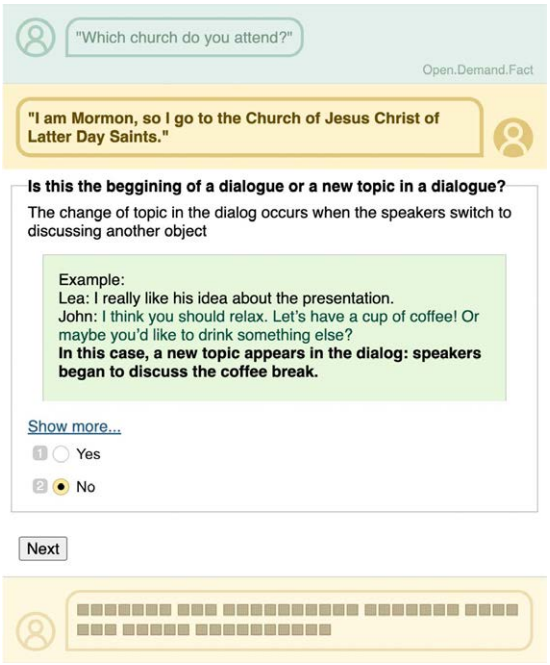


Figure 5: Guidelines for annotators

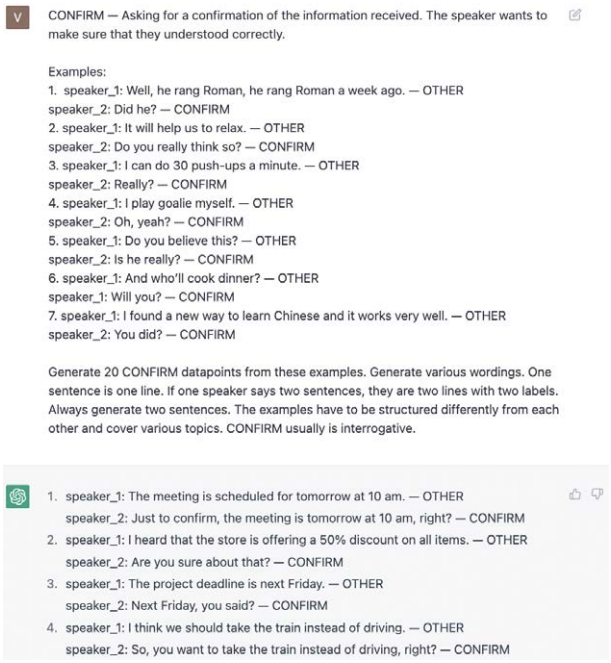


Figure 6: An example of a prompt and generation results

D Dataset statistics

	Class name	Synthetic dataset	Original dataset
Upper level	INTERROGATIVE	790	
	DECLARATIVE	738	
	MISCELLANEOUS	655	
	COMMAND	630	14
Declarative classes	AGREE	980	49
	EXTEND	996	383
	AFFIRM	933	54
	DISAVOW	800	7
	DISAGREE	774	39
	ACKNOWLEDGE	688	9
	RESOLVE	583	103
	ELABORATE	571	91
	CONTRADICT	544	2
	REFUTE	588	-
	RECHALLENGE	530	2
	REBOUND	511	5
	ENHANCE	424	77
Miscellaneous classes	REGISTER	502	78
	DETACH	630	4
	ENGAGE	504	6
	ACCEPT	307	17
Interrogative classes	CLARIFY	564	162
	CHECK	665	14
	CONFIRM	591	23
	PROBE	574	39
	REBOUND	543	5
	RECHALLENGE	509	2

E Metrics

- **The Fleiss' kappa** statistic is used to examine the level of agreement among multiple assessors evaluating a categorical or nominal variable. It is calculated by comparing observed and expected agreement among raters. The range of Fleiss' kappa is 0 to 1 where 1 implies full agreement. A value of 0.6 or more is considered to be a good agreement.

$$Fleiss'kappa = \frac{P_o - P_e}{1 - P_e}$$

  - P_o is the observed agreement among the raters;
  - P_e is the expected agreement by chance, which is calculated based on the marginal frequencies of the categories being rated.
- **Accuracy** measures how accurately a model or classifier predicts the proper outcome or label for a dataset. The model or classifier's accuracy score is the percentage of correct predictions.

$$accuracy = \frac{number\ of\ correct\ predictions}{total\ number\ of\ predictions}$$

- **Weighted Recall and Precision** in classification tasks where the classes are imbalanced. Recall is an evaluation of a model's capacity to identify all relevant instances of a target class. Precision is a measure of a model's ability to identify only instances of a target class that are relevant. In both cases, the weights wi can be adjusted to the proportion of instances in each class or to a value based on class importance.

$$weighted\ precision = \frac{\sum_{i=1}^{N} w_i \cdot TP_i}{\sum_{i=1}^{N} w_i \cdot (TP_i + FP_i)}$$

  - N is the number of classes;
  - w_i is the weight of class i;
  - TP_i is the number of true positives for class i;
  - FP_i is the number of false positives for class i.

$$weighted\ recall = \frac{\sum_{i=1}^{N} w_i \cdot TP_i}{\sum_{i=1}^{N} w_i \cdot (TP_i + FN_i)}$$

  - N is the number of classes;
  - w_i is the weight of class i;
  - TP_i is the number of true positives for class i;
  - FN_i is the number of false negatives for class i.
- **Micro F1** is a dataset-wide F1 score. Precision, recall, and F1 scores are obtained by measuring the total number of true positives, false positives, and false negatives across all classes.

$$F_1^{micro} = \frac{2 \cdot TP_{total}}{2 \cdot TP_{total} + FP_{total} + FN_{total}}$$

  - TP_{total} is the total number of true positives;
  - FP_{total} is the total number of false positives;
  - FN_{total} is the total number of false negatives across all classes.
- **Macro F1** is calculated for each class and averaged. It weights each class equally regardless of dataset frequency.

$$F_1^{macro} = \frac{1}{N} \sum_{i=1}^{N} F_1^i$$

  - N is the number of classes;
  - F_1^i is the F1 score for class i.

Poly-predication in informal monological discourse (according to «What I saw» corpus)

Daria Panysheva
Russian State University for the Humanities, Moscow, Russia
panysheva97@gmail.com

Abstract

The article discusses the relationship between the mode of discourse and quantitative metrics of poly-predication. Based on the material of the corpus "What I Saw", oral and written versions of stories are compared according to the relative frequency of polypredicative constructions and the representation of certain types of polypredication, the features of semantics and grammatical labeling of such structures are described. Using the nonparametric Wilcoxon criterion, the absence of statistical significance between the density of poly-predication in the oral and written parts of the corpus is proved.

Keywords: spoken discourse, poly-predication, Russian language, discourse
DOI: 10.28995/2075-7182-2023-22-404-411

Полипредикация в неформальном монологическом дискурсе по данным корпуса «Что я видел»[1]

Дарья Панышева
РГГУ, Москва, Россия
panysheva97@gmail.com

Аннотация

В статье рассматривается взаимосвязь модуса дискурса и количественных метрик полипредикации. На материале корпуса «Что я видел» сравниваются устные и письменные версии рассказов по относительной частотности полипредикативных конструкций в них и представленности отдельных типов полипредикации, описываются особенности семантики и лексико-грамматического маркирования таких конструкций. С помощью непараметрического критерия Уилкоксона доказывается отсутствие статистической значимости между плотностью полипредикации в устной и письменной частях корпуса.

Ключевые слова: устная речь, полипредикация, русский язык, дискурс

1 Постановка исследовательской задачи

Во многих исследованиях отмечаются качественные и количественные различия в употреблении полипредикации в устном и письменном модусе. В ряде работ, посвящённых устному дискурсу в русском языке, отмечается более низкая плотность полипредикации в устной речи в сравнении с письменной [Сиротинина 1974; Лаптева 1976; Земская и др. 1981].

Основная исследовательская задача работы – проверка влияния канала передачи речи на сложность дискурса. В качестве основного параметра оценки синтаксической сложности текстов в данном случае взята плотность полипредикативных конструкций (ППК), предложенный в работах [Berman 2016; Подлесская 2019]. Помимо этого, в рассмотрение включаются и качественные характеристики: способы маркирования межклаузальной связи, линейный порядок клауз, грамматический класс и грамматическая форма предикатов главной и зависимой клаузы, семантика конструкций.

[1] Исследование выполнено при поддержке гранта РНФ № 22-28-00540

2 Материал исследования

Исследование проведено на материале корпуса «Что я видел», собранного в 2021–2022 гг., который представлен неподготовленными личными рассказами информантов в возрасте от 18 до 40 лет. От каждого информанта были получены рассказ о запомнившемся сновидении и об интересном случае из жизни. Оба сюжета записывались в устной и письменной версии с суточным промежутком. Таким образом, комплект материалов каждого рассказчика состоит из 4 записей. От части информантов во время первой сессии записи требовалось рассказать устные истории, от другой части – письменные.

Поставленная в исследовании задача требует исключить влияние на сложность дискурса каких-либо факторов кроме различия по модусу. Этого позволяют добиться следующие особенности корпуса: жанровая однородность текстов, отсутствие противопоставления рассказов по признаку моно- и диалогичности, формальности и неформальности дискурса.

Из общего объёма корпуса были взяты 27 пар текстов. Общая длина устных версий – 6615 слов, письменных – 3772.

3 Количественный анализ ППК

В рассмотрение включены ППК с грамматически маркированной (с помощью союзных средств и морфологических показателей в составе нефинитных глагольных форм) и семантически выраженной подчинительной связью. Просодические критерии при отборе ППК в рамках данного исследования не учитываются как основные. В связи с этим конструкции с сочинительной связью, для выделения которых в устной речи наличие союзных средств связи не является единственным достаточным признаком, не включаются в анализируемый массив ППК.

В основе классификации ППК лежат параметры, приведённые в [Gast 2012]: тип отношений между главной и зависимой клаузой, морфосинтаксические свойства зависимой клаузы, тип вершины, присоединяющей зависимую клаузу (именная или глагольная).

ППК, найденные в корпусе, таким образом были разделены на сентенциальные актанты (СА), сирконстанты (СС), определения (СО). Также в номенклатуру типов ППК включены конструкции с прямой речью как близкие к СА, однако обладающие меньшей степенью интегрированности зависимой и главной клаузы.

В абсолютных величинах общее количество ППК в устном и письменном подкорпусе составило 393 и 216 конструкций соответственно. Так как объём устной части корпуса, измеренный в словах (без учёта заполненных пауз, неречевых вокальных явлений и оборванных единиц), почти в 2 раза превышает объём письменной части, полученные числа были приведены к относительным величинам. Была измерена доля ППК в количестве конструкций на 100 слов текста в каждом отдельном рассказе и общая доля ППК на 100 слов каждого подкорпуса в целом. Таким же образом была вычислена относительная плотность каждого рассматриваемого типа полипредикации отдельно.

Различия в полученных количественных соотношениях были проверены на наличие статистической значимости с помощью непараметрического критерия Уилкоксона для парных выборок. Результаты приведены в таблице 1. Так как p-value для каждого параметра ниже критического значения 0,05, выборки можно считать однородными.

	плотнотность в устном корпусе	**плотность в письменном корпусе**	**критериальное значение W**	**p-value**
СА	3,36	3,29	174,0	0,731
СО	0,70	0,87	11,0	0,672
СС	1,12	1,27	127,0	0,346
прямая речь	0,77	0,29	25,0	0,519
общее количество ППК	**5,94**	**5,73**	**160,0**	**0,499**

Таблица 1: количество ППК в корпусе

4 Виды ППК и их особенности

Далее мы проверили, есть ли связанные с модусом дискурса особенности выбора говорящими средств связи, финитной или нефинитной формы зависимого предиката, линейной позиции клауз в рамках каждого отдельного вида ППК. Также были рассмотрены семантические различия конструкций.

4.1 Сентенциальные актанты

В эту группу отнесены финитные придаточные клаузы и зависимые инфинитивные обороты, заполняющие активную синтаксическую и семантическую валентность вершины

Так как семантика предиката главной клаузы во многом определяет требования к грамматической форме зависимой клаузы, количественные соотношения семантических классов вершины финитных и инфинитивных СА рассматриваются в отдельных подгруппах.

Распределение семантики вершинного предиката в конструкциях с СА в устной и письменной части корпуса относительно однородно. Некоторые различия есть в семантике конструкций с зависимым инфинитивом. Доля фазовых глаголов среди всех вершин, управляющих зависимым инфинитивом, в устном подкорпусе несколько выше, чем в письменном (33,7% в сравнении с 20,4%). Попарное сравнение устных и письменных версий историй показывает, что говорящие склонны в некоторых случаях по-разному описывать постепенно разворачивающиеся во времени процессы: глаголами совершенного вида в письменных текстах (1) и фазовым глаголом с инфинитивом в устных (2).

(1) DS_c07_dream-wr

Я упал, покалечился, дело было на одном из верхних этажей. Я покатился по склону из бетона.

(2) DS_c07_dream-sp

E004 /\упал,

p004 (0.11)

N004 (q 0.37)

E005 и начал /скользить по-о (q 1.05) отвесному-у (q 0.37) получается –склону,

Выбор средств связи СА с главной клаузой в основном одинаков и в устной, и в письменной речи. Единственным уникальным для устного дискурса маркером оказался нерасчленённый союз «то что», однако несмотря на общую тенденцию всё чаще встречаться в устной неформальной речи [Кибрик, Подлесская 2009], в корпусе «Что я видел» он не оказался употребительным (6,6% конструкций с сентенциальным актантом) и встретился в историях только одного рассказчика.

С точки зрения грамматического класса вершины, управляющей СА, интересны случаи, когда в вершине находится именная группа. Эти случаи немногочислены и в устной, и в письменной части корпуса, однако в устных текстах семантика именных вершин кажется более вариативной (3), чем в письменных, где именные вершины при финитных актантах в основном представлены словом «сон».

(3) DS_c04_story_sp

E015 и-и (0.39) закрыли это всё / ↓ диагнозом,

E016 что это неопознанная {sf 0.37} какая-то кластерная \боль.

4.2 Сентенциальные определения

В эту категорию входят относительные придаточные клаузы и нефинитные определения, выраженные причастиями.

В употребительности маркеров связи и положении зависимой клаузы относительно вершины различий между письменным и устным корпусом выявлено не было.

Причастная стратегия релятивизации оказалась менее предпочтительной, чем использование финитного относительного придаточного, как в устном, так и в письменном подкорпусе. Тем не менее, есть количественные различия, связанные с модусом: в устной части корпуса на долю причастных оборотов пришлось 8,5% от всех СО, в письменной – 20,6%. Также в пользу большей характерности нефинитных СО для письменного модуса говорит то, что такие конструкции встречаются в 6 письменных текстах и только в 2 устных, при этом записанных после письменной сессии. Можно предположить, что причастные обороты воспроизводятся «по памяти» при пересказе сюжета в устной версии (4, 5).

(4) DS_c05_story-wr

Мы переглянулись и побежали, я никогда не бегал так быстро, кажется я «подрезал» несколько замедляющихся перед светофором машин.

(5) DS_c05_story-sp

E022 Перебегая дорогу в неположенном /месте,

E023 кажется я подрезал (0.17) пару (0.16) тормозящих на светофоре (ɥ 0.57) /машин,

Приведённые результаты отличаются от соотношений, полученных в исследовании на материале НКРЯ, где в Основном подкорпусе 68,1% случаев причастной релятивизации подлежащего, а в Устном подкорпусе – 35,6% [Сай 2014]. Вероятно, выбор грамматической формы СО коррелирует не только с модусом, но и с другими параметрами дискурса: жанром текста, степенью формальности и др.

4.3 Сентенциальные сирконстанты

Эта группа представлена финитными клаузами с подчинительными союзами и нефинитными обстоятельствами (деепричастиями, целевыми инфинитивами). По семантике СС разделены на придаточные со значением времени, причины, уступки, цели, условия и сравнения.

Количественное соотношение этих семантических типов отличается в устной и письменной части корпуса. В письменном подкорпусе есть однозначное преобладание придаточных со значением временных отношений (61,4% от всех СС), в устном они составляют 27,9%. Возможно, это связано с тем, что в устном дискурсе говорящие чаще выражают отношения следования или предшествования во времени через линейное построение финитных конструкций.

Ещё одно различие состоит в доле причинных придаточных от всех СС. В устном корпусе таких придаточных 36,4%, в письменном значительно меньше – 11,4%. Есть несколько предположений о том, с чем связано такое распределение. При этом интересно, что только в устных текстах встречаются случаи иллокутивного или эпистемического употребления этих СС (6). В примере смысл высказывания можно представить так: «я поняла, что у меня сонный паралич, потому что не могла встать».

(6) DS_c02_dream-sp

E062 (ʔ 0.26) и-и (ʔ 0.33) у меня в тот момент был сонный – ↑ паралич,

E063 потому что я не могла ни –⟶встать,

E064 ни-и –кричать,

E065 ни /–говори-ить,,,

Неоднородное количественное распределение придаточных со значением причины может быть связано с несколькими факторами: 1) наличие слушателя при устном модусе повествования даже в монологическом режиме предполагает необходимость установления контакта, из-за чего появляется потребность сделать содержание истории понятным для слушателя и излагать события не исключительно в линейной последовательности; 2) экономия ресурса в режиме письменного повествования – стремление сделать рассказ компактным (почти в два раза меньший объём корпуса письменных текстов может служить свидетельством этой тенденции).

Сравним, как передаётся одинаковое содержание в устной (7) и письменной версии (8) одной истории. То, что описывается в устной записи через многоклаузальный комплекс с союзным маркированием связи частей, в письменной записи представлено линейной последовательностью предикаций.

(7) DS_c08_dream-sp

E013 \чувствовалось,

E014 что-о было какое-то очень раннее такое /утро,

E015 потому что {gp 0.82} в лесу ещё была такая (ɥ 0.48) свежесть от /росы,,,

(8) DS_c08_dream-wr

Мы вышли к полю, в воздухе пахло прохладой утренних трав, еще был виден туман, застилавший поле.

При сопоставлении финитной и нефинитной стратегий маркирования межклаузальной связи в группе СС было выявлено существенно преобладание доли деепричастий в письменном корпусе по сравнению с устным[2]. В письменных текстах 25% сирконстантов вводятся с помощью деепричастий, а в устных – 5,4%, причём эти вхождения, как и в случае с причастными оборотами, представлены текстами, которые вторичны по отношению к письменным. Также в примерах из устного корпуса можно заметить относительно продолжительные паузы, сопровождающие деепричастную конструкцию (9). Можно предположить, что эти паузы связаны с трудностями при порождении нефинитных обстоятельств, однако это предположение нуждается в проверке на большем объёме материала.

(9) DS_c08_dream-sp

E019 я периодически смотрела на какие-то \/травы,

E020 которые росли в \ ↑ по-оле,

N009 (ɥ 0.34)

E021 вспоминая /то,

p007 (0.52)

E022 как мы с /бабушкой часто ходили сюда собирать (ə 0.40) /зверобой,

В порядке расположения главной и зависимой клауз и в выборе союзных показателей синтаксической связи значимых различий между устным и письменным подкорпусом найдено не было.

[2] Мы благодарим анонимного рецензента за указание на то, что низкая частотность (дее)причастных оборотов может иметь, в числе прочих, и диахроническую причину - как категория, заимствованная из церковнославянского

4.4 Конструкции с прямой речью

В эту группу вошли все ППК, в которых при передаче чужой речи происходит индексальный сдвиг, в том числе конструкции со смешанным типом цитирования, где присутствует союзный маркер, характерный для косвенной речи (10).

(10) DS_c08_dream-sp

E021 и-и /разговаривает со мной таким /\споко-ойным разме-еренным /→го-олосом де-етским,,,

E022 что типа «/–ма-ам@,,,

E023 не /–пережива-ай,,,

E024 у меня вообще всё /–норма-ально...»

Несмотря на отсутствие статистически значимых различий в доле конструкций с прямой речью на 100 слов каждого подкорпуса, приведём ряд наблюдений, свидетельствующих, что в выборе между прямым и косвенным цитированием или цитированием и другими способами передачи содержания говорящие проявляют разные тенденции в зависимости от модуса речи. При попарном сравнении текстов выявляются примеры, где прямая речь используется в устном тексте, а в письменном тексте: а) тот же смысл передаётся косвенной речью, б) содержание чужой речи передаётся без ссылки на источник, в) выбирается другой синтаксический способ выражения содержания – например, глагол с зависимым инфинитивом с объектным контролем, г) вовсе отсутствует содержательно эквивалентный фрагмент.

(11) DS_c09_story-sp

E043 И водитель ко мне /подходит,

E044 \спрашивает:

E045 «\Девушка@,

E046 а-а вам-м \помощь /нужна?

N017 (ɥ 0.99)

E047 Что-то /случилось?»

(12) DS_c09_story-wr

Водитель выходит и спрашивает, нужна ли помощь.

(13) DS_c01_story-sp

E083 (ə 0.41) на что мне говорят «\А!,

E084 \ой!,

E085 –блин,

E086 я \забыла,

E087 я еду || я еду на /→дачу.»,,,

(14) DS_c01_story-wr

В день встречи выясняется, что она забыла, построила другие планы

(15) DS_c15_story-sp

E030 и –говорит:

N008 (q 0.24)

E031 «/\Ну-у,

E032 /\выбери любой /нож,

E033 (который тебе \понравится,)

p009 (0.14)

E034 и с этим /ножом (0.15) /я пойду /вас (0.18) \убивать.»

(16) DS_c15_story-wr

…выходит из квартиры мужик с набором кухонных ножей и предлагает моей подруге выбрать нож, которым он нас зарежет.

5 Заключение

Несмотря на распространённое предположение о меньшей синтаксической сложности устного дискурса в сравнении с письменным, проведённый количественный и статистический анализ показал, что, по крайней мере в рассматриваемом корпусе, плотность полипредикации не зависит от канала речепорождения. Так как этот параметр является важным, но не единственным критерием оценки синтаксической сложности, результаты исследования не позволяют напрямую сделать вывод о полной идентичности синтаксической сложности двух модусов, но, тем не менее, показывают, что устный дискурс проявляет не меньшую способность к порождению полипредикативных комлексов, чем письменный. Тем не менее, стратегии употребления ППК в разных модусах дискурса нельзя назвать одинаковыми. Материал корпуса «Что я видел» даёт основания считать, что различия присутствуют сразу по нескольким параметрам.

Отличается употребительность финитных и нефинитных клауз: деепричастные и причастные формы зависимого предиката более характерны для письменной речи, чем для устной. Однако, было обнаружено, что, несмотря на разницу в их количестве между двумя выборками, внутри каждого отдельного подкорпуса нефинитные стратегии одинаково не претендуют на позицию самого частотного способа оформления зависимой клаузы.

Описанные в разделах выше различия в распределении семантических типов ППК демонстрируют, что при необходимости передать похожее пропозициональное содержание говорящие могут предпочитать разные способы его синтаксического оформления в устной и письменной речи (например, выбирая между причинным обстоятельством и линейным соположением клауз).

Безусловно, многие частные особенности (например, более характерное для устной речи использование союзов со значением причины в иллокутивном значении) нуждаются в повторной проверке на большей по объёму выборке и не позволяют на данный момент распространять эти наблюдения на устный дискурс русского языка в целом. Однако, на данном корпусном материале можно сделать вывод, что письменный и устный дискурс отличаются не по плотности полипредикации, а по её качественным характеристикам.

References

[1] Berman R. A. Linguistic literacy and later language development // Written and Spoken Language Development across the Lifespan: Essays in Honor of Liliana Tolchinsky / Eds. Perera J., Aparici M., Rosado E., Salas N. New York, 2016. P. 181–200.
[2] Gast, Volker and Diessel, Holger. Clause Linkage in Cross-Linguistic Perspective: Data-Driven Approaches to Cross-Clausal Syntax, Berlin, Boston: De Gruyter Mouton, 2012.
[3] Kibrik A.A. Podlesskaya V.I. Stories about dreams. Corpora research of the Russian spoken discourse [Rasskazy o snovideniakh. Korpusnoye issledovanie ustnogo russkogo diskursa]. Moscow: Yazyki slavyanskikh kultur, 2009.
[4] Lapteva O.A. 1976. Russian spoken syntax [Russkiy razgovornuy sintaksis]. Moscow: Nauka
[5] Podlesskaya V.I., Ozolina V.A. Polypredication in Japanese oral narrative discourse comparing eith the written discourse: experience of the syntactic complexity corpora reseach [Polipredication v japonskom ustom narrativnom diskurse v sravnenii s pismennym: oput korpusnogo issledovaniya sintaksicheskoj slozhnosti]. Uralic-altai researches [Uralsko-altajskiye issledovaniya]. 2019, № 2 (33), 83-100.
[6] Saj S. Verbal [Prichastiye]. Materials for the corpora description project of the Russian grammar [Materialy dlya projekta korpusnogo opisaniya russkoi grammatiki] (http://rusgram.ru). On the rights of the manuscript. M. 2014.
[7] Sirotina O.B. 1974. Modern spoken discourse and its features [Sovremennaya razgovornaya rech' i ejo osobemmosti]. Moscow: Nauka.
[8] Zemskaya E.A., Kitajgorodskaya M.V., Shiryaev E.N. 1981. Russian spoken language. Common is-sues. Derivation. Syntax [Russkaya razgovornaya rech'. Obtshiye voprosy. Slovoobrazovaniye. Sintaksis]. Moscow: Nauka

References

[1] Berman R. A. Linguistic literacy and later language development // Written and Spoken Language Development across the Lifespan: Essays in Honor of Liliana Tolchinsky / Eds. Perera J., Aparici M., Rosado E., Salas N. New York, 2016. P. 181–200.
[2] Gast, Volker and Diessel, Holger. Clause Linkage in Cross-Linguistic Perspective: Data-Driven Approaches to Cross-Clausal Syntax, Berlin, Boston: De Gruyter Mouton, 2012.
[3] Земская Е.А., Китайгородская М. В., Ширяев Е. Н. 1981. Русская разговорная речь. Общие вопросы. Словообразование. Синтаксис. М.: Наука.
[4] Кибрик А.А., Подлесская В.И. Рассказы о сновидениях. Корпусное исследование устного русского дискурса. М.: Языки славянских культур, 2009.
[5] Лаптева О.А. 1976. Русский разговорный синтаксис. М.: Наука.
[6] Подлесская В. И., Озолина В. А. Полипредикация в японском устном нарративном дискурсе в сравнении с письменным: опыт корпусного исследования синтаксической сложности. Урало-алтайские исследования. 2019, № 2 (33), 83–100.
[7] Сай С. Причастие. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М. 2014.
[8] Сиротинина О.Б. 1974. Современная разговорная речь и ее особенности. М.: Наука.

Russian additive markers *takže* and *tože*: a synchronic and diachronic perspective

Olga E. Pekelis
Russian State University for the Humanities/Moscow, Russia
HSE University/ Moscow, Russia
opekelis@gmail.com

Abstract

It is well known that Russian additive markers *takže* and *tože* differ in terms of information structure: the scope of *takže* is focus, while the scope of *tože* is topic. Based on data of several corpora of Russian, this paper shows that in modern Russian, *takže* and *tože* are opposed on other language levels as well, namely syntactically (in terms of word order), lexically (a variant of *takže* that is synonymous with *tože* including at the level of the information structure, is going out of use), stylistically and as far as their involvement in grammaticalization processes is concerned (*takže* but not *tože* developed into a coordinate conjunction and a discourse marker). However, as evidenced by Russian National Corpus data, most of these contrasts were absent or less pronounced in the Russian language of the 18th-19th centuries. Thus, in the last two centuries *takže* and *tože* evolved toward their consistent differentiation.

Keywords: additive markers, Russian language of the 19th century, information structure, corpus studies
DOI: 10.28995/2075-7182-2023-22-412-420

Также и *тоже* в синхронии и диахронии

Пекелис О. Е.
Российский государственный гуманитарный университет, Москва, Россия
Национальный исследовательский университет
«Высшая школа экономики», Москва, Россия
opekelis@gmail.com

Аннотация

Известно, что аддитивные показатели *также* и *тоже* различаются коммуникативной сферой действия: *также* относится к реме, *тоже* – к теме. В настоящей статье с опорой на данные нескольких корпусов демонстрируется, что в современном языке *также* и *тоже* противопоставлены не только на уровне коммуникативной структуры, но и на других уровнях языка: синтаксически (в терминах порядка слов), лексически (из употребления уходит ударное *также*, синонимичное *тоже* с точностью до коммуникативной структуры), стилистически и по степени участия в процессах грамматикализации (*также*, но не *тоже*, стало источником грамматикализации для сочинительного союза и дискурсивного маркера). При этом, по данным Национального корпуса русского языка, в языке XVIII-XIX вв. эти различия в большинстве своем отсутствовали или были менее выражены. Тем самым, эволюция *также* и *тоже* в последние два века состояла в их последовательной дифференциации.

Ключевые слова: аддитивные маркеры, русский язык XIX века, коммуникативная структура, корпусные исследования

1 Введение

Слова *также* и *тоже* неоднократно исследовались [4], [5], [11], [15], [16] и др. Принято считать, что различие между ними лежит в области коммуникативной структуры. В [16: 314] это различие формулируется в терминах темы и ремы. Существенно упрощая, его суть можно описать так: и *тоже*, и *также* указывают на то, что утверждение истинно не только для объекта или признака, выражаемого составляющей, к которой *также/тоже* относится, но и для некоторого другого

объекта или признака того же рода; при этом *тоже* относится к теме, а *также* – к реме. Так, в (1) *тоже* вносит смысл ‘мы следим не только за исправностью машин, но и за чем-то еще, аналогичным’ (трезвостью водителей, как ясно из контекста), где *за исправностями машин* – тема.[1] В (2) *также* указывает на то, что власти рекомендуют использовать перчатки наряду с чем-то еще (с масками), и *перчатки* входит в состав ремы. В типологической литературе подобные показатели, выражающие включение некоторого элемента в множество подобных ему по некоторому признаку, называются аддитивными (см., например, [14: 33]). [2]

(1) *Если мы узнаем, что водитель был нетрезвым на работе, то его сразу увольняют. За исправностями машин тоже следим.* [Комсомольская правда, 2009.08]

(2) *Ношение масок <…> останется обязательным. Власти также рекомендуют использовать перчатки.* [Парламентская газета, 2020.05]

В [1] для описания *тоже* и *также* вместо понятий темы и ремы используется оппозиция данного и нового: *тоже* маркирует предшествующую ему часть высказывания как новое (ср. *исправности машин* в (1)), а последующую (*следим*) – как данное. *Также*, наоборот, маркирует предшествующую ему часть высказывания как данное (ср. *власти* в (2)), а последующую – как новое (*перчатки*).[3]

Поскольку отличие *тоже* от *также* – коммуникативное, а не семантическое, замена *тоже* на *также* и наоборот обычно допустима при изменении порядка слов, ср. (1) с (3) и (2) с (4):

(3) *Мы также следим за исправностями машин.*

(4) *Использовать перчатки власти тоже рекомендуют.*

До сих пор речь шла о *также*, коммуникативно противопоставленном *тоже*. Однако существуют и такие употребления *также*, в которых *также* синонимично *тоже* вплоть до коммуникативной структуры. Согласно [16: 312-313], два *также* различаются просодически: *также*, противопоставленное *тоже*, безударно (ср. (2)), тогда как *также*, синонимичное *тоже*, является носителем коммуникативно релевантного акцента (добавим – нисходящего акцента, ассоциируемого с ремой [22: 36-37]). Ср. (5), где *также* заменимо на *тоже* и относится к теме и новому (*отопление*):

(5) *<…> существенно сэкономить на плате за воду можно, установив счетчики. За отопление москвичи также* \ (≈*тоже*) *переплачивают.* [Ведомости, 2008.10]

В настоящей работе предпринята попытка проследить эволюцию *тоже* и *также* начиная с XVIII в. Данные Национального корпуса русского языка (НКРЯ) свидетельствуют о том, что в XVIII-XIX вв. соотношение между *тоже* и *также* отличалось от современного (раздел 2). Последнее определилось, по-видимому, лишь в XX в. Мы стремимся уточнить характер этого развития, опираясь на количественные (раздел 3), семантические и синтаксические наблюдения (раздел 4), и соотнести этапы эволюции *тоже* и *также* с более общими представлениями об аддитивных маркерах (раздел 5). Исследование продолжает проект Школы лингвистики НИУ ВШЭ, посвященный русскому языку XIX в. [18].

2 *Тоже* и *также* в XVIII-XIX вв.

В подкорпусе XVIII в. в составе НКРЯ слово *тоже* в аддитивном значении встречается редко. Среди 320 примеров с *тоже* (против 4486 с *также*) бóльшая часть соответствует современному сочетанию местоимения *тот* с частицей *же* (как в *то же время, то же сделала* и под.). Это дает основания предположить, что в XVIII в. аддитивное *тоже* было относительно новым словом. Косвенно это подтверждает [6: 389-390], указывая в качестве первых двух значений для *тоже*

[1] Здесь и далее, если не сказано иное, примеры с указанием источника заимствованы в Национальном корпусе русского языка.

[2] В русской традиции *также* и *тоже* в рассматриваемых здесь значениях считаются наречиями [9]. Но, поскольку нас интересует диахронический и, отчасти, типологический ракурс, мы абстрагируемся от частеречных характеристик *тоже* и *также*, называя их нейтрально – аддитивными показателями, или аддитивными маркерами (ср., например, [10]).

[3] Независимо от конкретного подхода, при описании значения *тоже* и *также* встает вопрос об определении того объекта или признака, с которым сопоставляется объект или признак, вводимый *тоже* или *также*. В [15], [16] этот вопрос решается на основе понятия ассоциативных связей. В круг задач настоящего исследования разработка этого вопроса не входит.

временные ‘тогда’ и ‘потом’, а близкое к аддитивному третье значение ‘также’ иллюстрируя примерами XVII в.

Основное отличие *тоже* и *также* в языке XVIII-XIX вв. от современных *тоже* и *также* состоит в том, что описанного коммуникативного противопоставления между ними не было – во всяком случае, оно не соблюдалось последовательно. На это указывает тот факт, что не только *также* могло употребляться как синоним современного *тоже* (как отмечено во Введении, такое встречается и сегодня), но и *тоже* могло использоваться в значении современного *также*. Ср. (6-8), где *тоже* относится к составляющей, которая расположена правее и входит в состав ремы и нового.

(6) *Тоже и в нынешних веках многие грады и страны от того раззорились.* [А. И. Богданов. Описание Санктпетербурга (1751)]

(7) *Очень хорошо тоже заняться всеобщею политикою, включая туда и естественное право.* [Н. И. Тургенев. Письма <…> (1811)]

(8) *Слышу тоже, что Ольга Сергеевна разъехалась с Павлищевым*. [А. Н. Вульф. Дневник (1830)]

Употребление *тоже* в условиях, свойственных современному *также*, окказионально встречается и в текстах XXв. Основного корпуса НКРЯ:

(9) *Говорили тоже, что старый хозяин был не то старовер, не то сектант.* [В. Ф. Кормер. Наследство (1987)]

Однако в Газетном корпусе – содержащем тексты СМИ конца XX-го – начала XXIв., которые, можно думать, меньше подвержены стилизации и отражают современную норму точнее, чем тексты художественной литературы, преобладающие в Основном корпусе – аналогичные примеры обнаружить не удалось. В как будто похожем на (6-8) примере (10) использование *тоже* можно объяснить отклонением от современного стандарта в порядке слов, а не в коммуникативной структуре, ср. более естественный порядок *и прокуратура, и следствие тоже ходатайствуют.* В (7) и (8) такая интерпретация маловероятна.

(10) *Обычно в этих случаях об изменении меры пресечения тоже и прокуратура, и следствие ходатайствуют. Крайне редко случается, когда только обвиняемый и его защита выступают за изменение меры пресечения* [Известия, 2018.03]

В (11), как в (7-8), *тоже* как будто относится к реме и новому (ср. *на вашу поддержку*). Однако такое употребление, по-видимому, оправдано тем, что антецедент местоимения *вашу* – именная группа *192 страны* – располагается левее *тоже* и входит в тему:

(11) *Вот 192 страны, члены организации федерации Красного Креста и Красного Полумесяца, надеемся тоже и на вашу поддержку, с тем чтобы обеспечивать продвижение нашей вакцины*. [Ведомости, 2021.12]

Существенно также, что, в отличие от приведенных примеров XVIII-XIX вв., (10), (11) и подобные окказионализмы в Газетном корпусе представляют собой прямую речь, которой, очевидно, небрежность свойственна больше, чем письменной речи.

3 *Тоже* и *также*: количественные данные

В настоящем разделе представлены данные о сопоставительной частотности *тоже* и *также* в трех подкорпусах НКРЯ (подкорпусе XVIII-XIXвв., XXв. и в Газетном корпусе) и двух веб-корпусах: Araneum Russicum Maius и Генеральном интернет-корпусе русского языка (ГИКРЯ), а именно, в его подкорпусах «Живой Журнал» (ЖЖ), «ВКонтакте» (ВК) и «Новости». Отдельно мы оценивали частотность среди всех вхождений *также* сочетания *а также*, к роли которого мы вернемся ниже. Как видно из таб. 1, подкорпусы Основного корпуса НКРЯ, а также ЖЖ и ВК в составе ГИКРЯ вместе демонстрируют рост доли *тоже* после XIX в. Обратим внимание, что частотность аддитивного *тоже* в подкорпусе XVIII-XIX вв. в действительности ниже, чем указано в таб.1, поскольку, как отмечено в предыдущем разделе, многие вхождения *тоже* в текстах этого периода не соответствуют аддитивному маркеру. Однако и та частотность, которая представлена в таблице, составляет значимую разницу с частотностью *тоже* в XX в. по данным

НКРЯ (χ^2, p<0,01). Между тем, данные остальных корпусов как будто противоречат гипотезе об экспансии *тоже* – в них более частотен *также*.[4]

	также	*а также*	доля *а также*	*тоже*	доля *тоже*
1701-1900 НКРЯ	39 721	4 247	0,1	41 843	0,5
1901-2000 НКРЯ	72 951	25 512	0,3	138 231	0,7
Газетный НКРЯ	1 237 530	535 444	0,4	241 796	**0,2**
Araneum	1 687 212	756 303	0,4	470 690	**0,2**
ГИКРЯ (ВК)	3 501 531	1 248 727	0,4	4 110 203	0,5
ГИКРЯ (ЖЖ)	3 778 597	1 448 095	0,4	7 760 196	0,7
ГИКРЯ (Новости)	1 740 184	662 887	0,4	100 000	**0,1**

Таблица 1: Частотность *тоже* и *также* по корпусным данным[5]

Разгадка этого противоречия, как кажется, кроется в разных стилистических предпочтениях *тоже* и *также*: *тоже* тяготеет к неформальным, а *также*, наоборот, к формальным типам текстов. В публицистических текстах преобладает формальный стиль, отсюда высокая доля *также* в Газетном корпусе и «Новостях» в составе ГИКРЯ. Корпус Araneum, хотя и является интернет-корпусом, содержит разнообразные типы текстов, в том числе публицистику. Напротив, подкорпусы ГИКРЯ ЖЖ и ВК, где более частотен *тоже*, включают тексты блогов и социальных сетей, т.е. преимущественно неформальные.

Предположение о стилистических расхождениях между *тоже* и *также* подтверждают и данные текстов разных жанров в составе Основного корпуса НКРЯ. Как видно из таб. 2, в неформальных текстах преобладает *тоже*, в формальных – *также* (различие статистически значимо, χ^2, p<0,01).

	также	*Тоже*	Всего	доля *также*
неформальные (ОБ+ЭК)	5 741	17 708	23 449	0,2
формальные (ОД+УН)	30 064	5 498	35 562	0,8

Таблица 2: Частотность *также* и *тоже* в обиходно-бытовых (ОБ) текстах, электронной коммуникации (ЭК), официально-деловых (ОД) и учебно-научных (УН) текстах (подкорпус 1950-2020 гг. Основного корпуса НКРЯ)

Таким образом, по сравнению с языком XVIII-XIX вв. сегодня можно предполагать экспансию *тоже* в текстах неформальных жанров. В формальных текстах сегодня преобладает *также*; существенно, однако, что по крайней мере отчасти распространение *также* можно объяснять не теми употреблениями, в которых *также* синонимичен *тоже*, а такими

[4]С данными в таб. 1 связана следующая проблема: поскольку речь идет о большом объеме данных, ручная фильтрация не могла быть проведена и не были отсеяны случаи неправильного (слитного) написания *то же* и *так же* (за исключением частотных ошибок *тоже мне, тоже самое* и *точно также*, учтенных в поисковом запросе). В то же время можно предполагать, что объем связанного с этим шума сопоставим в разных веб-корпусах, а также в разных подкорпусах НКРЯ, так что контрастные различия, например, между корпусом Araneum и ГИКРЯ или между Основным и Газетным корпусом НКРЯ в целом отражают реальную картину. Кроме того, по крайней мере в НКРЯ объем шума совсем небольшой: в первых 50 случайных вхождениях *тоже* и *также* в Газетном корпусе шум отсутствовал.

[5] Образцы запросов: -точно также, на расстоянии от 1 до 1 от Слова 1; тоже -самое & -мне, на расстоянии от 1 до 1 от Слова 1; [word!="точно|Точно"][word="также|Также"]; [word="тоже|Тоже"][word!="мне"&word!="самое"].

употреблениями, которые не имеют параллели с *тоже*. Это, во-первых, сочетание *также* с союзом *а*, которое приобрело свойство сочинительного союза и может соединять составляющие разных категорий – от именных групп (12) до клауз (13) [17: 254].

(12) *Салаты из сырых овощей или фруктов, блюда с добавлением мяса, птицы, копчёностей, а также* (**тоже*) *винегреты могут храниться 18 часов.*
[Парламентская газета, 2021.12]

(13) *Вакцина практически на 100% предотвращает смерть в результате этой болезни, а также* (**тоже*) *она предохраняет от тяжелого течения COVID-19.* [Парламентская газета, 2021.10]

Как демонстрируют данные в таб. 1, частотность *а также* после XVIII в. последовательно росла и в современных текстах составляет почти половину от всех употреблений *также*.

Во-вторых, растет употребление *также* в качестве своего рода дискурсивного маркера, т.е. единицы с метатекстовой функцией [21], близкой по значению вводному слову *кроме того*, как в (14) (к этому вопросу мы вернемся в разделе 4.2):

(14) *Также* (??*Тоже*)*, что немаловажно, большое внимание уделено драматизму работы Штирлица.* [Форум: 17 мгновений весны (2005-2010)]

Обратим внимание, что ни в (12-13), ни в (14) замена *также* на *тоже* невозможна. По-видимому, она была бы невозможна и при изменении порядка слов при условии сохранения компонента *а* в (12-13) и начальной позиции *тоже* в (14) (ср. в (12) ??*блюда с добавлением мяса, а винегреты тоже* и под.), что отличает такие употребления от собственно аддитивных (см. Введение).

4 *Тоже* и *также*: эволюция свойств

В настоящем разделе рассмотрены некоторые изменения в коммуникативных (раздел 4.1) и синтаксических (раздел 4.2) свойствах *также* и *тоже* после XIX в.

4.1 Коммуникативная структура

Один коммуникативный сдвиг, произошедший после XIX в., уже отмечен в разделе 2: современное коммуникативное противопоставление между *также* и *тоже* установилось лишь в XXв. Другой сдвиг касается употреблений *также*, в которых *также* синонимичен *тоже* вплоть до коммуникативной структуры, как в (15) (см. подробнее Введение):

(15) *В этой области проживают преимущественно сунниты <…>. Саддам Хусейн также (≈ тоже) был суннитом.* [Lenta.ru, 2003.11]

Частотность таких употреблений снижается от XVIII-XIX вв. к XXI-му. Об этом свидетельствуют данные в таб. 3, отражающие частотность двух употреблений *также* (‘также’ и ‘тоже’) среди первых 100 случайных вхождений *также* в трех подкорпусах НКРЯ – подкорпусе XVIII-XIX вв., подкорпусе XX в. и в Газетном корпусе – и в ГИКРЯ (в сегменте «ВКонтакте»). Можно видеть, что доля *также* в значении ‘тоже’ последовательно падает; различие между Газетным корпусом и обоими подкорпусами Основного корпуса, а также между ГИКРЯ и Основным корпусом статистически значимо (χ^2, $p<0{,}01$).

Данные ГИКРЯ особенно показательны. В Газетном корпусе, как продемонстрировано в разделе 3, частотность *тоже* низкая, поэтому низкую долю *также* в значении ‘тоже’ можно было бы объяснять редкостью самого значения ‘тоже’ (что, в свою очередь, можно было бы связывать, например, с редкостью коммуникативной структуры, задаваемой ‘тоже’, в публицистических текстах). Однако в сегменте ВК в составе ГИКРЯ, как мы убедились, *тоже* даже несколько более частотно, чем *также*, поэтому низкая частотность *также* в значении ‘тоже’ указывает на редкость не значения ‘тоже’, а слова *также* в этом значении.

	‘также’	‘тоже’	Всего	доля ‘тоже’
1701-1900 НКРЯ	17	83	100	0,8
1901-2000 НКРЯ	27	73	100	0,7
Газетный НКРЯ	87	13	100	0,1
ГИКРЯ (ВК)	71	29	100	0,3

Таблица 3: *Также* в значениях ‘также’ и ‘тоже’ по периодам[6]

В снижении частотности *также* в значении ‘тоже’ можно усматривать тенденцию к разграничению *тоже* и *также* – так же, как и в коммуникативном разграничении, произошедшем после XIX в.

4.2 Синтаксис

Для современного языка характерно располагать слово *также* не перед именной группой, к которой *также* относится по смыслу, а перед глаголом, предшествующим этой группе, т.е. порядок (16) (далее *S также V S*) более частотен, чем порядок (17) (далее *S V также S*). В языке XVIII-XIX вв., между тем, явного предпочтения одного из двух порядков не было.

(16) *Исключение также составляют закупки медицинских масок*. [Парламентская газета, 2021.12]

(17) *Исключение составляют также закупки медицинских масок*.

На это указывают данные в таб. 4: порядок *S также V S* более частотен, чем *S V также S* в Газетном корпусе НКРЯ и в Araneum Russicum Maius, тогда как в подкорпусах XVIII-XIX вв. и XX в. НКРЯ частотность порядков примерно одинаковая (различие статистически значимо, χ^2, $p<0,01$).

	S также V S	S V также S	всего	доля S V также S
1701-1900 НКРЯ	126	119	245	0,5
1901-2000 НКРЯ	174	220	394	0,6
Газетный НКРЯ	22 765	2 244	25 009	0,1
Araneum	5 220	1 234	6 454	0,2

Таблица 4: Частотность *также* до и после глагола по периодам[7]

Этот сдвиг в порядке слов кажется уместным интерпретировать как стремление современного *также* вводить «большие» коммуникативные составляющие – например, широкую рему, а не узкую. В этом отношении *также* снова противопоставлен современному *тоже*, которое, наоборот, обычно вводит «малые» составляющие. Проиллюстрируем это предположение на материале конструкций со связкой *быть* и творительным предикативным, как в (18-21).

(18) *Ящеры тоже были вегетарианцами.* [Комсомольская правда, 2013.07]

(19) *Овидий был тоже изгнанником.* [lenta.ru, 2016.03]

(20) *Ливанов также был режиссером и сценаристом ряда мультфильмов.* [Парламентская газета, 2021.07]

(21) *Янг был также музыкантом и автором песен.* [gazeta.ru, 2017.10]

[6] Образец запроса: (s | spro) & nom также, -amark, на расстоянии от 1 до 1 от Слова 1 v & indic, -amark, на расстоянии от 1 до 1 от Слова 2 (в НКРЯ); [pos="Pp.*n.*"|pos="N.*n.*"][word="также"][pos="V.*i.*"] (в ГИКРЯ).

[7] Образец запроса: Слово 1: s, first Слово 2: также, -amark на расстоянии 1 от Слова 1 Слово 3: v & indic, -amark на расстоянии 1 от Слова 2 Слово 4: s, -amark на расстоянии 1 от Слова 3 (в НКРЯ); [word="\."][tag="N.*" & word="[А-Я][А-Яа-я]*"][word="также"][tag="V.*"][tag="N.*"] (в Araneum).

В силу семантической нейтральности связки она может быть и частью темы, и частью ремы. Однако, как свидетельствуют данные в таб. 5, *также* свойственно включать связку в свою сферу действия (т.е. располагаться перед ней), а *тоже*, наоборот, не свойственно располагаться после связки.

	X_{nom} *также/тоже* быть Y_{ins}	X_{nom} быть *также/тоже* Y_{ins}	всего	доля употреблений с широкой сферой действия
Тоже	36	1	37	0,03
Также	83	12	95	0,87

Таблица 5: Частотность «узкой» и «широкой» сферы действия *также* и *тоже* в конструкции со связкой (Газетный корпус НКРЯ)[8]

Стремление к «большим» составляющим ярко проявляется при сближении *также* с дискурсивным маркером, как в (22) (см. также [17: 256]). Такое *также* занимает начальную позицию в предложении и может быть отделено запятой и паузой. [9] Кроме того, дискурсивное *также* может быть носителем ассоциируемого с темой восходящего акцента, отличаясь этим и от безударного аддитивного *также* в значении ‘также’, и от *также* в значении ‘тоже’, несущего нисходящий акцент (см. Введение).

(22) *Также/, омолаживающим эффектом обладают орехи и приправы.* [Парламентская газета, 2020.10]

Дискурсивное *также* выполняет метатекстовую функцию и включает в свою сферу действия целую пропозицию, не соотносясь ни синтаксически, ни семантически с отдельно взятой составляющей внутри предложения. Эти свойства считаются симптоматичными для дискурсивных слов [12: 78], [13: 4]. Как показывает рис. 1, частотность начального и отделенного запятой *также* последовательно растет по данным Газетного корпуса НКРЯ:

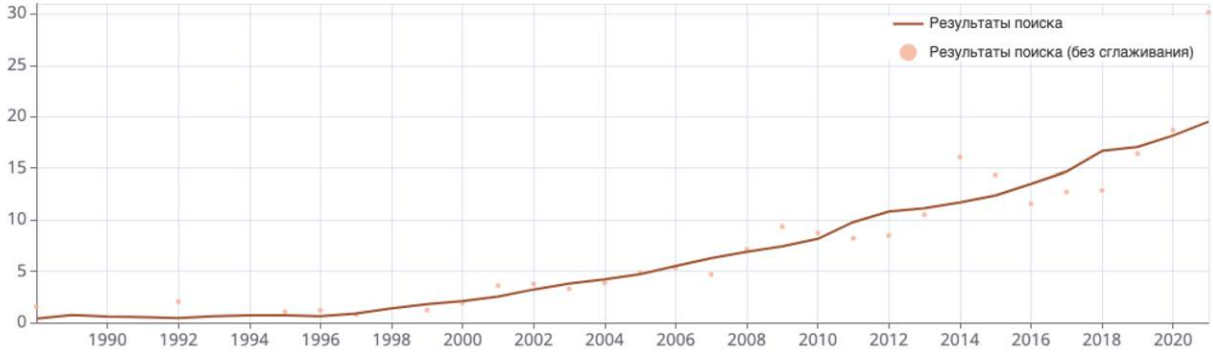


Рис. 1. Частотность *также* в начале предложения перед запятой по годам (Газетный корпус НКРЯ)

Заметим, впрочем, что далеко не всегда можно провести четкую границу между дискурсивным и аддитивным *также*. Так, в (23) перенос начального *также* в позицию перед глаголом как будто не влечет за собой ощутимого изменения смысла:

(23) *Также власти ограничили* (≈*Власти также ограничили*) *работу развлекательных центров, ресторанов <…> и других заведений.* [Парламентская газета, 2021.12]

Для сравнения, в (24) *также* не может быть перенесено вправо без изменения смысла; *также* вводит здесь даже не пропозицию, а речевой акт (подобно иллокутивно употребленному союзу, ср. [20]):

[8] Образец запроса: s & nom, first быть, v & indic, -amark, на расстоянии от 1 до 1 от Слова 1 также, -amark, на расстоянии от 1 до 1 от Слова 2 s & ins, на расстоянии от 1 до 1 от Слова 3. Примеры с *также* в значении ‘тоже’ не учитывались.

[9] В современных справочных изданиях по русскому языку такое употребление обычно признается ненормативным, ср., например, [7].

(24) *Также, сами подумайте* (≠ ??*Сами также подумайте*), *какие задачи у таких комплексов?* [Форум: Метро-2 (2008-2011)]

Тем самым, конвенционализация *также* в дискурсивный маркер, по-видимому, еще не завершена.

5 Некоторые обобщения (вместо заключения)

Приведенный материал свидетельствует о том, что в современном языке *тоже* и *также* противопоставлены на разных уровнях языка – коммуникативном, синтаксическом и стилистическом. В языке XVIII-XIX вв. такой дифференциации между ними еще не было.

Одно отличие *также* от *тоже* стоит особняком: *также*, но не *тоже*, стал источником грамматикализации для сочинительного союза *а также* и источником прагматикализации – для дискурсивного маркера *также*. По данным [10], направление этого развития отвечает типологическим ожиданиям: и функция сочинительного союза, и функция дискурсивного маркера (*conjunctional adverb*) свойственны аддитивным маркерам в языках мира. Объяснимо и то, почему эти функции возникли у *также*, а не у *тоже*: *также* соотносится с ремой, при этом и сочиненная клауза, вводимая *а также*, и независимая клауза, вводимая дискурсивным маркером *также*, должны получить собственную иллокутивную силу [3], а значит, и собственную рему.

Но эволюция *также* демонстрирует и некоторые неожиданные черты. Согласно [10], в языках, где аддитив используется как дискурсивный маркер (*conjunctional adverb*), он также получает функцию маркера контрастного топика, подобного русскому союзу *а* (ср. *Аня блондинка, а Нина брюнетка*). В русском языке этого не произошло.

Благодарности

Автор глубоко признателен Е.В.Рахилиной, П.А.Бычковой и И.И.Колесниченко за обсуждение некоторых фрагментов этой работы, а также анонимным рецензентам за советы и замечания. Исследование осуществлено в рамках Программы фундаментальных исследований НИУ ВШЭ.

References

[1] Apresyan Ju.D. Types of communicative information for an explanatory dictionary [Tipy kommunikativnoj informacii dlja tolkovogo slovarja] // Karaulov Ju.I. (ed.), Language: system and functioning [Jazyk: sistema i funkcionirovanie] – Moscow, 1988. – P. 10-22.

[2] Araneum Russicum Maius, available at: http://unesco.uniba.sk/aranea/index.html.

[3] Belyaev O. Systematic mismatches: Coordination and subordination at three levels of grammar. – Journal of Linguistics, 2015. – Vol. 51. – P. 267–326.

[4] Bogusławski A. On the issue of the secondary designation of a definite content in a Russian connected text [K voprosu o vtoričnom oboznačenii opredelennogo soderžaniya v russkom svyaznom tekste] // Scientific reports of higher school, Philological sciences [Naučnye doklady vysšey školy. Filologicheskiye nauki], 1969. – Vol.6.

[5] Bogusławski A. ALSO from ALL SO: On a set of particles in service of efficient communication. – Journal of pragmatics, 1986. – No. 10.

[6] Dictionary of the Russian Language of the 11th — 17th Centuries. – Vol.29. – Moscow, 2011.

[7] Djachkova N.A Also, never start a sentence with *takže* [Takže nikogda ne načinaj predloženie s *takže*]. Available at: https://uralsky-missioner.ru/doc/445/.

[8] General Internet-Corpus of Russian [General'nyj internet-korpus russkogo jazyka], available at: https://int.webcorpora.ru/drake/.

[9] Evgen'eva A. P. (ed.). Dictionary of Russian [Slovar' russkogo jazyka]. – Vol. 4. – Moscow, Russkij jazyk, 1988.

[10] Forker D. Toward a typology for additive markers. – Lingua, 2016. – Vol.180. – P. 69-100.

[11] Girke W. Zur Funktion von *i*, *tože*, *takže* // P. Hill, V. Lehmann (eds.), Slavistische Beitrage, Band 147, – München, Sagner, 1981. – P. 7-26.

[12] Günthner S., Mutz K. Grammaticalization vs. pragmaticalization? The development of pragmatic markers in German and Italian // W. Bisang, N. Himmelmann, B. Wiemer (eds.), What Makes Grammaticalization? A Look from its Fringes and its Components. – Berlin, Mouton de Gruyter, 2004. – P. 77–107.

[13] Heine B., Kaltenböck G., Kuteva T., Long H. The Rise of Discourse Markers. – Cambridge, Cambridge University Press, 2021.

[14] König E. The Meaning of Focus Particles: A Comparative Perspective. – Routledge, London, 1991.

[15] Paducheva E.V. *Tože* i *takže*: the relationship of information structure and associative links [*Tože* i *takže*: vzaimootnošenie aktual'nogo členenija i associativnyx svjazej] // Russian Language Institute. Preliminary publications [Institut russkogo jazyka, predvarital'nye publikacii – Vol.55. – Moscow, 1974.
[16] Paducheva E.V. *Tože* and *takže* twenty years later [*Tože* i *takže*, no dvadcat' let spustja] // M.Grochovski, D.Weiss (eds.). Words are physicians for an ailing mind. – München, Sagners Slavistische Sammlung, Band 17, 1991. – P. 311–322.
[17] Popkova N. *Takže* [Also] // O.Ju.Inkova (ed.), Semantics of connectors: a contrastive study [Semantika konnektorov: kontrastivnoye issledovaniye]. – Moscow, TORUS-PRESS, 2018. – P.240-266.
[18] Rakhilina E. V., Borodina M. A., Reznikova T. I. "Taman today": A corpus study of 19th-century Russian ["Taman' segodnya": korpusnoe issledovanie russkogo yazyka XIX veka] – Proc. of the V. V. Vinogradov Russian Language Institute, 2016. – Vol.10. – P.242–255.
[19] Russian National Corpus [Nacional'nyi korpus russkogo jazyka] (2003–2022), available at: http://www.ruscorpora.ru.
[20] Sweetser E. From Etymology to Pragmatics. Metaphorical and Cultural Aspects of Semantic Structure. – Cambridge, Cambridge University Press, 1990.
[21] Trask R.L. A Dictionary of Grammatical Terms in Linguistics. – London / N.Y., Routledge, 2012.
[22] Yanko T. E. Communicative strategies of Russian speech [Kommunikativnye strategii russkoj reči]. – Moscow, 2001.

The CoBaLD Annotation Project: the Creation and Application of the full Morpho-Syntactic and Semantic Markup Standard

Maria Petrova
A4 Technology
Moscow
g-fox-ive@mail.ru

Alexandra Ivoylova
RSUH
Moscow
a.m.ivoylova@gmail.com

Ilya Bayuk
A4 Technology
Moscow
ilya.bayuk@yandex.ru

Darya Dyachkova
RSUH
Moscow
d.dyachkova@bk.ru

Mariia Michurina
RSUH
Moscow
marimitchurina@gmail.com

Abstract

The current paper is devoted to the Compreno-Based Linguistic Data (CoBaLD) Annotation Project aimed at creating text corpora annotated with full morphological, syntactic and semantic markup. The first task of the project is to suggest a standard for the full universal markup which would include both morphosyntactic and semantic patterns. To solve this problem, one needs the markup model, which includes all necessary markup levels and presents the markup in a format convenient for users. The latter implies not only the fullness of the markup, but also its structural simplicity and homogeneity. As a base for the markup, we have chosen the simplified version of the Compreno model[1], and as data presentation format, we have taken Universal Dependencies.

At the second stage of the project, the Russian corpus with 400 thousand tokens (CoBaLD-Rus) has been created, which is annotated according to the given standard. The third stage is devoted to the testing of the new format. For this purpose, we have held the SEMarkup Shared Task aimed at creating parsers which would produce full morpho-syntactic and semantic markup. Within this task, we have elaborated neural network-based parser trained on our dataset, which allows one to annotate new texts with the CoBaLD-standard. Our further plans are to create fully annotated corpora for other languages and to carry out the experiments on language transfers of the current markup to other languages.

Keywords: Compreno, semantic markup, Universal Dependencies
DOI: 10.28995/2075-7182-2023-22-421-432

Проект CoBaLD: разработка и применение стандарта полной морфо-синтаксической и семантической разметки текстов

Петрова М.А.
A4 Technology
Москва, Россия
g-fox-ive@mail.ru

Ивойлова А.М.
РГГУ
Москва, Россия
a.m.ivoylova@gmail.com

Баюк И.С.
A4 Technology
Москва, Россия
ilya.bayuk@yandex.ru

Дьячкова Д.С.
РГГУ
Москва, Россия
d.dyachkova@bk.ru

Мичурина М.А.
РГГУ
Москва, Россия
marimitchurina@gmail.com

Аннотация

Данная работа посвящена проекту Compreno-Based Linguistic Data (CoBaLD), целью которого является создание корпусов с полной морфологической, синтаксической и семантической разметкой. Первой задачей проекта является создание стандарта полной универсальной разметки, включающей как морфо-синтаксический, так и семантический уровни. Реализация данной задачи требует, с одной стороны, наличия модели, предлагающей необходимые уровни разметки,

[1]The access to the Compreno data is provided according to the CC BY-NC 4.0 License which allows non-commercial use.

и, с другой стороны, возможности представить разметку в удобном для пользователя формате. Последнее требование предполагает не только полноту разметки, но также ее структурную простоту и однородность описания объектов. В качестве базы для подобной разметки мы выбрали упрощенную модель Compreno, в качестве формата представления данных - формат Universal Dependencies.

Вторым этапом проекта стало создание русскоязычного корпуса объемом 400 тысяч токенов - CoBaLD-Rus, размеченного по предложенному стандарту. Третий этап посвящен тестированию предложенной разметки, в рамках которого было проведено соревнование SEMarkup Shared Task. Задача состояла в создании парсеров, обученных на данном корпусе и позволяющих размечать новые тексты в соответствии с CoBaLD-стандартом. В качестве бейзлайна для соревнования мы также разработали нейросетевой парсер для решения поставленной задачи. В дальнейшем планируется создание аналогичных корпусов для других языков и проведение экспериментов по языковому переносу данной разметки на другие языки.

Ключевые слова: Compreno, семантическая разметка, Universal Dependencies

1 Introduction

In the given paper, we present the Compreno-Based Linguistic Data (CoBaLD) Annotation Project which is aimed at elaborating the general standard of the full text markup, including morphological, syntactic and semantic levels, and the creation of text corpora annotated according to the standard. The current work focuses on the following tasks:

(1) choosing the markup model, which is full enough and at the same time simple enough to be presented in the convenient format;

(2) choosing the format of the full markup presentation;

(3) elaborating the markup standard, including both morphosyntactic and semantic markup;

(4) creating the Russian corpus annotated according to the standard;

(5) conducting a shared task aimed at the creation of the automatic semantic markup in order to investigate the capabilities of the format (SEMarkup-2023 Shared Task);

(6) creating a baseline version of the parser trained on the annotated dataset which allows one to annotate new texts in the CoBaLD-standard.

Since the task of the linguistic markup is an important part of the NLP pipeline, a lot of efforts have been applied to create convenient markup formats.

As far as the formats of the morpho-syntactic markup are concerned, the most popular one is the Universal Dependencies (UD) project (De Marneffe et al., 2006). There are parsers created for the UD standard, such as UDPipe (Straka et al., 2016) (currently, for more than 100 languages including Russian), and, for the Russian language, - the Joint Morpho-Syntactic Parser (Anastasyev, 2020).

Concerning the semantic markup, there are several projects, most of which started with creating a machine translation algorithm. One of the oldest projects is the Universal Networking Language (Uchida and Zhu, 2001), which popularized the idea of using directed graphs for semantic descriptions. Among other well-known projects are Abstract Meaning Representations (AMR) (Banarescu et al., 2013), Universal Conceptual Cognitive Annotation (UCCA) (Abend and Rappoport, 2013), Prague Dependencies (Hajic et al., 2001), Discourse Representation Structures (DRS) (Groenendijk et al., 1984), and Universal Decompositional Semantics (UDS) (White et al., 2016). The Russian National Corpus (ruscorpora.ru) has recently included partial semantic markup, too.

These formats have significant differences with regard to their treatment of morphosyntax and semantics. For instance, UCCA and AMR ignore morphosyntactic data on purpose, while the Prague dependencies represent full three-level linguistic markup. The ETAP system (Boguslavsky, 1999) and the Compreno model (Anisimovich et al., 2012; Petrova, 2014) propose such integral labelling as well. Moreover, UDS, if joined with UD, could represent its semantic part. However, all these formats are rather complicated and difficult to work with. Therefore, there is no generally accepted standard up to now both for the semantic markup and for the full markup, which would include all three markup levels.

Thereby, our first purpose is to develop a standard, which would, on the one hand, include morphological, syntactic and semantic markup, and, on the other hand, be simple and convenient enough for the users to work with.

As for the format, we have chosen the UD model: it is concise enough and uses the CONLL (or CONLL Plus) format, which makes it convenient for scripting and automatic parsing purposes. However, UD lacks a semantical pattern. Therefore, we had to integrate it from some other model.

We have chosen the Compreno one, as the model is very simple from the structural point of view and suggests full semantic markup both for the word meanings and the relations between words.

Further, we will briefly describe the basic principles of the Compreno model and show the conversion process of the Compreno markup into the UD format. Afterwards, we will present our dataset annotated according to the CoBaLD-standard and focus on the SEMarkup-2023 Shared Task together with a baseline parser created for it.

In conclusion, we will sum up the results and discuss further perspectives of the work.

2 The Compreno Markup Format: Simplification and Conversion

2.1 Simplification of the Compreno Format

In Compreno, each word meaning is attributed to a semantic class (SC) - a semantic field denoting the word's meaning. The SCs are organized in a thesaurus-like hierarchy. All semantic links between words are expressed through the semantic roles, or slots (SS) corresponding to actant valencies (Agent, Experiencer, Addressee, etc.), adjuncts (Locative, Distance, Time, Condition, Concession, and so on), characteristics (for instance, evaluation, speed, price, form, or size), specifications and others. It allows one to annotate the semantical meanings of all words and to define all semantic relations of each word, both actant and circumstantial.

However, the model suggests a heavily detailed description: namely, it contains more than 200,000 SCs (which seems too much for a machine learning based parser trained on the dataset of our volume) and more than 330 SSs, which, in turn, does not seem necessary for most application tasks (except the task of building semantic sketches (Detkova et al., 2020)).

Therefore, we decided to reduce the number of categories. First, we have used not the terminal SCs, which denote the exact word meanings, but the hyperonym classes. That is, all words with motion semantics would now belong to the hyperonym class MOTION. Second, we have reduced the number of the SSs. For example, full Compreno markup suggests different roles for different characteristic dependencies, such as form, taste, sound, appearance, importance, genuineness, and so on - more than 60 characteristics in total. In the generalized variant, all such characteristics correspond to one characteristical slot. Or, full model contains several Instrument slots, which differ by the SCs each slot can include (see fig.1) - in the simplified variant, they are joined in one Instrument slot.

Instrument		to write [with a pen]
Instrument_Being	Instrument	to attack [with remaining army]
Instrument_Cognitive		to understand [with one's heart]
Instrument_Time		to be punished [by 30 years in prison]

Figure 1: Instrument slots in full and in reduced Compreno markup

As a result, the number of hyperonym SCs used in the markup was reduced to 1085 classes, and the number of the SSs - to 143 slots.

The semantic hierarchy of the hyperonym classes can be found on the Compreno Semantics Github[2]. The list and the description of the semantic roles are also available on the corpus page[3].

These simplified SCs and SSs are used in the final version of the markup in the UD format.

2.2 Annotation and Conversion

The Compreno markup can be obtained automatically or manually. For the current dataset, the markup includes the boundaries of the constituents, the SCs (their labels are marked with green below) and the

[2]https://github.com/compreno-semantics/semantic-hierarchy/blob/main/hyperonims_hierarchy.csv

[3]https://github.com/compreno-semantics/compreno-corpus/blob/main/semantic_slots.xlsx

SSs (their labels are marked with brown below) - see fig. 2.

Обычно бюджет ко второму чтению готовится непосредственно в Думе: депутаты корректируют правительственные планы. 'Usually the budget is prepared for the second reading directly in the Duma: the deputies update the government plans'.

```
#[[Time: Обычно"обычно:#frequentative_adverbs_adj:FREQUENTATIVE"] [Experiencer_Metaphoric:
бюджет"бюджет:бюджет:BUDGET"] [[ко"к:#preposition:PREPOSITION"] [OrderInTimeAndSpace:
второму"второй:TWO_ORDINAL"] Object_Situation: чтению "чтение:READING_OF_THE_DRAFT_LAW"] Predicate:
готовится"готовить:готовить:PREPAREDNESS" [[DegreeApproximative:
непосредственно"непосредственный:DIRECT_OBLIQUE"] [в"в_Prepositional:#preposition:PREPOSITION"] Locative:
Думе"дума:дума:DUMA"]#: [[Agent: депутаты"депутат:депутат:DEPUTY"] Specification_Clause:
корректируют"корректировать:корректировать:TO_CORRECT" [[Agent:
правительственные"правительство:правительство:GOVERNMENT"] Object_Situation:
планы"план:план:SCHEDULE_FOR_ACTIVITY"]]].
```

Figure 2: An example of the Compreno "bracket" format

The markup can also be provided with surface, or syntactic, roles (marked with $ sign - see Fig. 3 below), coreference and non-tree links, however, the purpose of the given dataset was only the semantic markup. [4]

As one can see, this format of markup representation does not contain morphological and other grammatical information. Nevertheless, after a sentence is annotated, the parser can build its parsing tree (see (Anisimovich et al., 2012)), where each token is provided with full grammatical and semantic data. Fig. 3 is an illustration of the Compreno parsing tree for the above given sentence, and fig. 4 being the fragment of the tree shows the morphological grammemes for the node "готовить:готовить:PREPAREDNESS".

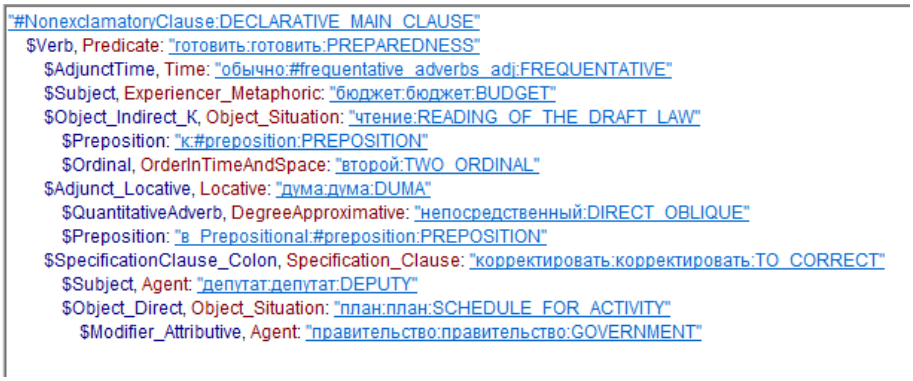


Figure 3: An example of the Compreno parse tree

The "bracket" format presented on fig. 2 is the one that the annotators work with to point out the information necessary for building the correct structure of a sentence, whereas the parsing tree is where full information about the sentence is stored (its syntactic and semantical structure, syntactic and semantic slots, SCs, grammatical features and information about coreference and non-tree links).

Unlike Compreno, UD stores all relevant information in the markup itself, presented in a table-view. Thereby, during the conversion of the Compreno markup into UD, the necessary data is taken from the parsing trees.

UD has its own morphology and syntax, therefore, the corresponding information in Compreno has to be converted into the UD format. Of course, there is a number of differences between the Compreno and the UD formats in this respect. Most significant distinctions concern POS-tagging, tokenization, lemmatization, asymmetry of mapping some grammatical features, ellipsis and copula description, coordination and dealing with punctuation. Besides, the UD format marks the tokens up with so called dependency

[4]The only surface slot mentioned in the markup is the $Dislocation slot – it is the slot for the constituents that syntactically depend on one core, while semantically – on the other core.

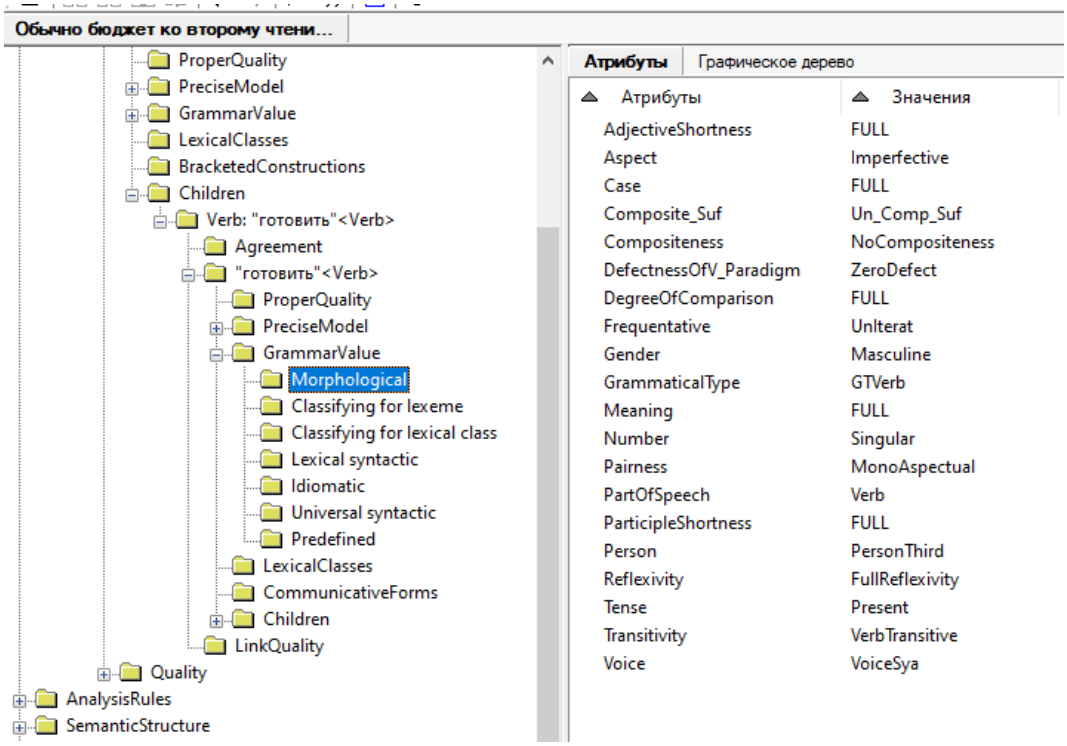


Figure 4: The grammemes for the node "готовить:готовить:PREPAREDNESS"

heads (each token gets the index of its head as a label) whereas the Compreno model operates with the boundaries of the constituents. During the conversion, the labeling of these heads was based on their boundaries. The conversion process is thoroughly discussed in (Ivoylova et al., 2023).

As far as the semantics is concerned, the UD format does not have the semantic level, so the information about the SCs and the SSs can be added to the UD markup in the way it is presented in Compreno (its simplified version).

After the conversion, the markup looks as in fig. 5 and includes morphology, syntax, and semantics.

```
# text = Обычно бюджет ко второму чтению готовится непосредственно в Думе : депутаты корректируют правительственные планы.
1   Обычно  обычно  ADV _   Degree=Pos  6   advmod  Time    CH_REFERENCE_AND_QUANTIFICATION
2   бюджет  бюджет  NOUN    _   Animacy=Inan|Case=Nom|Gender=Masc|Number=Sing   6   nsubj   Experiencer_Metaphoric  BUDGET
3   ко  ко  ADP _   _   5   case    _   PREPOSITION
4   второму второй  ADJ _   Case=Dat|Gender=Neut|Number=Sing    5   amod    OrderInTimeAndSpace
CH_REFERENCE_AND_QUANTIFICATION
5   чтению  чтение  NOUN    _   Animacy=Inan|Case=Dat|Gender=Neut|Number=Sing   6   obl Object_Situation
TO_PROCESS_INFORMATION
6   готовится   готовить    VERB    _   Aspect=Imp|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Mid  0
root    Predicate   READINESS
7   непосредственно непосредственно ADV _   Degree=Pos  9   advmod  Degree  CH_OF_CONNECTIONS
8   в   в   ADP _   _   9   case    _   PREPOSITION
9   Думе    дума    NOUN    _   Animacy=Inan|Case=Loc|Gender=Fem|Number=Sing    6   obl Locative    STATE_AUTHORITIES
10  :   :   PUNCT   _   _   12  punct   _   _
11  депутаты    депутат NOUN    _   Animacy=Anim|Case=Nom|Gender=Masc|Number=Plur   12  nsubj   Agent
PERSON_BY_SPHERE_OF_ACTIVITY
12  корректируют    корректировать  VERB    _   Aspect=Imp|Mood=Ind|Number=Plur|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
6   parataxis   Specification   TO_CORRECT
13  правительственные   правительственный   ADJ _   Animacy=Inan|Case=Acc|Degree=Pos|Number=Plur    14  amod    Agent
STATE_AUTHORITIES
14  планы   план    NOUN    _   Animacy=Inan|Case=Acc|Gender=Masc|Number=Plur   12  obj Object_Situation
SCHEDULE_FOR_ACTIVITY
15  .   .   PUNCT   _   _   6   punct   _   _
```

Figure 5: CoBaLD format example

3 Corpus Dataset

Our further goal was to obtain the Russian corpus annotated according to the CoBaLD standard.

For the corpus material, we have chosen news texts from the NewsRu.Com dataset, created during building the RuCoCo corpus (Dobrovolskii et al., 2022). The dataset contains 3 markup levels:

- morphological,
- syntactic,
- semantic.

We have labeled the CoBaLD-Rus dataset - a Compreno-Based Dataset of Russian. It is published on Github[5].

The volume of the corpus is around 400,000 tokens. As our next step is building the parser, the whole sample is divided into two parts:

360 000 training and validation sample,

40 000 test sample for quality evaluation.

The test data does not contain any categories which are not represented in the training data.

4 The Annotation of the Corpus

The annotation process was organized as follows. At the first stage, the corpus was automatically annotated with the Compreno semantic markup with the help of the Compreno parser and included the constituents boundaries, SCs and SSs. Afterwards, the markup and the correctness of the parsing trees were manually checked by a team of professional linguists.

The annotated corpus was converted into the UD format with the help of the Compreno-To-UD Converter presented in (Ivoylova et al., 2023). Finally, the simplification algorithm was applied, which changed the SCs and the SSs to their simplified correlates.

As the morphosyntactic part was converted automatically, about 10% of the conversion results were also human-checked. The percent of modified labels varies from 5 to 10%, which means that the total quality of the conversion is close to 95%.

To measure the ambiguity level of the markup, an experiment on the annotators' agreement has been carried out. 100 sentences have been annotated by two annotators independently. Afterwards, the comparison of the markups has been made, especially as far as the constituents borders, the SCs and the SSs are concerned. The results turned out to be as in the table 1:

	Heads diff.	SemSlots diff.	SemClasses diff.	Overall inter-annotator agreement
Original	0.93%	2.64%	2.72%	**93.71%**
Generalized	0.93%	2.49%	2.41%	**94.17%**

Table 1: Inter-annotator agreement

Most cases of disagreement between the annotators concern polysemy, that is, these are cases, where the sentence can be interpreted differently. For example:

Отметим, контактные линзы для собак и кошек с 2001 года продаются в Японии.
Token: Отметим
SemClass: TO_PERCEIVE / VERBAL_COMMUNICATION

Выявленный дефект во всех машинах будет устранен бесплатно.
Token: машинах
SemClass: APPARATUS / TRANSPORT

As one can see from Table 1, the generalized markup causes less disagreements, because in some cases it does not differentiate between the homonyms with closer semantics.

[5]https://github.com/compreno-semantics/compreno-corpus

5 SEMarkup-2023 Shared Task

To test the created markup format, we suggested the SEMarkup shared task - the task devoted to the creation of the automatic semantic markup. It presupposed creating a solution that would produce a simultaneous morpho-, syntactic and semantic markup. The competition was held on the CodaLab platform[6] and proposed to use the CoBaLD-Rus dataset for learning data. As a baseline, we created a neural networks based parser trained on the CoBaLD-Rus dataset, which allows one to annotate new texts with the CoBaLD standard.

Unfortunately, only one participant succeeded to present the final solution, however, both the baseline and the participant's model demonstrated promising results (see Table 2). Below, we discuss the baseline, the participant's model and our further experiments with the baseline solution.

	Total	Lemma	POS	Features	UAS	LAS	SemSlot	SemClass
baseline	92.2%	96.1%	98.2%	95.3%	90.0%	85.6%	87.8%	92.2%
postoevie	90.2%	94.2%	97.9%	94.5%	86.2%	81.1%	86.9%	90.3%

Table 2: Baseline and participant scores

5.1 Baseline

The baseline model for the competition is a multi-task tagger. It is based on Anastasyev's Joint Morpho-Syntactic Parser (Anastasyev, 2020) (a GramEval2020 winner) extended with semantic tags, and its structure is represented in the fig. 6.

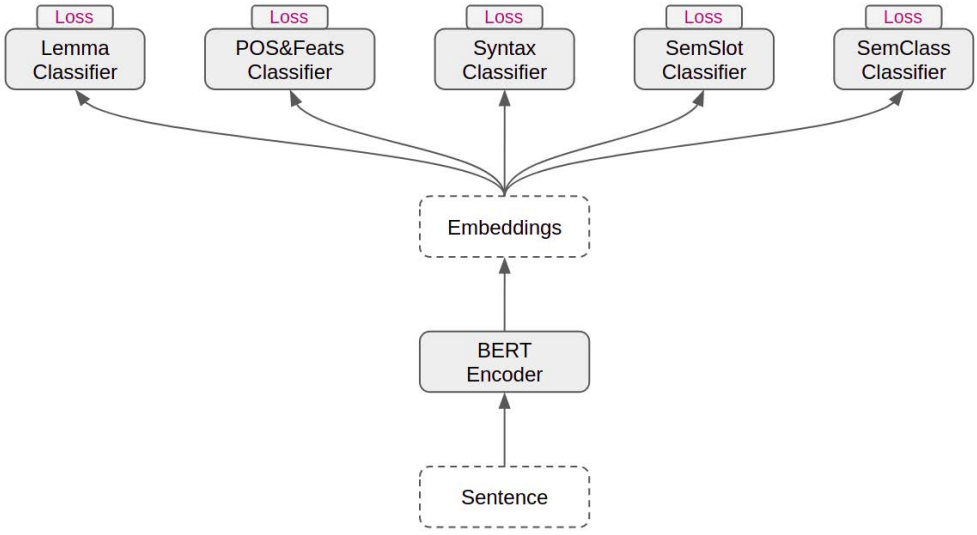


Figure 6: Baseline Architecture

As CoBaLD-Rus consists of multiple tags, the model itself has multiple heads.

Lemma classifier is a nonlinear feed-forward classifier predicting lemmatization rules. Lemmatization rule is a set of modification rules that have to be applied to a word to obtain its lemma. In our case, those are: "cut N symbols from the prefix of the word", "cut N symbols from the suffix" and "append a specific sequence of symbols to the suffix"[7].

POS & Feats classifier is a feed-forward classifier predicting joint POS and grammatical features tags[8].

[6]https://codalab.lisn.upsaclay.fr/competitions/10471
[7]See (Anastasyev, 2020) for details.
[8]See (Anastasyev, 2020) for details.

Syntax classifier is a biaffine dependency classifier (Dozat and Manning, 2016) predicting syntactic head and relation tags.

Semslot classifier and *Semclass classifier* are another two nonlinear feed-forward classifiers predicting SS and SC tags respectively.

The base dataset is split into train and validation parts so that train is 80% and validation is 20% of the base dataset size. The model is trained in a multi-task manner using slanted triangular learning rate scheduler along with gradual unfreezing and discriminative fine-tuning. The configuration is available on GitHub[9].

The model is implemented using the AllenNLP library and publicly available on our GitHub page[10].

For the base version of the parser, we used the pre-trained RuBERT-tiny[11] text encoder, which is 15 times smaller than the well-known DeepPavlov's RuBERT[12]. This exact version was submitted for the competition and set the baseline score, which can be observed in table 2.

We also experimented with the pre-trained Base XLM-RoBERTa[13] text encoder out of competition scope in order to evaluate the importance of embedding quality and the influence of language-specific features. The comparative quality for the variants can be seen in the table below.

	Total	Lemma	POS	Features	UAS	LAS	SemSlot	SemClass
RuBERT-tiny	92.2%	96.1%	98.2%	95.3%	90.0%	85.6%	87.8%	92.2%
XLM-R	**95.1%**	**97.3%**	**98.8%**	**96.8%**	**93.5%**	**89.8%**	**94.3%**	**94.8%**

Table 3: Baseline parser test scores using different encoders

The overall scores have not improved as much as we have expected. Nevertheless, there is a significant growth for SSs and some improvement for SCs scores. As the XLM-R model is multilingual, we can suspect that it could also positively influence the results, as well as its size.

5.2 Participant's model

Apart from the baseline, there is one model proposed for the competition. Generally speaking, it is close to the baseline, but has two new features added.

First, each non-linear feed-forward classifier head is accompanied with Linear Chain Conditional Random Field (CRF) (Huang et al., 2015). Although token embeddings are believed to contain some relevant information about all words in a sentence, feed-forward classifiers predict labels independently, and do not take other heads predictions into account. That is, for example, POS-tag of the last token in a sequence does not depend on the POS-tag of the first one. Chain CRFs are known to overcome this problem by explicitly utilizing tags relationships and modelling joint distribution of the whole sequence of tags throughout timeline, rather than that of a single tag at each timestep.

Second, the Label Attention Layer (Mrini et al., 2019) was introduced into the biaffine dependency classifier. The label attention is a modified version of self-attention, where each head is reasoned by a classification label, and not the other tokens of a sentence, as in the latter. The authors suggest that this mechanism allows the model to learn label-specific views of the sentence, and proves the technique improves the quality of biaffine dependency parser.

Unluckily, due to implementation issues, the proposed model did not manage to beat the baseline score, although, if implemented correctly, it would definitely have.

Now let's consider the evaluation metrics used for the estimation of the parser. Some of them represent the improved variants of the metrics used in GramEval2020 Shared Task (Lyashevskaya et al., 2020), the others had to be introduced specifically for the SSs and SCs.

[9]https://github.com/dialogue-evaluation/SEMarkup-2023/blob/main/parsers/configs/baseline.jsonnet
[10]https://github.com/dialogue-evaluation/SEMarkup-2023/tree/main/parsers
[11]https://huggingface.co/cointegrated/rubert-tiny
[12]https://huggingface.co/DeepPavlov/rubert-base-cased
[13]https://huggingface.co/xlm-roberta-base

5.3 Evaluation Metrics

The evaluation metric is an average of seven scoring functions. The latter can be divided into three categories: morphological, syntactic and semantic scores.

5.3.1 Morphology

Lemmatization score is a weighted true-false classifier, expressed as follows[14]:

$$ScoreLemma(test, gold) = LemmaWeight(gold_{POS}) * [Norm(test_{lemma}) = Norm(gold_{lemma})].$$

The weighting function depends upon a POS tag of a token. If the tag is one of *ADP*, *CCONJ*, *INTJ*, *PART*, *PUNCT*, *SCONJ*, *SYM* or *X*, the weight equals to 0.3. Otherwise, it equals to 0.7. The idea behind this is that we want immutable words to influence score less than mutable ones: normally, a dataset would have many more immutable words and this would make an overall score for lemmatization higher than it should actually be.

Function *Norm* makes input lowercase and replaces letter *ё* with letter *е*. For instance, the expression $[Norm(\text{Ёж}) = Norm(\text{еж})]$ equals to 1.

POS score is a true-false classifier:

$$ScorePOS(test, gold) = [test_{POS} = gold_{POS}].$$

Grammatical features of a token correspond to a set of pairs $(category,\ grammeme)$ where the category depends on the POS tag of a token and the grammeme depends on the category. Given a grammatical features of a token $token_{feats}$, denote a grammeme of a category cat as $token^{cat}_{feats}$. If features have no cat category, assume the notation equals to empty set.

Now, we can define *grammatical features score*:

$$ScoreFeats(test, gold) = Penalty(test_{feats}, gold_{feats}) * \frac{\sum_{(cat,\ gram)\ \in\ gold_{feats}} CatWeight(cat) * \left[gram = test^{cat}_{feats}\right]}{\sum_{(cat,\ gram)\ \in\ gold_{feats}} CatWeight(cat)}.$$

The left-hand multiplier penalizes test features for excessive length:

$$Penalty(f_{test}, f_{gold}) = \begin{cases} \frac{1}{1+(Size(f_{test})-Size(f_{gold}))} & \text{if } Size(f_{test}) > Size(f_{gold}), \\ 1 & \text{otherwise.} \end{cases}$$

This does not allow test features to contain too many categories. The latter is undesirable, for otherwise the model would gain higher scores by simply labelling a token with all possible categories.

The right-hand side is a weighted mean of true-false grammeme classifiers. The $CatWeight$ function accounts for category size, so that grammemes of a big category (which are harder to guess) are more valuable than those of a small one.

5.3.2 Syntax

We use Unlabeled and Labeled attachment scores as a measure of syntactic match quality:

$$UAS(test, gold) = [test_{head} = gold_{head}],$$
$$LAS(test, gold) = [test_{head} = gold_{head}] * [test_{deprel} = gold_{deprel}].$$

[14] $[x = y]$ is Iverson bracket notation

5.3.3 Semantics

Semantic slot score is a true-false classifier:

$$ScoreSemslot(test, gold) = [test_{semslot} = gold_{semslot}].$$

Semantic class score is calculated based on semantic hierarchy of hyperonym classes:

$$ScoreSemclass(test, gold) = \frac{1}{1 + Distance(test_{semclass}, gold_{semclass})},$$

where

$$Distance(u, v) = \begin{cases} PathLength(u, v) & \text{if } u \text{ and } v \text{ are in same tree,} \\ \infty & \text{otherwise.} \end{cases}$$

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

Averaging

Due to the weighting, some scores are strictly less than one, which means the score of ideal match is also less than one. To account for this issue, we divide the sum of test-gold scores by the sum of gold-gold scores. Now, a perfect match yields an accuracy of one.

Comparative evaluation

It would be interesting to compare the parser's quality with the quality of parsers, based on separate markup levels, namely, UD parsers and parsers aimed at the tasks of semantic labelling (such as UCCA (Hershcovich et al., 2019) or DRS (van Noord et al., 2020)), and to evaluate whether the integral approach makes the parsing process easier or not.

However, at the current stage, such comparison does not seem appropriate. We evaluate data of different corpora. The above mentioned semantic parsers do not suggest Russian parsing. Our metrics differ, as we made them stricter taking the word mutability into account and introducing penalty for excessive grammatical features.

Finally, it would be natural to compare our parser with the solutions for Word Sense Disambiguation task, as it can be solved with the help of the current dataset as well. For Russian, such work was conducted in 2020 (Bolshina and Loukachevitch, 2020). The best score was achieved on a fiction dataset with the use of a bi-LSTM model, and its f1 score is 95%. We have also calculated micro f1 (94%) and macro f1 (71%) scores for our baseline; the authors of the above-mentioned work haven't specified which type of f1 they used, unfortunately. As far as macro f1 is concerned, its lower score deals with SCs and SSc which are more rare and therefore poorly presented in the corpus. After analyzing such cases, we will enrich the corpus with the necessary data. Nevertheless, one should keep in mind that it is just a basic solution which can be seriously improved.

6 Results and Conclusion

First of all, we have simplified the full Compreno markup and made its usage easier. The markup has been converted into the UD format, which has been enriched with the semantic pattern. Therefore, we have elaborated the new standard, CoBaLD, for the full multi-level markup, which is the UD format including both morphosyntax and semantics.

Second, we have obtained the 400K Russian corpus CoBaLD-Rus annotated with the new standard. It is the first Russian corpus annotated in the format of this kind.

Third, we have tested the usage of the CoBaLD format during the SEMarkup-2023 Shared Task and created the integral three-level parser for this format based on neural networks.

Further plans concern several areas.

Currently, we are working on some optimizations of the labeling format, CoBaLD parser and the Compreno-to-UD converter, dealing mostly with ellipsis restoring and possibly adding other semantic information such as coreference. For that matter, we plan to move to the CONLL Plus format for better compatibility with UD.

Other important task is the creation of the English dataset annotated according to the CoBaLD standard. It would allow one to conduct comparative studies which can, inter alia, take semantic sketches into account (Detkova et al., 2020).

We are also considering the ability to hold a shared task on a "Lexical Sample" problem of WSD based on our markup standard.

Besides, we intend to experiment with the Language Transfer task which implies that the model trained on the donor language data can be applied to the data of the recipient language. The analysis of zero-shot transfer results may reveal a number of interesting details concerning the architecture of the parser itself and the qualities of the labelling format.

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.

KV Anisimovich, K Yu Druzhkin, KA Zuev, FR Minlos, MA Petrova, and VP Selegei. 2012. Syntactic and semantic parser based on abbyy compreno linguistic technologies. // *Proc Dialogue, Russian International Conference on Computational Linguistics*, P 91–103.

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.

Igor Boguslavsky. 1999. Translation to and from russian: the etap system. // *EAMT Workshop: EU and the new languages*.

Angelina Bolshina and Natalia Loukachevitch. 2020. All-words word sense disambiguation for russian using automatically generated text collection. *Cybernetics and Information Technologies*, 20(4):90–107.

Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, et al. 2006. Generating typed dependency parses from phrase structure parses. // *Lrec*, volume 6, P 449–454.

J Detkova, V Novitskiy, M Petrova, and V Selegey. 2020. Differential semantic sketches for russian internet-corpora. // *Computational Linguistics and Intellectual Technologies*, P 211–227.

Vladimir Dobrovolskii, Mariia Michurina, and Alexandra Ivoylova. 2022. RuCoCo: a new russian corpus with coreference annotation. // *Computational Linguistics and Intellectual Technologies*. RSUH.

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

Jeroen Groenendijk, Theo MV Janssen, and Martin Stokhof. 1984. *Truth, Interpretation and Information*. Foris Dordrecht.

Jan Hajic, Barbora Vidová-Hladká, and Petr Pajas. 2001. The prague dependency treebank: Annotation structure and support. // *Proceedings of the IRCS Workshop on Linguistic Databases*, P 105–114.

Daniel Hershcovich, Zohar Aizenbud, Leshem Choshen, Elior Sulem, Ari Rappoport, and Omri Abend. 2019. Semeval-2019 task 1: Cross-lingual semantic parsing with ucca. *arXiv preprint arXiv:1903.02953*.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

A Ivoylova, D Dyachkova, M Petrova, and M 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»*.

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.

Khalil Mrini, Franck Dernoncourt, Trung Bui, Walter Chang, and Ndapa Nakashole. 2019. Rethinking self-attention: An interpretable self-attentive encoder-decoder parser. *CoRR*, abs/1911.03875.

MA Petrova. 2014. The compreno semantic model: the universality problem. *International Journal of Lexicography*, 27(2):105–129.

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.

Hiroshi Uchida and Meiying Zhu. 2001. The universal networking language beyond machine translation. // *International Symposium on Language in Cyberspace, Seoul*, P 26–27.

Rik van Noord, Antonio Toral, and Johan Bos. 2020. Character-level representations improve drs-based semantic parsing even in the age of bert. *arXiv preprint arXiv:2011.04308*.

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.

HAlf-MAsked Model for Named Entity Sentiment analysis

Pavel Podberezko MTS AI
Andrey Kaznacheev MTS AI
Sabina Abdullayeva MTS AI
Anton Kabaev MTS AI

Abstract

Named Entity Sentiment analysis (NESA) is one of the most actively developing application domains in Natural Language Processing (NLP). Social media NESA is a significant field of opinion analysis since detecting and tracking sentiment trends in the news flow is crucial for building various analytical systems and monitoring the media image of specific people or companies.

In this paper, we study different transformers-based solutions NESA in RuSentNE-23 evaluation. Despite the effectiveness of the BERT-like models, they can still struggle with certain challenges, such as overfitting, which appeared to be the main obstacle in achieving high accuracy on the RuSentNE-23 data. We present several approaches to overcome this problem, among which there is a novel technique of additional pass over given data with masked entity before making the final prediction so that we can combine logits from the model when it knows the exact entity it predicts sentiment for and when it does not. Utilizing this technique, we ensemble multiple BERT-like models trained on different subsets of data to improve overall performance. Our proposed model achieves the best result on RuSentNE-23 evaluation data and demonstrates improved consistency in entity-level sentiment analysis.

Keywords: Roberta, Bert, Transformer, Ensemble, Sentiment analysis, text classification
DOI: 10.28995/2075-7182-2023-22-433-441

1 Introduction

Sentiment Analysis (SA) is a critical task in Natural Language Processing (NLP) that involves identifying the sentiment expressed in text. With the increasing amount of text data generated every day on various platforms such as social media, customer reviews, and news articles, sentiment analysis has become more important than ever. The goal of sentiment analysis is to automatically determine the emotional tone conveyed by a piece of text, which could be positive, negative, or neutral.

Over the years, several methods have been proposed for SA, ranging from traditional machine learning techniques to deep learning models based on neural networks. These methods have shown remarkable performance on different sentiment analysis tasks, such as document-level sentiment analysis, sentence-level sentiment analysis, and aspect-based sentiment analysis. In this competition, we faced one of the most significant and demanded SA problems called Named Entity Sentiment Analysis (NESA) which usually requires both identifying entities in text and determining their corresponding sentiment. However, in this case, we had to predict a sentiment label of the predetermined entity, so in this paper, we focus only on the second part.

Despite the significant progress made in sentiment analysis, there are still several challenges needed to be addressed. For example, handling sarcasm, irony, and figurative language in the text can be challenging, as these expressions may convey a sentiment opposite to their literal meaning. In addition, named entity sentiment may come from at least three different sources: author opinion, quoted opinion, and implicit opinion.

In this paper, we review our method for Named Entity Sentiment Analysis which achieves the best result on Dialog 2023 evaluation and discuss the challenges and opportunities in this field. We also present a comprehensive overview of approaches that have been tested including all the common and uncommon competition tricks. Finally, we identify some promising directions for future research in sentiment analysis, such as developing models that can handle linguistic nuances and context-dependent sentiment.

2 Related works

Modern Deep Learning contains a huge domain of tasks called Text Classification Tasks like topic or intent classification, spam and fraud detection, language identification, and many others. Methodologies of solving these problems were very similar over the years and worked quite well until (Pang et al., 2002) proposed to classify documents by sentiment. The authors found out that document sentiment analysis is a more challenging task to address and well-known at the time approaches were not that effective. So, researchers had to develop robust models capable of deciphering the intricate nuances of human language.

The Word2Vec model (Mikolov et al., 2013) revolutionized the field of natural language processing with its efficient training of word embeddings, which capture the semantic relationships between words. These embeddings have since become a fundamental component in many sentiment analysis models.

Recurrent Neural Networks (RNN) based approaches were dominant in the field of SA for a long time because they excel at modeling sequential data and capturing temporal dependencies in text. There are multiple improvements of this technique especially for sentiment analysis such as Recursive Neural Tensor Network (Socher et al., 2013) which computes compositional vector representations for phrases of variable length and syntactic type, CNN-BiLSTM model (Yoon and Kim, 2017) which combines high-level features of document extracted by CNN and the context considered by BiLSTM that capture long-term dependency, and generalization of the standard LSTM architecture to tree-structured network topologies named Tree-LSTM (Tai et al., 2015).

The introduction of the Transformer architecture (Vaswani et al., 2017) marked a major breakthrough in natural language processing. With its self-attention mechanism and parallel computation capabilities, the Transformer model has become the basis for many state-of-the-art sentiment analysis systems. The appearance of such a powerful tool blurred the distinction between varieties of Text Classification tasks to some degree. Fine-tuning BERT (Devlin et al., 2018) or its any further modifications has been a crucial method of addressing this task until recently. Although, there are plenty of tricks that can vastly improve model performance in such a specific domain. For a comprehensive understanding of this field of study, conference organizers offer an accurate and detailed observation (Golubev et al., 2023) of the various contributions to this domain.

3 Methodology

We explored and evaluated various approaches to tackle the given problem, aiming to identify the most effective technique. While we found HAlf MAsked Model (HAMAM) method to be the clear winner in terms of performance, other approaches still demonstrated notable results, deserving of honorable mention. This comprehensive assessment allowed us to not only establish the superiority of the winning technique but also gain valuable insights into the strengths and limitations of alternative methods within the context of the specific task.

3.1 Zero-shot NESA

The zero-shot named entity sentiment analysis method leverages pre-trained Masked Language Models (MLMs), such as BERT, to perform sentiment classification without the need for fine-tuning or labeled data specific to the task. The following steps and figure 1 provide a detailed overview of this approach:

1. Insert a [MASK] token right before the target entity in the sentence. This helps the model to focus on the context surrounding the entity.
2. Run the modified input sequence through the pre-trained MLM model, such as BERT. The model computes the probability distribution over the vocabulary for the [MASK] token based on the given context.
3. Create two lists of tokens representing "good" and "bad" sentiment, which will be used to compute average sentiment probabilities. For each list, extract the corresponding softmax output probabilities from the model for the tokens in that list.
4. Calculate the average probability of "good" and "bad" tokens. If the average probability of "good" tokens is higher than that of "bad" ones, classify the sentiment as positive. Otherwise, classify it as

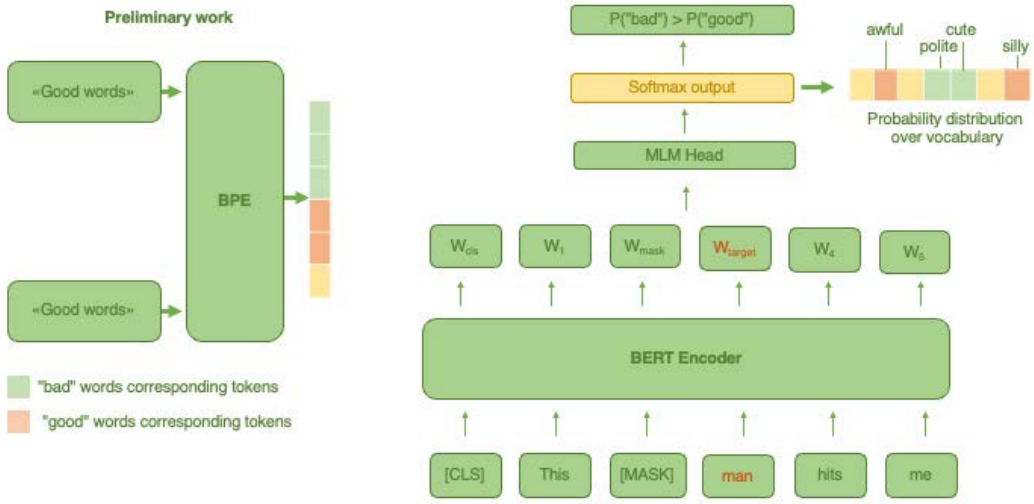


Figure 1: Scheme of zero-shot MLM-based approach.

negative.

This zero-shot method offers a straightforward approach to named entity sentiment analysis without the need for task-specific training. However, it may have limitations in handling more complex or nuanced sentiment expressions, as it relies solely on the pre-trained MLM's understanding of sentiment-related words. Moreover, it's tough to determine a neutral class because the difference between the average probabilities of "good" and "bad" tokens can be highly variable. Identifying a suitable threshold to distinguish neutral sentiment becomes difficult as the fluctuating difference makes it hard to establish clear boundaries of "approximately equal" probabilities.

3.2 Multi-sample dropout

In this section, we provide a detailed description of the multi-sample dropout technique, a regularization method presented by (Inoue, 2019). This advanced approach enhances the generalization capabilities of deep learning models by employing multiple dropout masks during training for the same mini-batch of input data.

In the original dropout technique, proposed by (Srivastava et al., 2014), a single dropout mask is generated for each input instance in a mini-batch during training. This mask is applied to deactivate a random subset of neurons (or features) with a certain probability (commonly between 0.2 and 0.5), helping to prevent the model from relying too heavily on any single neuron. After applying the dropout mask, the model performs a single forward and backward pass, updating its weights accordingly.

In contrast, the multi-sample dropout applies multiple dropout masks to each input instance in a mini-batch during training. As demonstrated in figure 2, for each input instance multiple forward passes are performed using different dropout masks, effectively exploring a broader range of neuron combinations. The outputs (logits or probabilities) from these multiple forward passes are then averaged, resembling an ensemble-like approach. The backward pass is performed using this averaged output, computing gradients and updating the model's weights. Therefore, this approach notably decreases the required number of training iterations.

In summary, while both the original dropout and multi-sample dropout techniques utilize dropout masks to improve generalization, the multi-sample dropout method extends this concept by employing multiple masks per input instance and averaging the resulting model outputs. This results in:

- accelerating training and improve generalization over the original dropout
- reducing a computational cost, as the majority of computation time is expended in the lower layers, while the weights in the upper layers are shared
- achieving lower error rates and losses

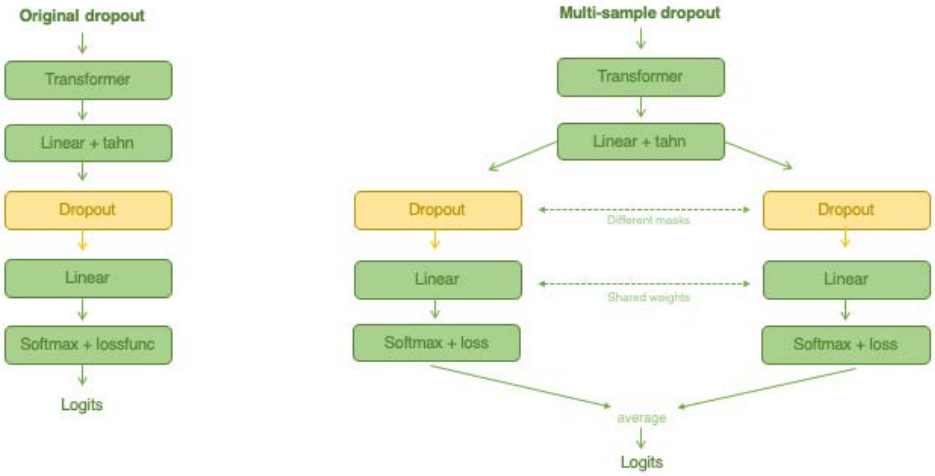


Figure 2: Overview of original dropout and multi-sample dropout.

3.3 Pooled Sentiment Model

This method for named entity sentiment analysis combines the use of special tokens, fine-tuning, regularization techniques, and cross-validation to create a comprehensive approach that addresses overfitting and improves the model's ability to accurately predict sentiment for specific entities in a given text.

We insert a special [SENTIMENT] token before the target entity in the input text. This token serves as an indicator for the model to focus on the context surrounding the entity when predicting sentiment. Further, we chose a transformer-based model and fine-tune it to extract sentiment from the embedding corresponding to the [SENTIMENT] token. In our experiments, we tested DistilBERT ((Kolesnikova et al., 2022)), BERT, and RoBERTa ((Liu et al., 2019)). During the experiments, overfitting emerged as a challenge. To address this, various techniques were employed, including weight decay, dropout, and the utilization of weights from models trained for different tasks (such as aspect-based sentiment analysis, sentiment analysis, and Named Entity Recognition (NER)). Also, some experiments were conducted using the Monte Carlo dropout approach at inference time, which involves applying dropout during the testing phase to create an ensemble-like effect, potentially improving generalization and uncertainty estimation.

The final model was trained using cross-validation, a technique that partitions the dataset into multiple folds, training and validating the ensemble of models on different subsets to ensure a more robust evaluation of its performance.

This approach proved to be quite robust, and had it not been for the superior method proposed, it would have secured the 2nd position on the leaderboard.

3.4 HAlf MAsked Model (HAMAM)

The model builds a contextualized representation of an entity and classifies it into three given classes "positive", "negative", and "neutral". As a backbone for building the representation, any transformer model can be selected. A transformer takes tokenized text as an input and produces vector representations $[h_1, ..., h_n]$ for each of the given tokens. Then two variants of entity representation are constructed:

- mean pooled $v_{mean} = (h_k + ... + h_m)/(m - k)$,
- max pooled $v_{max} = \text{Max}([h_k, ..., h_m])$, with taking maximum over the last dimension,

where k is the index of the first entity token and m is the index of its last token. Both v_{mean} and v_{max} are then passed through the classifier module, which consists of the following consecutive layers:

- linear transformation $[N \times N]$, where N is the size of the final hidden representation from transformer;
- hyperbolic tangent function;
- multi-sample dropout, described in the section 3.2;
- linear transformation from N-dimensional vector to 3-dimensional space of target classes.

The resultant three logits are averaged for cases of mean and max pooling: $l_{entity} = (l_{mean} + l_{max})/2$. But the values of l_{entity} are not used for the final prediction yet, because to avoid model overfitting to some particular words, another run of the model is performed at this point, but with the masked entity. The point is that in training data some entities might be overrepresented in one target class and underrepresented in any other, which may lead to bias in model predictions for such entities. Also while predicting any unseen entity, the model may utilize bias in the pre-trained representations of this entity. Masking the entity words (replacing them with '[MASK]' token) helps to mitigate this effect and forces the model to extract sentiment information from a context rather than prior knowledge of the entity itself.

The output of the masked run is a set of logits for three target classes l_{masked}, which are averaged with the l_{entity} before applying the argmax function to extract the predicted class. The complete architecture of the described approach is shown in figure 3. The intuition for keeping predictions from the model with an unmasked entity and averaging it with the masked run is that despite the mentioned problems about bias, the entity itself can contain useful information for creating accurate token representations by a transformer.

During training, the loss is calculated using final logits with the weighted cross-entropy function, where weights of 1 are assigned to the examples with positive and negative sentiments, while neutral examples have weights of 0.1. The lower weight of neutral examples is motivated by, firstly, the target competition metric, which concentrates on the quality of positive and negative predictions, and also by the fact that the neutral class dominates training data as there are approximately 2.5 times more neutral examples than there are positive and negative ones combined.

Another trick motivated by the target metric and which was tested with the HAMAM approach is the threshold on neutral class prediction. Instead of taking argmax of the final logits, we first apply softmax to get probabilities of each class and in cases when the neutral class has the highest probability, but its value is below some threshold, we select the most probable class from only "positive" and "negative".

Besides that, a well-known method for improving generalization – ensembling - was tested. In order to do that, we averaged the final logits from different transformer models trained on different subsets of training data. Specifically, the training dataset was split into five folds, each of which produces its own model trained on the four rest folds and the resultant models can be used for ensembling.

4 Experimental setup

Many experiments were carried out with different models and training setups. The final results for HAMAM were obtained with the following experimental setup.

The training dataset was split into 5 folds to perform cross-validation and eventually get 5 models, which can be ensembled for prediction on test data.

Training on each of the 5 parts of the initial dataset was conducted during 6 epochs with validation performed during each half-epoch. The checkpoint with the highest macro F1_{pn} on validation was selected to get a score on the dev and test sets.

Several transformers were tested as a backbone for HAMAM, namely, 'DeepPavlov/distilrubert-base-cased-conversational', 'sberbank-ai/ruBert-large', 'sberbank-ai/ruRoberta-large', etc. Final results were obtained with ensemble of 'sberbank-ai/ruRoberta-large', 'xlm-roberta-large' ((Conneau et al., 2019)), and 'google/rembert' ((Chung et al., 2021)).

The maximal learning rate for the backbone transformer model was set to 1e-5, while added weights

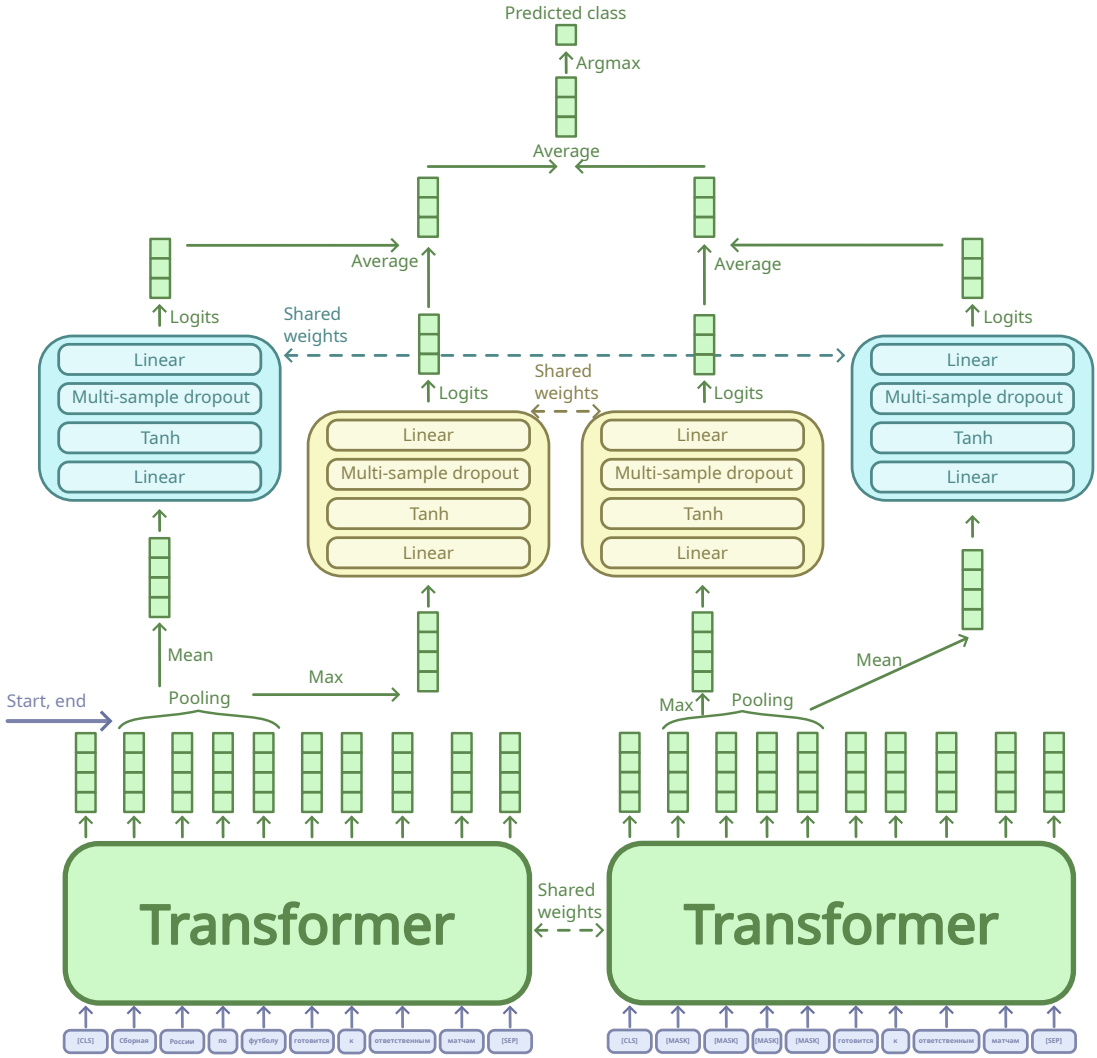


Figure 3: The architecture of HAMAM approach.

for classification were trained with the maximal learning rate of 1e-4. The initial learning rate was set to 0 for all weights and warmed up to its maximal value during the first tenth part of the total training steps number and then linearly decayed to 0.

The batch size of 8 was used, and the dropout rate in classification layers was set to 0.5. In the case of multi-sample dropout, the number of samples was set to 5.

5 Results

Table 1 shows results for models based on the various combinations of HAMAM parts. In the most basic form, such a model performs mean pooling of the given entity and then classifies it (first line in table). In the second line, we add class weights in the cross-entropy loss calculation. The third line also adds a multi-sample dropout to this configuration, but due to the worsening of the results on local cross-validation, we removed the multi-sample dropout in the fourth line and added an entity masking trick, after which the model can be marked as HAMAM. Here we can see a large increase in cross-validation score, so we assume that entity masking is indeed helpful in avoiding overfitting and increases the model's generalizing ability. The fifth line introduces a more sophisticated approach to entity vector pooling – a combination of 'mean' and 'max' poolings. Lines six and seven present another test of multi-sample dropout (this time model already has entity masking) both with mean and mean-max

pooling correspondingly. Based on the results of cross-validation alone it is hard to tell if the last two additions (mean-max pooling and multi-sample dropout) are really helpful for the model but based on the general considerations it was decided to use the full HAMAM model for test submission. Dev scores for configurations other than full HAMAM were not obtained during the development phase of the competition, so they did not influence the model selection for the test phase.

Model configuration	Macro $F1_{pn}$			
	Local cross-validation score, mean +/- std	Dev set score, mean +/- std	Dev set score from 5-fold ensemble	Test set score from 5-fold ensemble
Mean pooling	65.11 +/- 1.54	66.04 +/- 1.11	69.41	61.9
Mean pooling, class weights in loss	66.11 +/- 1.05	66.61 +/- 0.79	70.47	65.25
Mean pooling, class weights in loss, multi-sample dropout	65.31 +/- 1.23	66.72 +/- 0.66	70.85	62.84
Mean pooling, class weights in loss, entity masking	67.83 +/- 0.45	67.38 +/- 0.54	70.86	65.42
Mean-max pooling, class weights in loss, entity masking	67.63 +/- 1.07	67.45 +/- 0.24	69.14	65.73
Mean pooling, class weights in loss, entity masking, multi-sample dropout	67.57 +/- 1.20	66.99 +/- 0.78	70.49	65.67
Mean-max pooling, class weights in loss, entity masking, multi-sample dropout (full HAMAM)	67.73 +/- 1.22	67.20 +/- 0.31	69.52	66.25

Table 1: Macro $F1_{pn}$ score comparison from various configurations of HAMAM model based on 5-fold cross-validation.

Table 2 presents final results of our models on dev and test sets. HAMAM result with the threshold on the prediction of neutral class yielded a small increase in performance on the dev set, so this model configuration was used for the final submission on the test set, which gave our final test score of macro $F1_{pn}$ = 66.67.

6 Error Analysis

The first thing we found when manually analyzing errors is rather ambiguous labeling. Several such examples are shown in table 3.

Assessing human-level performance on this dataset could be intriguing. Typically, neutral sentiment tends to be mistaken for negative and positive ones, as anticipated. Instances where the model assigns a negative sentiment to a positive label or vice versa are highly uncommon and can be attributed to ambiguous labeling. It is evident that the model has overfitted for words with highly contrasting sentiments and when they are closely associated with the entity. For example: “полиция задержала двоих человек возле суда и одного — внутри.” The model returns "negative" for “полиция” entity.

Model configuration	Macro $F1_{pn}$	
	Dev set score	Test set score
Pooled Sentiment model (5 sberbank-ai/ruRoberta-large ensemble)	69.92	65.68
HAMAM (5 sberbank-ai/ruRoberta-large ensemble)	69.52	66.25
HAMAM (5 sberbank-ai/ruRoberta-large + 4 xlm-roberta-large + 2 google/rembert ensemble)	70.86	67.0
HAMAM (5 sberbank-ai/ruRoberta-large + 4 xlm-roberta-large + 2 google/rembert ensemble) + neutral class 0.55 threshold	70.94	66.67

Table 2: Macro $F1_{pn}$ score results from various models and their ensembles.

Sentence	Entity	Dataset true label	Predicted label
На момент смерти 54-летняя журналистка расследовала коррупцию в России и нарушения прав человека в Чечне, где ранее федеральное правительство подавило попытки сепаратистов создать исламистское государство.	правительство	negative	neutral
58-летний Чавес одержал в октябре победу над Каприлесом с большим численным перевесом, завоевав еще один шестилетний срок на посту президента.	Чавес	negative	positive
Это был первый случай, когда сирийская армия обстреляла предполагаемых повстанцев в Ливане, который старается соблюдать нейтралитет в гражданской войне в Сирии	Ливане	positive	neutral

Table 3: Examples of wrong predictions by HAMAM and of ambiguity in labeling.

7 Conclusion

In this paper, we studied different approaches for solving named entity sentiment classification task in the RuSentNE-23 competition. We presented the zero-shot technique, and also thoroughly investigated fine-tuning approach finding out that overfitting to the sentiment of certain entities is its main drawback. We described several attempts at mitigating overfitting, among which replacing entity with '[MASK]' tokens showed the best result. Using this trick, we developed a new approach, which after ensembling several transformer models scored macro $F1_{pn}$ = 66.67 and reached first place in the competition.

References

Hyung Won Chung, Thibault Févry, Henry Tsai, Melvin Johnson, and Sebastian Ruder. 2021. Rethinking embedding coupling in pre-trained language models. *// 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

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"*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *CoRR*, abs/1905.09788.

Alina Kolesnikova, Yuri Kuratov, Vasily Konovalov, and Mikhail Burtsev. 2022. Knowledge distillation of russian language models with reduction of vocabulary.

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. *CoRR*, abs/1907.11692.

Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. // Yoshua Bengio and Yann LeCun, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*.

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, July.

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, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, P 1631–1642. ACL.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.*, 15(1):1929–1958.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. // *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers*, P 1556–1566. The Association for Computer Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. // Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, P 5998–6008.

Joosung Yoon and Hyeoncheol Kim. 2017. Multi-channel lexicon integrated cnn-bilstm models for sentiment analysis. // Lun-Wei Ku and Yu Tsao, *Proceedings of the 29th Conference on Computational Linguistics and Speech Processing, ROCLING 2017, Taipei, Taiwan, November 27-28, 2017*, P 244–253. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Prosodic portrait of the Russian connector PRICHOM in the mirror of the multimedia corpus

Podlesskaya V. I.
Institute of linguistics, Russian Academy of Sciences;
Russian State University for the Humanities, Moscow, Russia
verapodlesskaya@gmail.com

Abstract

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses prosodic features of discourse fragments introduced by the connector *prichom* 'and besides'. The data of instrumental and perceptual analysis show that the fragment with *prichom* has communicative-prosodic autonomy: firstly, it has an internal thematic structure with an obligatory rheme and an optional theme; and secondly, there is a prosodic break before this fragment. The autonomy of the fragment introduced by *prichom* is preserved in a variety of contexts: (i) both in cases where this fragment is a complete clause and when it is a fragmented clause; (ii) both in those cases when the previous fragment is prosodically realized as final (projecting no continuation), and when it is realized as non-final (projecting continuation); (iii) both in those cases when the fragment introduced by *prichom* is an element of the main narrative chain, and when it is inserted parenthetically inside another fragment. In addition to the above, a fragment with *prichom* can form a separate turn in the conversation. Thus, the detected prosodic features of the fragment with *prichom* make it possible to objectify the idea earlier expressed in the literature (Kiselyova 1971, Vinogradov 1984, Inkova 2018, inter alia): that structures with *prichom* are built in two "communicative steps", or that they are used to express "concomitance established at the level of speech acts ". Clauses connected by the relationship of syntactic subordination quite often lose their prosodic autonomy (Podlesskaya 2014 a, b), and vice versa, clauses in coordinated constructions tend to retain prosodic autonomy. Therefore, the prosodic autonomy of the components of the construction with *prichom*, retained in various contexts, speaks in favor of its coordinated status, while a number of syntactic tests proper speak of the opposite.

Keywords: clause combining, subordination and coordination, prosodic autonomy, spoken discourse
DOI: 10.28995/2075-7182-2023-22-442-451

Просодический портрет коннектора ПРИЧЕМ в зеркале мультимедийного корпуса

Подлесская В. И.
Институт языкознания РАН,
РГГУ, Москва, Россия
verapodlesskaya@gmail.com

Аннотация

На материале мультимедийного подкорпуса НКРЯ рассматриваются просодические свойства дискурсивных фрагментов, вводимых коннектором *причем*. Данные инструментального и перцептивного анализа показывают, что фрагмент с ПРИЧЕМ обладает коммуникативно-просодической автономностью: во-первых, он обладает внутренней тема-рематической структурой с обязательной ремой и опциональной темой; и во-вторых, перед этим фрагментом имеется просодический шов. Автономность фрагмента, вводимого коннектором ПРИЧЕМ, сохраняется в самых разных контекстах: (i) и в тех случаях, когда этот фрагмент представляет собой полную клаузу, и тогда, когда это фрагментированная клауза; (ii) и в тех случаях, когда предшествующий фрагмент реализуется с интонацией завершенности, и тогда, когда он реализуется с интонацией незавершенности; (iii) и в тех случаях, когда фрагмент, вводимый ПРИЧЕМ, является элементом основной нарративной цепочки, и тогда, когда он вставлен внутрь другого фрагмента в статусе парентезы. В дополнение к перечисленному фрагмент с ПРИЧЕМ может формировать отдельную реплику в диалоге. Тем

самым, обнаруженные комммуникативно-просодические признаки фрагмента с ПРИЧЕМ позволяют объективировать высказывавшуюся в литературе (Киселева 1971, Виноградов1984, Inkova 2018, inter alia). идею о том, что конструкции с ПРИЧЕМ строятся в два «коммуникативных шага» или что они используются для выражения «иллокутивного сопутствования». Клаузы, связанные отношением синтаксического подчинения, достаточно часто утрачивают просодическую автономность (Подлесская 2014 а,б), и напротив, клаузы в составе сочинительных конструкций имеют тенденцию сохранять просодическую автономность. Поэтому продемонстрированная нами устойчивая просодическая автономность компонентов конструкции с ПРИЧЕМ, говорит в пользу ее сочинительного статуса, в то время как ряд собственно синтаксических тестов говорит об обратном.

Ключевые слова: полипредикация**,** сочинение и подчинение, просодическая автономия, устный дискурс

1. Постановка задачи

Предметом данного исследования являются конструкции с коннектором ПРИЧЁМ в русском языке. Этот коннектор может присоединять как полную клаузу, см. (1), так и фрагмент клаузы, который можно квалифицировать как результат сокращения любого материала клаузы, не исключая подлежащее и/или сказуемое, см. (2):

(1)
Петя приедет уже сегодня, причем он приедет с Машей / причем приедет с Машей / причем с Машей.
(2)
....причем ~~он~~ приедет с Машей
....причем ~~он приедет~~ с Машей

Конструкции с ПРИЧЕМ обычно обсуждаются в литературе в составе так называемых присоединительных конструкций, см. Кузнецова 1968, Киселева 1971, Виноградов 1984, Чжон 2003 inter alia. Частеречный статус этого коннектора окончательно не прояснен, в частности, нет убедительного ответа на вопрос, является ли этот коннектор союзом. В РГ-80 он осторожно именуется «союзным аналогом», см. РГ-80, Т.2, § 3151.

Нет убедительного ответа и на вопрос о том, является ли конструкция с ПРИЧЕМ сочинительной или подчинительной (ср. обсуждение этого вопроса с опорой на семантическую аргументацию в Weiss 1991). Набор стандартных синтаксических тестов на сочинение и подчинение (см., например, Пекелис 2015) дает противоречивый результат. С одной стороны, критерий морфосинтаксического локуса говорит в пользу сочинения – если, например, всю конструкцию поместить в контекст, требующий сослагательного наклонения, то это требование распространяется на оба компонента конструкции с ПРИЧЕМ:

(3)
[Я прошу, чтобы] Петя ~~приедет~~ приехал уже сегодня, причем [чтобы] он ~~приедет~~ приехал с Машей.

С пользу сочинения говорит и тот факт, что ни полная, ни фрагментированная клауза с ПРИЧЕМ не может выдвигаться в препозицию:

(4)
**Причем он приедет с Машей/ причем приедет с Машей / причем с Машей, Петя приедет сегодня,*

С другой стороны, ПРИЧЕМ, в отличие от сочинительных союзов, не обязательно располагается в начале вводимой им клаузы. НКРЯ массово выдает такие примеры:

(5)
со всех трибун клялись — абсолютно искренно — решить наконец-то продовольственную проблему, давнюю причём. [Анатолий Азольский. Лопушок // «Новый Мир», 1998]
(6)
Такая пустота настала/ ужасная причём пустота. [Д/ф из цикла «Письма из провинции» (ТК «Культура») (2014)]

Против сочинения говорит и тот факт, что фрагмент с ПРИЧЕМ может парентетически вставляться внутрь клаузы:

(7)

Видите ли/ в последние несколько лет/ просматривая газеты/ я убедился/ что некоторые методы уголовного мира/ причём самые грязные методы/ стали безнаказанно применяться в политике. [Игорь Масленников, Дойль Артур Конан. Приключения Шерлока Холмса и доктора Ватсона. Двадцатый век начинается, к/ф (1986)]

Неоднозначный синтаксический статус конструкции с ПРИЧЕМ можно рассматривать как проекцию особой дискурсивной функции этого коннектора. С одной стороны, фрагмент, который вводится коннектором ПРИЧЕМ, имеет меньший дискурсивный вес, чем присоединяющая его клауза: «сообщаемое во второй части подается в виде примечания к сказанному, служит уточняющим пояснением или поправкой к предшествующей части» (РГ-80, Т.2, § 3151). С другой стороны, говорящий намеренно выделяет фрагмент, вводимый ПРИЧЕМ, в самостоятельный дискурсивный шаг. Это решение сродни тому, что говорящий принимает, когда решает парцеллировать некоторый сегмент текста, который при иных обстоятельствах мог бы быть интегрированным с предшествующим сегментом в единый дискурсивный шаг (см. дискуссию о сходстве и отличиях парцелляции и присоединения в Виноградов1984). Например, предложение *Петя читает детективы на английском языке* говорящий может произнести за один дискурсивный шаг, а может выделить сообщение про английский язык в отдельный дискурсивный шаг: *Петя читает детективы. На английском языке.* Причины для такого решения могут быть различными – например, при развертывании дискурса говорящий мог с запозданием обнаружить, что информация об английском языке нужна, и добавляет ее постфактум; или, наоборот, уже в исходной точке говорящий решает, что на эту информацию нужно обратить специальное внимание слушающего и сообщение об этом формируется как отдельное высказывание. Меньший дискурсивный вес вводимого фрагмента и, одновременно, его дискурсивная автономность и составляют основное своеобразие присоединительного ПРИЧЕМ. Это своеобразие в той или иной форме отмечалось исследователями. Так, Н. П. Киселева (1971: 4) следующим образом характеризует данную особенность: «[присоединение] проявляется в прерывистой, двухактной реализации грамматической модели предложения в речи, обусловленной тем, что формально-грамматическая основа предложения организована не одним коммуникативным заданием, а двумя – основным и добавленным». Близкую точку зрения находим у О.Ю. Иньковой, которая предлагает усматривать между компонентами конструкции с ПРИЧЕМ отношение сопутствования на уровне речевых актов («иллокутивное сопутствование», Inkova 2018).

Вместе с тем, пока нет полной ясности в вопросе о том, какие языковые факты могли бы «объективировать» дискурсивное отношение иллокутивного сопутствования между компонентами присоединительной конструкции, или, иначе говоря, какие данные могли бы убедительно свидетельствовать о том, что компоненты конструкции с ПРИЧЕМ обладают дискурсивной автономией. Цель данной работы – частично восполнить этот пробел.

В подтверждение дискурсивной автономии фрагмента, вводимого коннектором ПРИЧЕМ, мы предъявим ряд просодических аргументов. Мы постараемся показать, что компоненты конструкции с ПРИЧЕМ образуют две коммуникативно-просодических составляющих (в терминах Янко 2008), или две элементарных дискурсивных единицы (ЭДЕ, в терминах Кибрик, Подлесская 2009). Для этого будут проанализированы коммуникативно-

значимые акценты в конструкции, направление движения тона в акцентах и характер просодического шва между компонентами конструкции. Источником данных служит мультимедийный подкорпус НКРЯ, будут приведены результаты как перцептивного, так и инструментального анализа с использованием компьютерного анализатора речи PRAAT (Boersma, Weenink 2021). Этой проблематике посвящен раздел 2 – основной раздел работы. В разделе 3 суммируются общие итоги исследования.

2. Коммуникативно-просодическая автономия фрагмента с ПРИЧЕМ

Важнейшая особенность конструкций с ПРИЧЕМ состоит в том, что оба компонента конструкции имеют автономную коммуникативно-просодическую организацию; каждый из них имеет собственную рему, плюс – могут иметь и собственную тему. Так в следующем примере оба компонента конструкции имеют и тему, и рему[1]:

(8)
Те/ кто не справляются с этим/ увы/ гибнут/ причём здесь прежние заслуги/ как говорится/ не учитываются. [Сергей Филонович. Жизненный цикл организаций (2018)]

Те кто не /справля̲ются с этим,
/–ув̲ы-ы̲,
(ɥ 0.35)
(ə 0.23)
\г̲ибнут.
(ˀ 0.13)
\Причём (ə 0.38) /здесь прежние /заслуги как говорится не \уч̲итываются.

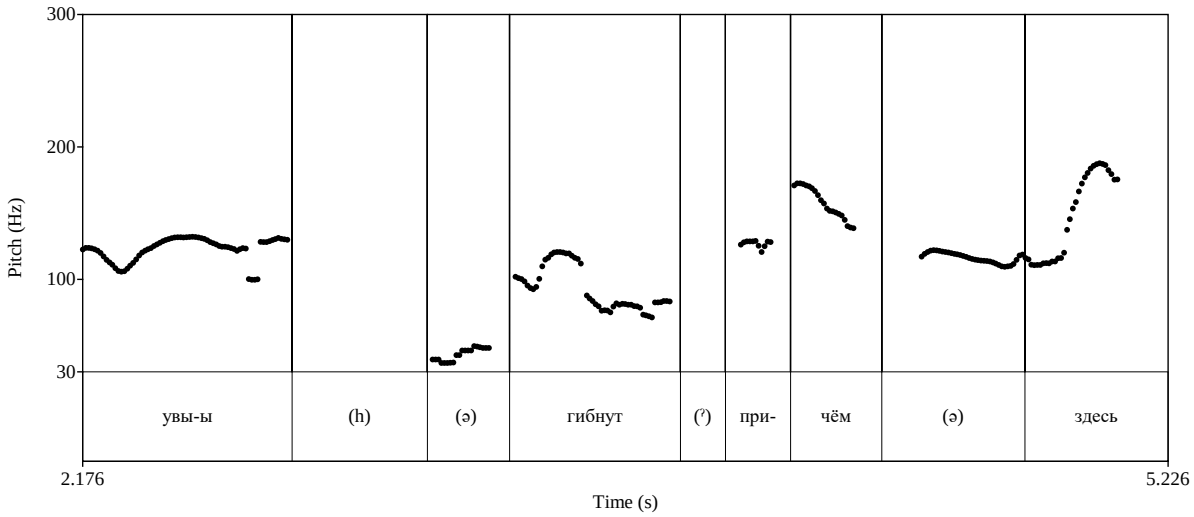


Рисунок 1. Интонограмма к примеру (8)

[1] Примеры приводятся в полной графической форме, как она дана в МУРКО, при необходимости приводится также просодическая транскрипция части примера с разметкой движения тонов и локализацией фразовых акцентов и интонограмма в формате анализатора PRAAT. Для указания на направление движения тона иконически используются знаки «/», «\» и «–». Ударный слог слова – носителя рематического акцента подчеркивается. О других деталях используемой системы просодической транскрипции см. Кибрик, Подлесская (ред.) 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.

В вышеприведенном примере в первом компоненте тематическая составляющая – *те, кто не справляются,* с подъемом тона по типу ИК3 в терминах интонационных конструкций (Брызгунова 1982) на ударном слоге глагола *справляются*, а рематическая – *гибнут*, с падением тона по типу ИК1. Во втором компоненте два тематических подъема тона – на словах *здесь* и *заслуги* – и рематическое падение на слове *учитываются*. Просодическая автономность компонентов в данном случае поддерживается также следующими необязательными, но часто встречающимися способами углубления просодического шва между ними: (а) заполненная пауза между компонентами (в данном случае реализуется как скрип); (б) ресет частоты основного тона – второй компонент начинается не с той частоты, на которой завершен первый, а с типичной для данного говорящего средней частоты; (в) акцентирование самого коннектора – ПРИЧЕМ произносится с выраженным падением тона на ударном слоге (термин «просодический шов» как удачный аналог английского prosodic break был введен а работах О.Ф.Кривновой и ее коллег, см., например, Князев, Кривнова, Моисеева 2016).

Существенно то, что просодическая автономия второго компонента обязательна и в тех случаях, когда он – не полная клауза, а фрагмент клаузы. Так, в следующем примере полномерную тема-рематическую структуру имеет не только первый компонент, представляющий собой полную клаузу, но и второй – представляющий собой определение к эллиптированной именной вершине:

(9)
Бихевиоризм использовал в качестве своей методологии позитивизм/ причём понимаемый очень вульгарно и примитивно. [Иван Иванчей. Изучение сознания в когнитивной психологии (2017)]

/Бихевиоризм использовал в качестве своей /методологии \позитивизм.
Причём /понимаемый очень /вульгарно и \примитивно.

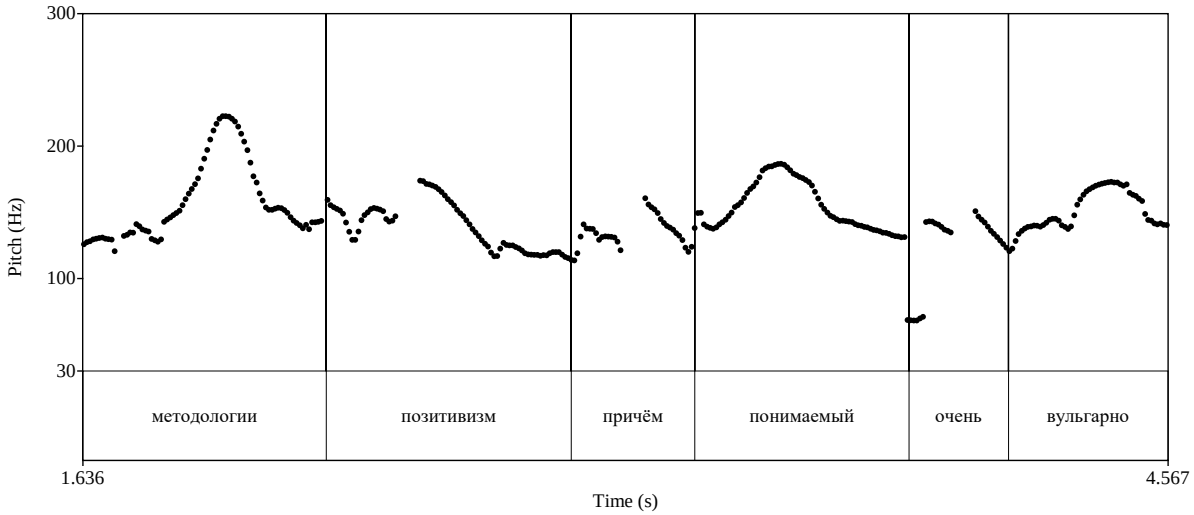


Рисунок 2. Интонограмма к примеру (9)

В вышеприведенном примере в первом компоненте тематические подъемы тона – на словах *бихевиоризм* и *методологии*, а рематическое падение тона – на слове *позитивизм*. Во втором компоненте тематический подъем тона на слове *понимаемый* и *заслуги* и сложная рема – сочиненная группа с подъемом тона на первом члене (*вульгарно*) и падением – на втором (*примитивно*).

Коммуникативно-просодическая автономность фрагмента с ПРИЧЕМ подкрепляется тем, что он может являться сферой действия правил расстановки фразовых акцентов, связанных с контрастом. В частности, во фрагменте с ПРИЧЕМ действует правило переноса акцента на согласованное определение в контексте контраста, см. следующий пример:

(10)
Гражданин Зубов! Бумаги бумагами/ а… ведь… главное – это согласие девочки. И причём охотное её согласие. [Искра Бабич, Валентин Михайлов. Мужики, к/ф (1981)]

/гла̲вное,
надо согласие \де̲вочки!
(0.35)
И /причём \охо̲тное её согласие.

Здесь во фрагменте с ПРИЧЕМ происходит коммуникативно обусловленный сдвиг рематического акцента. Сообщение строится в два коммуникативных шага. На первом шаге рема – *согласие девочки*, акцентоноситель – *девочки*; в соответствии с синтаксическими правилами расстановки акцентов в рематической составляющей, акцентносителем является несогласованное определение, (см. Ковтунова 1976, Янко 2008). На втором – автономном – шаге вводится новая рема, контрастная, и по общим правилам акцент располагается на согласованном определении *охотное*.

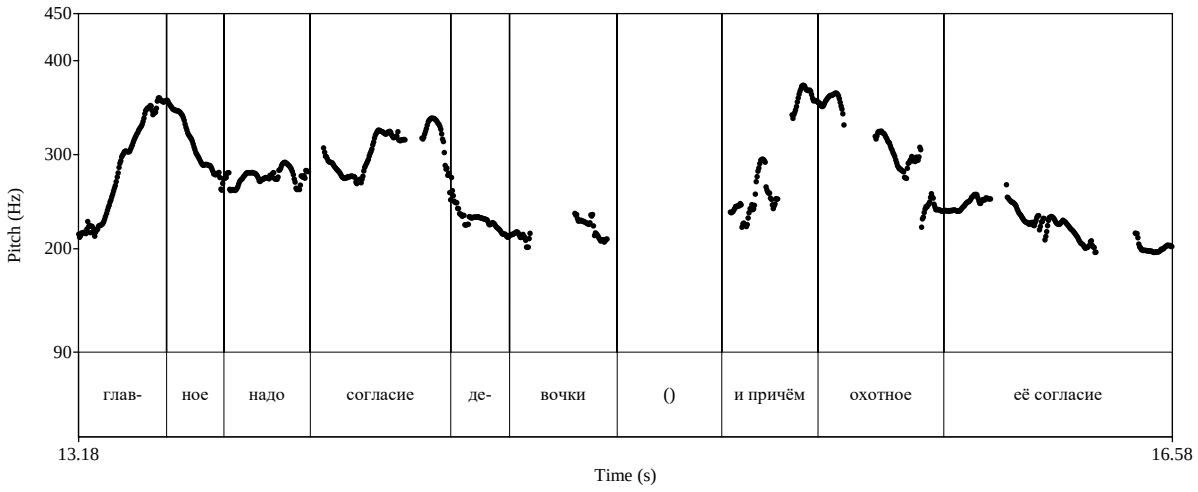


Рисунок 3. Интонограмма к примеру (10)

В приведенных выше примерах (8)-(10) первый компонент произносится с интонацией завершенности перед ПРИЧЕМ, а именно, как сказано выше – с падением тона по типу ИК1 в терминологии интонационных конструкций на словах, которые являются акцентоносителями ремы (*гибнут, позитивизм* и *девочки*, соответственно). Интересно, однако, что просодическая автономизация компонентов сохраняется и в тех случаях, когда первый компонент реализуется с интонацией незавершенности. Оба компонента и в этом случае сохраняют каждый свою тема-рематическую структуру, в том числе, каждый имеет свою рему. Так, в следующем примере, перед ПРИЧЁМ имеется выраженный просодический шов – пауза и тональный ресет, т.е. фрагмент с ПРИЧЁМ начинается с типичной для данного говорящего средней частоты основного тона. Сам же этот фрагмент имеет и собственную тему – акцентоноситель темы, словоформа *картинка* реализуется с типичным тематическим подъемом тона, и собственную рему –

акцентоноситель ремы, словоформа *агента* реализуется с типичным рематическим падением тона. При этом фрагмент, предшествующий ПРИЧЕМ, реализуется с интонацией незавершенности – восходящим движением тона на слове *мира*:

(11)
У меня/ у вас/ у каждого из нас есть цельная картинка окружающего мира/ причём эта картинка включает в себя агента/ субъекта/ который наблюдает за внешним миром. [Иван Иванчей. Изучение сознания в когнитивной психологии (2017)]

У каждого из /нас есть (ə 0.67) /цельная картинка окружающего /мира,
(0.27)
причём эта /картинка включает в себя \агента.

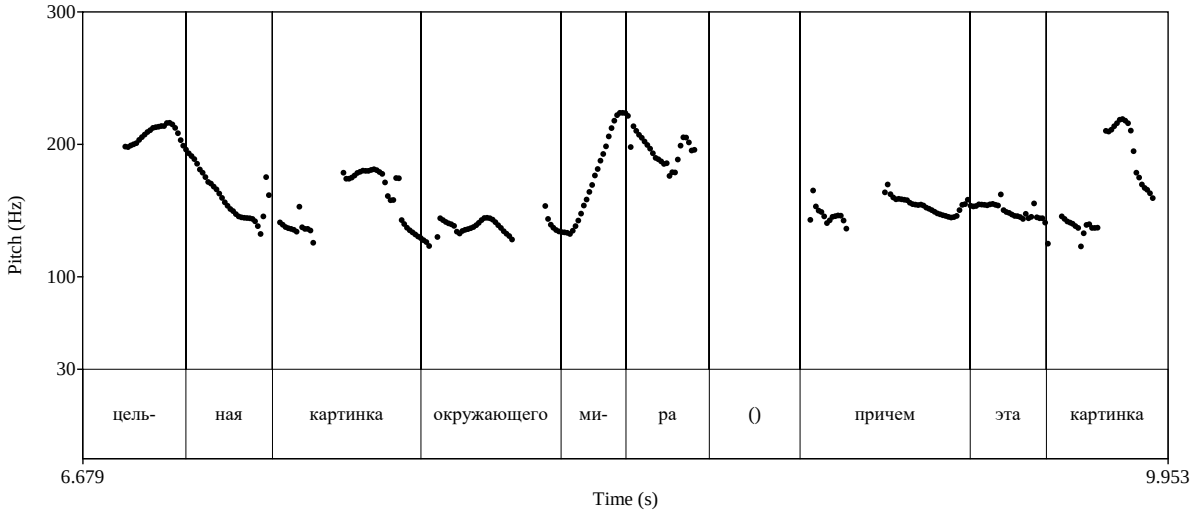


Рисунок 4. Интонограмма к примеру (11)

Фрагмент, вводимый ПРИЧЕМ, сохраняет свою просодическую автономность даже в тех случаях, когда он является вставкой внутри клаузы. Так, в следующем примере вставка *причём авторы там уважаемых журналов типа журнала «Вайрд»* размещается между подлежащим и сказуемым, при этом она сама имеет неэлементарную структуру с тремя последовательными подъемами тона на словоформах *авторы*, *журналов* и *Уайрд*. Самый высокий подъем по типу ИК3 на завершающей вставку словоформе *Уайрд* является дублем подъема тона на словоформе *люди* – подлежащем обрамляющей клаузы. Этот дублирующий подъем тона и является сигналом того, что фрагмент, введенный ПРИЧЕМ является уточняющим примечанием – в соответствии со своей основной дискурсивной функцией:

(12)
И серьёзные люди/ причём авторы там уважаемых журналов типа журнала «Вайрд» писали огромные статьи на основе там кучи интервью с игроками в го [Лекция Андрея Себранта, директора по маркетингу компании «Яндекс» (2017)]

И-и серьёзные /люди —
причём-м /авторы там-м (0.20) уважаемых /журналов,
типа журнала /«Уайрд»,
— писали /**огро-омные** \статьи!

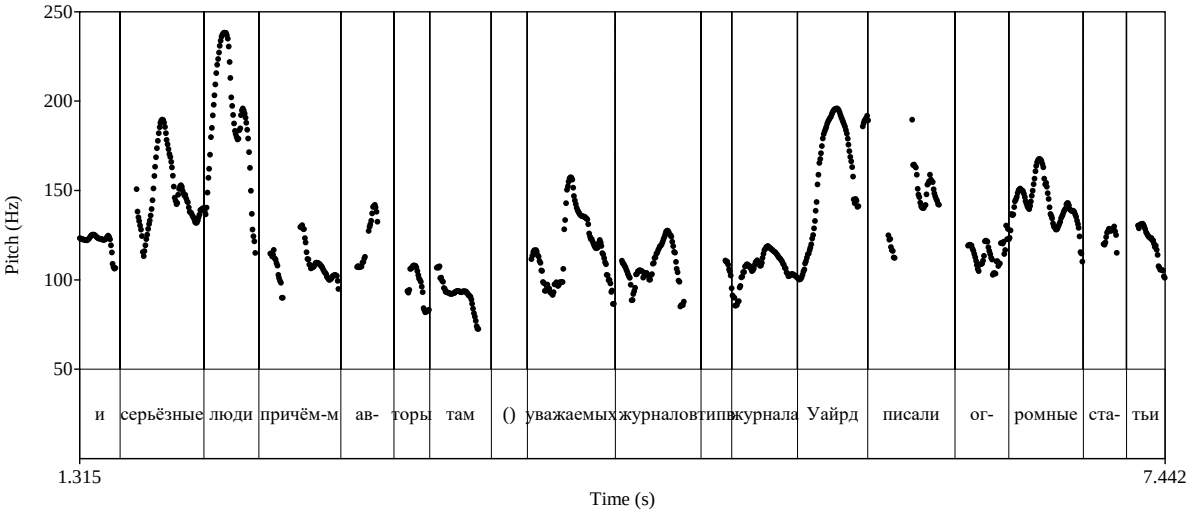


Рисунок 5. Интонограмма к примеру (12)

Наконец, просодическая автономность фрагмента с ПРИЧЕМ может дополняться дискурсивной автономностью: такой фрагмент может формировать отдельную реплику в диалоге. Так в следующем примере говорящий В начинает свою реплику с ПРИЧЕМ, даже не дожидаясь, пока его собеседник, говорящий М, закончит свою реплику. В транскрипции наложение части реплик показано квадратными скобками:

(13)
[Михеев Савва Михайлович, муж, историк] Это колоссальное/ конечно/ количество для [нрзб] одной церкви.
[Виноградов Андрей Юрьевич, муж, историк] Причём/ что очень интересно/ все почти восточнославянские/ то есть русские.
[Михеев Савва Михайлович, муж, историк] Да. Среди кириллическких записей только одна южнославянская.
[Алексей Гиппиус, Савва Михеев, Андрей Виноградов, Фекла Толстая. Беседа об эпиграфике в программе «Наблюдатель» (2016)]

М: Это ↑\колосс<u>а</u>льное конечно количество для-а [одной церкви!]
В: [\Причём что очень /и]нтер<u>е</u>сно,
/\все-е почти ↑\восточнослав<u>я</u>нские!
То есть \р<u>у</u>сские.

3. Итоги

Итак, мы продемонстрировали, что фрагмент, вводимый коннектором ПРИЧЕМ, обладает коммуникативно-просодической автономностью – (а) он обладает внутренней тема-рематической структурой с обязательной ремой и опциональной темой; и (б) перед этим фрагментом имеется просодический шов.

Коммуникативно-просодическая автономность фрагмента, вводимого коннектором ПРИЧЕМ, сохраняется в самых разных контекстах:

- и в тех случаях, когда этот фрагмент представляет собой полную клаузу, и тогда, когда это фрагментированная клауза;
- и в тех случаях, когда предшествующий фрагмент реализуется с интонацией завершенности, и тогда, когда он реализуется с интонацией незавершенности;

- и в тех случаях, когда фрагмент, вводимый ПРИЧЕМ, является элементом основной нарративной цепочки, и тогда, когда он вставлен внутрь другого фрагмента в статусе парентезы.

В дополнение к перечисленному фрагмент с ПРИЧЕМ может формировать отдельную реплику в диалоге. Тем самым, обнаруженные комммуникативно-просодические признаки фрагмента с ПРИЧЕМ позволяют объективировать высказывавшуюся в литературе (Киселева 1971, Виноградов1984, Inkova 2018, inter alia). идею о том, что конструкции с ПРИЧЕМ строятся в два «коммуникативных шага» или что они используются для выражения «иллокутивного сопутствования». Клаузы, связанные отношением синтаксического подчинения, достаточно часто утрачивают просодическую автономность (Подлесская 2014 а,б), и напротив, клаузы в составе сочинительных конструкций имеют тенденцию сохранять просодическую автономность. Поэтому продемонстрированная нами устойчивая просодическая автономность компонентов конструкции с ПРИЧЕМ, говорит в пользу ее сочинительного статуса, в то время как ряд собственно синтаксических тестов говорит об обратном.

Благодарности

Работа поддержана грантом РНФ № № 22-18-00120.

Литература

[1] Виноградов, А. А. (1984). Структура и функции присоединительных конструкций в современном русском литературном языке (Дисс. канд. филол. наук, Ужгородский государственный университет). Ужгород.

[2] Vinogradov, A. A. (1984). Struktura i funktsii prisoedinitel'nykh konstruktsii v sovremennom russkom literaturnom yazyke (Diss. kand. filol. nauk, Uzhgorodskii gosudarstvennyi universitet). Uzhgorod.

[3] Кибрик А. А., Подлесская В. И. (Ред.) (2009). Рассказы о сновидениях: Корпусное исследование устного русского дискурса. Москва: Языки славянских культур

[4] 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.

[5] Киселева, Н. П. (1971). Средства выражения присоединения между компонентами, связанными подчинительными отношениями в современном русском языке (Автореферат дисс. канд. филол. наук, МГПИ им. В. И. Ленина). Москва.

[6] Kiseleva, N. P. (1971). Sredstva vyrazheniya prisoedineniya mezhdu komponentami, svyazannymi podchinitel'nymi otnosheniyami v sovremennom russkom yazyke (Avtoreferat diss. kand. filol. nauk, MGPI im. V. I. Lenina). Moskva.

[7] Князев С.В., Кривнова О.Ф., Моисеева Е.В (2016). Исследования просодического членения звучащего текста на материале русского языка // Вестник Московского университета. Серия 9: Филология, Изд-во Моск. ун-та (М.), № 4, 17-44.

[8] Knyazev S.V., Krivnova O.F., Moiseeva E.V (2016). Issledovaniya prosodicheskogo chleneniya zvuchashchego teksta na materiale russkogo yazyka // Vestnik Moskovskogo universiteta. Seriya 9: Filologiya, Izd-vo Mosk. un-ta (M.), № 4, 17-44.

[9] Ковтунова И. И. (1976). Современный русский язык. Порядок слов и актуальное членение предложения. М.: Просвещение.

[10] Kovtunova I. I. (1976). Sovremennyi russkii yazyk. Poryadok slov i aktual'noe chlenenie predlozheniya. M.: Prosveshchenie.

[11] Кузнецова, О. Я. (1968). Сложное предложение с союзом «причем» (Автореферат дисс. канд. филол. наук, Ростовский государственный университет). Ростов-на-Дону.

[12] Kuznetsova, O. Ya. (1968). Slozhnoe predlozhenie s soyuzom «prichem» (Avtoreferat diss. kand. filol. nauk, Rostovskii gosudarstvennyi universitet). Rostov-na-Donu.

[13] Пекелис О.Е. (2015). Сочинение и подчинение. Материалы для проекта корпусного описания русской грамматики (http://rusgram.ru). На правах рукописи. М.

[14] Pekelis O.E. (2015). Sočinenie i podčinenie [Coordination and subordination]. Materials for the Russian corpus grammar project (http://rusgram.ru). Ms. Moscow.

[15] Подлесская В.И. (2014а). Просодия против синтаксиса в русских относительных предложениях // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН / Т. X. Ч. 2. Русский язык: грамматика конструкций и лексико-семантические подходы / Ред. тома С. С. Сай, М. А. Овсянникова, С. А. Оскольская. СПб.: Наука, 537-567.

[16] Podlesskaya V.I. (2014a). Prosodiya protiv sintaksisa v russkikh otnositel'nykh predlozheniyakh // Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovanii RAN / T. X. Ch. 2. Russkii yazyk: grammatika konstruktsii i leksiko-semanticheskie podkhody / Red. toma S. S. Sai, M. A. Ovsyannikova, S. A. Oskol'skaya. SPb.: Nauka, 537-567.

[17] Подлесская В.И. (2014б). О просодических симптомах интеграции в конструкциях с сентенциальными актантами // Acta Linguistica Petropolitana. Труды Института лингвистических исследований РАН / Т. X. Ч.3/ Ред. тома С. Ю. Дмитренко, Н. М. Заика. СПб.: Наука, 554-566.

[18] Podlesskaya V.I. (2014b). O prosodicheskikh simptomakh integratsii v konstruktsiyakh s sententsial'nymi aktantami // Acta Linguistica Petropolitana. Trudy Instituta lingvisticheskikh issledovanii RAN / T. X. Ch.3/ Red. toma S. Yu. Dmitrenko, N. M. Zaika. SPb.: Nauka, 554-566.

[19] РГ-80 (1980). Русская грамматика: В 2 т. / Под ред. Н.Ю. Шведовой. М.,

[20] RG-80 (1980). Russkaia grammatika [Russian grammar]: In 2 vol. Shvedova N. Yu. (ed.). Moscow, Nauka Publ. 1980. (In Russ.)

[21] Чжон, Х. Х. (2003). Присоединительные скрепы в современном русском языке: синтаксис и семантика (Дисс. канд. филол. наук, МГУ имени М. В. Ломоносова). Москва.

[22] Chzhon, Kh. Kh. (2003). Prisoedinitel'nye skrepy v sovremennom russkom yazyke: sintaksis i semantika (Diss. kand. filol. nauk, MGU imeni M. V. Lomonosova). Moskva.

[23] Янко Т. Е. (2008). Интонационные стратегии русской речи в сопоставительном аспекте. Москва: Языки славянских культур.

[24] 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.

[25] Boersma, Paul & Weenink, David (2021). Praat: doing phonetics by computer [Computer program]. Version 6.1.38, 2021. Access mode: http://www.praat.org/

[26] Lehmann, C. (1988). 'Towards a typology of clause linkage', in J. Haiman and S. A. Thompson (eds.), Clause combining in grammar and discourse, Amsterdam: Benjamins, 181 – 226.
Inkova, O. (2018). Коннекторы русского языка с формантом при: корпусное исследование. Russian Linguistics 42, 159–190

[27] Inkova O. (2018). Connectors in Russian with the element pri: a corpus-based study // Russian Linguistics 42, 159–190 (In Russ.)

[28] Weiss, D. (1991). Russisch pričem – eine Konnexion der dritten Art? In K. Hartenstein & H. Jachnow (Eds.), Slavistische Linguistik. Referate des XVI. Konstanzer Slavistischen Arbeitstreffens Bochum / Löllinghausen 19.–21.9.1990 (Slavistische Beiträge, 274, 301–325). M¨unchen.

HWR200: New open access dataset of handwritten texts images in Russian

Ivan Potyashin
Antiplagiat
potyashin@ap-team.ru

Mariam Kaprielova
Antiplagiat, FRC CSC RAS
kaprielova@ap-team.ru

Yury Chekhovich
Antiplagiat, FRC CSC RAS
chehovich@ap-team.ru

Alexandr Kildyakov
Antiplagiat
kildyakov@ap-team.ru

Temirlan Seil
Antiplagiat
seilov@ap-team.ru

Evgeny Finogeev
Antiplagiat
finogeev@ap-team.ru

Andrey Grabovoy
Antiplagiat, FRC CSC RAS
grabovoy@ap-team.ru

Abstract

Handwritten text image datasets are highly useful for solving many problems using machine learning. Such problems include recognition of handwritten characters and handwriting, visual question answering, near-duplicate detection, search for text reuse in handwriting and many auxiliary tasks: highlighting lines, words, other objects in the text. The paper presents new dataset of handwritten texts images in Russian created by 200 writers with different handwriting and photographed in different environment[1]. We described the procedure for creating this dataset and the requirements that were set for the texts and photos. The experiments with the baseline solution on fraud search and text reuse search problems showed results of results of 60% and 83% recall respectively and 5% and 2% false positive rate respectively on the dataset.

Keywords: OCR; handwriting; text reuse detection; computer vision; handwritten text recognition; HTR
DOI: 10.28995/2075-7182-2023-22-452-458

1 Introduction

There are many different tasks that require working with images of handwritten texts. For example, visual question answering (Mathew et al., 2020). Optical character recognition (OCR) for handwritten texts is an important (Nurseitov et al., 2021) and challenging problem(Yousef and Bishop, 2020; Coquenet et al., 2023), especially in languages where there is a lack of labelled data. An example of such a case can be any Cyrillic language. The solution to this problem can be applied in many areas: healthcare (Fogel et al., 2020), education (Yanikoglu, 2017; Bakhteev et al., 2021), digitization of historical documents (Wigington et al., 2018).

Standard datasets for the OCR problem are IAM (Marti, 2002) and Bentham (Gatos et al., 2014). The IAM-database is based on the Lancaster-Oslo/Bergen corpus. It consists of 13353 images with handwritten lines of text written in special forms by 657 people. The database is labeled at the sentence, line, and word levels. The total number of words in the collection is 115320. Bentham dataset contains over 6000 documents and over 25000 pages of text written by a philosopher Jeremy Bentham. An analogue of (Gatos et al., 2014) in Russian is Digital Petr (Potanin et al., 2021). It, like Bentham, consists of scanned historical documents written by a single person with line-level text segmentation. It contains about, 10000 image-text pairs corresponding to lines in historical documents.

The datasets (Potanin et al., 2021; Gatos et al., 2014) have a feature that all the texts are written by one person. Such data can be useful for solving problems of recognition of texts written in the same handwriting. Also, it can be helpful for researchers to compare different handwriting recognition models. But if one needs to develop a model that deals with different handwriting these datasets may not be the best choice.

Datasets of another type, which contain different handwritings are IDP-forms (idp,), HKR (Nurseitov et al., 2021), school_notebooks (sch,). When compiling the sber-idp-forms dataset, the assessors were asked to manually write the given words or phrases on special forms. In total, there are 5203 images of rectangles with written text and their annotations in this dataset. The collection may be used for text segmentation, handwriting recognition, writer identification and writer verification tasks. When creating

[1]Our dataset is available at: https://huggingface.co/datasets/AntiplagiatCompany/HWR200

the HKR collection, the same idea with form was used. The dataset includes 63000 phrases written by 200 people. The text is written in Russian and Kazakh where about 95% is presented in Russian. The datasets (idp, ; Nurseitov et al., 2021) consist of handwritten texts digitized in the same environment, while in some cases it is natural to work with handwritten texts converted to an image in different ways.

Another significant problem is reuse in handwritten texts such as essays in schools and universities (Wrigley, 2019). We separate this problem into two categories. The first one is text reuse and the second one is submitting the same writing photographed in a different environment which will be further referred to as *fraud*.

The collection (sch,) contains 1857 images of school notebooks with word-level polygonal markup. This dataset is quite small, and as in the previous datasets, the environment in which the photos are taken is the same.

Cyrillic languages are not so actively studied in the problem of handwritten text recognition (HTR). There is a small amount of marked up data in these languages. Thus for the Russian language there are no approaches with sufficient quality of handwriting recognition.

For a text reuse search task, even a small number of typos in the text leads to a degradation in the search quality. Thus it is important to have either high-quality OCR or a new approach that takes into account the imperfection of the OCR model.

To create new models, it is essential to have a highly diverse dataset of handwritten texts. We contribute to solving handwritten text recognition problem by introducing the HWR200 dataset: a collection of handwritten texts in Russian for HTR and the search for reuse in handwritten texts. This collection provides texts created by 200 different writers, photographed under different conditions. The peculiarity of this dataset is that the same texts were written by more than one assessors, and this information can be used when training a robust OCR model. Texts are written by 200 assessors and photographed in three different ways. In total there are 30030 images with handwritten texts in our dataset.

2 Description of the dataset

2.1 Text generation algorithm and markup structure

The basis of the dataset is 35 different unique texts further referred to as *originals*. They are used to generate most of the dataset: texts further referred to as *reuses*. The *reuses* consist of two types of sentences: sentences that appeared in *original* texts and unique sentences. The text generation algorithm is as follows:

1. We generate a number from 28 to 32 for the amount of sentences to be reused;
2. We randomly select one or two *originals* to be used;
3. We generate how many sentences will be taken from the first *original* text and how many from the second;
4. Unique sentences are added to the beginning, to the end, or both to the beginning and to the end.

In total, 2650 *reuses* are generated. In addition, there are 35 more unique texts further referred to as *fprs*. An example of json with metadata:

```
// for original texts:
{
    sentences: [{id: <id>, text: <sentence>}, ...],
    words_count: <word count>,
    full_text: <full text>
}

// for reuse texts:
{
    reuse_0: {
        sentences: [{id: <id>, text: <sentence>}, ...],
        id: <original text file name>
        intersection_score: <intersection_score>
```

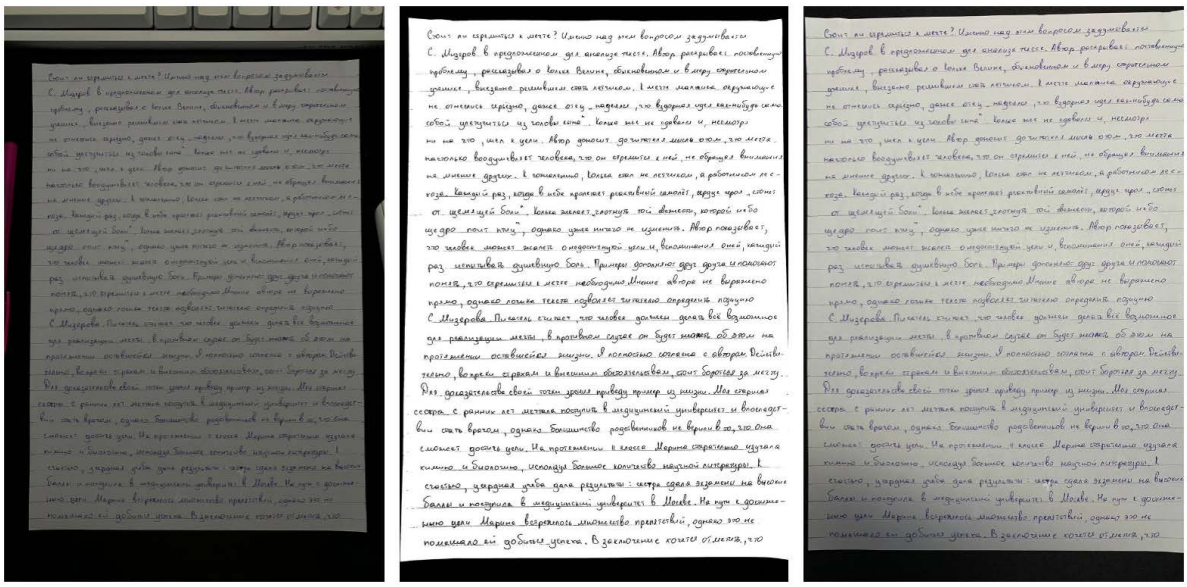


Figure 1: Three types of images: (left) photographed in poor light and with other objects, (center) scanned, (right) photographed in good light and without other objects.

```
    }
    reuse_1: {  // if exists
        sentences: [{id: <id>, text: <sentence>}, ...],
        id: <original text file name>
        intersection_score: <intersection_score>
    }
    start clear sentences: [<sentence>, <sentence>, ...]  // if exists
    end clear sentences: [<sentence>, <sentence>, ...]  // if exists
    words_count: <word count>
    full_text: <full text>
}

// for fpr texts:
{
    sentences: [{id: <id>, text: <sentence>}, ...],
    words_count: <word count>,
    full_text: <full text>
}
```

2.2 Image types

Each page of handwritten text had to be converted into an image in three different ways. First, it had to be scanned. Second, the assessor had to take a photo in good light. There should have been no glare, the page should not be cut off, extra objects should not fall into the frame. Third, the text had to be photographed in poor light. In this case, it was desirable that objects on the table fall into the frame, but the main part of the frame should have been occupied by the page. It was important that each page fits completely into the frame. See examples in Figure 1.

	Ours	**Bentham, Digital Petr**	**IAM**	**School notebooks**	**Sber-idp-forms, HKR**
Texts or phrases	texts	texts	phrases	texts	phrases
Word / line level markup	-	+	+	+	+
Different handwriting	+	-	+	+	+
Different environment	+	-	-	-	-

Table 1: Characteristics of datasets

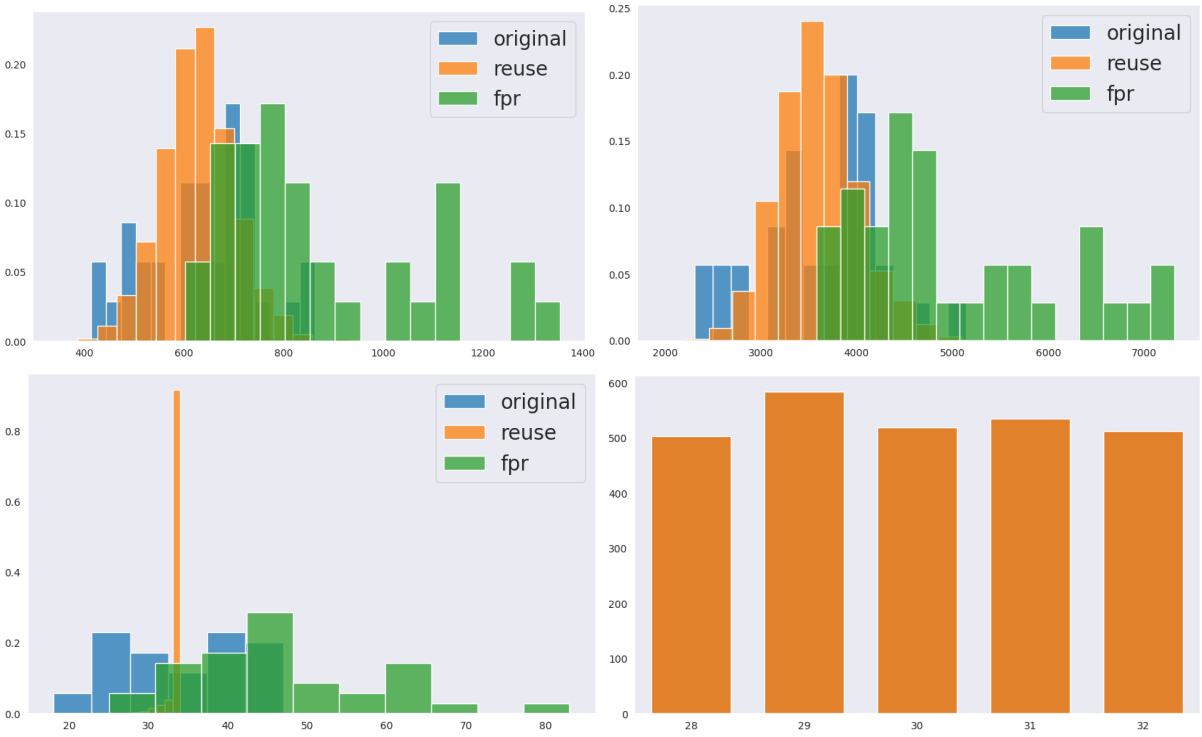


Figure 2: (top left) distribution of the number of words in texts. (top right) distribution of the lengths of texts. (bottom left) distribution of the number of sentences in texts. (bottom right) distribution of the number of duplicate sentences in *reuses*.

2.3 Distribution of texts by assessors

Each of 200 assessors wrote 15 texts. The first 175 assessors wrote one *original*, one *fpr* and 13 *reuses*, at that each *original* and each *fpr* are handwritten by 5 assessors. The rest 25 wrote 15 *reuses*. Thus 2650 *reuses* are handwritten once and 35 *originals* and 35 *reuses* are handwritten five times.

2.4 Characteristics of the dataset

The dataset contains 2720 handwritten texts with an average word count of 631, an average text length of 3617, and an average sentence count of 34. In addition, 47% of texts with duplicates have two *original* essays, the rest have only one. The distribution is show in Figure 2. The total number of images with text is 30030, so on average each text takes up 3.3 pages. A comparative table with the characteristics of various datasets can be seen in Table 1.

Dataset	Task	Recall@1	FPR
Ours	fraud detection	60%	5%
Bentham+IAM	fraud detection	80%	5%
Ours	reuse detection	83%	3%

Table 2: Results of our baseline solutions.

3 Experiments

We trialled the HWR200 dataset in two tasks: fraud search and text reuse detection. To evaluate our solutions we used recall@1 and fpr metrics. These tasks are actually binary classification tasks: is a given image a reuse or a fraud or not. Formally, metrics are defined as follows:

$$recall@1 = \frac{TP}{TP + FN}, \tag{1}$$

$$fpr = \frac{FP}{FP + TN}, \tag{2}$$

where TP is the number of true positive predictions, FN is the number of false negative predictions, FP is the number of false positive predictions, and TN is the number of true negative predictions. Results can bee seen in Table 2.

3.1 Fraud detection

As described above, every handwritten page is photographed three times in three different ways. We considered one of them as an original, and the other two as fraud. So, the task for each fraud page is to find the original page.

Our baseline solution for this task consists of three stages: embedding generation, candidate search and similarity estimation between query and candidates to find the closest one. We use a neural network to transform a handwritten document into embedding. For candidates search we use Faiss framework (Johnson et al., 2017) (the faiss index is filled with original photos) and similarity estimation is performed using deep learning approach inspired by (Sun et al., 2021).

This approach showed 60% recall@1 and 5% fpr. Similar approach on IAM and Bentham dataset showed 80% recall and 5% fpr. It should be taken into account that in that experiment, fraud images were generated from images in the dataset, whereas in our experiment, fraud images are part of the dataset.

3.2 Reuse detection

Every *reuse* contains some sentences from one or two *original* texts. The task for every *reuse* text is to find at least one original text.

The solution for this task also consists of three stages. First, the algorithm tries to recognize handwritten text. The input page is divided into lines, and text is extracted from each line using a deep learning OCR model optimized in supervised learning mode inspired by (Coquenet et al., 2020). Second, we split the text into bigrams and search for candidates based on them using a shingle index based on (Broder et al., 1997; Broder, 1997). Last, we compare the candidates with the input text and find the text with the highest reuse rate.

This approach showed 83% recall and 2% false positive rate.

4 Conclusion

We have introduced the dataset of handwritten texts in Russian. This collection contains texts written in different handwriting and photographed under various conditions. One of the key features of this

collection is that one text can be written by several assessors, which may be very useful for tasks where models have to be robust. Besides, the dataset can also be helpful for solving more specific tasks such as text reuse search or near-duplicate detection.

Acknowledgements

This research was supported by FASIE (Project No. 79068, application No. 208298).

References

Oleg Bakhteev, Dorodnicyn CC Antiplagiat, Rita Kuznetsova, Andrey Khazov, Aleksandr Ogaltsov, Kamil Safin, Tatyana Gorlenko, Marina Suvorova, Andrey Ivahnenko, Pavel Botov, et al. 2021. Near-duplicate handwritten document detection without text recognition.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web, Sep.

A.Z. Broder. 1997. On the resemblance and containment of documents.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. 2020. End-to-end handwritten paragraph text recognition using a vertical attention network. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:508–524.

Denis Coquenet, Clément Chatelain, and Thierry Paquet. 2023. Dan: a segmentation-free document attention network for handwritten document recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Sharon Fogel, Hadar Averbuch-Elor, Sarel Cohen, Shai Mazor, and Roee Litman. 2020. Scrabblegan: Semi-supervised varying length handwritten text generation. *// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Basilis Gatos, Georgios Louloudis, Tim Causer, Kris Grint, Verónica Romero, Joan Andreu Sánchez, Alejandro H. Toselli, and Enrique Vidal. 2014. Ground-truth production in the transcriptorium project. *// 2014 11th IAPR International Workshop on Document Analysis Systems*, P 237–241.

Idp-forms (2021). Available at: `https://github.com/ai-forever/htr_datasets/tree/main/IDP-forms`.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *IEEE Transactions on Big Data*, 7:535–547.

U.-V. Marti. 2002. The iam-database: an english sentence database for offline handwriting recognition. *International Journal on Document Analysis and Recognition*, 28(1):114–133.

Minesh Mathew, Dimosthenis Karatzas, R. Manmatha, and C. Jawahar. 2020. Docvqa: A dataset for vqa on document images. 07.

Daniyar Nurseitov, Kairat Bostanbekov, Daniyar Kurmankhojayev, Anel Alimova, Abdelrahman Abdallah, and Rassul Tolegenov. 2021. Handwritten kazakh and russian (hkr) database for text recognition. *Multimedia Tools and Applications*, P 1–23.

M. B. Potanin, Denis Dimitrov, A. Shonenkov, Vladimir Bataev, Denis Karachev, and Maxim Novopoltsev. 2021. Digital peter: New dataset, competition and handwriting recognition methods. *The 6th International Workshop on Historical Document Imaging and Processing*.

School_notebooks_ru (2021). Available at: `https://github.com/ai-forever/htr_datasets/tree/main/school_notebooks`.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. 2021. LoFTR: Detector-free local feature matching with transformers. *CVPR*.

Curtis Wigington, Chris Tensmeyer, Brian Davis, William Barrett, Brian Price, and Scott Cohen. 2018. Start, follow, read: End-to-end full-page handwriting recognition. *// Computer Vision – ECCV 2018"*, P 372–388, Cham. Springer International Publishing.

Stuart Wrigley. 2019. Avoiding 'de-plagiarism': Exploring the affordances of handwriting in the essay-writing process. *Active Learning in Higher Education*, 20(2):167–179.

Berrin Yanikoglu. 2017. Use of handwriting recognition technologies in tablet-based learning modules for first grade education. *Educational Technology Research and Development*.

Mohamed Yousef and Tom E. Bishop. 2020. Origaminet: Weakly-supervised, segmentation-free, one-step, full page textrecognition by learning to unfold. *// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June.

Simple Yet Effective Named Entity Oriented Sentiment Analysis

Leonid Sanochkin[1,2], Angelina Bolshina[1], Kseniia Cheloshkina[1,2], Daria Galimzianova[1,2], Aleksei Malafeev[1,2]
[1]MTS AI,
[2]HSE,
{l.sanochkin, a.bolshina, k.cheloshkina, d.galimzianova, a.malafeev}@mts.ai

Abstract

Sentiment analysis, i.e. the automatic evaluation of the emotional tone of a text, is a common task in natural language processing. Entity-Oriented Sentiment Analysis (EOSA) predicts the sentiment of entities mentioned in a given text. In this paper, we focus on the EOSA task for the Russian news. We propose a text classification pipeline to solve this task and show its potential in such tasks. Moreover, in general, EOSA implies labeling both named entities and their sentiment, which can require a lot of annotator labour and time and, thus, presents a major obstacle to the development of a production-ready EOSA system. To help alleviate this, we analyse the potential of applying an Active learning approach to EOSA tasks. We demonstrate that by actively selecting instances for labeling in EOSA the annotation effort required for training machine learning models can be significantly reduced.

Keywords: Aspect-based sentiment analysis, Entity-oriented sentiment analysis, sentiment analysis, Active Learning

DOI: 10.28995/2075-7182-2023-22-459-468

Классификационный подход к анализу тональности именованных сущностей в новостных текстах

Леонид Саночкин[1,2], Ангелина Большина[1], Ксения Челошкина[1,2], Дарья Галимзянова[1,2], Алексей Малафеев[1,2]
[1]МТС ИИ,
[2]НИУ ВШЭ,
{l.sanochkin, a.bolshina, k.cheloshkina, d.galimzianova, a.malafeev}@mts.ai

Аннотация

Автоматизированный анализ тональности текстов является одной из распространенных проблем автоматической обработки текстовой информации. В данной работе рассматривается оценка тональности по отношению к сущности в новостном тексте. Нами был предлоложен и протестирован подход, основанный на представлении данной задачи как задачи классификации. Кроме того, поскольку разметка данных для задач оценки тональности относительно сущности в тексте может быть трудоемким процессом, мы исследуем применимость активного обучения в данной задаче. Полученные результаты свидетельствуют о перспективности использования предложенного подхода в рамках активного обучения для задач оценки тональности относительно сущностей в тексте.

Ключевые слова: Анализ тональности текстов, тональность по отношению к сущности в тексте, активное обучение

1 Introduction

Nowadays, Aspect-based sentiment analysis (ABSA) is quite popular not only in the academic but also in the commercial sphere. Irrespective of the industry, it provides a fine-grained customer feedback analysis, offering valuable insights into the customer experience and helping to make data-driven decisions.

ABSA is a more fine-grained version of the classic sentiment analysis task that allows to obtain more detailed information from a text, which is more useful in real-life applications. The task of ABSA involves the extraction of various types of terms: 1) the aspect term (*a*); 2) the opinion term (*o*); 3) the

aspect category (c) corresponding to the aspect term; 4) the sentiment polarity (s) for a given aspect term (Gao et al., 2022). ABSA can be divided into several sub-tasks based on the combinations of the identified terms. This article proposes an approach to solve the Entity-Oriented Sentiment Analysis (EOSA), which can be also referred to an Aspect-Category Sentiment Analysis.

Since it is necessary to label both entities and their sentiment inside the text, the costs of data annotation for entity-oriented sentiment analysis can hinder the practical application of such systems. Thus, we analyse the applicability of an Active learning (AL) pipeline for this problem, as described in the section 5.2. The results obtained show that our approach can be used for the active selection of instances to label and, thus, can be helpful in solving the EOSA task in a low-resource setting.

To summarize our contribution:

- We demonstrated that the entity-oriented sentiment analysis task can be efficiently solved with a naïve text classification pipeline;
- We addressed the problem of data shortage for such tasks and showed that by actively selecting examples to label, we can achieve comparable performance to the model trained on full data with a significantly smaller amount of labeled data;

2 Related work

Despite its high demand, ABSA task suffers from data scarcity, like many other NLP research areas. The survey (Chebolu et al., 2022) presents a comprehensive overview of available datasets for ABSA.

As mentioned above, ABSA consists of several sub-tasks, namely, aspect term and category identification, opinion term identification, and aspect sentiment classification. These tasks can be solved either separately or jointly. The former approach considers only one task at a time, e.g. (Li et al., 2020), (Xu et al., 2021a), (Ma et al., 2018). More often, studies focus on several subtasks simultaneously. All approaches differ in the number of the subtasks they solve. For example, the studies (He et al., 2019), (Dai et al., 2020), (Zhao et al., 2020) are devoted to the extraction of pairs of terms. Some papers identify triples in a text (Xu et al., 2020), (Wu et al., 2021). The approach described in (Cai et al., 2021) aims at quadruple extraction.

ABSA can be treated as classification, sequence tagging, machine reading comprehension tasks, or a generative problem. (Hu et al., 2019), (Jiang et al., 2019), and (Zhang and Qian, 2020) tackle ABSA as a classification problem. Some approaches transfer subtasks to the sequence tagging problem: (Li et al., 2019), (Chen and Qian, 2020), (Wu et al., 2021), (Xu et al., 2021b). (Yu et al., 2021), (Mao et al., 2021), (Liu et al., 2022), and (Chen et al., 2021) proposed to solve ABSA as a machine reading comprehension task. Generative frameworks are also used to solve ABSA subtasks: (Gao et al., 2022), (Zhang et al., 2021), (Yan et al., 2021), (Hosseini-Asl et al., 2022).

(Luo and Mu, 2022) studies EOSA in the news texts and proposes a Negative Sentiment Smoothing Model to address the multiple entity sentiment analysis problem. In (Fu et al., 2022), the problem of EOSA is studied on noisy data, obtained from automatic speech recognition tools.

3 Proposed approach

To address the problem of EOSA, we propose a text classification pipeline with an additional information on the analysed entity. We provide the model with additional information on the analysed entity by adding the exact entity string to the input token sequence with the separation token. Our approach is highly motivated by the success of solving question answering tasks with a machine reading comprehension pipeline, such as in (Devlin et al., 2018) and by the previously mentioned papers that reported solving ABSA with machine reading comprehension (Yu et al., 2021; Mao et al., 2021; Liu et al., 2022; Chen et al., 2021). In the section 5.2 we show with ablation studies that concatenating entity string with the input sequence is the key component that contributes greatly to the overall performance of the model for the EOSA task.

4 Dataset analysis

We evaluate our approach on the RuSentNE dataset (Golubev et al., 2023) created for the first competition in targeted sentiment analysis on named entities in Russian news. In the dataset, the named entities are already recognized and classified into the following labels: PERSON, ORGANIZATION, PROFESSION, COUNTRY, and NATIONALITY. The task is, for every sentence in the dataset, to assign a given entity one of the three sentiment classes: "positive" ("1"), "negative"("-1") or "neutral"("0"). The sentences are not related, and there is always exactly one entity that needs to be labeled for sentiment. The dataset consists of three splits: training (6 637 examples, 15% negative / 72% neutral / 13% positive), validation (2 845 examples) and test (1947 examples). It is worth noting that, according to the survey (Chebolu et al., 2022), this dataset is one of the largest in terms of the number of entities.

As sentiment analysis is prone to be subjective, it is of interest here to investigate whether there are mislabeled examples or not. To get an understanding of how much data could be assigned wrong labels, we used the "Dataset Cartography" method (Swayamdipta et al., 2020), which was shown to be effective in detecting labeling errors. This model-specific method assumes that every example in a dataset can be automatically categorized as belonging to one of the following groups: easy-to-learn examples (consistently labeled correctly by the model with high confidence), hard-to-learn examples (consistently mislabeled) and ambiguous examples (of high variability). We applied this method to the training set and built its data map. Results are presented in Figure 1. It can be clearly seen that this map has a low-density region of hard-to-learn examples, which means that the dataset has high annotation quality.

Nevertheless, since it was demonstrated that hard-to-learn examples tend to be labeling errors, it is worth taking a closer look at them. There are such 97 hard-to-learn examples out of 6 637 (1.5%) with a strong predominance of the positive class: the class balance is 27% / 24% / 49% in this subsample ("-1" / "0" / "1"), although in the full training sample the proportions are 15% / 72% / 13%. An inspection of hard-to-learn examples reveals some labeling errors is presented in the Table 4.

labeled as positive (but looks like at least neutral):
Подозреваемыми оказались два **студента**, каждому из которых по 21 году. (The suspects were two **students**, each of whom is 21 years old.)
Власти Парагвая объявили трёхдневный траур в связи с гибелью **политика**. (The Paraguayan authorities have declared three days of mourning in connection with the death of the **politician**.)
Кеплен вспоминает, что в ходе следствия было несколько нестыковок и пытается выяснить правду... (**Keplen** recalls that there were several inconsistencies during the investigation and is trying to find out the truth...)
labeled as negative:
Во время выступления **прокурора** он молча сидел, скрестив ноги и работая со своим планшетным компьютером. (During the **prosecutor's** speech, he sat cross-legged in silence and worked with his tablet computer.)
Изучавший статую эксперт **Алессандро Мартелли** сказал: (The expert who studied the statue, **Alessandro Martelli**, said:)

Table 1: Examples of the label errors.

Thus, the dataset contains a small portion of mislabeled examples which were probably introduced by ambiguous annotation rules, as we further demonstrate in the model error analysis section.

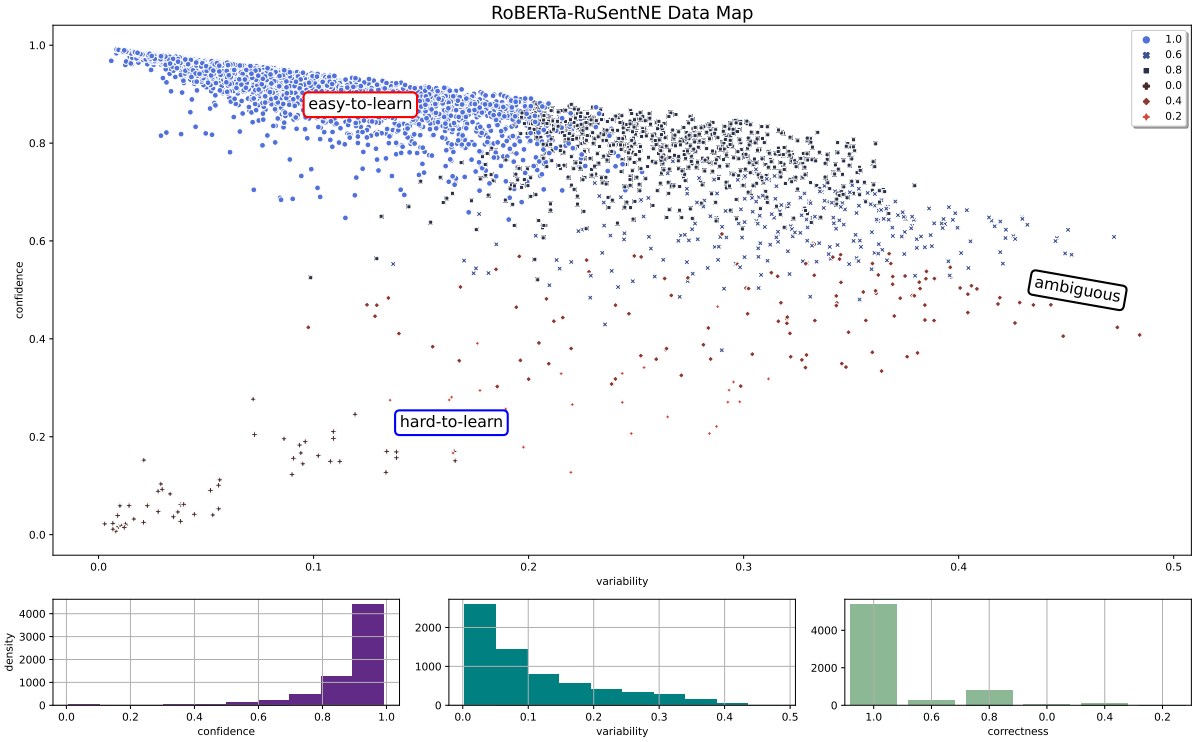


Figure 1: Dataset Cartography Map for RuSentNE.

5 Experiments

5.1 Experimental setup

The training data was randomly split into the training and validation parts in the 80/20 proportion. The provided results were computed on the test part of RuSentNE corpora via Codalab platform[1]. The competition uses a variant of a macro-$F1$ score ($F1_{pn}$), which is averaged over two sentiment classes: "positive" and "negative". The class "neutral" is excluded because it is more relevant to extract opinions and sentiments. The results are averaged over five random seeds in order to report the standard deviation of the scores.

For active learning experiments, we used the classic simulated active learning experiment design (Settles and Craven, 2008; Shen et al., 2017). We emulated the AL annotation cycle starting with sampling from the dataset randomly and using this small portion of data as a seed for the construction of the initial acquisition model. On each iteration, a fraction of the top informative instances is sampled from the unlabeled pool by some query strategy. The selected instances are labeled according to the gold standard, then they are added to the training dataset and removed from the unlabeled pool for the following iterations. We used the following query strategies to score the informativeness of the unlabeled instances: Least Confidence (LC) (Lewis and Gale, 1994), Breaking Ties (BT) (Luo et al., 2004), Prediction entropy (PE) (Roy and Mccallum, 2001), and Contrastive Active Learning (CAL) (Margatina et al., 2021). We have not used some of the modern AL strategies, such as Batch Active Learning by Diverse Gradient Embeddings (BADGE) (Ash et al., 2020) and Batch Active learning via Information maTrices (BAIT) (Ash et al., 2021) due to their low computational performance and the fact that they cannot outperform the baseline strategies (such as LC) for a significant margin on a vast amount of datasets (Margatina et al., 2021; Tsvigun et al., 2022). For the successor model, we used the same model as for acquisition. To report standard deviations of the scores, we repeat the whole experiment five times

[1]https://codalab.lisn.upsaclay.fr/competitions/9538

with different random seeds. We sampled 2% of all training data (132 samples) and selected the same amount from the unlabeled pool on each iteration. We performed AL for 20 iterations.

As backbone models for our experiments, we used pretrained transformer models for Russian language: ruBert-base[2] and ruRoberta-large[3].

5.2 Results and discussion

Ablation studies In this section, we investigate different options for highlighting the specific entity of interest in the model input to perform entity-oriented sentiment analysis. We compared the following approaches:

1. Adding entity type info: concatenate the full sentence and the entity type string with the [SEP] separator. Input: "sentence [SEP] entityType".
2. Without entity information. Input: sentence.
3. In-text demonstration: add the [SEP] token before and after the entity text inside the sentence. Input: "sentenceStart [SEP] entity [SEP] sentenceEnd".
4. Our proposed approach: concatenate the full sentence and the entity string with the [SEP] separator. Input: "sentence [SEP] entity".

The results of the study are shown in the Table 5.2. The proposed approach outperforms the ones without proper information about an entity by a significant margin. However, pointing the entity inside the text leads to results within the confidence interval for the score.

Model	Ours	Ablation 1	Ablation 2	Ablation 3
ruBert-base	53.336±1.557	43.936±1.859	37.572±2.193	53.068±0.380
ruRoberta-Large	61.400±1.033	49.324±3.734	42.683±1.178	62.834±0.997

Table 2: Model performance.

We also include the performance of the baseline model and the top-performing approach from the competition in the Table 5.2. It can be seen that our approach, despite its simplicity, is quite competitive for the task of EOSA and has been outperformed by the top solution by a small margin.

Method	F1
Ours	62.92
Baseline	40.92
Best model	66.67

Table 3: Comparison with other methods.

Error analysis To perform error analysis, we used validation set labels obtained from five different seeds of our model, and compared them with the ground truth annotations. We also measured two types of agreement with Krippendorff's Alpha, which is a reliability coefficient ranging from -1 to 1 that can be used for two or more raters and categories, is applicable to many types of data and measurement scales, and has a number of other advantages (Krippendorff, 2011). First, we measured the agreement between all five seeds, which was very high: 0.79. This is expected, but we wanted to make sure that the model variations learn similar facts about the task from the training data regardless of the seed. Second, we also calculated the pairwise agreement between each seed and the ground truth. These ranged from 0.49 to 0.51: fairly close between the seeds and moderately high agreement with the ground truth.

Let us consider a few specific categories of errors. Out of 2845 examples, in 337 cases (about 11.5%) all five variations of our model yielded the same label, but different from the ground truth. In 46 of these, all seeds gave the opposite answer, i.e. either 1 instead of -1 or -1 instead of 1. More distributional

[2]https://huggingface.co/ai-forever/ruBert-base
[3]https://huggingface.co/ai-forever/ruRoberta-large

details are given in the Figure 2. Darker colors correspond to greater quantities of examples. GT stands for "ground truth". All percentages given are relative to the total number of examples in the validation set (2845).

<table>
<tr><td colspan="5">all examples</td></tr>
<tr><td colspan="5">2845 (100%)</td></tr>
<tr><td>all seeds agree with GT</td><td colspan="4">at least 1 seed disagrees with GT</td></tr>
<tr><td>2013 (71%)</td><td colspan="4">832 (29%)</td></tr>
<tr><td></td><td>some agree with GT</td><td colspan="3">all disagree with GT</td></tr>
<tr><td></td><td>448 (16%)</td><td colspan="3">384 (13%)</td></tr>
<tr><td></td><td></td><td>different answers</td><td colspan="2">same answer</td></tr>
<tr><td></td><td></td><td>47 (1.5%)</td><td colspan="2">337 (11.5%)</td></tr>
<tr><td></td><td></td><td></td><td>not opposite to GT</td><td>opposite to GT</td></tr>
<tr><td></td><td></td><td></td><td>291 (10%)</td><td>46 (1.5%)</td></tr>
</table>

Figure 2: Agreement of the models, trained on different random seeds.

It is noteworthy that when all five seeds disagree with the ground truth, in about 88% of the cases (337 vs 47) they are unanimous, i.e. yield the same answer. This might indicate labeling inconsistencies between the training and the test sets, at least in some cases. Consider the following examples:

- Пиночет совершил ошибку, приказав убить Неруду», — говорит Арайя. Pinochet made the mistake of ordering the death of Neruda," says Araya.
  The ground truth label for the sentiment towards Neruda is questionably 1, while the five variations of the model unanimously suggest -1.

- Левая оппозиция желает проведения досрочных выборов, поскольку чувствует, что ветер успеха дует в ее паруса. The left opposition wants early elections because it feels that the wind of success is blowing in its sails. The ground truth label for "left opposition" is -1, while the model yields 1. Even if we accept that "positive" is a wrong answer, why is the ground truth answer not "neutral"?

These and other similar examples hint at the inherent difficulty and ambiguity of the targeted sentiment analysis task in the given setting. Indeed, the task description mentions that there are three possible sources of sentiment towards an entity: the author's opinion, a quoted opinion of a third party, and an implicit opinion (Golubev et al., 2023). This raises some methodological concerns:

1. What if the author's opinion and the quoted opinion are opposite, e.g. *They called my good friend Tom an idiot*. What is the sentiment towards Tom?
2. Is it possible to unambiguously define the implicit sentiment, when nothing but one sentence is given and we have no information about the author, the circumstances, etc.? For example, *Hitler came to power in 1933*. Should we consider the sentiment towards Hitler as negative because we know about his wrongdoings? But maybe the speaker is indeed pro-Hitler? Or is it a neutral context because the word choice is neutral? Or maybe "coming to power" by itself can be considered as slightly positive?

This is further aggravated by the distribution of the ground truth labels in the test set: 2045 neutral examples (72%), 438 negatives (15%) and 362 positives (13%). There are fewer than 30% examples with non-neutral sentiment, and even some of these are questionable, as manual error analysis of the mislabeled examples shows. It is hard to quantify exactly how many of the sentences in the test set are

mislabeled since there appears to be no obvious framework for unambiguous judgement on the 'correctness' of the labels, as discussed above.

On the Figure 3 is the confusion matrix for ground truth labels and model predictions aggregated by simple majority vote (there is always a majority since the number of seeds is greater than the number of possible labels and the number of seeds is odd).

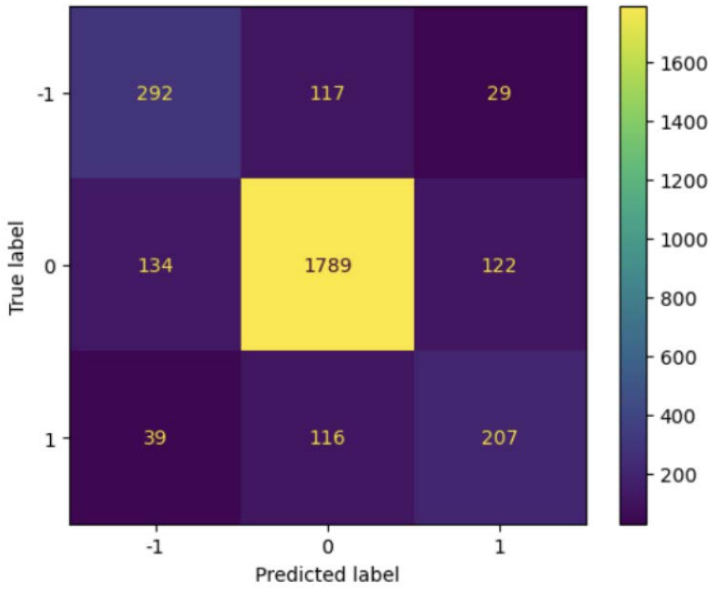


Figure 3: Confusion matrix.

As can be seen from the Figure 3, the model does not often confuse positive sentiment with negative (11% of all positive examples in the validation set) and negative for positive (6% of all negative examples in the validation set). However, there is a lot of confusion involving the neutral category (both type I and type II errors): 489 examples out of the total of 2845, or about 17%. This is understandable, as, firstly, the neutral category is the majority class, and secondly, it is easier to confuse neutral with positive / negative sentiment, rather than positive with negative or vice versa.

Active learning results The results of the best Active learning strategy are presented in Figure 4. It can be seen that the random sample selection baseline is outperformed by actively selecting samples according to an AL strategy. In our experiment, the best strategy for RuSentNE task was Breaking Ties, however, further research may be needed to determine the best query strategy and its hyperparameters for the EOSA tasks in general. Also, we plan to analyse the possibility of using smaller models as the acquisition model (without degrading successor performance) to make AL more efficient.

6 Conclusion

We analyzed the potential for solving EOSA tasks with a simple text classification pipeline and showed that our approach can be competitive in such tasks. Moreover, it can be easily adjusted to actively selecting instances for labeling. Our work demonstrates that active learning can be a promising approach for reducing the annotation effort in EOSA and improving the efficiency of the development of production-ready EOSA systems.

To further address the low-resource setting for EOSA tasks, we are looking forward to analysing the potential of applying few-shot methods for such tasks. Additionally, further research is needed on identifying the optimal hyperparameters of an AL pipeline.

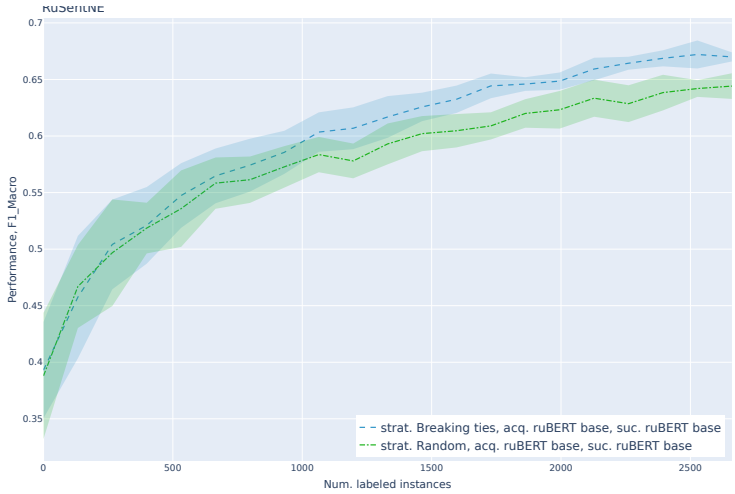


Figure 4: Active learning for RuSentNE.

References

Jordan T. Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. 2020. Deep batch active learning by diverse, uncertain gradient lower bounds. // *International Conference on Learning Representations*.

Jordan Ash, Surbhi Goel, Akshay Krishnamurthy, and Sham Kakade. 2021. Gone fishing: Neural active learning with fisher embeddings. // M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, *Advances in Neural Information Processing Systems*, volume 34, P 8927–8939. Curran Associates, Inc.

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. // *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 340–350, Online, August. Association for Computational Linguistics.

Siva Uday Sampreeth Chebolu, Franck Dernoncourt, Nedim Lipka, and Thamar Solorio. 2022. Survey of aspect-based sentiment analysis datasets. *arXiv preprint arXiv:2204.05232*.

Zhuang Chen and Tieyun Qian. 2020. Relation-aware collaborative learning for unified aspect-based sentiment analysis. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 3685–3694, Online, July. Association for Computational Linguistics.

Shaowei Chen, Yu Wang, Jie Liu, and Yuelin Wang. 2021. Bidirectional machine reading comprehension for aspect sentiment triplet extraction. // *Proceedings of the AAAI conference on artificial intelligence*, volume 35, P 12666–12674.

Zehui Dai, Cheng Peng, Huajie Chen, and Yadong Ding. 2020. A multi-task incremental learning framework with category name embedding for aspect-category sentiment analysis. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 6955–6965, Online, November. Association for Computational Linguistics.

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*.

Xue-Yong Fu, Cheng Chen, Md Tahmid Rahman Laskar, Shayna Gardiner, Pooja Hiranandani, and Shashi Bhushan TN. 2022. Entity-level sentiment analysis in contact center telephone conversations. *arXiv preprint arXiv:2210.13401*.

Tianhao Gao, Jun Fang, Hanyu Liu, Zhiyuan Liu, Chao Liu, Pengzhang Liu, Yongjun Bao, and Weipeng Yan. 2022. LEGO-ABSA: A prompt-based task assemblable unified generative framework for multi-task aspect-based sentiment analysis. // *Proceedings of the 29th International Conference on Computational Linguistics*, P 7002–7012, Gyeongju, Republic of Korea, October. International Committee on 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: papers from the Annual conference "Dialogue"*.

Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. 2019. An interactive multi-task learning network for end-to-end aspect-based sentiment analysis. // *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, P 504–515, Florence, Italy, July. Association for Computational Linguistics.

Ehsan Hosseini-Asl, Wenhao Liu, and Caiming Xiong. 2022. A generative language model for few-shot aspect-based sentiment analysis. // *Findings of the Association for Computational Linguistics: NAACL 2022*, P 770–787, Seattle, United States, July. Association for Computational Linguistics.

Mengting Hu, Shiwan Zhao, Li Zhang, Keke Cai, Zhong Su, Renhong Cheng, and Xiaowei Shen. 2019. CAN: Constrained attention networks for multi-aspect sentiment analysis. // *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 4601–4610, Hong Kong, China, November. Association for Computational Linguistics.

Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A challenge dataset and effective models for aspect-based sentiment analysis. // *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 6280–6285, Hong Kong, China, November. Association for Computational Linguistics.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

David D Lewis and William A Gale. 1994. A sequential algorithm for training text classifiers. // *SIGIR'94*, P 3–12. Springer.

Xin Li, Lidong Bing, Wenxuan Zhang, and Wai Lam. 2019. Exploiting BERT for end-to-end aspect-based sentiment analysis. // *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)*, P 34–41, Hong Kong, China, November. Association for Computational Linguistics.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 7056–7066, Online, July. Association for Computational Linguistics.

Shu Liu, Kaiwen Li, and Zuhe Li. 2022. A robustly optimized BMRC for aspect sentiment triplet extraction. // *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, P 272–278, Seattle, United States, July. Association for Computational Linguistics.

Manman Luo and Xiangming Mu. 2022. Entity sentiment analysis in the news: A case study based on negative sentiment smoothing model (nssm). *International Journal of Information Management Data Insights*, 2(1):100060.

Tong Luo, K. Kramer, S. Samson, A. Remsen, D.B. Goldgof, L.O. Hall, and T. Hopkins. 2004. Active learning to recognize multiple types of plankton. // *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, P 478–481 Vol.3.

Yukun Ma, Haiyun Peng, and Erik Cambria. 2018. Targeted aspect-based sentiment analysis via embedding commonsense knowledge into an attentive lstm. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), Apr.

Yue Mao, Yi Shen, Chao Yu, and Longjun Cai. 2021. A joint training dual-mrc framework for aspect based sentiment analysis. // *Proceedings of the AAAI conference on artificial intelligence*, volume 35, P 13543–13551.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. // *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, P 650–663, Online and Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Nicholas Roy and Andrew Mccallum. 2001. Toward optimal active learning through sampling estimation of error reduction. *Proceedings of the 18th International Conference on Machine Learning*, 08.

Burr Settles and Mark Craven. 2008. An analysis of active learning strategies for sequence labeling tasks. // *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, P 1070–1079, Honolulu, Hawaii, October. Association for Computational Linguistics.

Yanyao Shen, Hyokun Yun, Zachary Lipton, Yakov Kronrod, and Animashree Anandkumar. 2017. Deep active learning for named entity recognition. // *Proceedings of the 2nd Workshop on Representation Learning for NLP*, P 252–256, Vancouver, Canada, August. Association for Computational Linguistics.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 9275–9293, Online, November. Association for Computational Linguistics.

Akim Tsvigun, Leonid Sanochkin, Daniil Larionov, Gleb Kuzmin, Artem Vazhentsev, Ivan Lazichny, Nikita Khromov, Danil Kireev, Aleksandr Rubashevskii, Olga Shahmatova, Dmitry V. Dylov, Igor Galitskiy, and Artem Shelmanov. 2022. ALToolbox: A set of tools for active learning annotation of natural language texts. // *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, P 406–434, Abu Dhabi, UAE, December. Association for Computational Linguistics.

Chao Wu, Qingyu Xiong, Hualing Yi, Yang Yu, Qiwu Zhu, Min Gao, and Jie Chen. 2021. Multiple-element joint detection for aspect-based sentiment analysis. *Knowledge-Based Systems*, 223:107073.

Lu Xu, Hao Li, Wei Lu, and Lidong Bing. 2020. Position-aware tagging for aspect sentiment triplet extraction. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 2339–2349, Online, November. Association for Computational Linguistics.

Chi Xu, Hao Feng, Guoxin Yu, Min Yang, Xiting Wang, Yan Song, and Xiang Ao. 2021a. Discovering protagonist of sentiment with aspect reconstructed capsule network. // Christian S. Jensen, Ee-Peng Lim, De-Nian Yang, Wang-Chien Lee, Vincent S. Tseng, Vana Kalogeraki, Jen-Wei Huang, and Chih-Ya Shen, *Database Systems for Advanced Applications*, P 120–135, Cham. Springer International Publishing.

Lu Xu, Yew Ken Chia, and Lidong Bing. 2021b. Learning span-level interactions for aspect sentiment triplet extraction. // *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 4755–4766, Online, August. Association for Computational Linguistics.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. // *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 2416–2429, Online, August. Association for Computational Linguistics.

Guoxin Yu, Jiwei Li, Ling Luo, Yuxian Meng, Xiang Ao, and Qing He. 2021. Self question-answering: Aspect-based sentiment analysis by role flipped machine reading comprehension. // *Findings of the Association for Computational Linguistics: EMNLP 2021*, P 1331–1342, Punta Cana, Dominican Republic, November. Association for Computational Linguistics.

Mi Zhang and Tieyun Qian. 2020. Convolution over hierarchical syntactic and lexical graphs for aspect level sentiment analysis. // *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, P 3540–3549, Online, November. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. // *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. // *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, P 3239–3248, Online, July. Association for Computational Linguistics.

Is it possible to make the Russian punctuation rules more explicit?

Alexei Shmelev
Vinogradov Russian Language Institute
of the Russian Academy of Sciences /
Volkhonka 18/2, Moscow, Russia
shmelev.alexei@gmail.com

Abstract

This paper deals with some issues related to the Russian punctuation rules and their account in computer checkers and correctors (both "analytic" and "synthetic"). It also discusses variation of punctuation. The paper offers a critical assessment of reference books devoted to punctuation and makes special reference to certain verbs of propositional attitude and their parenthetical use (in particular, *dumat'* 'to think,' *videt'* 'to see,' and *slyshat'* 'to hear). It claims that the inherent characteristics of the verbs under consideration influence the punctuation, and therefore every verb deserves a detailed description (lexicographic portrait). In particular, *videt'* and *slyshat'* behave quite differently when used as parenthetical verbs. A step towards making the punctuation rules more explicit may consist in providing an index of words mentioned in the rules together with a subject index.

Keywords: punctuation; punctuation mark; colon; comma, parenthetical verb
DOI: 10.28995/2075-7182-2023-22-469-476

Возможна ли формализация правил русской пунктуации?

Алексей Шмелев
Институт русского языка
им. В.В. Виноградова РАН /
Волхонка 18/2, Москва, Россия
shmelev.alexei@gmail.com

1 Вступительные замечания

1.1 «Нагруженная» и «ненагруженная» вариантность пунктуации

Многие трудности учета правил русской пунктуации в компьютерных программах обусловлены недостаточной эксплицитностью самих этих правил, как они формулируются в различных пособиях и справочниках. Нередко эти формулировки оказываются туманными и допускающими различные толкования. Эта недостаточная эксплицитность накладывается на то, что русская пунктуация во многих случаях допускает вариантность. Иногда различие в пунктуации непосредственно связано с различием смыслов (такую вариантность можно назвать «нагруженной» вариантностью). Скажем, в предложении *Он пришел* в конце можно поставить точку, восклицательный знак, вопросительный знак или многоточие. Ср.:

(1) *Он пришел!*

(2) *Он пришел?*

Исследование выполнено за счет гранта Российского научного фонда № 22-28-01397, https://rscf.ru/project/22-28-01397.

Очевидно, что (1) представляет собою эмоциональное утверждение, а (2) — вопрос. Различие в пунктуации соответствует различию иллокутивной силы высказывания. Заметим, что в случаях такого рода различная пунктуация часто отражает различие интонации. Ср. также:

(3) *Он верно решил задачу.*

(4) *Он, верно, решил задачу.*

При естественном прочтении (3) сообщает, что решение задачи оказалось верным; слово *верно* представляет собою обстоятельство, и на него падает фразовое ударение. Напротив того, (4) выражает гипотезу — более или менее уверенное предположение, что субъект решил задачу; *верно* представляет собою вводное слово, а фразовое ударение при самом естественном прочтении падает на глагол *решил*[1].

Особый случай «нагруженной» вариантности представляют ситуации, когда пишущий может использовать знаки препинания для передачи каких-то факультативных дополнительных смыслов. Примером может служить использование кавычек. В большинстве случаев кавычки указывают на то, что происходит своего рода отклонение от «стандарта» семиотического акта [Зализняк 2007], однако пишущий может выбирать, требуется ли ему маркировать это «отклонение от стандарта».

Бывают и ситуации, когда при одном и том же (или почти одном и том же) понимании возможны разные пунктуационные решения и различие в пунктуации практически не влияет на интерпретацию высказывания, или, что то же самое, выбор знака препинания практически не зависит от смысла, который в высказывание хочет вложить пишущий (такую вариантность можно назвать «ненагруженной» вариантностью). Так, в «Полном академическом справочнике» ([Лопатин 2007]; далее — ПАС) содержится § 129, в котором говорится, что в бессоюзном сложном предложении ставится двоеточие, если вторая часть содержит пояснение, причину, обоснование того, о чем говорится в первой, или имеет изъяснительное значение; но далее к этому параграфу добавлено примечание, в котором говорится, что в бессоюзном сложном предложении «при обозначении пояснения, причины, обоснования, изъяснения допустимо употребление тире вместо двоеточия» (автор раздела «Пунктуация» — Н. С. Валгина). Собственно, возможность «ненагруженной» вариантности предусмотрена в преамбуле к разделу «Пунктуация», в котором говорится, что «в формулировки правил иногда включаются обороты типа *допускается употребление, может ставиться, возможен знак, допустимо употребление* и т. п.».

1.2 «Синтетические» и «аналитические» программы

Возможность вариантности пунктуации (как «нагруженной», так и «ненагруженной») обусловливает различие двух типов компьютерных программ, задающих пунктуацию: «синтетических» и «аналитических».

«Синтетические» программы для любого текста, поданного на вход и не содержащего знаков препинания (но, скорее всего, тем или иным образом разбитого на предложения), порождают множество «правильных» расстановок знаков препинания. При этом в «синтетических» программах на вход может подаваться письменный текст или звучащая речь. Если на вход подается письменный текст, то программа не может использовать интонацию для распознавания смыслового задания. Поэтому для фразы *Он пришел* такая программа должна породить по крайней мере варианты (1) и (2), а для фразы *Он верно решил задачу* — варианты (3) и (4). (Впрочем, контекст и презумпция связности текста может во многих случаях способствовать разрешению неоднозначности.) Если на вход подается звучащая речь, то программа, по существу, оказывается моделью такой довольно часто встречающейся деятельности, как запись под диктовку. Тем самым подобные программы могут иметь не только прикладное, но и теоретическое значение, поскольку моделируют распространенную человеческую деятельность.

[1] В этом случае правильно расставленные знаки препинания могут использоваться (и нередко используются) при анализе текста, в том числе в рамках автоматической обработки текста, как средство разрешения неоднозначности (см., в частности, [Бердичевский, Иомдин 2007]).

«Аналитические» программы оценивают уже готовую расстановку знаков препинания, поданную на вход, как «правильную» или «неправильную» и для предложений с исходной «неправильной» расстановкой предлагают возможные исправления. Такие программы имеют несомненное практическое значение и могу использоваться как изолированно, так и в составе программ-редакторов. При этом в основе таких программ может лежать метод «анализа через синтез»: программа использует результат работы «синтетической» программы, т. е. множество правильных расстановок знаков препинания для данного предложения (в данном контексте). После этого проверяемая расстановка сопоставляется с этими расстановками, и в случае несовпадения ее ни с одной из «правильных» расстановок этот факт отмечается как подлежащий дополнительной проверке (по-видимому, в интерактивном режиме).

При этом как «синтетические», так и «аналитические» программы используют правила пунктуации, приведенные в различных справочниках, из которых самым авторитетным может считаться ПАС. Однако использование справочников при создании программ затруднено тем, что сформулированные в них правила неполны и недостаточно эксплицитны (а иногда при буквальном применении даже могут привести к неправильной расстановке знаков препинания). Не вдаваясь в разбор технических особенностей указанных двух типов программ расстановки знаков препинания, рассмотрим некоторые случаи неполноты и неэксплицитности в ПАС и других справочниках и наметим пути, на которых правила можно сделать более полными и эксплицитными.

2 Лексическое наполнение пунктуационных правил

2.1 ПАС: «Указатель слов к разделу “Пунктуация”»

Во многих случаях для выбора пунктуационного оформления существенно наличие в предложении тех или иных лексических единиц. Указание на это обычно содержится в формулировке соответствующих правил; примером может служить целый ряд формулировок правил в ПАС. Для облегчения поиска правил, в которых упоминается то или иное слово, при пользовании бумажным вариантом правил в ПАС есть специальный «Указатель слов к разделу “Пунктуация”». Но мы с удивлением обнаруживаем, что в этом указателе содержатся далеко не все слова, упоминаемые в правилах. Так, в ПАС в § 129 формулируется правило, согласно которому в бессоюзном сложном предложении между частями ставится двоеточие, в частности, в тех случаях, когда «вторая часть бессоюзного сложного предложения имеет значение **изъяснительное**, что подчеркивается глаголами, помещенными в первой части предложения и предупреждающими о последующем изложении какого-либо факта». И далее правило конкретизируется: «Если имеются глаголы *видеть, понимать, слышать, смотреть, узнать, думать, чувствовать* и др., то между частями сложного предложения можно вставить союз *что*; если же имеются глаголы *выглянуть, взглянуть, оглянуться, прислушаться, посмотреть*, т. е. глаголы, не способные присоединять изъяснение непосредственно, то можно вставить сочетания *и увидел, что*; *и услышал, что; и почувствовал, что* и др.». Но ни одного из упомянутых глаголов в указателе нет! Кроме того, оба приведенных списка завершаются прямым указанием на то, что перечни не полны (*и др.*), но остаются совершенно неясными точные критерии отнесения конкретного глагола, не вошедшего в перечень, к данному множеству глаголов. Мы можем заметить, что первый список содержит глаголы несовершенного вида (кроме глагола *узнать*), а второй — глаголы совершенного вида. Относятся ли к нему видовые корреляты глаголов первого списка: *увидеть, понять, услышать, посмотреть, узнавать, подумать, почувствовать* (заметим, что глагол *посмотреть* включен во второй список) и видовые корреляты глаголов второго списка: *выглядывать, оглядываться, прислушиваться, смотреть* (снова обратим внимание на то, что глагол *смотреть* — видовой коррелят глагол *посмотреть* — включен в первый список)? И относятся ли к этому правилу синонимичные и квазисинонимичные глаголы *узреть, уразуметь, почуять, взглянуть, выяснить, решить, ощутить*?

Между тем некоторые из рассматриваемых глаголов существенны и для других правил пунктуации. Так, в ПАС приводятся списки вводных слов, которые следует выделять или отделять запятыми. В эти списки включены слова, которые «заключают в себе указание на источник сообщения», в том числе глаголы *вижу* и *думаю* (ПАС, § 91, примечание 1, пункт *г*). Приведенное

описание семантики этих слов не вполне точно[2], но дело даже не в этом. Глаголы приведены в форме первого лица единственного числа, но это не единственная форма, в которой глаголы *видеть* и *думать* употребляются как вводные слова. Ср. лишь несколько из многочисленных примеров употребления этих глаголов в функции вводных слов в Национальном корпусе русского языка (далее — НКРЯ):

(5) *…читал книги, читал, все прочел, а толку, видит, мало — собрал их в куль, да бросил в Волгу.* [Д. С. Мережковский]

(6) *Потом уже больше не мучили — все равно, видят, человек и сам помирает, — а отвезли его в Басов Кут и бросили…* [А. И. Куприн]

(7) *Мне сегодня Митрич, у которого я гостевал, рассказывал, как Рокоссовского прямо из лагеря в командармы произвели: стоял в барачной умывалке и портянки стирал, а за ним бегут: скорей! Ну, думает, портянок достирать не дали…* [Василий Гроссман]

(8) *Снял в келье люстру, подергал крюк — вроде крепкий, веревочку хорошенько намылил — мыло «Клубничное», пахучее такое, противное, начал уже стол пододвигать, смотрит: на столе лежит конфета. «Стратосфера», самая любимая его с детства, там еще розовые ракетки улетают в синий открытый космос. Конфеты дали вчера в трапезной на обед ради престольного праздника, а он сберег, да и забыл. Ладно, думает, съем, а там уж повешусь.* [Майя Кучерская]

2.2 Прямое и нарративное употребление глаголов в качестве вводных слов

Более детальное описание использования таких глаголов в качестве вводных слов потребует различения прямого и нарративного употребления. В нарративном режиме практически нет ограничений на формы лица и времени; однако глагольные формы употребляются без подлежащего либо с подлежащим в постпозиции; ср. примеры из НКРЯ:

(9) *Странная вещь, думала я, мое отношение к Мите.* [И. Грекова]

(10) *Умные люди, думал он, примирились с высшей исторической необходимостью…* [Юрий Давыдов]

(11) *А в самом деле, думал он, почему бы не прийти вечером.* [Виктория Токарева]

Напротив того, в прямом режиме используется почти исключительно первое лицо и при этом нередко внутри вводного оборота в препозиции имеется подлежащее (местоимение *я*), напр.:

(12) *Именно на этой территории, я думаю, происходит все самое главное.* [Сергей Довлатов]

(13) *Кто-то, я думаю, это учел и использовал…* [Вера Белоусова]

Необходимо подчеркнуть, что поведение многих лексических единиц, упоминаемых в правилах пунктуации (в частности, в ПАС), заслуживает отдельного рассмотрения, и подача их в виде списка может вводить в заблуждение. Ограничимся одним примером. Среди вводных слов и сочетаний слов, которые «заключают в себе указание на источник сообщения», в ПАС приведены формы *вижу* и *слышу* (всего список включает 30 с лишним выражений, причем список явно не исчерпывающий). Как уже говорилось, глаголы *видеть* и *слышать* упомянуты также среди глаголов, после которых ставится двоеточие в бессоюзном сложном предложении. Может создаться впечатление, что поведение этих глаголов совершенно одинаково. Однако такое впечатление ложно. Форма *вижу* свободно используется как в бессоюзном сложном предложении, так и в

[2] Семантике и прагматике вводности на материале английских глаголов посвящена классическая статья [Urmson 1970].

качестве вводного слова, в том числе в прямом режиме, причем в этом последнем случае при этой форме может быть или не быть подлежащее *я*. Ср. примеры из НКРЯ:

(14) *Вдруг вижу: идет мужчина с портфелем…* [Григорий Горин] — бессоюзное сложное предложение.

(15) *Он вот и с вами, вижу, коньяк выпивал…* [Василий Шукшин] — вводное слово.

(16) *Вот вы, я вижу, считаете это дело законченным.* [Вера Белоусова] — употребление в составе вводного оборота.

Напротив того, для формы *слышу*, свободно используемой в бессоюзных сложных предложениях, употребление в прямом режиме качестве вводного слова не очень характерно. Зато чрезвычайно часто встречается употребление качестве вводного слова оборота *я слышал(а)*, тогда как совершенно не характерно такое употребление для оборота *я видел*. Ср. примеры из НКРЯ:

(17) *Иду по коридору и слышу: во всех кабинетах, на всех этажах включено радио.* [Светлана Алексиевич] — бессоюзное сложное предложение.

(18) *Коньяк, я слышал, требует больших бокалов, не так ли?* [И. Грекова] — употребление в составе вводного оборота.

2.3 Пунктуационный словарь-указатель как неотъемлемая часть правил пунктуации

Все сказанное иллюстрирует общий тезис: отсутствие рассмотренных глаголов в «Указателе слов к разделу “Пунктуация”» в ПАС приходится признать упущением, которое препятствует формализации правил пунктуации. Разумеется, само по себе пополнение указателя не сделало бы правила более полными и эксплицитными. Однако оно могло бы помочь выявить лакуны и неточности в формулировке правил.

Здесь уместно обратить внимание на различие в кодификации орфографических и пунктуационных норм. В «Указатель слов к разделу “Орфография”» в ПАС включены все без исключения слова, упоминаемые в правилах (т. е. все слова, выделенные в тексте правил курсивом). Иногда такой сугубо формальный подход приводит к забавным казусам. Так, в указателе мы находим совершенно однотипные имена собственные *Ван Дейк* и *Ван Гог*, но отсылают они к разным параграфам, поскольку Ван Дейк упоминается в параграфе, посвященном слитному, раздельному и дефисному написанию, а Ван Гог — в параграфе, посвященном употреблению прописных букв. Но в целом «словоцентричный» подход к кодификации орфографии, с которым отчасти связана тщательность в составлении указателя слов, представляется совершенно оправданным. В соответствии с этим подходом «правильность» написания в конечном счете определяется орфографическим словарем, а «общие правила получают статус более или менее эффективных приемов, позволяющих предсказать (но всё же без полной гарантии), какое написание даст орфографический словарь» [Зализняк 2002: 587] (ср. также [Шмелев 2021: 16]). Собственно, и программы автоматической проверки орфографической правильности ориентируются именно на словарь.

В отличие от орфографических норм, пунктуационные нормы не могут быть исчерпывающим образом быть представлены в форме словаря. Тем не менее представляется, что словарь-указатель должен быть неотъемлемой частью правил пунктуации. В этом отношении важным шагом было создание словаря-справочника [Пахомов, Свинцов, Филатова 2012], в котором русские пунктуационные нормы впервые в истории были оформлены как словарь. Однако существенный недостаток этого справочника состоит в том, что составители последовали традиции составления указателей слов к правилам пунктуации и не включили в словарь многие единицы, непосредственным образом влияющие на пунктуацию (в частности, нет в справочнике словарных статей *видеть*, *думать*, *слышать*).

3 Предметный указатель

Помимо двух названных выше указателей слов, в ПАС есть еще «Предметный указатель к разделу “Пунктуация”». Необходимость в таком указателе в том виде, в каком он содержится в ПАС, представляется несколько сомнительной: по существу, это не указатель с собственном смысле слова, а скорее расширенное оглавление, поскольку включенные в него «предметы» следуют друг за другом в том же порядке, в каком идут параграфы справочника. Значительно полезнее был бы предметный указатель, устроенный совершенно по-другому. В этот указатель могли бы быть включены все термины, используемые в правилах, причем в бумажной версии они могли бы следовать друг за другом в алфавитном порядке; при каждом термине были бы указаны параграфы и пункты, в которых этот термин упоминается, а также по возможности дано более или менее формализованное определение того, что за данным термином кроется (если такое определение дать не удается, это может быть симптомом того, что соответствующий термин в правилах лучше не использовать или использовать с осторожностью). Примерами терминов, которые могли бы быть включены в такой указатель, могут служить выражения *бессоюзное сложное предложение*, *вводное слово*, *дополнение*, *однородные члены предложения*, *прямая речь*. Попытки точнее определить значения этих и других подобных терминов в определенных случаях могут способствовать прояснению того, в каких случаях некоторое рассматриваемое правило следует считать императивным, а в каких его применение остается на усмотрение пишущего. В последнем случае соотношение между пунктуацией и смыслом высказывания может оказаться обратным тому, как оно сформулировано в правиле. Так, между согласованными определениями в препозиции к определяемому слову при отсутствии союза положено ставить запятую, если определения однородные, и обходиться без запятой, если определения неоднородные (ПАС, § 37). Как узнать, нужна ли запятая в сочетании *сладкие (,) сдобные булочки*? По букве правила запятая нужна, если *сладкие* и *сдобные* — однородные определения, и не нужна, если *сладкие* и *сдобные* — неоднородные определения. Но фактически вопрос о наличии / отсутствии запятой здесь остается на усмотрение пишущего: если он поставил запятую, то, скорее всего, употребил эти определения как однородные, а если обошелся без запятой, то употребил определения как неоднородные. Это типичный пример «нагруженной» пунктуации[3].

Коснемся еще термина «прямая речь». Это один из немногих терминов, получивших если не определение, то краткое неформальное описание в ПАС. Термин объясняется так: «речь другого лица, включенная в авторский текст и воспроизведенная дословно». Само это объяснение несколько условно. Ведь так же, как прямая речь, может оформляться, напр., передача мыслей другого лица, когда трудно провести границу между «дословной» и не «дословной» передачей. Ср.:

(19) *Идя с допроса, он <u>думал</u>: «Нет, надо было бы ему все-таки рассказать про Эдинова».*
[Ю. О. Домбровский]

(20) *Мы <u>думали</u>: «Опубликуемся на Западе, и все узнают, какие мы гениальные ребята».*
[Сергей Довлатов]

Проблема в том, что предложение с так определенной прямой речью трудно отделить от некоторых смежных явлений, в частности от бессоюзных сложных предложений и предложений с вводными словами. Ср. пример сложного предложения с бессоюзной связью:

(21) Целую дитя у дверей —
беги, не опаздывай в школу, —
и <u>думаю</u>: из лагерей
вернут ли Руденко Мыколу?..
[Ю. П. Мориц]

[3] Ср. также замечания в [Левонтина 2021: 261–264] относительно коллизий, связанных с необходимостью отличать междометие *ну* от частицы *ну*.

Ясно, что последние две строчки могли бы получить и другое пунктуационное оформление — в соответствии с правилами оформления прямой речи:

и думаю: «Из лагерей
вернут ли Руденко Мыколу?..»

Относительно предложений с вводными словами следует упомянуть, что в ПАС к §§ 133–136, в которых описываются правила пунктуационного выделения прямой речи при ее подаче в строку (в подбор), дано примечание, согласно которому прямая речь не выделяется кавычками, «если прямое указание на источник сообщения оформлено как вводная конструкция», и приведен пример: *Статья ученого, сообщает критик, вызвала большой интерес общественности*. С другой стороны, рассмотренные выше примеры (8)–(11), в которых глаголы употреблены в нарративном режиме, могли бы быть оформлены и как предложения с прямой речью: *«Ладно, — думает, — съем, а там уж повешусь»*; *«Странная вещь, — думала я, — мое отношение к Мите»*; *«Умные люди, — думал он, — примирились с высшей исторической необходимостью»*; *«А в самом деле, — думал он, — почему бы не прийти вечером»*.

Встречается и «промежуточное» оформление предложений с «дословной» или почти «дословной» передачей слов или мыслей другого лица. Так, в следующем примере поставлены кавычки, как при прямой речи, но «слова автора» не отделены от прямой речи при помощи тире, как этого требуют правила:

(22) *Ходил, ходил, ездил в автобусах, объясняясь по преимуществу мычанием, и, наконец, проголодался, как зверь, заехал куда-то, черт его знает куда. «Дай, думает, зайду в ресторанчик, перекушу». [М. А. Булгаков]*

Такие примеры довольно часто обнаруживаются в НКРЯ, и, хотя пунктуация в них противоречит букве правила ПАС, они не воспринимаются как написанные с ошибками. При этом выбор пунктуационного оформления в них едва ли можно считать «нагруженным».

4 Заключительное замечание

Разумеется, приведенные соображения не только не исчерпывают проблем, связанных с недостаточной полнотой и эксплицитностью правил пунктуации, содержащихся в справочниках, но охватывают лишь ничтожно малую их часть. Совсем за пределами рассмотрения остались проблемы, которые часто вообще игнорируются справочниками или рассматриваются в них кратко и поверхностно: сочетание и возможное поглощение знаков препинания, а также их расположение в тех случаях, когда правила требуют постановки в одном и том же месте предложения разных знаков, взаимодействие знаков препинания в сложных конструкциях (указанные две проблемы обсуждаются в ПАС, но заведомо неполно), проблема пунктуации в предложениях, когда главная клауза сложноподчиненного предложения как бы «вставляется» внутрь придаточной, парцелляция, знаки во вставных конструкциях, специфика пунктуации в бумажной и электронной переписке[4] и т. п.

Все сказанное выше носит скорее иллюстративный характер и направлено на то, чтобы обосновать главный тезис: компьютерные программы автоматической расстановки знаков препинания или проверки пунктуации более всего страдают от нечеткости формулировок, рассчитанных на человека.

[4] Как показало замечание одного из рецензентов «Диалога», упоминание специфики пунктуации в бумажной и электронной переписке требует некоторого пояснения. Речь, разумеется, идет не о пунктуационном узусе (едва ли можно выяснить, имеется ли специфика пунктуации в бумажной переписке, в силу отсутствия репрезентативных корпусов, а в электронной переписке специфика узуса иногда сводится к низкому уровню грамотности, а в неформальной переписке еще и к намеренному игнорированию тех или иных пунктуационных норм). Имеются в виду представления грамотных носителей языка об эпистолярных пунктуационных нормах (подчас выраженные эксплицитно). Какой знак следует ставить после обращения перед основным текстом письма? Нужна ли запятая после заключительной формулы (напр., *с глубоким уважением*) перед подписью? Некоторые участники электронной переписки считают, что в цепочке сообщений, когда обращения и подписи уже не нужны, в конце каждого из звеньев цепочки не следует ставить точку, которая воспринимается как знак конца всей цепочки.

References

[1] Berdichevsky Alexander, Iomdin Boris (2006). The role of punctuation in disambiguation [Rol' punktuatsii v razreshenii neodnoznachnosti], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2006”], Bekasovo, pp. 44–49.

[2] Levontina I. B. (2021) Problems of punctuation of discourse words [Problemy punktuatsionnogo oformleniya diskursivnykh vyrazhenii] // Proceedings of the V. V. Vinogradov Russian Language Institute [Trudy Instituta russkogo yazyka im. V. V. Vinogradova], 2021, № 3, pp. 260–270.

[3] Lopatin V. V. (ed.) (2007). Rules of Russian Spelling and Punctuation: Complete Academic Reference Book [Pravila russkoi orfografii i punktuatsii. Polnyi akademicheskii spravochnik]. Moscow: Eksmo Publ.

[4] Pakhomov V. M., Svintsov V. V., Filatova I. V. (2012). Trudnye sluchai russkoi punktuatsii: Slovar'-spravochnik — Moscow: Eksmo Publ.

[5] Shmelev A. D. (2021). Russian orthography code: an outline plan in light of integrality principle [Prospekt “Svoda pravil russkoi orfografii” v svete printsipa integral'nosti] // Words, Constructions, and Texts in the History of the Russian Written Language [Slova, konstruktsii i teksty v istorii russkoi pis'mennosti]. — Moscow; Saint-Petersburg: Nestor-Istoriya. pp. 7–18.

[6] Urmson James Opie (1970). Parenthetical verbs. // Caton Ch. E. (ed.). Philosophy and ordinary language. — Urbana, Univ. of Illinois Press, 1970. — P. 220-240.

[7] Zaliznyak A. A. (2002). Old Russian graphic system with ъ — *о* and ь — *е* variation [Drevnerusskaya grafika so smesheniem ъ—*о* i ь—*е*] // Russian Nominal Declension. With an Appendix Containing Selected Papers in Russian and General Linguistics [Russkoe imennoe slovoizmenenie. S prilozheniem izbrannykh rabot po russkomu yazyku i obshchemu yazykoznaniyu]. Moscow: Yazyki slavyanskoi kul'tury Publ., pp. 577–612.

[8] Zalizniak Anna A. (2007), The semantics of inverted commas [Semantika kavychek], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006” [Komp’yuternaya Lingvistika i Intellektual’nye Tekhnologii: Trudy Mezhdunarodnoy Konferentsii “Dialog 2006”], Bekasovo, pp. 188–193.

Литература

[1] Urmson J. O. Parenthetical verbs. // Philosophy and ordinary language. Caton Ch. E. (ed.). Urbana, Univ. of Illinois Press, 1970, p. 220-240. [Русский перевод: Урмсон Дж. Парентетические глаголы // Новое в зарубежной лингвистике. Вып. 16: Лингвистическая прагматика. — М.: Прогресс, 1985, с. 196–216].

[2] Бердичевский А. С., Иомдин Б. Л. Роль пунктуации в разрешении неоднозначности // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2007. С. 44–49.

[3] Зализняк А. А. Древнерусская графика со смешением ъ—*о* и ь—*е* // Зализняк А. А. «Русское именное словоизменение» с приложением избранных работ по русскому языку и общему языкознанию. — М.: Языки славянской культуры, 2002. — С. 577–612.

[4] Зализняк Анна А. Семантика кавычек // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог 2007» (Бекасово, 30 мая – 3 июня 2007 г.) / Под ред. Л.Л. Иомдина, Н.И. Лауфер, А.С. Нариньяни, В.П. Селегея. – М.: Изд-во РГГУ, 2007. С. 188–193.

[5] Левонтина И. Б. Проблемы пунктуационного оформления дискурсивных выражений // Труды Института русского языка им. В. В. Виноградова, 2021, № 3. С. 260–270.

[6] Лопатин В. В. (ред.). Правила русской орфографии и пунктуации. Полный академический справочник. — Москва: Эксмо, 2007.

[7] Пахомов В. М., Свинцов В. В., Филатова И. В. Трудные случаи русской пунктуации: Словарь-справочник — М.: Эксмо, 2012.

[8] Шмелев А. Д. Проспект «Свода правил русской орфографии» в свете принципа интегральности // Слова, конструкции и тексты в истории русской письменности. — М.; СПб: Нестор-История, 2021. С. 7–18.

The role of Indicators in Argumentative Relation Prediction

Sidorova Elena
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
lsidorova@iis.nsk.su

Akhmadeeva Irina
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
i.r.akhmadeeva@iis.nsk.su

Kononenko Irina
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
irina_k@cn.ru

Chagina Polina
A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia
p.chagina@gmail.com

Abstract

The article presents a comparative study of methods for argumentative relation prediction based on a neural network approach. The distinctive feature of the study is the use of argumentative indicators in the preparation of the training sample. The indicators are generated based on the discourse marker dictionary. The experiments were carried out using an annotated corpus of scientific and popular science texts, including 162 articles available on the ArgNet-Bank Studio web platform. A set of all argumentative relations is described by internal connections of arguments and include the conclusion and the premise. In the first stage of training set construction, fragments of text that included two consecutive sentences were examined. In the second stage, indicators were retrieved from the corpus texts and, for each indicator, statements presumably corresponding to the premise and conclusion of the argument were extracted. In total, 4.2 thousand indicator-based training contexts and 13.6 thousand pairs of sentences were obtained from the corpus with annotation of the presence of an argumentative relation. Based on this training sample, four classifiers were built: without indicators, with marking indicators in sentences using tags, taking into account segmentation of text based on indicators, with segmentation and tags. The results of the experiments on argumentative relation prediction are presented.

Keywords: argument mining; text corpus; argument annotation of text; argumentation indicator; argument scheme; argumentative relation prediction

DOI: 10.28995/2075-7182-2023-22-477-485

Исследование роли индикаторов при извлечении аргументативных отношений

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

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

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

Чагина Полина
Институт систем информатики им. А.П. Ершова СО РАН, Новосибирск, Россия
p.chagina@gmail.com

Аннотация

В статье проводится сравнительное исследование методов извлечения аргументативных отношений на основе нейросетевого подхода. Особенностью исследования заключается в использовании индикаторов аргументации при подготовке обучающей выборки. Индикаторы сгенерированы на основе словаря дискурсивных маркеров и задаются набором лексико-синтаксических шаблонов. Для экспериментов использовался размеченный корпус научных и научно-популярных текстов, включающий 162 статьи, размещенные на веб-платформе ArgNetBank Studio. Множество всех аргументативных отношений описываются внутренними связями аргументов и включают заключение и посылку. Построение обучающей выборки проходило в два этапа. На первом этапе рассматривались фрагменты текста, включающие два подряд идущих предложения, и отмечалось наличие или отсутствие аргументации. Считалось, что аргументация присутствует, если фрагмент включал заключение и хотя бы одну посылку одного и того же аргумента из разметки. На втором этапе осуществлялся поиск индикаторов и для каждого индикатора извлекались утверждения, предположительно соответствующие посылке и заключению аргумента. Каждый такой набор размечался аналогично по наличию аргументативного отношения в аннотации. Всего на основе корпуса было получено около 4,2 тысяч обучающих контекстов на основе индикаторов и 13,6 тысяч пар предложений с разметкой наличия аргументативной связи. На основе данной обучающей выборки было построено четыре классификатора: без учета индикаторов, с разметкой индикаторов в предложениях с помощью тегов, с учетом сегментации текста на основе индикаторов, с сегментацией и тегами. Приведены результаты экспериментов по извлечению аргументативных отношений.

Ключевые слова: анализ аргументации; корпус текстов; аргументативная разметка текста; индикатор аргументации; схема аргумента; извлечение аргументативных отношений

1 Introduction

Over the past two decades, the study of argumentation involves, in particular, describing the structure of a text in the form of statements connected by relations of support or conflict. Argument Mining is a field of computational linguistics, which has been actively developing during the last decade. Its goal is to automatically extract arguments represented by a sequence of statements ("premises") leading to a certain conclusion ("thesis") from texts. Automating the extraction of arguments from texts became a priority area only a few years ago [8].

The analysis of argumentation presented in a natural language text requires not only the extraction of arguments and argument chains supporting or disproving a thesis (abstract argumentation), but also the exploration of the structure of each argument and its role and relevance to the argument as a whole (structural argumentation). Models or schemes of arguments are used to describe different ways of reasoning [16]. The best known compendium of structured argumentation that has found application in practical systems of argument analysis is that of D. Walton [18]. It contains about 60 argumentation schemas, based on which an ontology of argumentation (AIF-ontology) was constructed in [11].

One of the main conditions for the development of this field is the creation of corpuses of texts with argumentative annotation. The best known resource with argumentation annotation is the AIFdb database, formerly the Araucaria corpus [7], which contains news articles, records of parliamentary and political debates, etc. — a total of 170 corpora of varying size and quality in 14 languages. However, the main research languages are still English and, to a lesser extent, German, and the data themselves have different annotation schemes, making them impossible or very difficult to use combined. There are very few such resources for the Russian language. The annotated corpus of sentences with annotation of the presence of argumentation ("for" or "against") was developed as part of the RuARG-2022 competition [6]. In [2] a web-based resource for the analysis of argumentation in popular science discourse is presented. The annotation model is based on the ontology of argumentation and D. Walton's argumentation schemes [18].

An important linguistic aspect of the study of discourse is the registration of discourse markers - linguistic instruments of structuring discourse, which play a key role in the process of its understanding. Thus, indicators of argumentation simplify the identification and reconstruction of the steps of argumentation that are carried out in an argumentative dialogue or text [5]. The aim of our work is to investigate the role of indicators in detecting argumentative relations and evaluating their effectiveness. The main research tools are annotated text corpora and dictionaries of indicators of argumentative relations.

2 Related works

The solution of Argument Mining (AM) task involves solving the following subtasks, which can be formulated as classification problems:

1. Detection of text fragments containing argumentation (Argument Detection);
2. Classification of statements according to the used argumentation scheme (Argument Component Classification);
3. Identification of relationships between argument components (Argumentative Relation Prediction);
4. Classification of arguments according to the classes presented in the ontology of argumentation (Argument Classification).

According to the multiple reviews [8, 13-14, 17, 19] it is clear that modern pre-trained Deep Learning models (DL), such as BERT, have shown good results on many AM related tasks and they are currently one of the main tools in the field of AM. The subtask that is called either Edge Prediction or Relation Prediction is considered the most difficult part of Argument Mining. Currently there are not so many papers dedicated to the applying of modern NLP techniques to the Relation Prediction problem. The results demonstrated by modern DL models, however, remain comparable to the results of classical models, such as, for example, SVM. It was shown in [3] that, despite having a superior performance on the Argument Component Classification problem with F1 score = 0.86 against 0.79 as the best of the other models, the BERT-based (BERT-base-uncased) model was inferior in performance on the Argumentative Relation Prediction (ARP) task. Trained on the CDCP corpus it obtained F1 score = 0.15 against 0.34 of the LSTM-based model [9] and 0.27 of the SVM with GloVE embeddings as an input.

Lexical features are applied when teaching classical ML models in Argument Detection and Argument Component Classification tasks [14]. When analyzing argumentation in Russian-language texts, it is necessary to study the composition, structure, and role of both primary and secondary connectors of the Russian language [15] used as indicators of argumentation.

From the analysis of recent works we can conclude that the problem of Argumentative Relation Prediction is far from being solved, and, depending on the data and their annotation, a broad range of modern techniques can be applied: from traditional ML models with various features to DL models, and lexical and syntactic features are an important part of the training of classical models. Also, works on Argument Mining do not pay enough attention to the role of lexical features, such as indicators of argumentation (markers) and n-grams, genre segmentation of the text, and the possibility to apply knowledge of the rhetorical structure of the text. For Russian, this problem is even more relevant due to the small amount of annotated data.

3 Corpus of texts with argumentation markup

For this study we used the annotated corpus of scientific and popular science texts, including 162 articles available on the ArgNetBank Studio web platform (https://geos.iis.nsk.su/arg). Each text was annotated according to the AIF (Argument Interchange Format) standard [4], by constructing an oriented connected graph (see Fig. 1) with two types of vertices: information vertices, which correspond to the statements (rectangular blocks), and relation vertices, which indicate the connections between the statements (oval blocks).

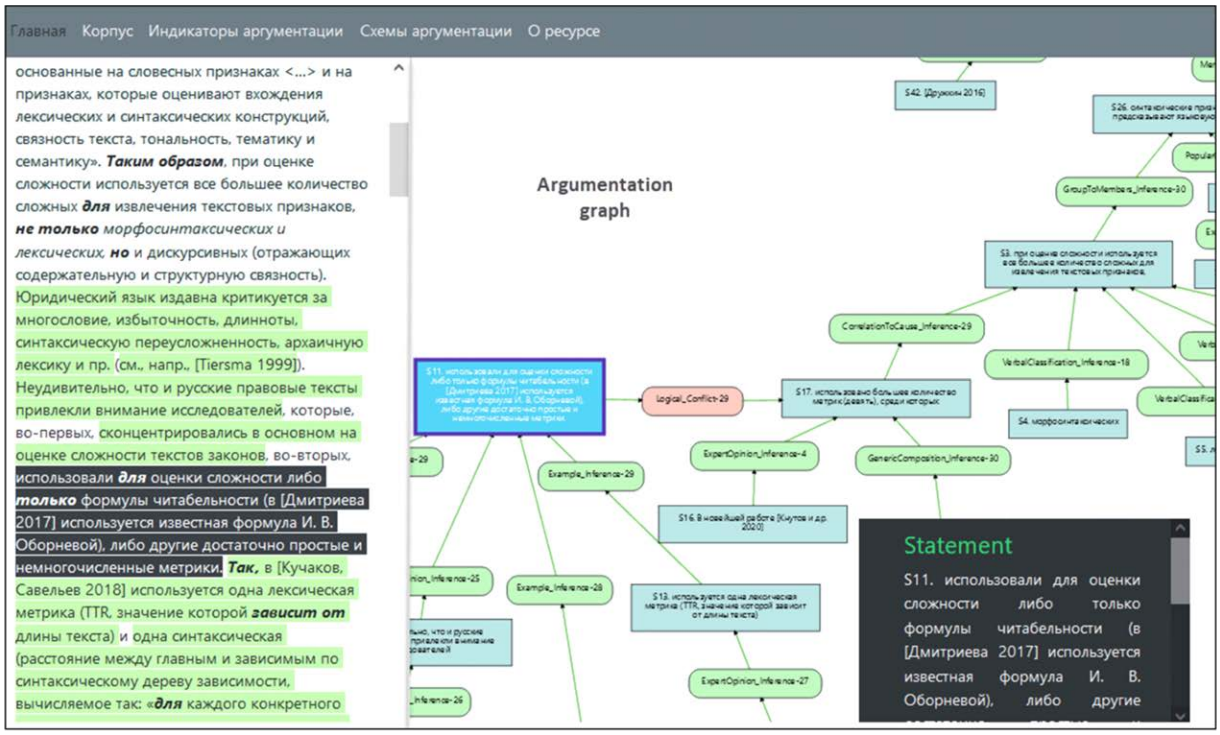


Figure 1: Argumentative text markup on the ArgNetBank Studio platform

A set of all argumentative relations in a graph is described by internal connections of arguments according to the typical models (schemes) of reasoning. Each argument has a conclusion and a premise (or several premises) in the case of support or conflict. The annotated corpora can be downloaded in json format and used in further research. The corpus articles belong to different subject areas and are relatively small (3,500 words on average). A higher-quality, expert-tested sub-corpus was selected from the corpus, from which texts were taken for testing. Training was performed on the entire corpus.

Argumentation indicators play an important role in the research. On the ArgNetBank Studio platform to bring the user's attention to the arguments presented explicitly in the text and to assist in highlighting the boundaries of statements in the text and in choosing the scheme of argumentation, a preliminary linguistic processing of texts is performed, which reveals specific cues in the text expressed as various kinds of verbal cliches, which are potentially indicators of argumentation (Fig. 1). The list of significant words which can act as indicators or anchor words of indicator constructions is heterogeneous and fundamentally incomplete. Therefore, the platform contains additional tools for finding, exploring and creating indicators (Fig. 2).

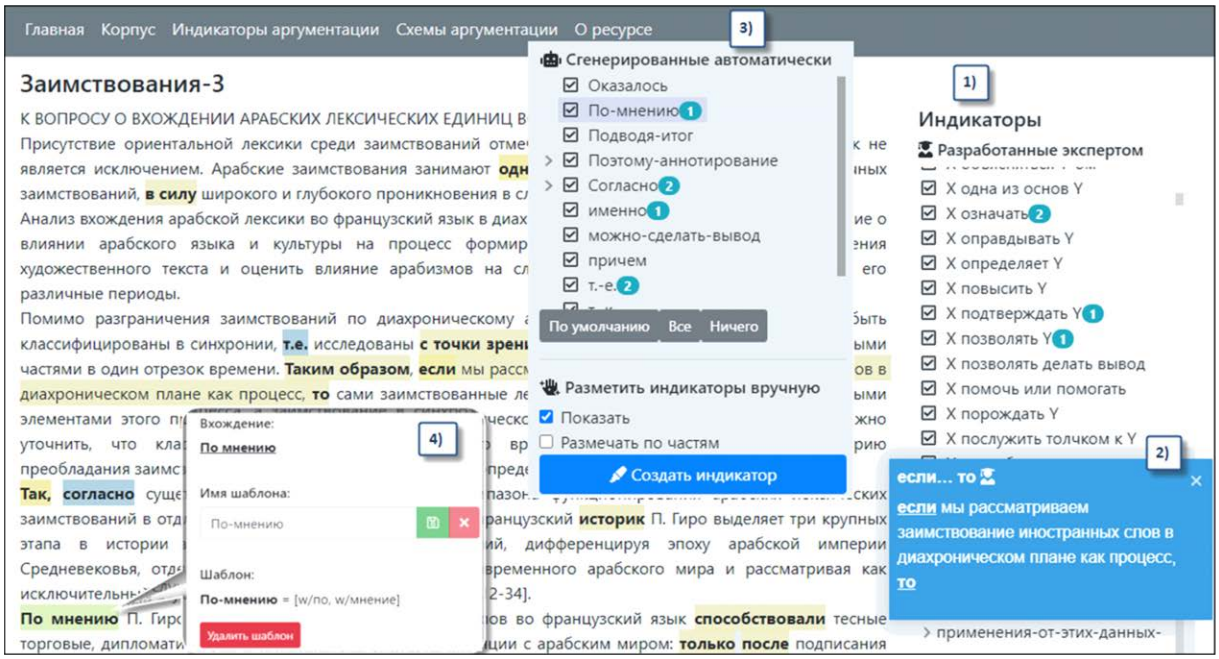


Figure 2: Argumentation indicators annotation on the ArgNetBank Studio platform

Developed tools provide loading dictionaries of lexico-syntactic patterns for search of indicators of argumentation (1), indicators annotation in the text (2), and concordance creation. On the basis of the fragment selected by a user (3), a pattern is automatically generated (4), in which the structure and normalized form of lexical units, punctuation marks and gaps (if the user has highlighted a splitted fragment) are captured.

4 Indicators of argumentation analysis

Indicators of argumentation are words and constructions used in a discourse that indicate the presence of an argument in the text. They help identify the presence of arguments and their components, identify the boundaries of statements in the text, reconstruct the relations between statements, and relate the argument to a certain scheme of reasoning (a form of deduction that expresses the relationship between premises and conclusions).

The indicator can signal different pragmatic aspects of argumentation [12]:

1. the degree of confidence the author has in the statement: *по-видимому 'seemingly'*, *уверен 'sure that'*;
2. the relation of inference between two statements (presence of argumentation): *следует что 'it follows that*, *если…то 'if…then'*;
3. the type of argumentative relation: *поскольку 'due to'* (support) *vs. хотя 'although'* (conflict);
4. the role of the statement in the inference: *потому что 'because of'* (premise) *vs. поэтому 'that's why'* (conclusion);
5. the semantic-ontological relation on which the typical scheme of reasoning is based in this case: *по причине 'by reason of'*, *X вызывает Y 'X causes Y'* (causation), *в частности 'in particular'*, *например 'for example'* (hyper-hyponymy), *похожий 'similar'* (analogy);
6. the structure of argumentation (multiple vs. sequential argumentation) *к тому же 'besides'*, *не говоря уже о 'not to mention'* (multiple argumentation) vs. *в конце концов 'eventually'* (sequential argumentation).

The original list of discourse markers contained 294 items, from which a list of 143 markers was manually selected. This list was also extended with previously developed indicators for expert opinion reasoning extraction.

Indicators are described in a formal pattern language that allows the use of tokens, arbitrary character sequences, auxiliary patterns, alternatives and gaps.

All markers and their contexts of use were extracted from the corpus of texts in order to study the indicators. For each marker, the context of its use was divided into three statements: the main statement, which included the indicator, and the right and left contexts. For each statement, its role in the structure of the argument was identified. Thus, the data have the following representation:

pattern | *main* | *left* | *right* | *main_arg* | *left_arg* | *right_arg* | *same_arg* | *text* | *sent_n*

where:

- pattern name (*pattern*) - name of the anticipated indicator;
- main statement (*main*) - the sentence containing the marker;
- left context (*left*) - part of the sentence preceding the indicator entry; if the marker is close to the beginning of the sentence, the sentence preceding the main statement (if any) is also taken;
- right context (*right*) - the sentence following the main statement;
- argumentation parameters for the main (*main_arg*), left (*left_arg*) and right (*right_arg*) statements - presence of argumentation and roles in the argument structure, which take values: 0 no argumentation, 1 the premise of the argument, 2 the conclusion of the argument, 3 the premise in one argument and the conclusion in another;
- binary argumentative relation parameter (*same_arg*) - the presence in the main statement and the left context of the premise and conclusion of the same argument (in any order), which indicates the presence of an argumentative relation and implicitly means that the marker is a true indicator of argumentation;
- *text* - reference to the text where the marker was encountered;
- sentence number (*sent_n*) - reference to the sentence in which the marker was encountered

A total of 4,207 patterns and their contexts were obtained from the corpus. Of these, 972 cases contained an argumentative relation. In other words, in only 23% of the cases the marker was a connector between the premise and the conclusion of an argument.

In addition, there were 1,496 cases of simultaneous occurrence of a premise and a conclusion in the same statement, which corresponds to the situation of sequential argumentation, i.e., when the statement is an intermediate (non-leaf) vertex in the argumentation graph.

In terms of identifying the boundaries of argumentative structures, indicators can be divided into the following functional groups:

1. Patterns that break a single sentence, containing a premise and a conclusion, into parts and specify the boundaries of statements.
   Эти варианты различны для разных видов контаминированной речи, ***например*** *воспроизведение английской или русской речи немца не похоже на передачу речи китайца.*
   'These variants are different for different types of contaminated speech, ***for example****, the reproduction of English or Russian speech by a German is not similar to the transmission of speech by a Chinese.'*
2. Patterns that are on the edge of a sentence and signal that the nearest sentence is part of the argument structure.
   Свою семантическую значимость пропозиция обретает только в рамках высказывания. ***Поэтому*** *необходимо обратиться к вербальным способам актуализации пропозиций победы в исследуемых текстах.*
   'A proposition acquires its semantic significance only within the framework of an utterance. ***Therefore****, it is necessary to turn to verbal ways of actualizing the propositions of victory in the texts under study.'*
3. Patterns with a gap that contains either a conclusion or a premise within it.
   Но тот факт, что радионуклид был выявлен на такой обширной территории, говорит о том, что активность в выбросе была весьма высокой.
   'But the fact that the radionuclide was detected over such a vast area suggests that the activity in the release was very high.'

An analysis of the relative positions of premise and conclusion with respect to the marker showed that most indicators (about 90%) allow us to accurately indicate which of the context statements will play the role of premise or conclusion in the case of argumentation detection. Thus, the use of the indicator looks promising, both for improving the quality of argumentative relations extraction, and for postprocessing involving the identification of the roles of statements in the argument structure.

5 Argument relation prediction

The training set construction consisted of two stages. In the first stage, fragments of text that included two consecutive sentences were examined and the presence or absence of argumentation was noted. Argumentation was considered to be present if the fragment included a conclusion and at least one supporting or refuting premise of the same argument from the annotation. In the second step, indicators were retrieved from the corpus texts and, for each indicator, statements presumably corresponding to the premise and conclusion of the argument were extracted. Each such set was annotated similarly by the presence of an argumentative relation in the annotation. In total, 4,207 indicator-based training contexts and 13,655 pairs of sentences were obtained from the corpus with annotation of the presence of an argumentative relation. Thus, the data for the experiments included about 18 thousand examples, of which 2,617 were positive examples and about 15,5 thousand were negative examples.

The ruRoberta (ai-forever/ruRoberta-large) model was used to represent the Russian text, where the two contexts are provided as input separated by the special token [SEP]. We use encoding output for the [CLS] token as the relation representation between two contexts. Then a fully connected neural network consisting of two linear layers with a ReLU activation function and a dropout layer between them is applied to the representation. Finally, a Softmax function was used to obtain the probability distribution of the argumentative relation. We used the following configurations to construct classifiers that predict the presence of argumentative relations.

1. Independent classification (**simple-model**): the classifier is applied to embeddings of sentences obtained by a sliding window of 2 sentences.
2. Independent classification (**simple-indicator-model**): the difference from simple-model is that additionally the argumentation indicators are marked with a special punctuation mark (`*`) similar to the work [1].
3. Classification taking into account the segmentation based on indicators (**context-model**): the classifier is applied to the statements obtained as the left and main indicator contexts; in the absence of an indicator, the partitioning is performed on sentences.
4. Classification with marking indicators (**context-indicator-model**): the difference from context-model is that additionally the argumentation indicators are marked with a special punctuation mark (`*`).

We carried out 5-fold cross validation over our dataset, with the same parameters used for all models in the process: learning rate = 3e-7, batch size = 4, epochs = 5. The results of the experiments are presented in Table 1.

Classifier	Precision	Recall	F1
simple-model(1)	19.87	51.94	28.55
simple-indicator-model(1)	19.85	46.04	27.63
context-model(1)	20.70	**65.90**	31.30
context-indicator-model(1)	**21.32**	65.25	**31.95**
simple-model(2)	41.38	53.13	46.31
simple-indicator-model(2)	41.20	54.14	46.60
context-model(2)	43.47	**66.65**	52.29
context-indicator-model(2)	**44.33**	66.48	**52.86**

Table 1: The results of the experiments

Analysis of the results of experiment (1) reveals problems with the quality of the corpus annotation (a high-quality annotation constitutes only about half of all the annotated data) and the problem of disagreement between different annotators. For the part of the corpus annotated by several experts, the agreement was 0.78 for annotating argumentative statements and 0.55 for annotating argumentative relations. Compared to the results of other studies [8] (for non-experienced annotators k = 0.58, and for experts k = 0.83) the data give worse results, which seems to be related both to the complexity of the annotation scheme and to the studied genre itself.

To solve this problem, the dataset was further processed to remove "badly" annotated texts: texts with abnormally low argumentation coverage were removed. The results of experiment (2) show a stable improvement in the quality of all classifiers.

Overall, the experimental data show that on this corpus, the use of indicators improves the quality of the classifiers performance on all three metrics. And segmentation based on indicators is more effective than simply marking indicators.

6 Conclusion

In this paper we continued our investigation of the role of indicators in argument extraction. While previously we considered only the problem of sentence detection, in this study the focus was on identifying the argumentative connection between two statements. The distinctive features of the applied approach include a) the study of Russian-language texts of scientific and popular science genre, b) the use of a corpus annotated according to one of the most difficult for automatic processing standards of argumentative annotation, c) the construction of one universal classifier instead of a chain of classifiers used consistently to solve the problem [8], d) the integration of the indicator approach with deep learning methods. Additionally, we have taken into account the drawback associated with the exclusion from consideration of text fragments that do not contain indicators.

Thus, further research will be related to the study of the following issues: a) improving the quality of annotation by developing annotation methodology for texts of scientific and popular science genres; b) enriching and refining the vocabulary of argumentation indicators; c) developing independent classifiers that identify whether a marker is an indicator in a given context; d) exploring the role of indicators for classifying argumentation schemes.

Acknowledgements

The work was funded by Russian Science Foundation according to the research project no. 23-21-00325.

References

[1] Alibaeva K., Loukachevitch N. Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”, 2022. – P. 8–17.

[2] Ilina Daria, Kononenko Irina, Sidorova Elena. On Developing a Web Resource to Study Argumentation in Popular Science Discourse. In Computational Linguistics and Intellectual Technologies // Proceedings of the International Conference “Dialog-2021”. 2021. — P. 318–327.

[3] Chen Ting, BERT Argues: How Attention Informs Argument Mining (2021). Honors Theses. 1589.

[4] Chesñevar C.I., McGinnis J., Modgil S., Rahwan I., Reed C., Simari G., South M., Vreeswijk G., Willmott S. Towards an argument interchange format. — The knowledge engineering review, 21(4), 2006. — P. 293-316.

[5] Eemeren F.H. VAN, Houtlosser P., Snoeck Henkemans F. Argumentative Indicators in Discourse: A Pragma-Dialectical Study. — Dordrecht: Springer. 2007.

[6] Kotelnikov E., Loukachevitch N., Nikishina I., Panchenko A. RuArg-2022: Argument Mining Evaluation // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialogue 2022”. 2022. — P. 1–16.

[7] Lawrence John, Chris Reed. AIFdb corpora // In Proceedings of the Fifth International Conference on Computational Models of Argument (COMMA 2014). — Pitlochry, 2014. — P. 465–466.

[8] Lawrence J., Reed C. Argument mining: A survey. — Computational Linguistics, 45(4), 2019. — P. 765-818.

[9] Niculae V., Park J., Cardie C. Argument mining with structured SVMs and RNNs // In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017. — Vol. 1: Long Papers. — P. 985–995.
[10] Pimenov I.S. Compatibility of Arguments from Different Functional Groups in Scientific Texts [Sochetaemost` argumentov raznykh funktcional`nykh grupp v nauchnykh tekstakh] // Philology. Theory & Practice. [Filologicheskie nauki. Voprosy` teorii i praktiki] Tambov: Gramota, 2022, vol. 11. pp. 3672-3680.
[11] Rahwan I, Banihashemi B, Reed C, Walton D, Abdallah S. Representing and classifying arguments on the semantic web. — The Knowledge Engineering Review, 26(4), 2011. — P. 487-511.
[12] Sidorova E.A., Ahmadeeva I.R., Kononenko I.S., Chagina P.M. (2022), Argumentation Extraction Based on Indicator Approach [Izvlechenie argumentatcii na osnove indikatornogo podhoda], Proceedings of the 20-th Russian Conference on Artificial Intelligence RCAI-2022 [Trudy 20-oi natcionalnoi konferentcii po iskusstvennomu intellektu KII-2022], Moscow, pp.219-233.
[13] Schaefer Robin, Stede Manfred. Argument mining on twitter: A survey // it - Information Technology. — Vol. 63(1). — P. 45–58.
[14] Stede Manfred. Automatic argumentation mining and the role of stance and sentiment // Journal of Argumentation in Context, 2020. — Vol. 9(1). — P. 19–41.
[15] Toldova S., Pisarevskaya D., Vasilyeva M., Kobozeva M. The cues for rhetorical relations in Russian: "Cause-Effect" relation in Russian Rhetorical Structure Treebank // Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference “Dialogue”, 2018. — P. 747–761.
[16] Toulmin S. The Uses of Argument. – Cambridge: Cambridge University Press, 2003.
[17] Vecchi Eva Maria, Falk Neele, Jundi Iman, Lapesa Gabriella. Towards argument mining for social good: A survey // Proceedings of the 59th Annual Meeting of the 9th Annual Meeting of the Association for Computational Linguistics, 2021. — P. 1338–1352.
[18] Walton D., Reed C., Macagno F. Argumentation schemes. — Cambridge University Press, 2008.
[19] Zhang G., Nulty P., Lillis D. A decade of legal argumentation mining: Datasets and approaches // International Conference on Applications of Natural Language to Information Systems. – Springer, Cham, 2022. — P. 240-252.

Text VQA with Token Classification of Recognized Text and Rule-Based Numerical Reasoning

Surkov V. O.
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
surokpro2@gmail.com

Evseev D. A.
Moscow Institute of Physics and Technology
Dolgoprudny, Russia
dmitrij.euseew@yandex.ru

Abstract

In this paper, we describe a question answering system on document images which is capable of numerical reasoning over extracted structured data. The system performs optical character recognition, detection of key attributes in text, generation of a numerical reasoning program, and its execution with the values of key attributes as operands. OCR includes the steps of bounding boxes detection and recognition of text from bounding boxes. The extraction of key attributes, such as quantity and price of goods, total etc., is based on the BERT token classification model. For expression generation we investigated the rule-based approach and the T5-base model and found that T5 is capable of generalization to expression types unseen in the training set. The proposed architecture of the question answering system utilizes the structure of independent blocks, each of which can be enhanced or replaced while keeping other components unchanged. The proposed model was evaluated in the Receipt-AVQA competition and on FUNSD dataset.

Keywords: visual question answering, optical character recognition, receipt images, token classification, numerical reasoning

DOI: 10.28995/2075-7182-2023-22-486-496

Ответ на вопросы по тексту на изображениях с помощью классификации токенов распознанного текста и численных рассуждений на основе правил

Сурков В. О.
Московский физико-технический институт
Долгопрудный, Россия
surokpro2@gmail.com

Евсеев Д. А.
Московский физико-технический институт
Долгопрудный, Россия
dmitrij.euseew@yandex.ru

Аннотация

В данной работе описывается система для ответа на вопросы по изображениям с текстом с возможностью численного рассуждения по извлеченным структурированным данным. Система выполняет распознавание текста на изображении, определение ключевых атрибутов в тексте, генерацию выражения для численного рассуждения и его выполнение с ключевыми атрибутами в качестве аргументов. Распознавание текста включает в себя следующие этапы: определение областей с текстом на изображении и последующий перевод их в текст. Извлечение ключевых атрибутов, таких как количество и цена товаров, сумма и т. д. выполняется моделью классификации токенов на основе BERT. Для генерации выражений были исследованы подход на основе правил и модель T5-base и установлено, что T5 способен к обобщению на типы выражений, не встречающиеся в обучающей выборке. Архитектура вопросно-ответной системы реализована в виде набора независимых блоков, каждый из которых может быть заменен или улучшен при сохранении остальных компонентов неизменными. Предложенная модель была применена в соревновании Receipt-AVQA и протестирована на датасете FUNSD.

Ключевые слова: ответ на вопросы по изображениям, распознавание текста, изображения товарных чеков, классификация токенов, численные рассуждения

1 Introduction

Visual Question Answering (VQA) is the task of finding an answer given an image and a question in natural language. Text VQA is the subfield of VQA which involves reading text on images such as signboards, receipts, documents etc. Answering to the questions about text on images requires performing optical character recognition and fusion of text and image representations.

One of the first approaches to VQA (Kazemi and Elqursh, 2017) was based on processing of an image with CNN, a question with RNN, attention between question and image representations and classification of possible answers. A similar approach to Text VQA (Singh et al., 2019) includes recognition of text on images and obtaining scene text representations. Pretraining of Transformers (Vaswani et al., 2017) on images and recognized text (Yang et al., 2021), (Li et al., 2021b), (Biten et al., 2022) with the objective functions of masked language modeling, masked image modeling and word-patch alignment improves the quality of question answering.

In our paper we describe the system for extraction of structured information from document images and subsequent question answering. The system can be applied to understanding receipt or form images which has many applications in industry. For example, information, extracted from receipts, is useful to keep track of customers' expenses or to optimize the supply chain of companies. The system is capable of numerical reasoning over extracted key attributes during answer generation. The question answering system includes the following components: building a numerical reasoning expression for the question, extraction of structured information from the image and execution of the expression with extracted values as operands. This pipeline-based approach enables replacement of any component for more elaborate one and makes the process of answer generation interpretable. The model was trained and evaluated on Receipt-AVQA dataset which contains receipt images, text and questions and FUNSD (Jaume et al., 2019) dataset of form images. The proposed system scored MASE of 0.1164 on QA track and 0.2331 on VQA track of Receipt-AVQA competition and achieves competitive performance (F1=78.4) on FUNSD dataset.

2 Related Work

Text VQA. Question answering on images with textual content, such as signboards, recepts, invoices etc. (Singh et al., 2019), (Biten et al., 2019), (Mishra et al., 2019), has been an active area of research in last years. TextVQA (Singh et al., 2019) is one of the first datasets which contains questions related to text on images. The authors of the dataset proposed the LoRRa model, which is based on fusion of question, image and OCR text representations, and subsequent classification on the vocabulary words. The model (Mishra et al., 2019) performs text block extraction and defines which of the blocks contains the answer. Unlike the approaches of late fusion of image and text representations, obtained with CNNs (Lin et al., 2017), (Simonyan and Zisserman, 2014) and LSTM (Hochreiter and Schmidhuber, 1997), M4C (Hu et al., 2020) is a multimodal Transformer which takes as input the embeddings of question words, detected words and OCR tokens. The answer is generated in autoregressive way with dynamic pointer network. M4C outperforms previous approaches on TextVQA dataset.

Pretraining of language models on images and recognized text leads to further improvements in the task of Text VQA, because it gives better joint representations than a sole objective toward correct answer. In Text-Aware Pre-training (Yang et al., 2021) embeddings of text words, visual objects and scene text are fed into the multi-modal Transformer, pretrained with masked-language modeling (MLM), relative position prediction and image-text matching objectives. Layout Transformer (Biten et al., 2022) is pretrained on text with spatial cues (coordinates of the text region) on denoising task. In SelfDoc (Li et al., 2021b) the Transformer takes as input sentence embeddings of the text from the document and embeddings of object proposals and is pretrained with MLM objective. ERNIE-Layout (Peng et al., 2022) adopts a reading order prediction task in pre-training and spatial-aware disentangled attention mechanism. LayoutLMv3 (Huang et al., 2022) is pretrained with unified text and image masking and word-patch alignment to learn cross-modal alignment. LayoutLMv3 achieves SOTA performance on text-centric and image-centric VQA tasks.

Sequence Tagging. In our system relevant numerical values are extracted from receipt OCR text using

sequence tagging method, which involves matching categorical labels to sequence items. Its classical examples are Part-of-speech tagging and Named Entity Recognition. The common approach to sequence tagging involves encoding of text tokens with BiLSTM (Lample et al., 2016), CNN (Ma and Hovy, 2016) or pretrained language models (Devlin et al., 2018a) (Bao et al., 2020) and subsequent classification of hidden states or Conditional Random Field layer (Lafferty et al., 2001).

Generation of expressions for numerical reasoning Questions in Receipt-AVQA require numerical reasoning, which is commonly performed with encoder-decoder architecture. ELASTIC (Zhang and Moshfeghi, 2022) encodes a task text with RoBERTa (Liu et al., 2019) and separately generates operators and operands for the final mathematical expression. In the work of (Cobbe et al., 2021) GPT3 (Brown et al., 2020) models generate a chain of reasoning and verify it to validate reasoning correctness.

3 Task and data

Receipt-AVQA is a question answering task that requires answering a quantitative question related to a given receipt instance. The task comprises two tracks: Visual Question Answering and Question Answering. In the VQA track, the receipts' instances are provided as images, while in the QA track, participants are given all text tokens from receipts along with their coordinates.

There are three types of questions: *amount*, *count*, and *ratio*, which denote the expected answer type. Each receipt uses one of two currencies: *Malaysian ringgit* and *Indonesian rupiah*, which have different scales. Participants have access to question types and currencies, as well as lists of operations required to calculate the answer (e.g., subtraction, sorting).

The participants' solutions are evaluated using the metric, based on MASE score. Specifically, all questions are divided into six groups based on question type and currency, and MASE values are calculated for each group, the scores are then averaged. The task and evaluation method pose difficulties, as answers can lie in a wide range.

The dataset consists of 21,837 questions (16,611 in the training subset, 2,302 in the development subset, and 2,924 in the testing subset) and 1,957 receipts (1,537 in the training subset, 210 in the development subset, and 210 in the testing subset).

4 Method

The scheme of the proposed approach is depicted in Figure 1. Optical character recognition (4.1) is used to transform a photo of a receipt into textual information, thereby reducing the VQA task to a QA task. The Attribute Extractor (4.3) extracts numerical information and structures it. The Question Processor (4.2) accepts a question and generates a mathematical expression. Finally, the Answer Generator (4.4) produces an answer from the receipt contents (constructed by the Attribute Extractor) and the expression.

4.1 Image-to-Text Extraction

Text recognition in our model is performed in the following steps:

- Detection of regions with text on the image;
- Cropping the text regions and feeding them into the model, which generates text;
- Splitting text regions into lines and sorting by the line number from top to bottom and within the line from left to right (using the coordinates of detected text regions).

The text detection component utilizes the PP-OCRv3 (Li et al., 2022) architecture. PP-OCRv3 includes the Path Aggregation Network (PAN)(Liu et al., 2018) for the calculation of feature maps and the Feature Pyramid Network(Lin et al., 2017) for object detection (regions on the image with text in our case). We trained PP-OCRv3 on bounding boxes with the text from the train split of the Receipt-AVQA-2023 dataset. The model was trained in three epochs with a batch size of 8, learning rate of 0.001, and achieved precision=0.899, recall=0.905 on the dev split. An example of detected text regions can be seen in Figure 2. Text recognition in detected regions is based on the Transformer encoder-decoder TrOCR model (Li et al., 2021a).

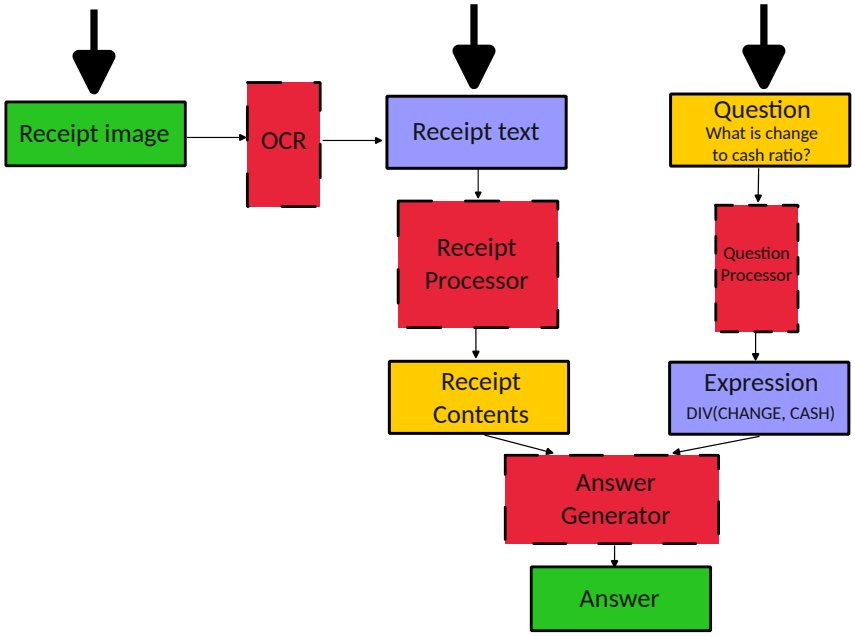


Figure 1: Model scheme

Figure 2: An example of text detection using PP-OCRv3.

4.2 Question Processor

Question Processor transforms questions in English into mathematical expressions. Expressions provide exhaustive information on how to generate an answer provided all variables. We analyzed two approaches to question processing: a rule-based approach and a generative model. Description of expression structure can be found in Appendix A.1.

4.2.1 Rule-Based Question Processing

We divided questions into 50 groups, each with its own expression. To figure out which group a question belongs to, each group of questions is matched against a regular expression, which represents the group. For instance, questions «What is the average price of a position?» and «What is the mean price of a position?» belong to the same group with expression `DIV(SUM(AMOUNTS),COUNT(AMOUNTS))` and regex `'What is the (average|mean) price of a position?'`. Then, all numbers from the question are extracted, and they will be used to substitute `NUM1` and `NUM2` later (if `NUM1` and `NUM2` are needed).

This approach is sufficient for the competition as participants have access to questions and new types of questions can be added manually.

4.2.2 T5 Question Processing

Since the rule-based approach does not generalize to new questions, we decided to develop an approach based on a generative model. We generated expressions for all questions from the train subset and fine-tuned T5-base (Raffel et al., 2020) to yield an expression given a question. The T5-base was trained on 1 epoch with a batch size of 32, AdamW optimizer (Loshchilov and Hutter, 2017), a learning rate of 1.5×10^{-4}, weight decay of 0.01, $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$.

4.3 Attribute Extractor

The purpose of Attribute Extractor is extracting and structuring necessary numerical data of a receipt. This structured data is referred to as *Receipt Contents* in the model scheme 1. For each receipt we keep its general values (e. g. total, tax), and for each good we keep its unit price, quantity and total price. The example of receipt contents is shown in Figure 3.


```
{
  "card": null,
  "cash": null,
  "change": null,
  "discount": 19000.0,
  "goods": [
    {
      "price": 59000.0,
      "quantity": 1.0,
      "total": 59000.0
    },
    {
      "price": 190000.0,
      "quantity": 1.0,
      "total": 190000.0
    },
    {
      "price": 10000.0,
      "quantity": 1.0,
      "total": 10000.0
    }
  ],
  "round": null,
  "service": 9600.0,
  "subtotal": null,
  "tax": 52416.0,
  "total": 302016.0
}
```


Figure 3: `train/receipt_00003` image and corresponding contents in json format

4.3.1 Line Breaking

The textual information of a receipt comprises a set of words along with the coordinates of the rectangles containing them. To facilitate further text processing, the set of rectangles is divided into subsets of lines. A greedy algorithm is used for line splitting, which prioritizes pairs of rectangles with large intersections. Since the receipt is split into lines, coordinates are no longer required.

4.3.2 Rule-Based Approach

Given a sequence of receipt lines, a rule-based attribute extractor produces structured information about the receipt. The algorithm is divided into two parts: parsing goods and parsing general information.

In the first part, the rule-based attribute extractor creates a list of goods by searching for the unit price, quantity, and total price for each position. In the second part, the extractor finds general values (such as change or service fee), more details are given in Appendix A.2.

However, this approach has several flaws. Firstly, the set of strategies is not exhaustive and the model cannot handle novel formats of goods. Secondly, it cannot handle receipts with a non-unified format of goods. Lastly, it cannot parse lines containing two or more general values.

4.3.3 BERT Approach

First, we describe the process of constructing the training dataset for our BERT approach. We used the rule-based method mentioned above to generate receipt contents for the training subset. We considered the rule-based approach to have produced the correct receipt contents for a receipt if, when using this content, the entire model produced the correct answers to all questions for that receipt. We then ruled out incorrect receipts, and this formed the training dataset for our BERT approach.

To generate receipt contents, we used two BERT-base models (Devlin et al., 2018b) referred to as $BERT_{labels}$ and $BERT_{goods}$. Both models were used for a tagging problem, where the sequence to be tagged is the concatenation of a receipt's lines. $BERT_{labels}$ predicted tokens containing general values (e.g., tokens `B-TOTAL` and `O-TOTAL`) or information about a particular good (tokens `B-POSITION` and `O-POSITION`). Similarly, $BERT_{goods}$ predicted tokens containing values related to goods (e.g., tokens `B-PRICE` and `O-PRICE`). We adjusted the rule-based approach to yield these tags, and using all information about the tags, we could unambiguously identify all general values and a list of goods, and form the receipt contents.

Both BERT models were trained for 30 epochs with a batch size of 20, using the AdamW optimizer (Loshchilov and Hutter, 2017), a learning rate of 2×10^{-5}, weight decay of 10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-6}$.

With tags obtained from $BERT_{labels}$ and $BERT_{goods}$, we identified the tokens containing numbers for the values of the receipts. To extract a number from a token, we removed any letters and other symbols unrelated to the number and replaced the decimal point with a comma if it was represented as a comma.

4.3.4 Pruning

The MASE metric highly penalizes large errors, even in a small part of the sample, unlike the accuracy metric. Therefore, we decided to prune large answers. Specifically, for each pair *(currency, expression)*, we calculated the median m, average a, and standard deviation s on the corresponding subset of the training dataset. If an answer exceeded $a + 3s$, we replaced it with m. We chose the median as the replacement value because it minimizes the MAE.

4.4 Answer Generator

The answer generator uses an expression based on the question and receipt contents to yield an answer to the task. First, any missing information on the receipt contents is filled in. For example, if there is no information about the unit price of a position in a receipt, it is calculated by dividing the total price by the quantity. After that, all variables in the expression are replaced with their respective values from the receipt contents. The resulting expression consists only of procedures and numbers, which are then evaluated. The final value obtained from evaluating this expression is the answer to the task.

5 Experiments and Analysis

5.1 Results on Receipt-AVQA dataset

The T5 model for question processing achieved an absolute quality score of 100% on both the development and test subsets, indicating that its performance was flawless and there were no errors or inaccuracies in its processing of the questions. To explore what questions the model can handle and to what extent it can generalize, we tested it against a pre-prepared list of questions. The results are presented in Table 4.

The model sometimes succeeds in generating correct expressions for reformulated questions and unknown expressions, but it is not reliable for very complex novel structures and wordings as it tends to imitate known expressions.

The rule-based approach for receipt processing generated all correct answers for 68% of the 1041 receipts in the train subset and for 70% of the 147 receipts in the development subset. These receipts were used as the training and validation datasets for the BERT approach. The results of both approaches are presented in Table 1.

Model	Total	Amount	Count	Ratio	Accuracy 10%
Rule-Based	0.2230	0.1338	0.3707	0.1645	84.99%
BERT	**0.1164**	**0.0844**	**0.1020**	**0.1627**	**91.45%**
OCR+Rule-Based	0.3073	**0.2952**	0.4106	0.2161	75.41%
OCR+BERT	**0.2331**	0.3427	**0.1573**	**0.1994**	**81.91%**

Table 1: Results on the test set of Receipt-AVQA (MASE metrics)

The BERT approach outperforms the rule-based one in almost all metrics. This showcases that BERT is able to generalize its knowledge about the structure and contents of receipts and overcome some of the disadvantages of the rule-based approach.

Additionally, we provided the time and memory performance of some components A.3.

5.2 Results on FUNSD dataset

The pipeline of our model can be applied to structured information extraction from any kind of document images (not only receipts). Transformer-based Attribute Extractor component was trained and tested on FUNSD dataset, which contains form images and corresponding annotations: recognized text, coordinates of regions with text and tags of entities ("header", "question", "answer", "other"). The coordinates of the boxes (text regions) were used to split the boxes list into lines. The special token "<ln>" was inserted at the beginning of each line, the special token "<box>" – at the beginning of each box.

BERT-base model was replaced by Longformer-base (Beltagy et al., 2020) to enable processing of long texts in forms. The model was trained for 30 epochs with a batch size of 20, using the AdamW optimizer (Loshchilov and Hutter, 2017), a learning rate of 2×10^{-5}, weight decay of 10^{-6}, $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\varepsilon = 10^{-6}$. Attribute Extractor achieves competitive performance (F1=78.4) on FUNSD dataset (Table 2).

Model	F1
UniLMv2-base (Bao et al., 2020)	68.9
UniLMv2-large (Bao et al., 2020)	72.6
Our model	78.4
LayoutLMv2-base (Xu et al., 2020)	82.8
LayoutLMv3-large (Huang et al., 2022)	92.1
ERNIE-Layout-large (Peng et al., 2022)	93.1

Table 2: Results on FUNSD dataset

6 Conclusions

In this paper, we present a question answering system on document images which is capable of numerical reasoning over extracted structured data. The proposed architecture of the question answering system utilizes the structure of independent blocks, each of which can be enhanced or replaced while keeping other components unchanged. The system includes the following components: OCR, Attribute Extractor, which finds values of key attributes in text, Question Processor, which defines a numerical reasoning expression, and Answer Generator. Text recognition is performed using the TrOCR model which generates text from bounding boxes detected by PP-OCRv3. The Attribute Extractor is based on BERT for token classification. In the Answer Generator component we applied a rule-based approach and a T5-based model.

The proposed model achieves competitive performance on FUNSD dataset. Also, the model was evaluated in the Receipt-AVQA competition, the version with the BERT receipt processor scored MASE

of 0.1164 on the QA track and MASE of 0.2331 on the VQA track. Additionally, while this is not reflected in the competition score, we found that T5 is capable of generalization to expression types unseen in the training set, making the whole scheme more resilient to new question types.

References

Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Xiaodong Liu, Yu Wang, Jianfeng Gao, Songhao Piao, Ming Zhou, et al. 2020. Unilmv2: Pseudo-masked language models for unified language model pre-training. // *International conference on machine learning*, P 642–652. PMLR.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Ali Furkan Biten, Ruben Tito, Andres Mafla, Lluis Gomez, Marçal Rusinol, Ernest Valveny, CV Jawahar, and Dimosthenis Karatzas. 2019. Scene text visual question answering. // *Proceedings of the IEEE/CVF international conference on computer vision*, P 4291–4301.

Ali Furkan Biten, Ron Litman, Yusheng Xie, Srikar Appalaraju, and R Manmatha. 2022. Latr: Layout-aware transformer for scene-text vqa. // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, P 16548–16558.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018a. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018b. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for textvqa. // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 9992–10002.

Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. 2022. Layoutlmv3: Pre-training for document ai with unified text and image masking. // *Proceedings of the 30th ACM International Conference on Multimedia*, P 4083–4091.

Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. Funsd: A dataset for form understanding in noisy scanned documents. // *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, volume 2, P 1–6. IEEE.

Vahid Kazemi and Ali Elqursh. 2017. Show, ask, attend, and answer: A strong baseline for visual question answering. *arXiv preprint arXiv:1704.03162*.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. 2021a. Trocr: Transformer-based optical character recognition with pre-trained models. *arXiv preprint arXiv:2109.10282*.

Peizhao Li, Jiuxiang Gu, Jason Kuen, Vlad I Morariu, Handong Zhao, Rajiv Jain, Varun Manjunatha, and Hongfu Liu. 2021b. Selfdoc: Self-supervised document representation learning. // *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, P 5652–5660.

Chenxia Li, Weiwei Liu, Ruoyu Guo, Xiaoting Yin, Kaitao Jiang, Yongkun Du, Yuning Du, Lingfeng Zhu, Baohua Lai, Xiaoguang Hu, et al. 2022. Pp-ocrv3: More attempts for the improvement of ultra lightweight ocr system. *arXiv preprint arXiv:2206.03001*.

Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. // *Proceedings of the IEEE conference on computer vision and pattern recognition*, P 2117–2125.

Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. 2018. Path aggregation network for instance segmentation. // *Proceedings of the IEEE conference on computer vision and pattern recognition*, P 8759–8768.

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. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Anand Mishra, Shashank Shekhar, Ajeet Kumar Singh, and Anirban Chakraborty. 2019. Ocr-vqa: Visual question answering by reading text in images. // *2019 international conference on document analysis and recognition (ICDAR)*, P 947–952. IEEE.

Qiming Peng, Yinxu Pan, Wenjin Wang, Bin Luo, Zhenyu Zhang, Zhengjie Huang, Teng Hu, Weichong Yin, Yongfeng Chen, Yin Zhang, et al. 2022. Ernie-layout: Layout knowledge enhanced pre-training for visually-rich document understanding. *arXiv preprint arXiv:2210.06155*.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, P 8317–8326.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, et al. 2020. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. *arXiv preprint arXiv:2012.14740*.

Zhengyuan Yang, Yijuan Lu, Jianfeng Wang, Xi Yin, Dinei Florencio, Lijuan Wang, Cha Zhang, Lei Zhang, and Jiebo Luo. 2021. Tap: Text-aware pre-training for text-vqa and text-caption. // *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, P 8751–8761.

Jiaxin Zhang and Yashar Moshfeghi. 2022. Elastic: Numerical reasoning with adaptive symbolic compiler.

A Appendix

A.1 Question Processor metadata

An expression consists of variables and procedures. A variable can denote a number (e.g., `TOTAL` and `CHANGE` account for the total amount of purchase and change, respectively) or a list of numbers (e.g., `PRICES` designates a list of unit prices of goods in the same order as they appear in the receipt). There are two special variables `NUM1` and `NUM2` for the first and the second number in a question. A procedure designates an operation that should be performed on its arguments. For example, the expression `SUM(FIRST_POSITIONS(NUM1,PRICES))` means «get the first `NUM1` elements of the `PRICES` list and calculate their sum». The lists of variables and procedures are presented in Table 3.

Variable	Explanation
TOTAL	Total amount of purchase
SUBTOT	Total excluding taxes. Used if it is explicit in the receipt. May not coincide with the subtotal inscription in the receipt
CASH	Cash used for payment
CARD	Payment by card
TAX	Tax amount
CHANGE	Change amount
DISCOUNT	Amount discounted
ROUND	Rounding value
SERVICE	Service fee
PRICES	List of unit prices (the same order as in a receipt)
QUANTITIES	List of quantities (the same order as in a receipt)
AMOUNTS	List of total prices of positions (the same order as in a receipt)
NUM1	First number in the question
NUM2	Second number in the question

Procedure	Explanation
ADD	$a + b$
SUB	$a - b$
MUL	ab
DIV	$\frac{a}{b}$
INTDIV	$\left[\frac{a}{b}\right]$
COUNT	List length
FIRST_POSITIONS	First n elements of a list
IS_ZERO	$I\{L_i = 0\}$
MIN, MAX, SUM	Minimum, maximum element or sum of elements of a list
AMIN, AMAX	Position of a minimum or maximum element
LARGER_THAN, SMALLER_THAN	Only values larger (smaller) than threshold remain in a list
LARGER_EQ_THAN, SMALLER_EQ_THAN	Only values larger (smaller) or equal to threshold remain in a list

Table 3: Lists of variables and procedures

A.2 Rule-based approach to numerical data extraction from receipts

At the first stage, the attribute extractor searches the unit price, quantity, and total price for each position. As different receipts can have different positions, a set of strategies has been developed for finding the positions of the quantities. For example, one of the strategies detects a good in a line if the line has the format `[QUANTITY] [PRICE] [NAME OF PRODUCT] [TOTAL] [OPTIONAL TAIL]`, where each expression in brackets represents a substring in the line. Some strategies handle goods that occupy two or three lines. The processor iterates over the strategies from more to less specific, and attempts to find goods using the strategy. If it finds one or more goods, it terminates the iteration and yields parsed goods.

At the second stage, the processor finds general values (such as change or service fee). For each such value, it iterates over lines and determines whether the lines contain the value. If so, it returns the parsed quantity. If a line is already occupied by another value or good, it is not considered.

A.3 Model Performance

We measured average time for processing one instance of dataset and GPU memory consumption for TrOCR, Receipt Processor based on BERT and Question generator based on T5. The results are present in the Table 5.

Question	Generated expression	In train	Correct	Comment
How much is tax to total amount ratio?	`DIV(TAX,TOTAL)`	Yes	Yes	-
Divide tax to total amount, please?	`DIV(TAX,TOTAL)`	No	Yes	The same task with different wording
How much is tax divided by total?	`TAX`	No	No	Failed to handle different wording
What is the change to cash ratio?	`DIV(CHANGE,CASH)`	Yes	Yes	-
What is the cash to change ratio?	`DIV(CASH,CHANGE)`	No	Yes	Successfully dealt with argument order
How many goods in the 2nd position can be bought for 300?	`INTDIV(NUM2,FROM_TOP(NUM1,PRICES))`	Yes	Yes	-
With 300, How many goods in the 2nd position can be bought?	`INTDIV(NUM2,FROM_TOP(NUM1,PRICES))`	No	No	Should have swapped `NUM1` and `NUM2`
What is the ratio between price of the 1st position and price of the 2nd position?	`DIV(FROM_TOP(NUM1,PRICES),NUM2)`	No	No	Tried to mimic to the format of the known expressions
What is the sum of prices of the first 4 positions?	`SUM(FIRST_POSITIONS(NUM1,AMOUNTS))`	Yes	Yes	-
What is the sum of prices of the last 4 positions?	`SUM(LAST_POSITIONS(NUM1,AMOUNTS))`	No	Yes	Invented the unknown `LAST_POSITIONS` operator
How many positions are with price lower than 300?	`COUNT(SMALLER_THAN(NUM1,PRICES))`	Yes	Yes	-
What is the average price of positions with prices lower than 300?	`DIV(SUM(SMALLER_THAN(NUM1,PRICES)), COUNT(SMALLER_THAN(NUM1,PRICES))`	No	Yes	Managed to combine known operators for a novel query

Table 4: Examples for expression generation with T5. *In train* column indicates whether the question (possibly, with other numbers) is present in the train set.

Component	GPU	Instances per second	GPU memory
TrOCR	Tesla P100	0.25	8.5 Gb
BERT Receipt Processor	Tesla P100	248	4.8 Gb
T5 Question Generator	Tesla P100	6.3	1.9 Gb

Table 5: Model components performance

Scalar structure for *polu-* ‘half’

Tatevosov S. G.
Lomonosov Moscow State University
Interdisciplinary School «Preservation of the World Cultural and Historical Heritage», 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 paper explores restrictions on the distribution of *polu-* ‘half’ in combination with adjectival stems in Russian. Relying on the literature on degree semantics, we analyze *polu-* as a degree modifier that specifies the degree to which the adjective maps an individual as ½ of the maximal degree. This correctly predicts that *polu-* can only combine with upper closed scales. We argue that unlike *half* in English, *polu-* does not require a scale be lower closed.

Keywords: scalarity, gradable adjectives, degree modification, accommodation
DOI: 10.28995/2075-7182-2023-22-497-506

Полу- и скалярная структура

Татевосов С. Г.
МГУ имени М. В. Ломоносова / НОШ «Сохранение мирового культурно-исторического наследия»
Москва, Россия
tatevosov@gmail.com

Киселева К. Л.
ИРЯ им. В. В. Виноградова РАН,
Москва, Россия
xkisseleva@gmail.com

Аннотация

В статье обсуждаются ограничения на дистрибуцию элемента *полу-* в составе сложных прилагательных. Опираясь на наработки авторов, работающих в парадигме степенной семантики, мы предлагаем анализ *полу-* как скалярного модификатора, который помещает степень обладания параметрическим свойством в окрестность ½ соответствующей максимальной степени. Это предсказывает основное семантическое требование, которое *полу-* предъявляет прилагательному: последнее должно быть привязано к закрытой сверху шкале.

Ключевые слова: скалярность, параметрические прилагательные, степенная модификация, аккомодация

1. Основной контраст

Цель этих заметок — изложить несколько наблюдений и эскиз семантического анализа адъективов, которые содержат элемент *полу-*. Следуя русской грамматической традиции, мы рассматриваем такие адъективы как продукт словосложения, однако никакие дальнейшие рассуждения не опираются на допущения о морфологическом статусе *полу-*. (Следует также отметить, что в литературе *полу-* рассматривается преимущественно именно как морфологическая проблема («аффиксоид», etc.); в качестве примечательных образцов семантического обсуждения см. [1], [2], [3].)

Исходный контраст в дистрибуции *полу-*, который мы хотим объяснить, иллюстрируется в (1):

(1) Ограничение на дистрибуцию *полу-* с прилагательными:
   а. *полупустой, полуоткрытый, полуголый*
   б. ??*полубыстрый,* ??*полудлинный,* ??*полукрасивый*

Как показывает (1), прилагательные неоднородны с точки зрения возможности присоединения *полу-*. (1а) — однозначно допустимые единицы. Единицы в (1б) по крайней мере вне специального контекста воспринимаются как аномальные.

Суждения, которые отражены в (1), — часть языковой компетенции носителей русского языка, которая тем самым должна быть представлена в языковых моделях, использующихся при его автоматической обработке. Соответственно, мы надеемся, что сказанное ниже будет небесполезно и для исследователей, работающих в области теоретической семантики, и для специалистов, занимающихся созданием систем обработки естественного языка.

Наши наблюдения опираются на материал, доступный в Национальном корпусе русского языка (НКРЯ). Мы хотели бы подчеркнуть, что не беремся осветить в этой статье все употребления *полу-* (около 40 тыс. токенов) и не планируем количественный анализ представленных в НКРЯ данных.

Дальнейшее изложение организовано следующим образом. В разделе 2 мы сформулируем основную гипотезу, объясняющую контраст в (1). Раздел 3 посвящен важным нюансам этой гипотезы и делает более эксплицитным предлагаемое нами теоретическое решение. В разделе 4 мы обсудим, опираясь на корпусной материал, как гипотеза расширяется на случаи, менее тривиальные, чем те, которые иллюстрируются в (1).

2. *Полу-* и скалярная семантика

Основная гипотеза, которая объясняет различия в приемлемости в (1а-б), сформулирована в (2).

(2) Гипотеза о скалярной структуре
а. Дистрибуция *полу-* ограничена элементами, содержащими скалярную структуру в своем семантическом представлении или способными к контекстной аккомодации такой структуры.
б. Скалярная структура должна опираться на закрытую сверху шкалу.

Поясним используемые в (2) понятия.

Скалярная структура — это тройка элементов вида (3):

(3) $<S_O, R, \Delta>$,
где S — шкала, O — один из четырех структурных типов в (7), R — отношение упорядочивания, а Δ — скалярный параметр.

Прототипические языковые выражения, содержащие скалярные структуры в своей семантике, — **параметрические прилагательные**. Например, прилагательное *широкий* в своем основном значении привязано к скалярной структуре в (4):

(4) *Широкий*: $<S_{]0,1[}, \geq,$ ПРОТЯЖЕННОСТЬ В ГОРИЗОНТАЛЬНОМ ИЗМЕРЕНИИ >

Шкала — линейно упорядоченное множество точечных **степеней**, абстрактных объектов, используемых для выражения количественности. *Широкий* отображает любого индивида в интервал на шкале с соответствующим скалярным параметром — ПРОТЯЖЕННОСТЬ В ГОРИЗОНТАЛЬНОМ ИЗМЕРЕНИИ.

Отношение «≥» в (4) означает, что ширина индивида тем больше, чем больше его протяженность в горизонтальном измерении. Прилагательное *узкий* привязано к той же шкале, но предполагает противоположное отношение «≤»: узость индивида тем больше, чем меньше его протяженность в горизонтальном измерении.

Соответственно, основной компонент семантики прилагательных — это функция, которая сопоставляет индивиду его проекцию на шкале, например, ШРК в (5)-(6).

Прилагательные в сравнительной степени соотносят проекции двух индивидов, как показано в (5):

(5) *Дверь шире окна.*
ШРК(**дверь**) > ШРК(**окно**)

Прилагательные в положительной степени сопоставляют проекцию индивида со стандартом сравнения:

(6) *Дверь широкая.*
ШРК(**дверь**) ≥ STND(ШРК)

Таков в общих чертах скалярный подход к параметрическим прилагательным, практикуемый представителями степенной семантики ([4], [5], [6], [7], [8], [9], [10], [11]).

Важнейшая характеристика скалярной структуры — тип шкалы, который в (3), вслед за [6], обозначен как О.[1] Тип шкалы определяется тем, есть ли у нее минимальное и максимальное значения. Таких типа четыре:

(7) Типология шкал для параметрических прилагательных
а. Открытая шкала (изоморфна интервалу действительных чисел]0,1[): нет ни минимальной, ни максимальной степени
б. Закрытая сверху шкала (изоморфна интервалу]0,1]): есть максимальная степень, нет минимальной степени
в. Закрытая снизу шкала (изоморфна интервалу [0,1[): есть минимальная степень, нет максимальной степени
г. Закрытая с двух сторон шкала (изоморфна интервалу [0,1]): есть минимальная и максимальная степени

Определить, закрыта ли сверху шкала, привязанная к конкретному прилагательному, помогают модификаторы типа *совершенно*, *абсолютно*, *полностью*, *стопроцентно* и т.п., функция которых — указание на максимальную положительную степень обладания параметрическим свойством. Предложения (8а-в) показывают, что прилагательное *безопасный* работает со степенями на закрытой сверху шкале, а (9а-б) — что шкалы для прилагательных *длинный* и *изогнутый* открыты сверху.[2]

(8) а. *Этот путь абсолютно безопасен.*
б. *Этот путь совершенно безопасен.*
в. *Этот путь полностью безопасен.*

(9) а. ??/#*Эта дорога абсолютно длинная.*
б. ?/#*Эта линия стопроцентно изогнута*.

Увидеть, закрыта ли шкала снизу, несколько сложнее. Ни в русском языке, ни, насколько нам известно, в других языках нет модификаторов наподобие несуществующего **нульпроцентно*, которые указывали бы на минимальную степень.[3] Поэтому, чтобы диагностировать наличие

---

[1] В [6] можно также найти обсуждение вопроса о том, как скалярные характеристики прилагательных соотносятся с понятиями, известными в русской традиции как оппозиция качественных и относительных прилагательных.

[2] Значительно меньше подходит для определения типа шкалы наречие *совсем*, на первый взгляд стоящее в одном ряду с *абсолютно*, *совершенно* и *полностью*. По-видимому, это наречие допускает, кроме 'в максимальной степени', также интерпретацию 'в очень высокой степени'. В этом качестве оно намного охотнее сочетается с прилагательным открытой шкалы типа *длинный*, ср. (i), где очевидно не предполагается, что 5 км — это дистанция максимально возможной/доступной длины:

(i) [Спортсменка о беге на разные дистанции:] Папа хочет еще на «тройку» меня вытащить. Да и на «пятерку»… Но это уж **совсем длинная** дистанция. [https://matchtv.ru]

Это уточнение обязано своим появлением в тексте комментарию анонимного рецензента «Диалога».

[3] Анонимный рецензент отмечает, что для диагностики минимальной степени, возможно, следует использовать «наречия (в том числе, отрицательные) с отрицательной формой того же прилагательного: *Отнюдь / никоим образом не изогнутая*».

Мы полностью согласны с тем, что эта возможность требует серьезного рассмотрения. Обстоятельства *никоим образом* или *ни в малейшей степени*, если они интерпретируются под сентенциальным отрицанием, действительно должны давать желаемый результат: 'такая, что неверно, что она изогнута хотя бы в минимальной степени'. Необходимо,

минимальной точки на шкале, следует воспользоваться прилагательными противоположной полярности. Минимальная степень для прилагательного типа *изогнутый*, если она есть, одновременно выступает максимальной степенью для прилагательного *прямой*. Определить, есть ли максимальная степень у прилагательного *прямой*, можно, как мы уже видели, при помощи модификаторов типа *абсолютно* или *совершенно*:

(10) *Эта линия абсолютно / совершенно / стопроцентно прямая.*

(10) показывает, что шкала содержит максимальную степень прямости, которая одновременно выступает минимальной степенью изогнутости. Тем самым шкала изогнутости **закрыта снизу**.

Аналогичный прием дает для прилагательных *безопасный* и *длинный* отрицательный результат: они открыты снизу. Поскольку нельзя быть максимально опасным или коротким, нельзя быть и минимально безопасным или длинным:

(11) а. ??*Этот путь абсолютно / совершенно опасен.*
б. ??*Эта дорога абсолютно / совершенно короткая.*

Примечательно, что тип шкалы, привязанный к прилагательному, не полностью предсказывается его семантикой. Прилагательные *низкий / высокий* или *короткий / длинный* описывают пространственную протяженность объекта. В физическом смысле протяженность может быть нулевой, но несмотря на это, шкала, привязанная к таким прилагательным, открыта снизу, ср. ??*абсолютно / совершенно / стопроцентно короткий / низкий*. Если выражение типа *стопро-центно низкий* и можно как-то интерпретировать, оно не значит ‘имеющий нулевую протяженность в вертикальном измерении’. Аналогично прилагательное *кривой*, будучи (квази)синонимичным прилагательному *изогнутый*, в отличие от него привязано к закрытой шкале, ср. корпусной пример *Гипс самостоятельно отвалился, и мы увидели* ***абсолютно кривой*** *палец* [19rus.info].[4]

Важный вопрос, связанный с типами в (7), который поднимает анонимный рецензент «Диалога», — вопрос о количественном соотношении реализующих их прилагательных. Насколько нам известно, таких количественных данных на данный момент не собрано ни для одного из языков, обсуждаемых в связи с типологией в (7).

Имея эти выкладки, мы можем сформулировать семантику для элемента *полу-* следующим образом:

(12) *Полу-* соединяется с параметрическим прилагательным G и создает предикат над индивидами x (= множество индивидов) такой, что
1. имеется степень *d*, лежащая в контекстно-зависимой окрестности ε_C точки, которая представляет собой половину от максимальной степени на шкале S_G для прилагательного *G*;
2. любой индивид *x* обладает параметрическим свойством, описываемым прилагательным *G*, в степени *d*.
Более формально: $\| \text{полу} \|^C = \lambda G.\lambda x.\exists d[\, d \in \varepsilon_C(1/2\ \mathbf{max}(S_G)) \wedge G(d)(x)]$

В соответствии с (12), *полупустой x* — это такой *x*, степень пустоты которого составляет примерно половину от максимальной, причем характер «примерности» определяется контекстом.

Семантика в (12) объясняет контраст в (1) непосредственным образом. Она предписывает вычислить половину от максимальной степени на шкале, привязанной к параметрическому свойству. Чтобы это было возможно, шкала должна содержать максимальное значение, то есть быть закрытой сверху.

однако, убедиться, что отрицание в такой конфигурации действительно является сентенциальным и имеет более широкую сферу действия, чем обстоятельства. Проработку этой возможности мы оставляем на будущее. Что касается *отнюдь*, то предварительные наблюдения показывают, что это наречие, вероятно, имеет дистрибуцию, не ограниченную типом шкалы, ср. *отнюдь не длинный, отнюдь не умный* и т.п.

[4] Мы признательны анонимному рецензенту «Диалога», указавшему на важность сопоставить *изогнутый* и *кривой* в контексте текущего обсуждения.

Контраст между прилагательными в (1а-б) — это в точности контраст по закрытости шкалы сверху, как иллюстрируют примеры в (13):

(13) а. *абсолютно пустой, полностью открытый, совершенно голый*
б. ??*абсолютно быстрый*, ??*полностью длинный*, ??*совершенно красивый*

Словосочетания в (13а) приемлемы в нулевом контексте. В (13б), напротив, представлены случаи, которые вне контекстов, предполагающих семантический сдвиг для прилагательного или наречия, воспринимаются как аномальные.

Подчеркнем, что семантика в (12) задает необходимое, но не достаточное условие для соединения *полу-* и прилагательного. Чтобы сложное слово было семантически корректным, шкала для прилагательного должна быть закрыта сверху. Мы предсказываем — как кажется, верно, — что сложные слова с *полу-* образуются от таких прилагательных с достаточной регулярностью и продуктивностью. Мы, однако, не предсказываем, что любое закрытое сверху прилагательное должно образовывать композиты с *полу-*, ср. *пустой* и *полный*: оба закрыты сверху, но соединяется с *полу-* без затруднений только первое. Чтобы определить, с чем связана невозможность *полуполный* (или по крайней мере значительное снижение приемлемости по сравнению с *полупустой*) требуется дополнительное исследование, которое должно, во-первых, выявить круг лексем, показывающих такое ограничение, а во-вторых, определить, что объединяет их в естественный класс. Мы не пытаемся предпринять такое исследование в пределах этой статьи.

Важный аргумент в пользу анализа, увязывающего приемлемость *полу-* с характером шкалы, дают многозначные прилагательные, которые в одном из значений привязаны к закрытой сверху шкале, а в другом — к открытой. Пример такого прилагательного — *темный*, имеющий по меньшей мере два значения: 'лишенный света, погруженный во тьму' и 'по цвету близкий к черному, не светлый' (Ожегов, Ушаков). Можно заметить, что максимальное значение у шкалы есть в первом значении, но не во втором:

(14) а. *Комната была абсолютно темная.*
б. ??*Пальто было абсолютно темное.*[5]

Рассмотрим сочетаемость прилагательного *полутемный*. В (15) выписаны существительные, с которым *полутемный* образует словосочетания в порядке убывания их частотности в НКРЯ:

(15) Сочетаемость *полутемный* (первые 50 существительных в выдаче):
комната, коридор, зал, угол, помещение, передняя, сени, лестница, комнатка, подвал, столовая, прихожая, кабинет, кухня, гостиная, каморка, камера, спальня, улица, зала, вестибюль, вагон, церковь, горница, бар, изба, сенцы, квартира, коридорчик, проход, барак, каюта, осветить, переход, подъезд, храм, лавка, комнатушка, палата, переулок, фойе, номер, двор, салон, уголок, холл, арка, будуар, дом, закуток

Сочетаемость исходного прилагательного *темный* показана в (16). Пересечения со списком в (15) выделены в (16) курсивом.

[5] Анонимный рецензент «Диалога» отмечает: «методологическая проблема состоит … в оценке соотношения между лексической семантикой и прагматикой, в частности, возможности употребления выражений с *полу-* и *абсолютно* в контекстах с семантическим сдвигом… Как кажется, фраза (14б) приемлема в контексте изменения значения признака: *После стирки пальто было абсолютно темное (= потемнело)*». Хотя мы не вполне разделяем это семантическое суждение, мы полностью согласны, что *абсолютно* подвержено семантическим сдвигам. Например, в подслушанном одним из авторов разговоре в коридоре прозвучало предложение *Эта юбка мне абсолютно длинна*. Очевидно, в этом случае не имеется в виду, что юбка достигает абсолютного максимума длины (которого попросту нет). Речь идет о максимуме длины, приемлемой для ношения конкретным индивидом. Аналогичную реинтерпретацию можно предполагать и в обсуждаемом рецензентом предложении — в той степени, в которой такая реинтерпретация оказывается контекстно доступной.

(16) Сочетаемость *темный* (первые 50 существительных в выдаче):
ночь, глаз, *угол*, пятно, сила, *комната*, лес, *коридор*, волос, очки, небо, вода, фигура, сторона, человек, *улица*, масса, цвет, туча, окно, дело, полоса, царство, лицо, силуэт, место, глубина, фон, личность, зелень, *лестница*, время, платье, аллея, тень, стена, *переулок*, бровь, облако, материя, круг, стекло, *сени*, дерево, *уголок*, костюм, вечер, слух, точка, *двор*

Благодаря списку в (15) хорошо заметно, что в комбинацию с *полу-* вступают исключительно названия пространственных объектов — от комнаты до закутка. Именно в таких комбинациях представлено первое значение *темный* — когда отсутствие света может достигать максимальной степени, ср. *абсолютно темная/ый комната / коридор / зал* и т.д. Элементы списка в (16), не пересекающиеся с (15) — это, напротив, случаи, когда *темный* представлено в других значениях, в первую очередь как цветообозначение (*темный/ая/ые глаза / волосы / туча / силуэт / фон / платье / стена* и т.д.) или дескрипция качеств, имеющих негативные коннотации (*темный/ая/ые сила, человек, дело, царство, личность*) и т.д. Несколько выпадает из ряда самое частотное существительное *ночь* с сомнительным *?полутемная ночь*, при том что в *темная ночь* прилагательное описывает, как и в (15), отсутствие света. Можно, впрочем, заметить, что в отличие от (15) *ночь* не является названием объекта, имеющего четкую пространственную локализацию, с чем по-видимому связана затрудненность закрытой шкалы в этом случае, ср. *?абсолютно / совершенно темная ночь*.

Таким образом, предположение, которое отражено в (12), — *полу-* нуждается в прилагательных, закрытых сверху, то есть привязанных к шкале с максимальным значением, представляется эмпирически оправданным — по крайней мере на том материале, который мы только что рассмотрели.

В следующем разделе мы приведем одну важную альтернативу, обсуждаемую в литературе, и выскажем соображение, почему она не подходит для анализа *полу-*.

3. *Полу-* и альтернативный анализ

В [6] обсуждается семантика английского *half* в предложениях типа (17):

(17) а. *The glass is half / mostly full.*
б. *Her eyes were half / most of the way closed.*
в. *These images are half / mostly invisible.*

Эмпирическое обобщение К. Кеннеди и Б. Левин состоит в том, что *half* предъявляет прилагательному, с которым соединяется, более строгие условия, чем предполагает анализ в (12). А именно: *half* нуждается не просто в шкале, закрытой сверху, а в шкале, закрытой с двух сторон. Действительно, все прилагательные в (17) удовлетворяют этому свойству, ср. *fully visible / invisible* и т.п.

Соответственно, семантика для английского *half* задается с опорой не только на максимальную, но и на минимальную степень на релевантной шкале.

(18) *Half* соединяется с параметрическим прилагательным *G* и создает предикат над индивидами *x* (= множество индивидов *x*) такой, что
1. Имеется степень *d* такая, что разность между *d* и минимальной степенью на шкале S_G совпадает с разностью между максимальной степенью на этой шкале и *d*.
2. = (12.2)

Соответственно, *half visible* и *half invisible* указывают на одну и ту же степень, которая лежит посередине между максимальной и минимальной (то есть нулевой) видимостью/наблюдаемостью.

Выбор между вариантами анализа в (12) и в (18) для русского *полу-* кажется несколько эмпирически менее очевидным, чем для английского. Решающее значение для этого выбора имеет поведение прилагательных, закрытых сверху, но открытых снизу. Анализ в (12), где достаточно максимального значения, предсказывает, что *полу-* способен соединяться с такими прилагательными в осмысленное целое. Анализ в (18), когда требуется также и минимальное значение, напротив, предсказывает невозможность *полу-* в таком сочетании.

Некоторые факты как будто указывают на то, что русское *полу-* подчиняется таким же — более строгим — ограничениям, как английское *half*. Например, прилагательное *прямой* связано со шкалой, закрытой сверху (ср. *совершенно/абсолютно прямая линия*) и открытой снизу ([?/??/#]*совершенно/абсолютно кривая линия*). Для тех носителей, кто считает выражения типа *абсолютно кривой* приемлемыми, они выступают описаниями высокой, но не максимальной степени — именно ввиду отсутствия у кривизны максимума. При этом прилагательное в (19), по-видимому, так же аномально, как и в (1б) (если исключить из рассмотрения математический термин *полупрямая*).

(19) [??]*полупрямая линия*

С другой стороны, прилагательные типа *голый*, также открытые снизу (ср. [??/#]*совершенно/абсолютно одетый*), допускают *полу-* без ограничения.

Решающее эмпирическое соображение, склоняющее нас к анализу в духе (12), а не (18), связано с дистрибуцией *полу-* в контексте прилагательных с префиксом *без-*. Такие прилагательные всегда закрыты сверху, поскольку описывают отсутствие каких-либо проявлений параметрического свойства. Их закрытость снизу определяется закрытостью исходного прилагательного. В (20)-(22) показаны несколько корпусных примеров:

(20) *Я велел Файке идти в другой магазин через дорогу, а сам пока покупал кисель — килограмм за 1.23 и какое-то* ***полубесплатное*** *яблочное повидло за 63 коп. банка.* [Н. Н. Козаков. Дневник (1962)]

(21) *Как отдыхающий в секс-туре — в Таиланде или в Праге… много хуже того — как пресыщенный,* ***полубессильный*** *старичок с насекомыми инстинктами вместо мозгов и вшивным моторчиком вместо сердца.* [С. А. Самсонов. Аномалия Камлаева (2006-2007)]

(22) *Таков этот* ***полубезвестный****, но могущественный временщик, выходец из дер [евни] Сопляки.* [А. Т. Твардовский. Рабочие тетради (1963) // «Знамя», 2000]

Во всех этих случаях имеется максимум бесплатности, бессильности и безвестности, что позволяет говорить о закрытости этого конца шкалы. Противоположный конец, по-видимому, открыт — ввиду отсутствия минимума бесплатности, бессильности и безвестности (он же максимум платности, сильности и известности). Как сама возможность образования композитов с *полу-*, так и отсутствие в (20)-(22) каких-либо нетривиальных семантических эффектов склоняет нас к выбору варианта анализа в (12), изложенного в разделе 2.

Сказанное выше предполагает, что семантика единиц типа *half-* / *полу-* подвержена межъязыковому варьированию. В этой связи анонимный рецензент «Диалога» отмечает, что «соответствие подобной шкале — чисто семантический, а потому универсальный параметр, и в этом случае логично было бы предположить отсутствие языковых варьирований в данной области». Нам представляется, что в этом вопросе следует различать грамматические значения как таковые и семантическое наполнение конкретных грамматических показателей в конкретных языках. Первые в типологической литературе часто предполагаются универсальными, тогда как вторые могут быть подвержены варьированию. В нашем случае универсальной следует признать типологию в (11), допустив возможность, что грамматические элементы типа *half-* и *полу-* могут различаться тем, где в этой типологии должно находиться допустимое для них зависимое прилагательное.

4. *Полу-* и аккомодация шкалы

Рассмотренный выше небольшой материал и набросок анализа предполагают единственное и притом достаточно простое ограничение на дистрибуцию *полу-*. Полная картина, однако, намного сложнее и интереснее. В НКРЯ (корпус со снятой омонимией) для композитов с *полу-* представлено немногим менее 40 тыс. токенов, группируемых в несколько сотен лемм. Полный анализ этого массива данных еще только предстоит. В этом разделе мы отметим два наиболее интересных случая.

Это, во-первых, случай когда *полу-* соединяется с **непараметрическим прилагательным**. Во-вторых, это композиты, где *полу-* представлен в комбинации с единицами, которые по крайней мере внешне выглядят как **прилагательные с открытой шкалой.**

Первая возможность иллюстрируется примерами в (23)-(24):

(23) *На нем было дешевое, враспашку,* ***полубумажное*** *пальтишко, суконная рубаха с массой мелких пуговок.* [Л. М. Леонов. Русский лес (1950-1953)]

(24) *Литва, Латвия и Эстония объявили согласие «мирно переговариваться» с большевиками. Хотят, однако, не нормального мира, а какого-то* ***полубрестского****, с «нейтральными зонами».* [З. Н. Гиппиус. Дневники (1914-1928)]

Очевидно, что ни прилагательное *бумажный*, обозначающее материал, ни прилагательное *брестский*, производное от топонима, не являются параметрическим. Об этом свидетельствует невозможность сравнительной степени *более бумажный* или *более брестский* в нулевом контексте.

Означает ли это, что примеры вида (23) или (24) нарушают обобщение, сформулированное в (2a) как первая часть нашей исходной гипотезы? Мы даем отрицательный ответ на этот вопрос.

И (23) и (24) в действительности имеют скалярную интерпретацию. Наиболее естественное понимание (23) — ‘пальто, наполовину (хлопчато)бумажное’. (24) сообщает о том, что вводимый в рассмотрение мирный договор обладает примерно половиной характеристик мирного договора, подписанного в 1918 году в Бресте.

Таким образом, в обоих случаях мы имеем дело с **аккомодацией шкалы**, то есть приписыванием прилагательному скалярной структуры, исходно отсутствующей в его семантическом представлении. Можно предположить, что механизм аккомодации имеет коэрсионную природу: он запускается ввиду необходимости приписать комбинации непараметрического прилагательного и *полу-* когерентную семантическую интерпретацию.

(23)-(24) иллюстрируют два распространенных типа шкал, возникающих в результате такой аккомодации, которые можно условно назвать **количественными** и **интенсиональными**. Естественно, для аккомодации подходят только закрытые сверху шкалы, поскольку только они удовлетворяют семантическим потребностям *полу-*.

Количественный тип в (23) — это шкалы, упорядочивающие пропорции описываемых прилагательным сущностей в составе других сущностей. В нашем случае речь идет о пропорции хлопчатобумажной ткани в изделиях швейной промышленности. Точно такие же шкалы аккомодируются всегда, когда названия материала комбинируются со степенными выражениями, характерными для параметрических прилагательных. Для предложений типа *Стол более деревянный, чем шкаф* единственная возможная интерпретация — ‘стол содержит большую пропорцию дерева, чем шкаф’. Естественно, шкалы такого типа имеют максимальное значение, описываемое, например, как *Стол полностью деревянный.*

Интенсиональный тип представлен в (24). В этом случае шкала упорядочивает свойства, которыми обладают сущности, обозначаемые исходным прилагательным типа *брестский*. Один конец шкалы означает обладание полным набором таких свойств, другой — нулевым набором. Между ними расположены промежуточные случаи, один из которых и описывается прилагательным *полубрестский.*[6]

Мы предполагаем, что примерно то же самое происходит при образовании многочисленных композитов с *полу-*, присоединяемых к прилагательным с сортовой интерпретацией типа *полусухое шампанское* или *полукопченая колбаса* (с той оговоркой, что денотатом прилагательных в этом случае выступают не обычные индивиды, а типы, kinds). Сортовые прилагательные не являются параметрическими и так же аккомодируют шкалу, упорядочивающую пропорции свойств того или другого сорта.

[6] Мы рассматриваем как возможный, но не обсуждаем в этой статье и другой вариант анализа употреблений типа (24). При таком анализе шкала имеет металингвистический характер: на ней упорядочиваются степени соответствия описываемых сущностей той дескрипции, которую обозначает исходное прилагательное. *Полубрестский мир* — это мир, который подходит под описание *брестский мир* примерно наполовину.

Мы констатируем, таким образом, что аккомодация шкалы — это механизм, который позволяет непараметрическим прилагательным выполнить семантическое обязательство, которое возлагает на них обобщение в (2).

Второй интересующий нас случай — прилагательные с открытой шкалой, которые тем не менее способны соединяться с *полу-* вопреки (2б). По всей видимости, они допускают те же виды аккомодации, что и непараметрические прилагательные. *Полугрустный* и *полунизкий* (ср. ??*совершенно / абсолютно низкий / грустный*) иллюстрируются в (25)-(26):

(25) *Лет пять-шесть спустя, Ремизов говорил со мной о смерти. Разговор был* ***полугрустный****, полушутливый.* [Ю. П. Анненков. Дневник моих встреч (1966)]

(26) *Темный вход; довольно большая* ***полунизкая*** *передняя, из которой несколько маленьких дверей в крошечные приемные.* [А. И. Спиридович. Записки жандарма (1925)]

(25) — явный пример аккомодации количественной шкалы. (25) сообщает, что степень представленности того, что описывается как *грустный*, во вводимом в рассмотрение разговоре составляет половину максимальной. Вторая половина отведена тому, что описывается как *шутливый*. С точностью до лексических единиц это та же интерпретация, что и в случае с *полубумажный* в (23). Такое прочтение в целом очень характерно для конфигураций, когда для одной и той же сущности предлагаются две дескрипции с *полу-*, как в (25), ср. также известные *полуфанатик-полуплут* и *полумонахиня-полублудница* (пусть даже в этой случае мы имеем дело с существительными, а не с прилагательными).

В (26) речь, очевидно, не идет о половинной степени максимальной низкости — ввиду невозможности для соответствующей шкалы иметь максимум (**абсолютно/совершенно низкий*). Наиболее естественная интерпретация примеров типа (26) состоит в том, что *полунизкая передняя* описывает сущности, характеризуемые (примерно) половинным набором свойств объектов, для которых подходит дескрипция *низкая передняя*. Соответственно, перед нами интенсиональная аккомодация.

Дополнительным аргументом, подтверждающим реальность описанных типов аккомодации, выступает интерпретация сложных глаголов с *полу-* (гораздо менее многочисленных, чем прилагательные). *Полупрыгать* и *полупрочитать* иллюстрируются в (27)-(28):

(27) *Григоращенко в течение десяти минут* ***полупрочитал***, ***полурассказал*** *то, что было отпечатано на двух страничках с моих черновиков.* [ru-ecology.info]

(28) *Что за насекомое, похоже очень на муху, но больше и такое серо-коричневое? оно ещё так странно* ***полупрыгает*** *по потолку и стенам...* [otvet.mail.ru]

В (27) пропорции чтения и пересказа в изложении того, «что было отпечатано на двух страничках», составляют (примерно) поровну. Это аккомодация количественной шкалы — ровно такая же, как в (23) и (25) с прилагательными.

Единственно возможная интерпретация (28) состоит в том, что движения, производимые неизвестным насекомым, обладают примерно половиной свойств, которые характеризуют ситуации, описываемые как *прыгать*.

Следует отметить один мыслимый тип аккомодации шкалы, который не встречается в наших данных. Это аккомодация такой шкалы, которая идентична открытой шкале исходного прилагательного, но к которой добавляется контекстно-заданный максимум. Если бы такое было возможно, мы бы ожидали приемлемости предложений типа (29).

(29) **В Фоминском дома высокие, а в Терехово полувысокие.*

В (29) первая клауза задает контекстно-зависимый стандарт высоты по отношению к классу сравнения ‘дома в Фоминском’. Если бы этот стандарт можно было внедрить в шкалу для прилагательного во второй клаузе в качестве максимума, она бы интерпретировалась как ‘высота домов в Терехово составляет примерно половину стандартной высоты домов в Фоминском’. Это, однако, невозможно.

5. Вместо заключения

Сюжет нашего исследования далеко не исчерпан. Напротив, мы находимся в начале пути.

Материал композитов с *полу-* огромен, и в этой статье освещается лишь весьма незначительная его часть. Поэтому авторы просят читателей, которые не нашли здесь обсуждения известных им важных и интересных примеров с *полу-*, быть снисходительными.

Кроме того, следует иметь в виду, что структура шкалы представляет собой имплицитный параметр интерпретации, а результатом аккомодации выступают компоненты значения, вовсе отсутствующие в семантическом представлении. Обычная в таких случаях ситуация — значительное варьирование в суждениях носителей языка по части приемлемости и допустимой интерпретации изучаемых выражений.

Отчасти это компенсируется данными корпусов, которые дают положительный материал. Отрицательный материал, однако, — это серьезная проблема, которая, возможно, требует решения экспериментальными методами.

Отметим, наконец, что предметом этой статьи выступали почти исключительно прилагательные. Мы надеемся, исходя из примеров типа (27)-(28), что предложенные обобщения распространяются и на глаголы, а также, возможно, на существительные (*полусон, полубред, полувзгляд, полузащита, полумгла, полулюбитель*), которые остались целиком за пределами рассмотрения.

Завершая изложение, мы хотели бы выразить сдержанный оптимизм. Предложенные нами предварительные обобщения позволяют, как кажется, объяснить значительное количество характеристик дистрибуции *полу-* и избежать при этом серьезных эмпирических затруднений.

Благодарности

Исследование поддержано грантом РНФ 22-18-00285, реализуемом в МГУ имени М.В. Ломоносова. Авторы признательны анонимным рецензентам «Диалога» за комментарии и критику, которые побудили нас внести в текст многие существеннее уточнения. Любые недочеты, которые читатель обнаружил выше, остаются на совести авторов.

Литература

[1] Пирога Н.Г. Семантические, словообразовательные и орфографические особенности слов с *пол-*, *полу* // Вісник Дніпропетровського університету імені Альфреда Нобеля. Серія «Філологічні науки». — 2013. — № 1.

[2] Гапонова Ж.К. Семантика с префиксом *полу-* в русском языке (историко-лексикологический аспект) // Верхневолжский филологический вестник. — 2019. — № 2 (17).

[3] Гапонова Ж.К. Лексика с префиксоидом *полу-* (на материале ярославских говоров) // Лексический атлас русских народных говоров (материалы и исследования). — СПб., 2019.

[4] Caudal Patrick, Nicolas David. Types of degrees and types of event structures // C. Maienborn and A. Wollstein-Leisten (eds.) Event arguments: Foundations and applications. — Tubingen: Niemeyer, 2004.

[5] Hay Jennifer, Kennedy Christopher, Levin Beth. Scale structure underlies telicity in ‘degree achievements’. Semantics and Linguistic Theory. — 1999. — Vol. 9. — P. 134-151.

[6] Kennedy Christopher, McNally Louise. Scale structure and the semantic typology of gradable predicates // Language — 2005. —Vol. 81. — P. 345–381.

[7] Kennedy, Christopher and Beth Levin. Measure of Change: The Adjectival Core of Degree Achievements. In L. McNally, C. Kennedy (eds.) Adjectives and Adverbs: Syntax, Semantics and Discourse. Oxford: Oxford University Press, pp. 156-182.

[8] Kennedy Christopher. Vagueness and grammar // Linguistics and Philosophy. — 2007. — Vol. 30. — P. 1–45.

[9] Kennedy Christopher. The composition of incremental change // V. Demonte, L. McNally (eds.) Telicity, Change, State: A Cross-Categorial View of Event Structure. — Oxford: Oxford University Press, 2012. — P. 103–121.

[10] Piñón Christopher. Aspectual composition with degrees // L. McNally, C. Kennedy (eds.) Adjectives and Adverbs: Syntax, Semantics and Discourse — Oxford: Oxford University Press, 2008. — P. 183-219.

[11] Rotstein Carmen, Winter Yoad. Total adjectives vs. partial adjectives: Scale structure and higher-order modifiers // Natural Language Semantics. — 2004. — Vol. 12. — P. 259–88.

Text simplification as a controlled text style transfer task

Tikhonova Maria
HSE University, SberDevices
Moscow, Russia
m_tikhonova94@mail.ru

Fenogenova Alena
SberDevices
Moscow, Russia
alenush93@gmail.ru

Abstract

The task of text simplification is to reduce the complexity of the given piece of text while preserving its original meaning to improve readability and understanding. In this paper, we consider the simplification task as a subfield of the general text style transfer problem and apply methods of controllable text style to rewrite texts in a simpler manner preserving their meaning. Namely, we use a paraphrase model guided by another style-conditional language model. In our work, we perform a series of experiments and compare this approach with the standard fine-tuning of an autoregressive model.

Keywords: text simplification, natural language processing, machine learning, text style transfer
DOI: 10.28995/2075-7182-2023-22-507-516

Задача симплификации текста как задача управляемого переноса стиля

Тихонова Мария
НИУ ВШЭ, SberDevices
Москва, Россия
m_tikhonova94@mail.ru

Феногенова Алена
SberDevices
Москва, Россия
alenush93@gmail.ru

Аннотация

Задача автоматического упрощения текста состоит в том, чтобы уменьшить сложность подаваемого текста с целью улучшения удобочитаемости и понимания, но при этом сохраняя первоначальный смысл. В данной статье мы рассматриваем упрощение текста как задачу переноса стиля (style transfer). Мы исследуем методы управляемой генерации при переносе стиля текста для автоматической генерации упрощенных текстов. А именно, мы используем исходную модель перефразирования текста и дополнительный стилевой дискриминатор (GeDi-classifier), который контролирует выход и направляет генерацию модели в нужный стиль "упрощения"текста. В работе мы проводим серию экспериментов и сравниваем этот подход со стандартным дообучением авторегрессионной модели.

Ключевые слова: автоматическое упрощенние текстов, обработка естественного языка, текстовый стайл трансфер, перенос стиля, генеративные модели

1 Introduction

The goal of text simplification (or TS, in short) is to reduce the linguistic complexity of the given text fragment to improve its readability and to make it easier to understand. Text complexity depends on the presence of participial and adverbial constructions, complex grammatical structures, infrequent and ambiguous words, and subordinate sentences. Thanks to its numerous applications, the TS problem has received significant attention in Natural Language Processing (or NLP). For instance, it may simplify communication for non-native speakers and people with cognitive disorders such as aphasia or dyslexia. In addition, text simplification can improve language model performance on such NLP tasks as semantic role labeling, summarization, information extraction, machine translation, etc.

One standard approach to solving this task is to fine-tune a pre-trained language model on a large text corpus containing aligned complex and simplified sentences.

In this paper, we step aside from this paradigm and consider TS as a text style transfer task, regarding the "simplicity of the text" as a particular style. For this purpose, we use methods of controllable text generation. Namely, the GeDi algorithm proposed in (Krause et al., 2020) and further developed in (Dale et al., 2021). Following their methodology we use a paraphrase model (the main model) guided by another language model conditioned for the "simple" style (or GeDi-classifier). The choice of such an approach was motivated by its several advantages compared to standard fine-tuning of the pre-trained language model. First, it does not change the main language model. The trained GeDi-classifier can be used with different main models (for example, rewriter based on RuT5-Large, rewriter based on RuT5-XL, summarizer based on RuT5-Large, summarizer based on RuT5-Large, etc.), which gives more freedom for its usage. Thus, it simplifies the fine-tuning process as the classifier should only be trained once and then can be used in combination with various main models. Second, we can train several GeDi-classifiers with different target styles (sentiment, simplification, toxicity, etc.) and use them with any of the main language models we have. Thus, we only need to fine-tune M main models and N GeDi-classifiers instead of fine-tuning $N * M$ models for each combination.

In this work, we perform a series of experiments on the simplification dataset from the RuSimpleSentEval-2021 Shared Task (Sakhovskiy et al., 2021). We compare the controllable text style transfer approach with standard fine-tuning of autoregressive language models and show that GeDi-based approach of controllable text style transfer achieves quality comparable with standard fine-tuning.

The rest of the paper is structured as follows: first, in section 2 we overview the papers related to the field of TS and a paraphrase task, which can be regarded as its generalization, as well as the methods for controllable style generation. Next, in section 3 we discuss the controllable text style transfer approach we use. Then, section 4 describes the experimental setup. Section 5 presents evaluation results. Finally, section 6 concludes the paper.

2 Related Work

The task of text simplification is a popular generation task in NLP, useful in many applications: from pre-processing for machine translation to assistive technology for people with cognitive disorders. The systems of TS improve text readability and simplify text understanding while retaining its original information content as much as possible. The automation of this process is a complex problem which has been explored from many points of view. Several good extensive surveys cover the datasets and most of the classical methods for TS problem (Shardlow, 2014), (Al-Thanyyan and Azmi, 2021).

The interest and the development of TS systems for the Russian language rapidly increased with the RuSimpleSentEval Shared Task (Sakhovskiy et al., 2021), for which the authors presented the dataset and baselines. In addition, other Russian datasets exist for TS, among which are ruBTS (Galeev et al., 2021) and the aligned parallel TS dataset from language learner data (Dmitrieva et al., 2022).

The TS task can be considered the sub-task of the paraphrase task due to the similarity of the task definition and criteria of the generated text: the format should be changed while preserving the original text content. For the Russian language, several paraphrase models in the open source are commonly used, for example, paraphrased library (Fenogenova, 2021), or models by David Dale [1]. These models work on the sentence level. In addition, there exist a model from Sber [2] that rewrites extensive texts, which can contain many sentences.

For the evaluation of paraphrase tasks, the standard natural language generation (NLG) metrics are commonly used. There are surface-based metrics such as variations of BLEU, ROUGE, CHRF+; and BERT-base metrics such as LABSE (Feng et al., 2020) and BertScore (Zhang et al., 2019). For instance, their combinations are presented in the GEM benchmark (Gehrmann et al., 2021). Besides, for the TS task, special metrics such as SARI (Xu et al., 2015), included in the EASSE [3] package and Lens (Maddela

[1] https://huggingface.co/cointegrated/rut5-base-paraphraser
[2] https://sbercloud.ru/ru/datahub/rugpt3family/demo-rewrite
[3] https://github.com/feralvam/easse

et al., 2022), were proposed.

The controllable text style transfer approach has received considerable attention in recent years. One of the pioneers in this field was (Keskar et al., 2019), where authors use conditioned controlled codes for guided text generation.

GeDi (Krause et al., 2020) uses a small external language model classifier (or simply GeDi-classifier) to guide the generation of the main language model, re-weighting next token probabilities and, thus, increasing the probabilities of words in the given style. ParaGeDi (Dale et al., 2021) adopts this idea to the paraphrasing task by applying the GeDi approach in combination not with the standard language model but with the paraphraser fine-tuned to rephrase the original text preserving its original meaning.

In (Liu et al., 2021) the authors proposed DExperts. Their approach uses two extra language models conditioned towards and against the desired style (or topic), which are used to re-weight the probabilities of the next tokens predicted by the main language model.

(Yang and Klein, 2021) explores the usage of text classifiers for controllable text generation with FUDGE. This idea is further developed in (Sitdikov et al., 2022), where authors proposed CAIF sampling, which is a method for controllable text generation based on re-weighting logits with a free-form classifier.

Thus, while most methods for controllable text style transfer concentrate on controllable text generation in a given style, we focus on the task of paraphrasing the original text in a given style, preserving the meaning and applying the ideas from the ParaGeDi method for text simplification, regarding the simplicity of the text as a specific style. It should also be noted that while the work ParaGeDi uses GPT-2 language models, we use RuT5-Large based models. In other words, both components are derived from the same pre-trained language model version. Such an approach avoids problems with the difference in the vocabulary in the process of fine-tuning.

In addition, we perform our research for the Russian Language, which distinguishes our work from the papers mentioned above, which concentrate on English.

3 Method

Besides the standard approach of fine-tuning a pre-trained language model used as a baseline for the style-transfer experiments, we consider several versions of controlled text generation models based on the GeDi algorithm proposed in (Krause et al., 2020). In it a language model performs text generation guided by another language model conditioned for the specific topic or style or topic. More precisely, in our work, we adopt the extension of this method presented in (Dale et al., 2021), where the authors enable the model not only to generate but to paraphrase the input text. Below, a brief description of the method is given.

3.1 GeDi

In the original GeDi algorithm, the whole model consists of two parts. The first component is a generation autoregressive model. The second model is an autoregressive discrimination model, trained on sentences labeled with a specific style or topic, which we will further refer to as **GeDi-classifier**. Thus, in the process of training GeDi-classifier learns the word distributions conditioned on a particular label. At each generation step, the distribution of the next token predicted by the main language model P_{LM} is adjusted using the Bayes rule and an additional class-conditional language model P_D:

$$P(x_t|x_{<t}, c) \propto P_{LM}(x_t|x_{<t})P_D(c|x_t, x_{<t})$$

Here, x_t is the current token, $x_{<t}$ is the prefix of the text, and c is the desired style (e.g. simplicity or sentiment) — one of C classes. The first term in the formula is predicted by the main language model P_{LM}. The second term is calculated using GeDi-classifier P_{DC} via the Bayes rule. As a result the tokens which are more likely to appear in a text of the chosen style receive a higher probability:

$$P_D(c|x_t, x_{<t}) \propto P(c)P_{DC}(x, x_{<t}|c)$$

In the original paper, GeDi was successfully used for guided text generation with GPT-2 language model making the generation of the less toxic texts.

3.2 ParaGeDi

In our work, we adopt the approach of ParaGeDi, where the authors enable GeDi to preserve the meaning of the input text. For this, they replace the language model with a paraphraser. Thus, ParaGeDi models the following probability:

$$P(y_t|y_{<t}, x, c) \propto P_{LM}(y_t|y_{<t}, x)P(c|y_t, y_{<t}, x) \approx P_{LM}(y_t|y_{<t}, x)P_D(c|y_t, y_{<t})$$

where x is the original text, y is the generated text of length T, and c is the desired style.

The last transition in the equation above is an approximation which allows us to decouple the paraphraser from the GeDi-classifier model. As a result, the paraphraser and the GeDi-classifier can be trained independently in such a formulation.

As for the training process, ParaGeDi loss $\mathcal{L}_{ParaGeDi}$ consists of two components: the generative loss $\mathcal{L}_G$ used in language model training and the discriminative loss $\mathcal{L}_D$ which further pushes different classes away from one another.

$$\mathcal{L}_G = -\frac{1}{N}\sum_{i=1}^{N}\frac{1}{T_i}\sum_{t=1}^{T_i}\log P(y_t^{(i)}|y_{<t}^{(i)}, c^{(i)})$$

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N}\log P(c^{(i)}|y_{1:T_i}^{(i)})$$

$$\mathcal{L}_{ParaGeDi} = \lambda\mathcal{L}_D + (1-\lambda)\mathcal{L}_G$$

where $\lambda \in [0, 1]$ is the weight of the discriminative loss.

Besides, to improve the preservation of the original content and to increase the style transfer accuracy, the following heuristics are used:

First, the conditional language model probability is raised to the power $w > 1$, which biases the discriminator towards the correct class in the process of generation:

$$P(y_t|y_{<t}, x, c) \propto P_{LM}(y_t|y_{<t}, x)P_{CC}(c|y_t, y_{<t})^w$$

Second, probabilities are smoothed by adding a small $\alpha > 0$ to all probabilities from the conditional language model:

$$P_\alpha(c|x_t, x_{<t}) = \frac{\alpha + P(c)P_{CC}(x, x_{<t}|c)}{\sum_{c' \in C}(\alpha + P(c')P_{CC}(x, x_{<t}|c'))}$$

Such a heuristic discourages the generation of tokens with low probability conditional on all classes.

Third, for class-conditional corrections, asymmetric lower and upper bounds (l and u) are used :

$$P_{\alpha,l,u}(c|x_t, x_{<t}) = \max(l, \min(u, P_\alpha(c|x_t, x_{<t}))).$$

This discourages the insertion of new tokens, as opposed to prohibiting existing tokens.

4 Experiments

4.1 Data

We perform a series of experiments on the dataset RuSimpleSentEval-2021 Shared Task (Sakhovskiy et al., 2021). This simplification dataset contains parallel pairs of sentences: complex – their corresponding simplified versions. Below, a sample from the dataset is presented.

Example from the dataset:

Source sentence:

"Климат Казани – умеренно континентальный , сильные морозы и палящая жара редки и не характерны для города"

Simplified paraphrases:

1. *"В Казани редко бывают и сильные морозы, и жаркая летняя погода"*
2. *"В Казани зимой не слишком холодно, а летом не слишком жарко"*
3. *"В Казани зимой не очень холодно, а летней жары почти не бывает"*

The organizers of the RuSimpleSentEval-2021 shared task constructed the dataset using automatic translation and post-processing WikiLarge corpus (Zhang and Lapata, 2017). The resulting dataset was split into the train, dev and two test sets (public and private). And while the train set was not filtered or verified, the organizers validated the dev, public and private test sets via crowd-sourcing using Yandex.Toloka [4] and filtered them. In this work, we evaluate the results on official public and private test sets. We additionally filtered the train part, which contains inappropriate examples due to its original automatic construction. For its cleaning, we used the following procedure: exclude examples with less than two lemmas in the intersection between the lemmatized source and target sentences (lemmatization was done via pymorphy2 [5] tagger (Korobov, 2015)); discard examples where the source sentence is a sub-string of the target one and the length is bigger than of the source one. Besides training and validation, we also use extra dev set filtered by the organizer.

4.2 Models

We conduct experiments and compare the results of the following models:

- **Golden testset**. We evaluate the golden references (first answer) from the fixed RuSimpleSentEval-2021 test sets (public/private);
- **Paraphraser**. We use a paraphrase model [6] trained on 7000 examples from different sources of various domains: 1) text level: literature domain, prose; back translation (with ru-en translation model [7]) of the texts from different domains filtered with Bertscore Rouge-L); 2) sentence level: Russian version of Tapaco corpus (Scherrer, 2020) and filtered ParaphraserPlus (Gudkov et al., 2020) corpus.
- **Fine-tuned paraphraser**. We additionally fine-tune the paraphrase model on the train set to check the hypothesis of the capabilities combinations that the model learn (both paraphrasing and simplification);
- **Fine-tuned ruT5-Large** [8]. We fine-tune the row ruT5-Large model on the simplification train set.
- **ParaGeDi**. We train GeDi-classifier on the train part of the RuSimpleSentEval-2021 set and use the paraphrase model described above as the main model for ParaGeDi controllable approach.

In our work, all models we use are derived from the pre-trained RuT5-Large[9] model, which is a T5 model (Raffel et al., 2020) pre-trained for the Russian language. The fact that we derive both components from the same model allows us to avoid problems with the difference in the model vocabulary.

As for the GeDi-classifier model, we fine-tune RuT5-Large on the RuSimpleSentEval-2021 Shared Task train set. We use the Adam optimizer with the learning rate $1e-4$, three epochs, and the weight of the discriminative loss $\lambda = 0.3$.

We evaluate several style power coefficients ($w = 5, 10, 15, 20$). It should also be noted that we do not evaluate $w = 0$ as, in this case, the influence of the GeDi-calssifier is neglected, and the result is equal to the original paraphrase model, which is our baseline. To avoid randomness, we use the following generation parameters:

[4] https://toloka.ai/tolokers/

[5] https://github.com/pymorphy2/pymorphy2

[6] `https://habr.com/ru/company/sberdevices/blog/667106/`

[7] `https://huggingface.co/Helsinki-NLP/opus-mt-en-ru`

[8] https://huggingface.co/sberbank-ai/ruT5-large

[9] `https://huggingface.co/sberbank-ai/ruT5-large`

- $do\_sample = False$,
- $num\_returned\_sequences = 1$,
- $max\_len = 128$.

4.3 Metrics

We evaluate the model on public and private test sets of RuSimpleSentEval-2021 Shared Task using the following metrics:

- **BertScore**(Zhang et al., 2019), which is computed between the original (complex) sentences and model predictions.
- **SARI** (Xu et al., 2016), which is commonly recognized as a metric for evaluating automatic text simplification systems. The metric compares the model predictions against the references and the original (complex) sentences.
- **BLEU score**(Papineni et al., 2002), which in our case is computed between the reference answers and predictions
- **iBLEU** (Sun and Zhou, 2012) which is computed as follows:

$$iBLEU = a * BLEU(preds, refs) + (1 - \alpha) * BLEU(preds, source),$$

  where α is the parameter responsible for the balance between adequacy and dissimilarity. In our work, we follow the methodology from the original paper and use $\alpha = 0.9$.
- **Diversity** We report a degree of diversity measured using the mean number of distinct n-grams, normalized by the length of text (Li et al., 2015). We report dist-1, dist-2, and dist-3 scores for distinct uni-, bi-, and trigrams, respectively.

5 Results

Results on public and private test sets are presented in Tables 1 and 2, respectively. The results reveal that the GeDi-based approach with style power coefficients of 5 and 10 shows quality comparable with the standard fine-tuning approach. Larger values of the style power coefficient lead to a decrease in quality as the classifier influence becomes too strong, which negatively affects the generated output. Thus, the ParaGeDi-based approach can be considered a good alternative to standard fine-tuning. In addition, as long as it does not change the initial model and can be used with different main models, it gives more freedom for its usage.

Model	BertScore	SARI	BLEU	iBLEU 0.9	dist 1	dist 2	dist 3
Golden testset	0.816874	**66.106573**	1.0	0.916141	0.971855	**0.940157**	**0.882364**
Paraphraser	0.925663	41.004799	0.314653	0.342387	0.964854	0.923054	0.855773
FT paraphraser	**0.970198**	41.594171	0.367276	0.412937	0.974326	0.932282	0.866955
FT ruT5-Large	0.969541	41.819602	**0.369884**	**0.415395**	**0.974098**	0.931853	0.866066
ParaGeDi (sp 5)	0.914065	40.792974	0.310180	0.332548	0.965152	0.919561	0.848917
ParaGeDi (sp 10)	0.888886	40.501325	0.295284	0.307751	0.969362	0.911230	0.831918
ParaGeDi (sp 15)	0.826108	38.539389	0.256159	0.255457	0.882723	0.815006	0.731320
ParaGeDi (sp 20)	0.659992	33.045052	0.081489	0.075360	0.401245	0.356622	0.307940

Table 1: The results on the public test set of the RuSimpleSentEval-2021. ParaGeDi is evaluated with different Style Power coefficients (sp in shortly). *FT* stands for fine-tuned. Detailed metrics descriptions are given in subection 4.3.

In addition, we compared our results with the top-3 solutions of the RuSimpleSentEval-2021 competition (Sakhovskiy et al., 2021), which include *qbic* solution based on Multilingual Unsupervised Sentence Simplification (Martin et al., 2020) and fine-tuned GPT-based solutions by *orzhan*, *ashatilov*, and *alenusch*. To complete the picture, we also included mBART-based (Liu et al., 2020) baseline presented by the organizers. Results are presented in Table 3. First, it can be seen that all our solutions (which are RuT5-based) surpass the baseline. Second, most of them, including the ParaGeDi method with reasonable style power coefficient of 5 and 10, outperform competition winners (mostly GPT-based) showing

Model	BertScore	SARI	BLEU	iBLEU 0.9	dist 1	dist 2	dist 3
Golden testset	0.816874	**66.106573**	1.0	**0.967823**	0.940655	0.883676	**0.882364**
Paraphraser	0.92467	40.418701	0.301265	0.330843	0.961526	0.922913	0.857691
FT paraphraser	**0.968782**	41.643578	**0.358353**	0.404432	0.968473	**0.931082**	0.866247
FT ruT5-Large	0.965881	41.517535	0.357556	0.402777	**0.969426**	0.929413	0.863115
ParaGeDi (sp 5)	0.912825	40.859850	0.300608	0.324721	0.961111	0.918092	0.848473
ParaGeDi (sp 10)	0.887088	40.240902	0.274954	0.289805	0.960448	0.907891	0.830453
ParaGeDi (sp15)	0.824515	38.249361	0.255155	0.255730	0.873924	0.810920	0.730028
ParaGeDi (sp 20)	0.668402	33.238699	0.098595	0.091794	0.432894	0.389271	0.339774

Table 2: Simplification results on the private test set. ParaGeDi is evaluated with different Style Power coefficients (sp in shortly). *FT* stands for fine-tuned. Detailed metrics descriptions are given in subection 4.3.

higher SARI scores. Such results can be regarded as another proof of the quality of the ParaGeDi approach. In addition, such results indicates that RuT5 is a better backbone for the text simplification task than the GPT-based models. We observe the same trends on the TS task in the GEM benchmark [10]. The T5-small model shows the best performance on the analogous datasets for English, among which are wiki auto, asset turk, and test turk datasets (Xu et al., 2016)).

Model	SARI
Golden testset	**66.106**
FT ruT5-Large	**41.819**
FT paraphraser	**41.594**
Paraphraser	**41.004**
ParaGeDi (sp 5)	**40.792**
ParaGeDi (sp 10)	**40.501**
№1 orzhan	40.233
№2 alenusch	38.870
№3 ashatilov	38.843
ParaGeDi (sp 15)	38.539
ParaGeDi (sp 20)	33.045
BASELINE	30.152

Model	SARI
Golden testset	**66.106**
FT paraphraser	**41.643**
FT ruT5-Large	**41.517**
Paraphraser	**40.418**
ParaGeDi (sp 5)	**40.859**
ParaGeDi (sp 10)	**40.240**
№1 qbic	39.689
№2 orzhan	39.279
№3 ashatilov	38.491
ParaGeDi (sp 15)	38.249
ParaGeDi (sp 20)	33.238
BASELINE	-

Table 3: Simplification results comparison with the top-3 solutions of RuSimpleSentEval-2021 on public (left) and private (right) test sets respectively. Following the original competition approach, we compare results using the SARI metric. *sp* stands for style power. *FT* stands for fine-tuned. *BASELINE* stands for the mBART-based baseline from RuSimpleSentEval-2021 competition.

Analyzing model's performance on the concrete examples, we explored the behavior of Fine-tuned paraphraser, which showed best performance on the private test set (see Table 4). It can be seen that the model preserves the original meaning of the sentence and does not change facts. However, it only slightly simplifies sentences replacing individual words rather than rewriting the whole sentence in a more simplistic style. Thus, there is still room for improvement.

As for the limitations of the ParaGeDi method, with the growth of the style power coefficient, ParaGiDi starts to lose coherence and forget the original sentence's meaning. While the generations with the style power coefficients 5 or 10 yields reasonable quality, sentences generated with the coefficient 20 often lack

[10] https://gem-benchmark.com/results

Original sentence	Prediction
Дорожка следов у волка ровнее и образует почти ровную строчку, а у собак — извилистую линию.	Дорожка следов у волка ровная, а у собак — извилистая линия.
В том же году компания открыла офис в Казани; главная страница портала, некоторые службы и поиск стали доступны как на русском, так и на татарском языке.	В том же году компания открыла офис в Казани. Многие службы и поиск стали доступны как на русском, так и на татарском языке.
В связи с появившейся необходимостью медицинского обследования врачи приняли решение направить его в одну из московских клиник.	В связи с возникшей необходимостью медицинского обследования врачи отправили его в одну из московских клиник.
В рационе древних египтян присутствовали горох, бобы и нут, огурцы, в больших выращивался салат-латук.	В рационе древних египтян были горох, бобы и огурцы, в большом количестве выращивался салат-латук.
Атлантические течения, разогретые Гольфстримом, приносят мягкие зимы; иногда зимой и ранней весной здесь бывают снегопады, хотя снег обычно лежит недолго.	Атлантические течения приносят мягкие зимы, и иногда зимой и ранней весной здесь бывают снегопады, хотя снег обычно лежит недолго.

Table 4: Fine-tuned paraphraser examples from the test set.

meaning. In addition, as long as the ParaGeDi approach uses two language models, it works slower and requires more computational resources during the inference stage compared to the fine-tuned language models.

6 Conclusion

In this paper, we dealt with the text simplification problem regarding it as a special case of text style transfer task. We adopted the ParaGeDi method, which uses the idea of controlled text style transfer. We used the combination of two RuT5-Large models (paraphrase model and GeDi-classifier) to solve this task. In the experiments, that approach proved quite promising; the results are comparable to fine-tuning for the single style class. The ruT5-based simplification models surpassed the best results on the RuSimpleSentEval-2021 shared task.

As a part of future research, we plan to consider the reverse problem of making the text more complex and official. Thus, we plan to explore the capabilities of the models, which can work in both directions: simplifying the text or making it more complex and official.

References

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: a survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

David Dale, Anton Voronov, Daryna Dementieva, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Text detoxification using large pre-trained neural models. *arXiv preprint arXiv:2109.08914*.

Anna Dmitrieva, Antonina Laposhina, and Maria Yuryevna Lebedeva. 2022. Creating a list of word alignments from parallel russian simplification data. *Frontiers in Artificial Intelligence*, 5:984759.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Alena Fenogenova. 2021. Russian paraphrasers: Paraphrase with transformers. // *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, P 11–19, Kiyv, Ukraine, April. Association for Computational Linguistics.

Farit Galeev, Marina Leushina, and Vladimir Ivanov. 2021. rubts: Russian sentence simplification using back-translation. // *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, P 259–267.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezudo, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. // *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, P 96–120, Online, August. Association for Computational Linguistics.

Vadim Gudkov, Olga Mitrofanova, and Elizaveta Filippskikh. 2020. Automatically ranked russian paraphrase corpus for text generation. *arXiv preprint arXiv:2006.09719*.

Nitish Shirish Keskar, Bryan McCann, Lav R Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation. *arXiv preprint arXiv:1909.05858*.

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.

Ben Krause, Akhilesh Deepak Gotmare, Bryan McCann, Nitish Shirish Keskar, Shafiq Joty, Richard Socher, and Nazneen Fatema Rajani. 2020. Gedi: Generative discriminator guided sequence generation. *arXiv preprint arXiv:2009.06367*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A Smith, and Yejin Choi. 2021. Dexperts: Decoding-time controlled text generation with experts and anti-experts. *arXiv preprint arXiv:2105.03023*.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2022. Lens: A learnable evaluation metric for text simplification. *arXiv preprint arXiv:2212.09739*.

Louis Martin, Angela Fan, Éric De La Clergerie, Antoine Bordes, and Benoît Sagot. 2020. Muss: multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. // *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, P 311–318.

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. *The Journal of Machine Learning Research*, 21(1):5485–5551.

Andrey Sakhovskiy, Alexandra Izhevskaya, Alena Pestova, Elena Tutubalina, Valentin Malykh, Ivan Smurov, and Ekaterina Artemova. 2021. Rusimplesenteval-2021 shared task: evaluating sentence simplification for russian. // *Proceedings of the International Conference "Dialogue*, P 607–617.

Yves Scherrer. 2020. Tapaco: A corpus of sentential paraphrases for 73 languages. // *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Askhat Sitdikov, Nikita Balagansky, Daniil Gavrilov, and Alexander Markov. 2022. Classifiers are better experts for controllable text generation. *arXiv preprint arXiv:2205.07276*.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. // *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, P 38–42.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Kevin Yang and Dan Klein. 2021. Fudge: Controlled text generation with future discriminators. *arXiv preprint arXiv:2104.05218*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

An attempt to determine a preposition and delimit the class of derived prepositions in Russian

Elena Uryson
Russian Language Institute RAS
Moscow, Volkhonka, 18/2, 119019, Russia
uryson@gmail.com

Abstract

The object of the paper are Russian words traditionally described as derived prepositions. The problem is that there is no formal definition of preposition in theoretical or applied linguistics. Non-derivative, or primitive prepositions are given in grammar by the closed list, so strictly speaking there is no need to define this class of words. However. we must have criteria for determining derived prepositions. I suggest a set of necessary conditions that a preposition must satisfy. I demonstrate that so called adverbial prepositions in Russian do not satisfy them and should be described as adverbs. Similarly, some Russian verbal prepositions, and some Russian denominative prepositions should not be described as prepositions.

Key words: government, parts of speech, prepositions, adverbs, Russian, semantics, valency.
DOI: 10.28995/2075-7182-2023-22-517-524

К определению предлога и уточнению списка русских производных предлогов

Елена Урысон
Институт русского языка им. В.В. Виноградова РАН,
Москва, Волхонка 18/2, 119019
uryson@gmail.com

Аннотация

Объект работы – русские слова, относимые к разряду предлогов. В академической грамматике считается, что список предлогов открыт – он постоянно пополняется за счет производных предлогов. Однако в лингвистике отсутствует сколько-нибудь строгое определение предлога как части речи, следовательно неясны и основания, по которым та или иная единица причисляется к предлогам. Цель предлагаемой работы – во-первых, сформулировать необходимые условия, при соблюдении которых слово может быть отнесено к разряду предлогов; во-вторых, показать, что многие единицы, трактуемые академической грамматикой как производные предлоги, этим требованиям не удовлетворяют, так что отнесение их к разряду предлогов некорректно.

Ключевые слова: предлог; наречие; производный предлог; активная синтаксическая валентность; пассивная синтаксическая валентность; реализация валентности.

1. Определение предлога в академической грамматике и необходимые требования к предлогу

Объект работы – русские слова, относимые к разряду предлогов. В академической грамматике считается, что список предлогов открыт – он постоянно пополняется за счет производных предлогов. Однако в лингвистике отсутствует сколько-нибудь строгое определение предлога как части речи, следовательно неясны и основания, по которым та или иная единица причисляется к предлогам[1]. Цель предлагаемой работы – во-первых, сформулировать необходимые условия, при соблюдении которых слово может быть отнесено к разряду предлогов; во-вторых, показать, что

[1] Один из результатов такого положения дел – концепция М.В. Всеволодовой и ее коллег [Vsevolodova 2010], по которой в разряд предлогов (или предложных единиц) попадают тысячи слов и оборотов. Это противоречит общему представлению о служебных словах: в языке, по определению, служебных слов гораздо меньше, чем знаменательных, – так же как морфем-аффиксов гораздо меньше, чем корневых морфем.

многие единицы, трактуемые академической грамматикой как производные предлоги, этим требованиям не удовлетворяют, так что отнесение их к разряду предлогов некорректно.

Русская академическая грамматика делит предлоги на два класса: первообразные и производные (непервообразные). Ядерную группу русских предлогов образуют т. н. первообразные, или непроизводные предлоги, *в, на, у, к, из, о, с, до, по* и т.п.; сюда же относятся двойные предлоги, ср. *из-за, из-под* и т.п. Первообразные предлоги представляют собой закрытую немногочисленную группу «простейших слов» [Russian Grammar 1980: 706]. Эта группа закрыта в том смысле, что она не пополняется. Иными словами, непроизводные предлоги представляются списком.

Производные предлоги выделяются в языке по аналогии с первообразными. К классу производных предлогов относят такие единицы, которые связаны словообразовательно с какими-то словами, не являющимися предлогами, но сами ведут себя как предлоги, ср. *ввиду, в течение, благодаря* и т.п. Академическая грамматика не различает синхронную связь единицы с мотивирующим словом (ср. *ввиду, благодаря* и т.п.) и такую связь, которую можно обнаружить только в диахронии, ср. *кроме, ради*. Для наших целей это тоже несущественно.

Предлог — это морфологически неизменяемое слово, и данная часть речи определяется прежде всего через ту синтаксическую функцию, которую она выполняет в предложении. [LES 1990] дает следующее определение предлога: это «разряд служебных, морфологически неизменяемых слов, выражающих различные отношения между зависимыми и главными членами словосочетания и осуществляющих подчинительную синтаксическую связь» [LES 1990: статья «Предлог»]. Современная академическая грамматика дополняет это определение указанием на семантическую функцию предлога: это служебное слово не только оформляет подчинительную связь, но и выражает определенные отношения между объектом (ситуацией), обозначаемым главным членом словосочетания, и тем объектом (ситуацией), который обозначается зависимым членом словосочетания [Russian Grammar 1980] (в терминах когнитивной семантики это отношение между фоном и фигурой).

Такое определение не является операционным и, по-видимому, не поддается формализации: в его основе лежит понятие «служебная часть речи», которое тоже интуитивно ясно, но не определено сколько-нибудь строго.

В традиционной грамматике служебные части речи – к ним относятся предлог, союз и частица – противопоставлены знаменательным (это, прежде всего, глагол, существительное, прилагательное и наречие). Школьная грамматика различает знаменательные и служебные слова по следующему операционному критерию. В словосочетании (предложении) к знаменательному слову можно подставить вопрос от другого слова; ср. *ехать домой медленно*: *ехать* (куда?) *домой, ехать* (как?) *медленно*. (В предложении такой вопрос позволяет также определить, каким членом предложения является данное слово. Ср. *Ехали домой медленно*: *домой* – обстоятельство места, *медленно* – обстоятельство образа действия.) Что касается служебного слова, в частности предлога, то к нему подставить вопрос нельзя. Ср. *ехать* (куда?) *в Париж, жить* (где?) *в Париже, уехать* (откуда?) *из Флоренции*. При этом предлог может повторяться в вопросе, ср. *бороться с противником – бороться* (с кем?) *с противником*. Эта «неотделимость» предлога от существительного отражается в традиционной терминологии: предложно-падежную группу (ср. *в Москву, из города, с другом*) называют также предложно-падежной формой существительного.

Данный критерий легко усваивается детьми в начальной школе, однако очевидно, что он основан на языковой интуиции, причем она до сих не эксплицирована. Одна из задач теоретической лингвистики состоит в экспликации таких интуитивно ясных базовых понятий. (Такова, в частности, была задача определения падежа, при том что понятие падежа использовалось в грамматике и считалось вполне ясным сотни лет [LES 1990, статья «Падеж»]). Естественно предположить, что служебные слова, в том числе предлоги, отличаются от знаменательных слов коммуникативным статусом в высказывании, Однако эта гипотеза не разработана, а потому не может быть ни верифицирована, ни опровергнута. Мы попытаемся решить гораздо более скромную задачу – исходя из достаточно четко определенных синтаксических понятий, сформулировать необходимые требования, которым удовлетворяет предлог.

Легко видеть, что предлог обладает следующими синтаксическими свойствами.

(I) Он имеет две синтаксические валентности – пассивную (по этой валентности он синтаксически подчиняется другому предикату, ср. *ехать → в Москву, играть → на лугу, поездка → в Москву, игры → на лугу; полный → до краев*) и активную, по которой он управляет зависимым от него словом, ср. *в → Москве, на → свете*.

Поясним понятия «активная и пассивная синтаксические валентности». Активная синтаксическая валентность – это способность слова синтаксически подчинять себе другое слово. Пассивная синтаксическая валентность – это способность слова быть синтаксическим зависимым другого слова [LES 1996, статья «Валентность»]. Заметим, что на семантическом уровне мы имеем дело с предикатом и его семантическим актантом, так что понятий активная и пассивная валентности на семантическом уровне нет. Что касается синтаксического уровня, то предикат обычно синтаксически подчиняет себе обозначение того или иного семантического актанта, ср. *любить → мороженое, купить → за сотню*; таковы, в частности, предикаты-глаголы. Однако некоторые предикаты на синтаксическом уровне сами подчиняются обозначению своего семантического актанта; таковы многие прилагательные – одноместные предикаты, ср. *красивый, хороший, плохой, холодный, деревянный* и т.п. Семантический актант этих предикатов – объект, которому предицируется данный признак; этот актант обычно выражается существительным, причем это существительное синтаксически подчиняет себе обозначение предиката-характеристики. Для наших целей важно различать активную и пассивную синтаксические валентности предлога; при этом предложная группа может быть обозначением как актанта своего синтаксического хозяина (*ехать → в Москву, поездка → в Москву*), так и его сирконстанта (*играть → на лугу, игры → на лугу*). О понятиях актант и сироконстант см. [Testelets 2001].

(II) Активная синтаксическая валентность предлога практически обязательно должна быть заполнена.

(III) Активная синтаксическая валентность предлога заполняется существительным, в языках с именным словоизменением определенной падежной формой, ср. *в → Москву, до → краев*.

(IV) Предлог не может иметь никаких других синтаксических зависимых, кроме этого единственного управляемого им синтаксического актанта.

(V) Предлог линейно располагается перед управляемым существительным, точнее – непосредственно перед именной группой, в вершине которой стоит данное существительное.

(Условия (I) – (IV) являются, по-видимому, общими для предлогов и послелогов. Условие (V) отличает предлог от послелога. Обсуждение послелогов выходит за рамки предлагаемой работы.)

Свойство (I) не нарушаемо. Свойства (II), (IV) и (V) допускают специально оговариваемые исключения,

Известное исключение из (II) представляют собой противопоставительные (часто экспрессивные) контексты, ср. – *Вам кофе с лимоном или без? – Лучше без; Сумка не НА кровати, а ПОД, ПОД!*

Исключение из (IV): некоторые предлоги могут подчинять себе наречие со значением ‘без существенного пространственного или временно́го промежутка’, ср. *Детскую площадку устроили непосредственно ← перед домом, Взрыв был зафиксирован непосредственно ← после вспышки, Веник стоит сразу ← за дверью*.

Свойством (V) обладают не все единицы, относимые к предлогам: как известно, некоторые предлоги могут располагаться как перед, так и после зависимого слова, ср. *ради Христа – Христа ради*. Таких предлогов очень мало, и они задаются списком. (На первый взгляд, такие единицы можно называть послелогами. Однако это некорректно, т.к. послелог по определению всегда располагается после своего синтаксического зависимого, т.е. данные единицы опять окажутся исключением. Но русскому языку послелоги как минимум несвойственны, и описание этих единиц как послелогов с нестандартным поведением окажется весьма неэкономным. Столь же некорректно будет назвать эти единицы предлогами-послелогами, т.к. по определению предлог, как и послелог, обладает фиксированной позицией относительно управляемого слова.)

2. Экскурс: О количестве семантических актантов предлога

Остановимся на количестве семантических актантов предлога. У большинства предлогов их два: одному актанту соответствует пассивная синтаксическая валентность, а другому – активная. Ср. *идти (А2) с приятелем (А1)*: выражение актанта А1 («второе действующее лицо» ситуации)

заполняет активную синтаксическую валентность предлога, а обозначение актанта А2, т.е. самой ситуации, реализует пассивную валентность предлога.

Но некоторые предлоги имеют три семантических актанта. Таков, например, предлог *через* во временно́м значении, ср. *встретиться (А2) через год (А1) после развода (А3)*. Предлог *через* указывает здесь на временной интервал А1 (год), разделяющий события А2 (встретиться) и А3 (развод). Однако на синтаксическом уровне у этого предлога всего две валентности. Активную валентность заполняет выражение актанта А1, т.е. временно́го интервала (*через → год*), а пассивную валентность – выражение актанта А2, т.е. более позднего события (*встретиться → через*). Что касается предшествующего события А3, то его выражение (*развод*) синтаксически подчиняется слову *после*, а не *через*. При этом вся группа *после А3* синтаксически подчиняется тому же предикату, что и группа *через год*, т.е. предикату *встретиться*: *встретиться → после → развода.* Тем самым, на синтаксическом уровне обозначение более раннего события А3 подчиняется не предлогу *через*, а другому слову.

Аналогичным образом устроен и предлог *за* во временно́м значении, ср. *встретиться (А2) за год (А1) до войны (А3)*. Этот предлог также указывает на временной интервал между двумя событиями (А2 – встретиться, А3 – война). При этом *за*, подобно *через*, управляет обозначением временно́го интервала (*за → год*), но, в отличие от *через*, подчиняется обозначению более раннего, а не более позднего события: *встретиться → за → год*. (Тем самым, предлоги *через* и *за* временно́м значении являются конверсивами.) Что касается обозначения более позднего события (*война*), то оно подчиняется не самому предлогу *за*, а предлогу *до*. Таким образом, обозначение актанта А1 (временно́го интервала) этого предлога и обозначение его актанта А3 (более позднего события) оказываются синтаксически соподчиненными – они оба синтаксически зависят от предиката *встретиться* (через предлоги: *встретиться → до → войны, встретиться → за → год*). При этом предлог *за*, в отличие от *через*, накладывает еще и определенные ограничения на выражение актанта А3 — этот актант должен быть выражен обязательно: неприемлемо **Они встретились за год* (ср. нормальное *Они встретились через год*).

Аналогичную структуру с соподчинением глаголу двух предлогов мы усматриваем в случаях типа *остановится в трех метрах от дерева*. Предлог *в* (в данном значении) имеет три семантических актанта: А1 – расстояние (*три метра*), А3 – ориентир, относительно которого оценивается расстояние (*дерево*), А2 – описываемая ситуация (*остановиться*). На синтаксическом уровне предлоги *в* и *от* (со своими именными группами) соподчиняются предикату *остановиться*.

Допускаем, что по три семантических актанта имеют и некоторые другие предлоги.

Подчеркнем, что требования (I) – (V) являются необходимыми, но не достаточными: данными свойствами теоретически может обладать не только предлог, но и какое-нибудь другое неизменяемое слово, например, деепричастие. Тем не менее, список этих требований полезен при анализе «кандидатов» в производные предлоги, т.к. на его основании можно сразу отсечь некоторые спорные единицы.

3. О некоторых других формальных требованиях к русским первообразным предлогам, а также наречиям

Кроме перечисленных свойств, которые присущи, по-видимому, всем предлогам (исключения для русского языка оговорены выше), русские первообразные предлоги обладают еще тремя формальными особенностями [Es'kova 1996]. Эти особенности таковы.

(а) После первообразного предлога употребляются «особые формы местоимений-существительных *он, она, оно, они* с начальным *н*: *него, неё, них* и т.п. (так и именуемых – «припредложными формами»)» [Там же: 458]. Ср. *у него, из нее, к нему, к ней, о ней, в них* и т.п.

(б) Для первообразного предлога обязательна интерпозиция между компонентами местоимений с начальным *ни-* или *не-*: *никто, ничто, никакой, ничей; некого, нечего*. Ср. *ни у кого, ни о ком, ни в чем, ни из чьего, не о ком, не в чем* и т.п. Кроме того, для первообразного предлога допустима интерпозиция между компонентами местоимений с начальным *кое (кой-)*: *кое-кто, кое-что, кое-какой, кое-чей*. Ср. *кое у кого, кое над чем, кое с каким, кое о чьем* и т.п., причем «строго нормативной считается интерпозиция предлога <…>, но эта норма достаточно

часто нарушается, причем конструкции с препозицией предлога употребляются и хорошими авторами <…>. Например: *на кое-кого* (В. Набоков), *для кое-кого* (В. Кардин), *на кое-какие вопросы* (Ю. Домбровский), <…> *к кое-каким новинкам* (В. Шаламов), *с кое-какими средствами* (Е. Носов)» [Там же: 460-461]. Наконец, интерпозиция первообразного предлога обязательна при сочетании его с *друг друга*, ср. *друг без друга, друг о друге, друг у друга, друг за другом, друг к другу* и т.п.

(в) Между первообразным предлогом и управляемой им группой невозможна вставка частиц, ср. **в же доме, *до ли войны, *из-то Москвы*; нормально *в доме же, до войны ли, из Москвы-то.*

Эти формальные особенности русского первообразного предлога не вытекают из основных свойств предлога (I)-(V) и логически никак с ними не связаны. Поэтому теоретически не исключено, что та или иная единица может обладать одной или более из особенностей (а)-(в), но при этом не иметь свойств предлога (I)-(V). Очевидно, что свойства (I)-(V) являются основными – хотя бы потому, что ими обладает предлог как минимум во всех славянских языках. Поэтому естественно считать, что если единица не обладает хотя бы одним из этих свойств, то она и не входит в разряд предлогов. Требуется выяснить, как свойства (а)-(в) коррелируют с основными свойствами предлога (I)-(V).

С этой точки зрения особый интерес представляет свойство (а) – употребление «*n*-форм» при данной единице: этим свойством действительно могут обладать не только первообразные предлоги. Свойства (б) и (в) присущи только первообразным предлогам, поэтому мы их рассматривать не будем.

Утвердилось мнение, что «надежный формальный признак предлога – употребление после него n-форм местоименных слов» [Es'kova 1996: 460]. Этот признак положен в основу первого критерия определения предлога в работе [Sichinava 2018]. Однако этот подход при всей своей привлекательности не верен.

Во-первых, *n*-формы местоимений употребляются не только после предлогов, но и после некоторых компаративов, ср. *лучше нее, хуже него, больше него, меньше нее* и т.п. [OD 1983]. Следовательно, употребление *n*-формы местоимения после какого-нибудь слова в общем случае не может свидетельствовать, что это слово является предлогом.

Во-вторых, в некоторых случаях этот признак вступает в конфликт с основным признаком (II) предлога – обязательностью заполнения его активной синтаксической валентности. Значит, признак (а) в целом неинформативен для определения предлога.

Продемонстрируем это, опираясь на работы [Uryson 2014; 2017]. Возьмем, например, слово *позади* в контекстах типа *Полиция шла позади колонны демонстрантов – Колонна демонстрантов приближалась к мэрии, полиция шла позади*. Слово *позади* требует *n*-формы местоимения, ср. *позади него, позади нее, позади них*. Является ли слово *позади* предлогом?

Очевидно, что слово *позади* в приведенных контекстах выступает в одном и том же значении. Очевидно также, что это слово имеет семантический актант 'ориентир', а на синтаксическом уровне – синтаксическую валентность, заполняемую обозначением этого актанта. В обоих примерах семантический актант 'ориентир' слова *позади* выражен словом *колонна*. Основное различие между контекстами – синтаксическое. В случае *позади колонны* данная синтаксическая валентность реализуется формой родительного падежа существительного *колонна*, т.е. слово *позади* управляет словом *колонна*. Во втором случае та же синтаксическая валентность остается нереализованной – слово *колонна*, выражающее семантический актант 'ориентир' слова *позади*, находится в предтексте и синтаксически от него не зависит.

Как видим, слово *позади* имеет активную синтаксическую валентность, реализуемую падежной формой существительного, и следовательно, обладает свойством (I) предлога. Однако эта синтаксическая валентность реализуется необязательно (причем безотносительно к эллипсису), т.е. это слово не обладает свойством (II) предлога. Тем самым, слово *позади* некорректно относить к разряду предлогов. Слово *позади* естественно считать наречием, имеющим активную синтаксическую валентность, которая реализуется факультативно.

Аналогичным образом устроена достаточно большая группа русских наречий. Ср. *Она шла впереди – Она шла впереди группы; Вокруг костра сидели рабочие – Горел костер, вокруг сидели рабочие*; *На фото она сидит в кресле, он стоит рядом – На фото она сидит в кресле, он стоит рядом с ней*. Все эти наречия требуют *n*-форм местоимений, ср. *впереди него <нее, них>, вокруг него <нее, них>*.

В академической грамматике и лексикографии наречия, подобные *впереди, вокруг, позади*, принято «раздваивать» на две единицы – собственно наречие и производный (наречный) предлог. В контексте с управляемой падежной формой существительного такая единица признается предлогом, а при отсутствии такой формы – наречием. Неэкономность такого подхода очевидна. При нашем подходе все такие единицы считаются наречиями с факультативной синтаксической валентностью. Подчеркнем, что необязательность реализации синтаксической валентности – явление, широко распространенное в языке. Так, у русского глагола необязательна реализация субъектной валентности, однако независимо от того, заполнена эта валентность или нет, перед нами бесспорно один и тот же глагол, ср. *Я люблю ее – Люблю ее, Опять ты играешь в эти компьютерные игры – Опять ты играешь* и т.п. Очевидно, что если у слова (в данном значении) в каком-то контексте не реализована синтаксическая валентность, то это не значит, что в данном контексте оно представлено особой единицей, относящейся к другому грамматическому разряду.

Тем самым, мы считаем, что в русском языке нет наречных предлогов типа *позади, впереди, вокруг,* но есть наречия, которые способны, но, в отличие от предлога, не обязаны управлять падежной формой (или предложно-падежной группой, ср. *рядом с кем/чем-л.*). Такие наречия в [Uryson 2014; 2017] выделяются в особый класс – предлогообразные наречия. (Тем самым, развивается подход, впервые намеченный Д.Н. Овсянико-Куликовским [Ovsyaniko-Kulikovskij 1902] и затем поддержанный Е.Т. Черкасовой [Cherkasova 1967]). При таком подходе упростится (с логической точки зрения) автоматическая разметка текста. Действительно, при академическом подходе для того чтобы правильно разметить в тексте единицы типа *позади*, требуется сначала установить, есть ли у такой единицы управляемая падежная форма: если она есть, то единица признается предлогом, а если нет, то наречием. При нашем подходе подобная единица размечается по словарю, без обращения к ее синтаксическим связям.

В русском языке есть еще один класс управляющих наречий – предикативные наречия, или предикативы; ср. *страшно (Ей страшно), жаль (Ему жаль ее), стыдно (Им стыдно за детей)* и т.п. Таким образом, способность управлять отнюдь не чужда русскому наречию.

Однако в русской академической грамматике считается, что наречие вообще не способно управлять. Последовательное проведение этой точки зрения приводит к весьма неэкономному представлению материала. Во-первых, предикативные наречия выводятся из разряда наречий и выделяются в особый разряд – «категорию состояния», при том что данный разряд отсутствует в списке частей речи и их подклассов (эта непоследовательность в академической грамматике не оговаривается). Во-вторых, неоправданно увеличивается количество значений наречий типа *позади* – каждая такая единица дается в словаре дважды, причем эти «подзначения» семантически тождественны, а различаются единицы лишь выражением семантического актанта.

Вернемся к формальным свойствам предлога. Мы убедились, что формальную особенность (а) первообразного предлога, т.е. употребление после него *n*-форм местоимений, некорректно считать критерием определения предлога (две другие особенности присущи только первообразным предлогам и поэтому не рассматривались). Употребление после *позади*, *впереди* и подобных наречий *n*-форм местоимений свидетельствует лишь о том, что *n*-формы требуются не только предлогом, но и другими группами слов. Заметим, что компаративы, требующие *n*-форм местоимений (*лучше нее, больше него* и т.п.), до сих пор не предлагалось считать производными предлогами.

В заключение отметим, что употребление *n*-форм местоимений обязательно лишь в пределах современной литературной нормы. В массе русских говоров эти формы неизвестны или факультативны [DARL 1996: 166]. Тем самым, свойство (а) предлога не является универсальным даже в пределах русского языка.

2.0. К уточнению списка русских производных предлогов

Академическая грамматика относит к производным, а именно – наречным, предлогам достаточно большую группу наречий, способных управлять; ср. *позади, впереди, вокруг, рядом* и т.п. Выше было показано, что эта группа не обладает свойством (II) предлога, и поэтому ее естественно отнести к классу наречий. Правда, тогда придется признать, что русское наречие способно управлять, ср. *позади → дома*. Однако способностью управлять обладают и т.н. предикативные наречия, ср. *Им ← страшно → за детей, Ей ← холодно* и т.п. Поэтому предлагаемое описание грамматики наречий экономнее, чем раздваивание единиц типа *вокруг* на две (наречие vs. предлог), причем семантически тождественные единицы. Эта проблема

подробно обсуждается в работах [Uryson 2014; 2017], и мы на ей сейчас не останавливаемся. Перейдем к другим производным предлогам.

2.1. *Кроме*. Слово кроме единодушно принято считать предлогом: действительно, оно имеет активную и пассивную синтаксические валентности, причем активная валентность реализуется падежной формой существительного. Однако это необязательно: активная синтаксическая валентность кроме может заполняться словом как, ср. *Нигде кроме как в Моссельпроме; Такого нет ни у кого, кроме как у нас*. Тем самым, кроме не обладает свойством (II) предлога и относить это слово к предлогам некорректно.

2.2. Некоторые единицы, относимые к отглагольным предлогам: *спустя, погодя, не доходя до*. Единица *спустя* ‘через промежуток времени A1’, относимая академической грамматикой к отглагольным предлогам, представлена в контекстах типа *Он встретились спустя год / год спустя*. Эта единица, в отличие от «классического» предлога, может располагаться как в препозиции, так и в постпозиции к управляемому слову, т.е. не удовлетворяет требованию (V). Однако этому требованию не удовлетворяет, например, и слово *ради*, которое тем не менее относят к предлогам. Казалось бы, для признания *спустя* предлогом достаточно расширить список исключений из условия (V), включив в него и данную единицу.

Но *спустя* не обладает гораздо более важным предложным свойством: активная синтаксическая валентность этого слова может реализоваться не только падежной формой существительного, но и наречием *немного*. Ср. *Спустя немного к столику подсели еще литераторы, и скоро образовалось непринужденное и веселое общество* (К. Вагинов); *Спустя немного прибежал ко мне и Петров поздравить меня* (Ф. М. Достоевский). Аналогичные примеры с другим порядком слов: *Она улыбнулась и немного спустя уже сама заговаривала со мной* (И. С. Тургенев); *Мы вошли, поздоровались с ним и разговорились. Немного спустя он предложил кое-что прочесть* (Б. Пастернак).

Наречие *немного* может выступать в качестве вершины группы *немного времени*. Ср. *Немного времени спустя все сидели на лужайке, выпивали и закусывали* (А. Новиков-Прибой); *— Это твоя жена тебя во Франкфурт за покупками посылала? — спросил Санин спустя немного времени* (И. С. Тургенев). В этом случае наречие *немного* управляет существительным *время*, т.е. активная синтаксическая валентность *спустя* реализуется наречием *немного*, но не существительным, как того требует предлог.

Данная синтаксическая валентность слова *спустя* может заполняться и существительным, ср. *год спустя, какое-то время спустя*. Это, однако, не меняет дела: слово *спустя* не обладает свойством (III) предлога. Тем самым, относить *спустя* к предлогам некорректно.

Единица *погодя* ‘через промежуток времени A1’ не обладает сразу двумя основными свойствами предлога (II) и (III). Ее активная синтаксическая валентность реализуется необязательно (нарушение требования (II)). Ср. *Он услышал имя Штерн, потом, погодя, несколько раз Нейман и застыл в бессильной злобе* (Ю. Домбровский); *Топаешь целый день, — заметил он погодя, — и дела будто не делаешь, а устанешь как собака и проголодаешься* (В. Богомолов). При этом активная синтаксическая валентность *погодя* может заполняться наречием, а не падежной формой существительного (нарушение требования (III)). Ср. *Ну, хватит, Верочка… — сказал он, немного погодя… — подите высморкайтесь и глотните чаю* (Дина Рубина*); Стояло ясное, солнечное утро, где-то далеко-далеко, точно на краю земли, надрывался паровозный гудок, и чуть погодя ему отвечали два или три с разных сторон, еще более отдаленные и тонкие* (Ю. Домбровский). Следовательно, считать *погодя* предлогом некорректно.

Единица *не доходя*, трактуемая академической грамматикой как составной предлог *не доходя до*, представлена в контекстах типа *Не доходя до станции есть большой супермаркет*. Эта единица, хотя и управляет падежной формой существительного (ср. *не доходя → до станции*), но не удовлетворяет основным предложным свойством (IV) – *не доходя* имеет «лишний» синтаксический актант. Ср. *Знаешь, немного не доходя до рынка, там мороженщица всегда стоит* (В. Панова); *Там, чуть не доходя до главного здания КГБ, маленький хозяйственный магазин* (В. Войнович). Этот синтаксический актант соответствует семантическому актанту ‘величина расстояния до объекта A1’ предиката *не доходя*. Кроме того, *не доходя* не обладает

основным свойством (II) предлога: активная синтаксическая валентность этой единицы, реализуемая падежной формой, может оставаться незаполненной. Ср. *Да мы же рядом живем, вы в конце переулка, а я не доходя, наискосок* (В. Драгунский).

2.3. Некоторые единицы, относимые к отыменным предлогам: *за исключением, по причине, в области, по поводу, по случаю*. У всех этих единиц есть активная синтаксическая валентность, реализуемая падежной формой существительного. Ср. *за исключением → Петрова, по причине → дождей, в области → математики, по поводу → юбилея, по случаю → выходного дня*. Однако активная синтаксическая валентность этих единиц реализуется необязательно, и в этих случаях существительное в составе единицы имеет определение, т.е. такое синтаксическое зависимое, которое не заполняет его синтаксическую валентность. Ср. *Приглашали всех за единственным исключением – никогда не звали Иванова, Елку устраивали всегда, за редкими исключениями; Рейс отменили по причине сильных дождей – По этой причине рейс отменили; достижения в области математики – достижения в этой области; По поводу юбилея устроили торжества – По этому поводу устроили торжества; По случаю выходного занятия отменили – По такому поводу занятия отменят*. Следовательно, данные единицы не имеют сразу двух свойств предлога – (III) и (IV). В соответствии с нашим подходом, ни одна из них не является предлогом.

Очередная задача заключается в определении грамматического статуса рассмотренных единиц.

References

[1] Cherkasova E.T. (1967) Shift of content words to prepositions [Perekhod polnoznachnykh slov v predlogi], Moscow, Nauka.

[2] DARL (1989) Dialectological atlas of Russian. The center of the European part of the USSR [Dialektologicheskij atlas russkogo jazyka], N. 2: Morhpology. Bromlei S. V. (ed.), Moscow, Nauka.

[3] Es'kova N. A. (1996) Primitive and non-primitive prepositions: Formal aspect [Pervoobraznye i nepervoobraznye predlogi: Formalnyj aspekt], Russian, Slavonic, and Indo-European Studies [Rusistika. Slavistika. Indoevropeistika]. Nikolaeva T. M. (ed.). Moscow, Indrik, pp. 458–464.

[4] LES (1990) The linguistic encyclopaedic dictionary [Lingvisticheskij slovar' russkogo jazyka], Moscow, Sovetskaja entsiklopedija.

[5] Russian grammar (1980) [Russkaya grammatika], Shvedova N. Yu. (ed.), Moscow, Nauka.

[6] Ovsyaniko-Kulikovskij D.N. (1902) Syntax of Russian [Sintaksis russkogo yazyka],]. St. Petersburg, D. E. Zhukovsky.

[7] OD (1983) Ortographical dictionary of Russian. Pronunciation. Stress. Grammatical forms [Orfograficheskii slovar' russkogo yazyka. Proiznoshenie. Udarenie. Grammaticheskie formy], Moscow, Russkij jazyk.

[8] Sichinava D.V. (2018) Preposition [Предлог], Moscow, 2018, http://rusgram.ru/

[9] Testelets Ya. G. (2001) Introduction to General Syntax [Введение в общий синтаксис], Moscow, RGGU.

[10] Uryson E.V. (2014) On derivative prepositions: Adverbial prepositions [О производных предлогах: наречные предлоги], Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialog 2014" [Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог»], 13(20), v. 1, pp. 695–707.

[11] Uryson E.V. (2017) Adverbial prepositions as a subclass of adverbs (Predlog ili narechije: chasterechnyj status narechnykh predlogov], Questions of Linguistics [Voprosy jazykoznanija], 5, pp. 36–55.

[12] Vsevolodova M. V. (2010) Grammatical properties of the prepositional units in Russian: Typology, structure, syntagmatics, and syntactical modifications [Grammaticheskije aspekty russkikh predlozhnykh edinits: tipologija, struktura, sintagmatika i sintaksicheskije modifikatsii], Questions of Linguistics [Voprosy Jazykoznanija], 4, pp. 3–26.

Estimating cognitive text complexity with aggregation of quantile-based models

Arseniy Veselov
Lomonosov Moscow State University
Moscow, Russia
arseniy.veselov@yandex.ru

Maksim Eremeev
New York University
New York, USA
eremeev@nyu.edu

Konstantin Vorontsov
Moscow Institute of Physics and Technology
Moscow, Russia
vokov@forecsys.ru

Abstract

In this paper, we introduce a novel approach to estimating the cognitive complexity of a text at different levels of language: phonetic, morphemic, lexical, and syntactic. The proposed method detects tokens with an abnormal frequency of complexity scores. The frequencies are taken from the empirical distributions calculated over the reference corpus of texts. We use the Russian Wikipedia for this purpose. Ensemble models are combined from individual models from different language levels. We created datasets of pairs of text fragments taken from social studies textbooks of different grades to train the ensembles. Empirical evidence shows that the proposed approach outperforms existing methods, such as readability indices, in estimating text complexity in terms of accuracy. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

Keywords: cognitive complexity of texts, language levels, ensemble learning
DOI: 10.28995/2075-7182-2023-22-525-538

Оценивание когнитивной сложности текста с помощью агрегирования моделей, основанных на квантилях

Веселов А.С.
Московский государственный
университет им. М. В. Ломоносова
Москва, Россия
arseniy.veselov@yandex.ru

Еремеев М.А.
Нью-Йоркский
университет
Нью-Йорк, США
eremeev@nyu.edu

Воронцов К.В.
Московский физико-технический
институт
Москва, Россия
vokov@forecsys.ru

Аннотация

В данной работе описывается подход к оцениванию когнитивной сложности текста на разных уровнях языка: на фонетическом, морфемном, лексическом и синтаксическом. В его основе лежит определение токенов с аномальной частотой их сложностей. Частоты определяются по эмпирическим распределениям, построенным на основе референтного корпуса текстов, в качестве которого используется русскоязычная Википедия. Из отдельных моделей с разных уровней языка создаются агрегированные модели. Для их обучения мы создали выборки пар фрагментов текстов, взятых из учебников по обществознанию разных учебных классов. Проведённые в работе эксперименты показывают у предлагаемого подхода более высокую точность ранжирования текстов по сложности в сравнении с индексами удобочитаемости. Целью проведения данного исследования является создание одного из важных компонентов системы рекомендации научно-образовательного контента.

Ключевые слова: когнитивная сложность текстов, уровни языка, ансамблевое обучение

1 Introduction

Many readability indices have been developed for the task of estimating the complexity of the text. Most of them are a linear combination of some trivial statistical parameters of the text based on the number of letters, syllables, words, and sentences. In this paper, we continue the research and improvement of the generalised quantile-based approach to the estimation of the cognitive complexity of the text at different levels of the language (phonetic, morphemic, lexical, and syntactic). The idea of such an approach was first presented by Eremeev M.A. and Vorontsov K.V. in (Eremeev and Vorontsov, 2019). It is based on the detection of tokens with an abnormal frequency of their complexity scores. We use the reference corpus of texts, which is the Russian-language Wikipedia, to construct the empirical distributions for this purpose. This paper is devoted to the study of the aggregation of individual quantile-based models in order to take information from different levels of the language into account, and this is its novelty. We train aggregated models on datasets of pairs of text fragments, which we created on the basis of social studies textbooks of different educational grades. In this paper, we conduct experiments to compare the accuracy of our models with adapted readability indices, including the comparison of accuracy over each pair of educational grades. The analysis of the contribution of individual components to the aggregated model (ablation study) and the analysis of the dependence of the ranking accuracy on the average length of a text fragment in a dataset are also carried out. The experiments conducted in the paper demonstrate that the proposed approach has a higher accuracy of ranking texts in terms of cognitive text complexity compared to readability indices. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

2 Readability indices review

Historically linguists use readability indices for estimating text complexity of the educational literature. Many of them were initially developed for the US education system and were therefore adapted for the English language.

The automated readability index (ARI) was developed by R.J. Senter and E.A. Smith in 1967 (Senter and Smith, 1967). It approximates a representation of the US grade level required to understand the analysed text. For a document d written in English ARI has the following calculation formula:

$$\mathrm{ARI}(d) = 4.71 \times \frac{C}{W} + 0.5 \times \frac{W}{S} - 21.43,$$

where C is the number of letters and digits, W is the number of words, and S is the number of sentences in the text of the document d.

Läsbarhetsindex (LIX) was developed by Swedish scientist Carl-Hugo Björnsson in 1968 (Björnsson, 1968). Index value monotonically increases with respect to text complexity. LIX does not take into account the language in which the text is written and is calculated as follows:

$$\mathrm{LIX}(d) = \frac{A}{B} + 100 \times \frac{C}{A},$$

where A is the number of letters, B is the number of sentences, and C is the number of words longer than 6 letters in the text of the document d.

In 1969 G. Harry McLaughlin developed the Simple Measure of Gobbledygook (SMOG) (McLaughlin, 1969). This readability index produces an approximate number of years of study needed to comprehend the text. SMOG is calculated for the document d written in English with the following formula:

$$\mathrm{SMOG}(d) = 1.0430\sqrt{A \times \frac{30}{B}} + 3.1291,$$

where A denotes the number of polysyllabic words (3 and more syllables in English), and B is the number of sentences.

Coleman–Liau index (CLI), developed in 1975 by Meri Coleman and T.L. Liau (Coleman and Liau, 1975), approximates a representation of the US grade level necessary to understand the given text. For the document d written in English CLI has the following calculation formula:

$$\mathrm{CLI}(d) = 0.0588 \times L - 0.296 \times S - 15.8,$$

where L denotes the average number of letters per 100 words, and S refers to the average number of sentences per 100 words.

In 1948 Rudolf Flesch developed the most popular measure of text complexity — the Flesch reading-ease score (FRES) (Flesch, 1948). The index value monotonically declines with respect to text complexity. FRES is calculated for the document d written in English as follows:

$$\mathrm{FRES}(d) = 206.835 - 1.015 \times \mathrm{ASL} - 84.6 \times \mathrm{ASW},$$

where ASL is the average sentence length in words, and ASW is the average number of syllables per word.

Flesch–Kincaid grade level (FKGL) was developed by J. Peter Kincaid in 1975 (Kincaid et al., 1975). This readability index approximates a representation of the US grade level. FKGL has the following formula for calculation for the document d written in English:

$$\mathrm{FKGL}(d) = 0.39 \times \mathrm{ASL} + 11.8 \times \mathrm{ASW} - 15.59.$$

The Estonian linguist Juhan Tuldava proposed in 1975 his own readability index (Tuldava, 1975), which we refer to in our article as the Tuldava index (TI). TI does not take the language of the text into account and is calculated as follows:

$$\mathrm{TI}(d) = \mathrm{ASW} \times lg(\mathrm{ASL}).$$

In this paper, we estimate the complexity of Russian texts. Therefore, we use adapted versions of indices for comparison with the proposed quantile-based approach.

Irina Oborneva made a significant contribution to the development of the readability formulae for texts in Russian by adapting the FRES and FKGL indices in 2005 (Oborneva, 2005):

$$\mathrm{FRES}_{\mathrm{ru}}(d) = 206.835 - 1.3 \times \mathrm{ASL} - 60.1 \times \mathrm{ASW},$$

$$\mathrm{FKGL}_{\mathrm{ru}}(d) = 0.5 \times \mathrm{ASL} + 8.4 \times \mathrm{ASW} - 15.59.$$

Later, the results of the adaptation of the readability formulae for automated analysis of texts in Russian were presented by Ivan Begtin in 2014 (Begtin, 2014). These implementations were collected in the Python library ruTS by Sergey Shkarin in 2021 (Shkarin, 2021). We utilise this library to reproduce baseline results for this paper. We have extended it by adding the Tuldava index and correcting the wrong coefficients in the Coleman–Liau index. See the formulae for ARI_{ru}, $SMOG_{ru}$, and CLI_{ru} readability indices adapted for the Russian language below (the variables that are not explained below are the same as for the formulae for English):

$$\mathrm{ARI}_{\mathrm{ru}}(d) = 6.26 \times \frac{C}{W} + 0.2805 \times \frac{W}{S} - 31.04.$$

$$\mathrm{SMOG}_{\mathrm{ru}}(d) = 1.1\sqrt{A \times \frac{64.6}{B}} + 0.05,$$

where A denotes the number of polysyllabic words (4 and more syllables in Russian).

$$\mathrm{CLI}_{\mathrm{ru}}(d) = 0.055 \times L - 0.35 \times S - 20.33.$$

Text complexity estimates have many applications. For example, Arina Dmitrieva describes the methods of analysing legal documents in Russian based on readability indices (Dmitrieva, 2017). The FKGL readability index was developed in order to compile the texts of instructions for the use of weapons or technical means, and the SMOG index was used to study the text complexity of instructions for medicines and preparations. Many indices are used to estimate the comprehensibility of textbooks offered to students of different ages. The use of text complexity estimation can be helpful for predicting the time spent processing regulations, documents, and educational literature.

3 Generalised text complexity model

Let d be an arbitrary document of length n consisting of tokens $x_1, \ldots, x_n$ from a fixed finite alphabet A_h, where h denotes the level of the language: phonetic, morphemic, lexical or syntactic. In this paper, we consider letters, syllables, words, or sentences (or structures describing a part of speech and the syntactic function of words) as tokens, depending on the level of the language. Suppose that every token x_i of the document d has its own processing complexity c_i caused by its context or by its internal structure. Also assume that each token $a \in A_h$ has its usual processing complexity, which is a result of the language evolution within a historical and cultural environment. If the current processing complexity of the token $x_i = a$ in the analysed text turns out to be abnormally high compared to the usual processing complexity of token a, then we will assume that the token x_i carries an excessive difficulty of perception. The information about usual complexity of tokens can be retrieved from a *reference collection* denoted by K, which is a large union of texts of medium complexity. In order to determine if the token $x_i \in d$, $x_i = a$ is abnormally complex we need to construct an empirical distribution of complexity scores $\hat{c}_j$ of every token $\hat{x}_j \in K$ such that $\hat{x}_j = a$. The token x_i is considered as abnormally complex if its complexity score is greater than the γ-quantile $C_\gamma(x_i)$ of the constructed distribution for token (see Figure 1).

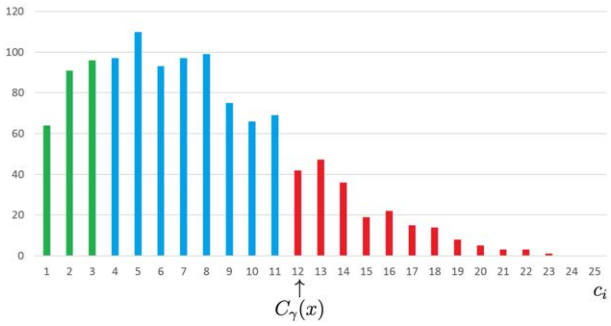


Figure 1: Histogram for empirical distribution of complexity scores and its γ-quantile

In Figure 1 the red zone corresponds to an abnormally high complexity. The green zone corresponds to a low complexity. The blue zone indicates the usual complexity of the token.

We shall call the nonlinear sum of weights w_i of tokens with abnormal complexities the *document complexity score* and denote it by $S(d)$.

$$S(d) = \sum_{i=1}^{n} w_i^p \left[c_i > C_\gamma\left(x_i\right)\right], \tag{1}$$

where $[\ ]$ is the Iverson bracket (i.e. $[\text{true}] = 1$, $[\text{false}] = 0$), p is a positive integer.

The weight w_i is a non-negative value that does not decrease with increasing complexity c_i. Complexity c_i is defined up to an arbitrary increasing function.

Table 1 shows several examples of possible weights.

w_i	Meaning of w_i
1	number of complex tokens
$1/n \times 100\%$	percentage of complex tokens
c_i	total complexity
c_i/n	mean complexity
$c_i - C_\gamma\left(x_i\right)$	excessive complexity
$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	mean excessive complexity

Table 1: Examples of weights w_i

4 Token complexity functions

4.1 Distance-based complexity function

Let r_i be a distance from the previous occurrence of the token x_i to its current occurrence in the text:

$$\dots \boxed{x_{i-r_i} = a} \quad \underbrace{x_{i-r_i+1} \quad x_{i-r_i+2} \quad \dots \quad x_{i-2} \quad x_{i-1} \quad \boxed{x_i = a}}_{r_i} \dots$$

$$r_i = \min_{1 \leqslant j < i} \{i - j \mid x_i = x_j\}.$$

In the first occurrence of the token a in the text, at the position i, the distance r_i is undefined. In that case r_i is redefined such that the sum of all distances r_j for this token $x_j = a$ equals to the document length n.

To obtain a frequency model of complexity as a special case of the generalised model, the parameters c_i are defined as some decreasing function of r_i, for example:

$$c_i = -r_i \tag{2}$$

4.2 Counter-based complexity function

In the counter-based approach, as in the special case of the generalised approach, it is assumed that the alphabet A_h consists of a single token $A_h = \{a\}$, i.e. we distinguish not the tokens themselves, but only their complexity. The complexity of tokens is determined by their linguistic properties, and each token has exactly one possible complexity value. Thereby, just one empirical distribution of token complexities is constructed over all tokens from the reference collection. In that case, $C_\gamma(x_i) = C_\gamma$.

5 Considered models

In this section, we describe individual models at different levels of language in terms introduced when considering a generalised model above, i.e. by specifying the alphabets of tokens and complexity functions. The available means of morphological, lexical, and syntactic analysis can be used to form alphabets of tokens and characteristics of their complexity.

5.1 Phonetic level

We consider individual letters as tokens here. For this type of the models we use the distance-based approach. The name of the model implemented in that way is $letter\_dist\_model$.

5.2 Morphemic level

There are two possible ways to form tokens: either take the original syllables, or rearrange the letters in them in alphabetical order so that the order of the letters is not taken into account. Therefore, we consider two distance-based models, which we refer to as $syllab\_dist\_model$ and $syllabsort\_dist\_model$.

5.3 Lexical level

The tokens here are individual words. For models at this level (except the $lexical\_len\_model$) we consider different forms of one word to be equal and use the lemmatization of words as a preprocessing.

Distance-based model at this level is called $lexical\_dist\_model$.

Word length counter-based model considers the length of the word as its complexity score. To implement such a model, we construct an empirical distribution of lengths of all words in the reference collection. We refer to this model as $lexical\_len\_model$.

Counter-based model at lexical level is based on the assumption that the rarer a word is encountered in the reference corpus, the more specific and difficult it is. In the experiments, the following complexity function is used:

$$c_i = -\operatorname{count}(x_i), \tag{3}$$

where $\operatorname{count}(x_i)$ is the number of token x_i occurences in the reference collection. We refer to this model as $lexical\_cnt\_model$.

5.4 Syntactic level

The tokens here are sentences or structures describing the part of speech and the syntactic function of words in the sentence. In this paper, we use the UDPipe library to divide the text into sentences and extract the syntactic dependencies and parts of speech (Straka and Straková, 2017).

Counter-based model at this level uses the maximum length of the syntactic dependency in the sentence as a complexity score of this sentence. We refer to this model as $syntax\_len\_model$.

Distance-based model considers a sentence as a set structures describing a part of speech and the syntactic function of words in the sentence. Each such structure corresponds to one word. The word itself is ignored, but information about its part of speech and syntactic role in the sentence is considered. We refer to this model as $syntaxpos\_dist\_model$.

6 Experiments

6.1 Reference collection and datasets

We use the Russian Wikipedia (1.5 million articles) as a reference collection for our experiments. The ruwiki-latest-pages-articles.xml.bz2 archive was processed by the WikiExtractor parser. After the additional preprocessing it was translated into a format where each article corresponds to its own TXT document.

As a dataset we use the sets of social studies textbooks, prepared in (Solovyev et al., 2018): textbooks by L.N. Bogolyubov for 6, 7, 8, 9, 10, 10+, 11+ grades («+» denotes a version with in-depth study) and textbooks by A.F. Nikitin for 5, 6, 7, 8, 9, 10, 11 grades. In this dataset, each document contains randomly shuffled sentences from the textbook. In order to create a dataset for the training and validation of models, we first combined the texts of the textbooks intended for the same grade and then cut them into pieces of similar length consisting of whole sentences. Afterwards, the fragments of texts of different grades were combined into pairs, where a piece of text from a textbook of a higher grade comes second: $D = \{(d,\ d') \mid d' \text{ more complex than } d)\}$. The complexity of the textbook is determined by its grade, which should be a fairly reliable characteristic to estimate the cognitive complexity of the text, since textbooks are created in accordance with educational standards.

Eight datasets with different number of pairs were prepared. They are available at this link. For this purpose, the length of one text fragment varied (see Table 2). Each dataset consists of all possible pairs in such a way that each text piece of one grade is compared with each text piece of each other grade.

Dataset name	Number of pairs of text fragments	Average number of symbols in one text fragment
D1	1027	94 100
D2	2532	59 850
D3	5001	42 650
D4	10 041	30 100
D5	45 058	14 200
D6	250 152	6000
D7	1 008 881	2950
D8	5 400 136	1250

Table 2: Datasets

We create such a number of datasets in order to investigate the dependence of ranking accuracy on the average length of a text fragment in the last experiment. In other experiments, only datasets D1, D2, D3, D4 are used, because their average lengths of a text fragment are large enough to provide as much information as possible to models and readability indices to estimate the complexity of text fragments. Moreover, we will focus more on D4 in further experiments since there are quite a lot of pairs of text fragments in this dataset, so that we can get more different possible values of the quality criterion as well as train aggregated models on a larger number of pairs.

6.2 Quality criterion

As a quality criterion we consider accuracy, i.e. the ratio of the number of correctly estimated pairs of text pieces to the total number of pairs:

$$\text{accuracy(S)} = \frac{\sum_{(d,d')\in D}\left[S(d') > S\left(d\right)\right]}{|D|}, \tag{4}$$

where S denotes a model (or readability index), which produces document complexity score.

6.3 Separate models

In experiments (see Table 3), the best parameters $(p,\ w_i,\ \gamma)$ (see the formula 1) of the models that maximized the quality criterion are selected on D3 and D4 datasets (since they have more pairs, and this means that it is potentially possible to get more different values of accuracy) or on similar-sized datasets based on a series of textbooks by only one of the authors.

The weights w_i are searched over the grid $\{1,\ c_i,\ c_i/n,\ c_i - C_\gamma\left(x_i\right),\ \left(c_i - C_\gamma\left(x_i\right)\right)/n\}$. The parameter γ is searched over the following grid with a step 0.05: $\{0.01\}\cup[0.05,\ 0.1,\ 0.15,\ \ldots,\ 0.9,\ 0.95]\cup\{0.99\}$. The parameter p is searched over the grid $[1,\ 2,\ 3,\ 4]$. In addition to the distance-based models with the complexity function (2), the experiments also estimated the quality of the models, which are based on the opposite hypothesis that the rarer the same tokens are found in the analysed text, the more difficult they are to comprehend, with a complexity function $C_i = r_i$. But the quality of such models was in the range of 30-60%, so they are not presented further.

№	Model name	Hyperparameters			Accuracy on dataset, %			
		w_i	p	γ	D1	D2	D3	D4
1	$\boldsymbol{letter\_dist\_0}$	c_i/n	1	0.10	79.45	77.49	77.70	76.36
2	$\boldsymbol{letter\_dist\_1}$	c_i/n	1	0.85	81.60	77.13	76.42	75.74
3	$letter\_dist\_2$	c_i	1	0.05	80.92	72.08	80.66	67.29
4	$syllab\_dist\_0$	c_i/n	1	0.01	63.49	76.46	78.08	78.23
5	$syllab\_dist\_1$	c_i	1	0.65	73.61	63.19	72.27	59.57
6	$syllabsort\_dist\_0$	c_i	1	0.05	79.07	67.65	82.04	76.25
7	$lexical\_dist\_0$	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.01	75.17	76.11	84.88	82.01
8	$lexical\_dist\_1$	c_i	1	0.99	82.38	76.94	85.64	76.17
9	$\boldsymbol{lexical\_len\_0}$	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.55	92.02	90.72	89.58	89.57
10	$\boldsymbol{lexical\_len\_1}$	c_i/n	1	0.85	88.41	87.52	87.00	86.86
11	$\boldsymbol{lexical\_len\_2}$	c_i/n	1	0.45	92.11	90.84	91.10	91.08
12	$\boldsymbol{lexical\_len\_3}$	$\left(c_i - C_\gamma\left(x_i\right)\right)/n$	1	0.30	**93.48**	**92.06**	91.70	**91.28**
13	$\boldsymbol{lexical\_len\_4}$	c_i/n	2	0.65	90.36	89.38	**92.76**	87.72
14	$lexical\_cnt\_0$	c_i/n	2	0.35	70.79	61.77	72.02	63.10
15	$lexical\_cnt\_1$	c_i/n	2	0.15	84.32	83.10	87.16	80.78
16	$lexical\_cnt\_2$	c_i	1	0.85	63.68	57.70	63.25	57.50
17	$lexical\_cnt\_3$	c_i	1	0.45	73.81	67.69	71.25	60.92
18	$\boldsymbol{syntax\_len\_0}$	c_i/n	2	0.01	88.61	83.77	86.14	83.95
19	$\boldsymbol{syntax\_len\_1}$	c_i/n	2	0.35	88.51	83.81	85.80	83.89
20	$syntaxpos\_dist\_0$	c_i	1	0.45	81.60	81.67	85.58	78.99
21	$syntaxpos\_dist\_1$	c_i	1	0.35	83.93	82.39	86.30	80.84

Table 3: Selected parameters and accuracy of individual models on D3 and D4. Bold lines separate different types of models. Models highlighted in bold show the greatest contribution to the ensemble in ablation studies

As a result, 21 models were selected to be used in aggregation experiments. The word length counter-based lexical models show the best quality amongst the individual models, surpassing all the readability

indices, whose accuracy on the same datasets is shown in the table 4. $\mathrm{FKGL_{ru}}$ and $\mathrm{FRES_{ru}}$ demonstrate the best accuracy amongst the readability indices. If we focus only on the D4 dataset, then the best index is $\mathrm{FRES_{ru}}$.

Index	Accuracy on dataset, %			
	D1	D2	D3	D4
$\mathrm{FKGL_{ru}}$	**91.04**	**90.00**	89.94	89.49
$\mathrm{FRES_{ru}}$	90.75	**90.00**	**90.30**	**90.50**
$\mathrm{CLI_{ru}}$	89.97	89.26	89.76	89.09
$\mathrm{SMOG_{ru}}$	90.26	88.63	88.24	87.80
$\mathrm{ARI_{ru}}$	90.36	89.69	90.14	89.64
LIX	90.65	89.22	89.44	88.79
TI	90.94	89.97	89.92	89.55

Table 4: Accuracy of readability indices on D1, D2, D3, and D4

6.4 Ensemble models

From the selected separate models (Table 3) the ensemble models are constructed. Due to the small size of the datasets, linear regression with non-negative weights is used for the ensembling.

$$S(d, \boldsymbol{\alpha}) = \sum_{k=1}^{K} \alpha_k S_k(d),\ \alpha_k \geqslant 0, \tag{5}$$

where vector $\boldsymbol{\alpha}$ is a solution of the following optimization problem:

$$\sum_{(d,\, d') \in D} \mathcal{L}(S(d', \boldsymbol{\alpha}) - S(d, \boldsymbol{\alpha})) + \lambda\, \mathrm{Reg}(\boldsymbol{\alpha}) \to \min_{\boldsymbol{\alpha}}, \tag{6}$$

where $\mathcal{L}(M)$ is a non-increasing function of margin M, and Reg is a regularizer. The ensemble models are trained on 80% of the dataset and validated on the remaining 20%.

We compare the ensemble models with and without regularization in experiments. For this purpose, L1, L2, or elastic net regularization with a mixing hyperparameter equal to 0.5 are used. The hyperparameter λ is optimized over the grid $[10^{-4}, 10^{-3}, 10^{-2}, 0.1, 1]$. The following loss functions $\mathcal{L}$ of margin M are used:

$\mathcal{L}_1(M) = (1 - M).clip(min = 0)$, $\quad \mathcal{L}_2(M) = |1 - M|$, $\quad \mathcal{L}_3(M) = (1 - M^2)$,
$\mathcal{L}_4(M) = \log(1 + e^{-M})$, $\quad \mathcal{L}_5(M) = \frac{1}{1+e^M}$, $\quad \mathcal{L}_6(M) = e^{-M}$.

Tables 5, 6 show for each loss function the ensembles of 21 separate models from Table 3) of the best validation accuracy on the datasets D4, D3, respectively.

The loss function $\mathcal{L}_6(M)$ had an overflow problem, so its results are not shown in Tables 5, 6. The following functions proved to be bad for our problem, thus their results are not presented in this paper: $\mathcal{L}_7(M) = -|M|,\ \mathcal{L}_8(M) = -M^2,\ \mathcal{L}_9(M) = 1 - M,\ \mathcal{L}_{10}(M) = (-M)^3$.

№	Loss function	Reg	λ	Acc. on D4 [val.], %
1	$\mathcal{L}_1$	L2	10^{-4}	**92.78**
2	$\mathcal{L}_2$	L1	10^{-2}	91.24
3	$\mathcal{L}_3$	L1	10^{-3}	92.14
4	$\mathcal{L}_4$	L1	10^{-3}	88.00
5	$\mathcal{L}_5$	L1	1	82.73

Table 5: Validation accuracy of ensembles of 21 separate models for each loss function on D4

№	Loss function	Reg	λ	Acc. on D3 [val.], %
1	$\mathcal{L}_1$	L2	10^{-3}	93.61
2	$\mathcal{L}_2$	L1	10^{-2}	93.31
3	$\mathcal{L}_3$	L2	10^{-3}	**94.51**
4	$\mathcal{L}_4$	L1	0.1	91.11
5	$\mathcal{L}_5$	L1	1	90.81

Table 6: Validation accuracy of ensembles of 21 separate models for each loss function on D3

The experiments have shown the loss functions $\mathcal{L}_1(M)$ and $\mathcal{L}_2(M)$ to consistently be of the highest quality, i.e. they are less sensitive to the selection of hyperparameters.

The accuracy of the readability indices on the same validation parts of the datasets D3, D4 is presented in Table 7.

Index	Acc. on D3, %	Acc. on D4, %
$\mathrm{FKGL_{ru}}$	89.71	88.40
$\mathrm{FRES_{ru}}$	**89.81**	**89.90**
$\mathrm{CLI_{ru}}$	89.11	87.90
$\mathrm{SMOG_{ru}}$	87.31	86.71
$\mathrm{ARI_{ru}}$	89.31	88.75
LIX	89.01	87.76
TI	**89.81**	88.60

Table 7: Accuracy of readability indices on validation part of D3 and D4

6.5 Accuracy over grade pairs

The accuracy of the best ensemble of 21 separate models (first in Table 5) is examined in more detail in the following section. Table 8 shows the values of the quality criterion (4) on every pair of grades separately.

Acc.	6	7	8	9	10	10+	11	11+
5	1	1	1	1	1	1	1	1
6	—	0.95	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.955	0.97	1	1	1
9	—	—	—	—	0.636	0.953	0.935	1
10	—	—	—	—	—	0.705	0.736	0.98
10+	—	—	—	—	—	—	0.591	0.984
11	—	—	—	—	—	—	—	0.98

Table 8: Validation accuracies of ensemble of 21 separate models on D4 over grade pairs

Table 8 demonstrates that the ensemble model accurately ranks by complexity the text pieces from grades that are more than one—two years apart. It is also noticeable that the lower the grades of both text pieces in a pair, the easier it is for the model to arrange them correctly. That looks logical, since the increase in the complexity of texts of middle school textbooks should be more dramatic than that of high school textbooks.

Table 9 shows the results for the $FRES_{ru}$ readability index, which demonstrated the best accuracy among other indices.

Acc.	6	7	8	9	10	10+	11	11+
5	1	1	1	1	1	1	1	1
6	—	0.8	1	1	1	1	1	1
7	—	—	0.975	1	1	1	1	1
8	—	—	—	0.736	0.993	1	1	1
9	—	—	—	—	0.882	0.915	0.871	0.991
10	—	—	—	—	—	0.524	0.491	0.967
10+	—	—	—	—	—	—	0.341	0.992
11	—	—	—	—	—	—	—	1

Table 9: Accuracy of $FRES_{ru}$ on validation part of D4 over grade pairs

6.6 Ablation study

In this experiment, we reduce the number of individual models in the ensemble model so as not to degrade, but even to improve the quality.

For this purpose, we examine the vector $\boldsymbol{\alpha}$, computed as a result of training ensemble of 21 separate models (first in Table 5). We sort its components in descending order: the weights corresponding to the individual models that make the greatest contribution to estimating the text complexity are the first (see Figure 2).

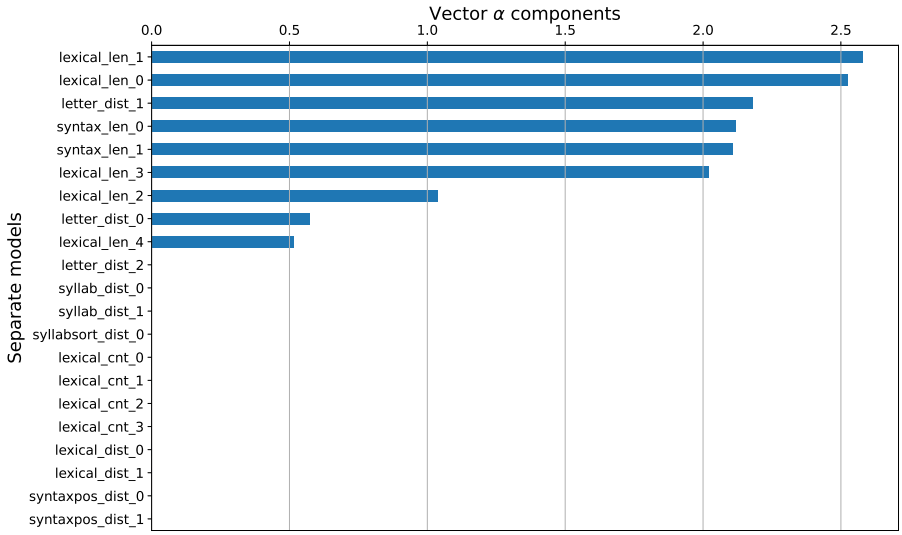


Figure 2: Importance of separate models

Further, a comparison of the validation accuracy on the dataset D4 of different ensembles with one removed block of separate models of one type has shown that deleting the block with word length counter-based lexical models or counter-based syntactic models leads to a significant loss of quality in all ensembles with the loss function $\mathcal{L}_1$ and regularization. We examine ensembles with the loss function $\mathcal{L}_1$ and regularization because this combination proved to be the best. Deleting the distance-based phonetic models block leads to a drop in accuracy on most of these ensembles. Deleting the distance-based syntactic models, counter-based lexical models, distance-based lexical models or distance-based morphemic models block almost does not lead to significant quality losses, and in some cases even increases it. Figure 3 shows how the accuracy of the best ensemble changes when one of the blocks is removed.

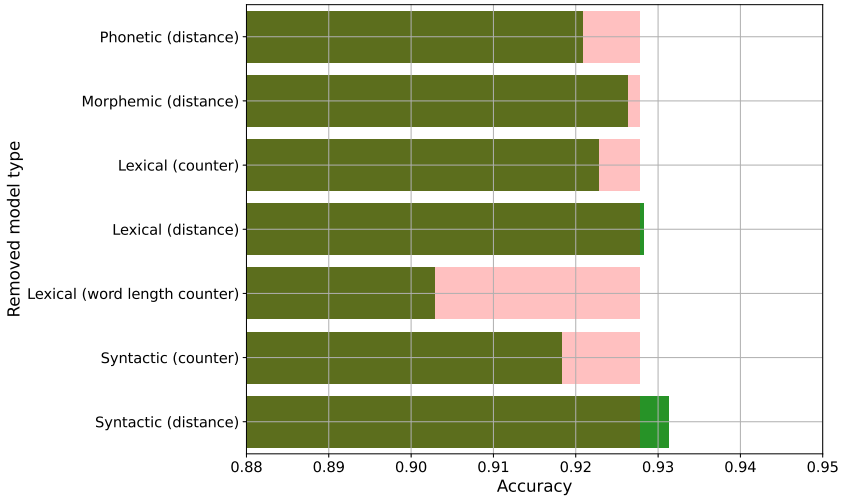


Figure 3: The change in validation accuracy on D4 when removing the block of models of one type: pink shows a decline in accuracy with respect to an ensemble of 21 models; bright green, respectively, shows an improvement

As a next step, the comparison of ensembles of different sets of blocks without separate models of the least importance (Figure 2) is carried out. As a result, the following ensemble model of nine separate models proved to be the best:

- Distance-based phonetic models: $letter\_dist\_0$, $letter\_dist\_1$;
- Word length counter-based lexical models: $lexical\_len\_0$, $lexical\_len\_1$, $lexical\_len\_2$, $lexical\_len\_3$, $lexical\_len\_4$;
- Counter-based syntactic models: $syntax\_len\_0$, $syntax\_len\_1$.

Tables 10, 11 show for each loss function ensembles of nine separate models of the best validation accuracy on the datasets D4, D3, respectively.

№	Loss function	Reg	λ	Acc. on D4 [val.], %
1	$\mathcal{L}_1$	elastic net	10^{-4}	**93.48**
2	$\mathcal{L}_2$	L2	10^{-2}	92.33
3	$\mathcal{L}_3$	L2	0.1	92.33
4	$\mathcal{L}_4$	—	0	93.23
5	$\mathcal{L}_5$	—	0	93.33
6	$\mathcal{L}_6$	L1	10^{-3}	93.23

Table 10: Validation accuracy of ensembles of 9 separate models for each loss function on D4

№	Loss function	Reg	λ	Acc. on D3 [val.], %
1	$\mathcal{L}_1$	—	0	94.91
2	$\mathcal{L}_2$	elastic net	10^{-2}	94.51
3	$\mathcal{L}_3$	—	0	**95.60**
4	$\mathcal{L}_4$	—	0	93.51
5	$\mathcal{L}_5$	L1	10^{-4}	94.61
6	$\mathcal{L}_6$	L2	10^{-4}	94.91

Table 11: Validation accuracy of ensembles of 9 separate models for each loss function on D3

The experiments have shown that of all the loss functions $\mathcal{L}_1(M)$, $\mathcal{L}_2(M)$, $\mathcal{L}_3(M)$ and $\mathcal{L}_6(M)$ consistently demonstrate a high accuracy with different values of hyperparameters. As for $\mathcal{L}_4(M)$, $\mathcal{L}_5(M)$, it is better not to use regularization at all, since with it the quality drops quickly. It is also noticeable that with a good set of separate models for ensembling, one can get an acceptable quality with almost any loss function.

Thus, the experiments show that using the loss function $\mathcal{L}_1(M)$ and any weak regularization with hyperparameter $\lambda = 10^{-4} \dots 10^{-3}$ (or without regularization at all) is the best option.

6.7 Dependence of accuracy on the text fragment average length

In this experiment, the analysis of the dependence of the ranking accuracy on the average length of a text fragment in a dataset is carried out. For this purpose, all built datasets based on textbooks are used (see Table 2).

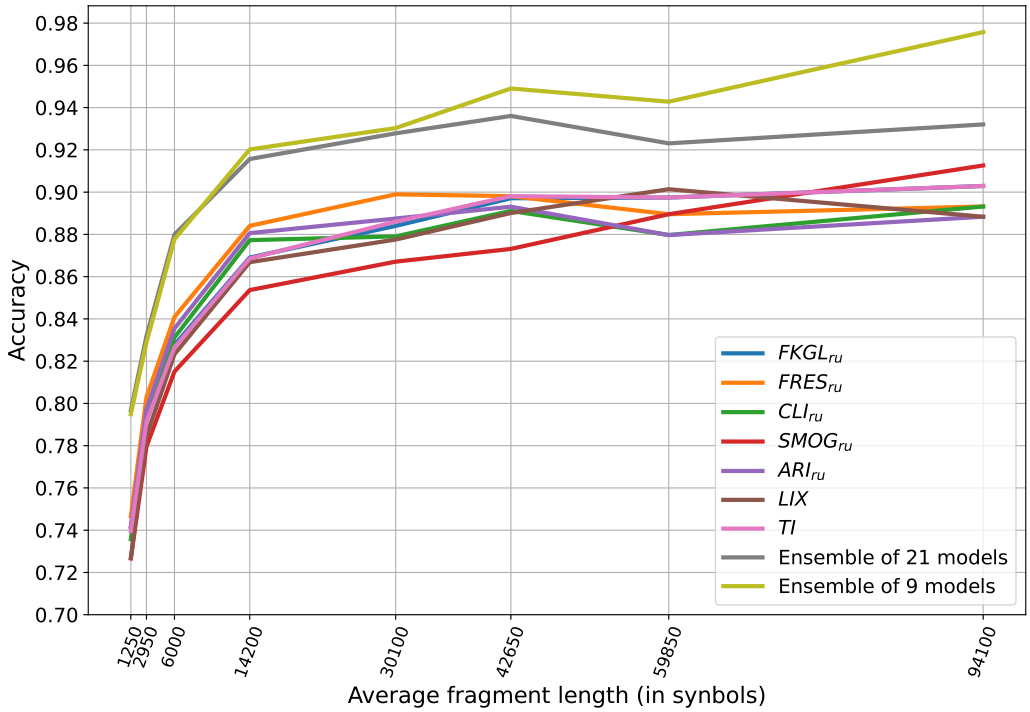


Figure 4:

Figure 4 demonstrates that the accuracy begins to decrease as the length of the text fragment decreases, both for models and for readability indices. A particularly sharp drop is noticeable, starting with a length of 14200 characters or less. While with lengths of more than 14200 symbols, many indices and models have a plateau in ranking accuracy. It is also clear from the figure that the aggregated models demonstrate a higher quality than the readability indices for all the lengths of text fragments, and the ensemble of 9 models shows higher accuracy than the ensemble of 21 models. For this experiment, an ensemble of 21 models with a loss function $\mathcal{L}_1(M)$ and with L2 regularization with hyperparameter $\lambda = 10^{-4}$, and an ensemble of 9 models with loss function $\mathcal{L}_1(M)$ without regularization were selected.

7 Conclusion

In this paper, a method of estimating the cognitive complexity of a text based on quantile-based models is investigated. In particular, models are implemented at the phonetic, morphemic, lexical, and syntactic levels of the language, as well as their ensembling. For the individual models the empirical distributions of tokens over the reference collection of Russian Wikipedia articles are calculated. Ensemble models are trained on the datasets formed from social studies textbooks for different grades. All the models considered are compared in accuracy with the readability indices adapted for the Russian language. Among the individual models, the word length counter-based lexical models have shown the best accuracy, surpassing all the readability indices. The ensemble of 21 best separate models of all types has even more significantly surpassed all the readability indices in terms of the accuracy of ranking pairs of text fragments. The results of analysis of its accuracy for each pair of grades separately are consistent with our ideas about the complexity of school textbooks. It is observed that the ensemble model accurately ranks the text pieces by complexity from grades that are more than one to two years apart. It is also noticeable that the lower the grades of both text pieces in a pair, the easier it is for the model to arrange them correctly. The selection of the best ensemble (ablation study) is carried out, as a result of which the ensemble of nine separate models shows further significant improvement in quality. It consists of models of the following types: distance-based phonetic model, word length counter-based lexical model, and counter-based syntactic model. The paper also analyzes the dependence of the ranking accuracy on the average length of a text fragment in a dataset. As a result, it turned out that the accuracy decreases as the average length of the fragment decreases. A particularly sharp drop begins when the number of symbols is less than 14200.

References

Ivan Viktorovich Begtin. 2014. Plain russian language. `https://github.com/infoculture/plainrussian`.

Carl-Hugo Björnsson. 1968. *Läsbarhet: Lesbarkeit durch Lix*. Liber, Stockholm, Sweden.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283–284.

Arina Viktorovna Dmitrieva. 2017. «iskusstvo yuridicheskogo pis'ma»: kolichestvenniy analiz resheniy konstitucionnogo suda rossiyskoy federacii. *Sravnitel'noe konstitucionnoe obozrenie*, 3(118):125–133. Online available at: `https://academia.ilpp.ru/wp-content/uploads/2021/10/SKO-3-118-2017-125-133-Dmitrieva.pdf`.

Maksim A. Eremeev and Konstantin V. Vorontsov. 2019. Lexical quantile-based text complexity measure. // *Proceedings of the International Conference on Recent Advances in Natural Language Processing*, P 270–275. Online available at: `https://aclanthology.org/R19-1031.pdf`.

Rudolf Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32:221–233.

J. Peter Kincaid, Robert P. Fishburne Jr, Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Naval Technical Training Command Millington TN Research Branch*, P 40. Online available at: `https://web.archive.org/web/20220409163459/https://apps.dtic.mil/sti/pdfs/ADA006655.pdf`.

G. Harry McLaughlin. 1969. Smog grading — a new readability formula. *Journal of reading*, 12(8):639–646. Online available at: `https://web.archive.org/web/20220119124738/https://ogg.osu.edu/media/documents/health_lit/WRRSMOG_Readability_Formula_G._Harry_McLaughlin__1969_.pdf`.

Irina Vladimirovna Oborneva. 2005. Avtomatizaciya ocenki kachestva vospriyatiya teksta. *Vestnik Moskovskogo gorodskogo pedagogicheskogo universiteta. Seriya: Informatika i informatizaciya obrazovaniya*, 5:86–91.

R. J. Senter and Edgar A. Smith. 1967. Automated readability index. *AMRL-TR*, 66(220). Online available at: `https://web.archive.org/web/20160305161235/http://www.dtic.mil/get-tr-doc/pdf?AD=AD0667273`.

Sergey Shkarin. 2021. ruts, a library for statistics extraction from texts in russian. `https://github.com/SergeyShk/ruTS`.

Valery Solovyev, Vladimir Ivanov, and Marina Solnyshkina. 2018. Assessment of reading difficulty levels in russian academic texts: Approaches and metrics. *Journal of intelligent & fuzzy systems*, 34(5):3049–3058. Online available at: `https://www.researchgate.net/publication/324583915_Assessment_of_reading_difficulty_levels_in_Russian_academic_texts_Approaches_and_metrics`.

Milan Straka and Jana Straková. 2017. Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe. // *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies*, P 88–99.

J.A. Tuldava. 1975. On measuring text difficulty. // *Proceedings of Tartu State University*, volume 345, P 102–119.

MaxProb: Controllable Story Generation from Storyline

Sergey Vychegzhanin
Vyatka State University
Kirov, Russia
vychegzhaninsv@gmail.com

Anastasia Kotelnikova
Vyatka State University
Kirov, Russia
kotelnikova.av@gmail.com

Alexander Sergeev
Vyatka State University
Kirov, Russia
sergeev.alexander0@gmail.com

Evgeny Kotelnikov
Vyatka State University
Kirov, Russia
kotelnikov.ev@gmail.com

Abstract

Controllable story generation towards keywords or key phrases is one of the purposes of using language models. Recent work has shown that various decoding strategies prove to be effective in achieving a high level of language control. Such strategies require less computational resources compared to approaches based on fine-tuning pre-trained language models. The paper proposes and investigates the method *MaxProb* of controllable story generation in Russian, which works at the decoding stage in the process of text generation. The method uses a generative language model to estimate the probability of its tokens in order to shift the content of the text towards the guide phrase. The idea of the method is to generate a set of different small sequences of tokens from the language model vocabulary, estimate the probability of following the guide phrase after each sequence, and choose the most probable sequence. The method allows evaluating the consistency of the token sequence for the transition from the prompt to the guide phrase. The study was carried out using the Russian-language corpus of stories with extracted events that make up the plot of the story. Experiments have shown the effectiveness of the proposed method for automatically creating stories from a set of plot phrases.

Keywords: text generation; decoding strategy; GPT
DOI: 10.28995/2075-7182-2023-22-539-553

MaxProb: Управляемая генерация историй на основе сюжетных линий

Вычегжанин С. В.
Вятский государственный университет
Киров, Россия
vychegzhaninsv@gmail.com

Котельникова А. В.
Вятский государственный университет
Киров, Россия
kotelnikova.av@gmail.com

Сергеев А. В.
Вятский государственный университет
Киров, Россия
sergeev.alexander0@gmail.com

Котельников Е. В.
Вятский государственный университет
Киров, Россия
kotelnikov.ev@gmail.com

Аннотация

Управляемая генерация историй по направлению к ключевым словам или выражениям является одной из целей использования языковых моделей. Недавние работы показали, что использование различных стратегий декодирования является эффективным подходом для достижения высокого уровня управления языком. Такие стратегии требуют меньше вычислительных ресурсов по сравнению с подходами, основанными на тонкой настройке предварительно обученных языковых моделей. В статье предложен и исследован метод управляемой генерации историй на русском языке *MaxProb*, работающий на этапе декодирования в процессе генерации

текста. Метод основан на использовании генеративной языковой модели для оценки вероятности ее токенов с целью смещения содержания текста к направляющему выражению. Идея метода заключается в генерации множества различных небольших по длине последовательностей токенов из словаря языковой модели, оценке вероятности следования направляющей фразы после каждой последовательности, и выборе наиболее вероятной последовательности. Метод позволяет оценить логичность последовательности токенов для перехода от затравки к направляющему выражению. Исследование проводились с использованием русскоязычного корпуса историй с выделенными событиями, составляющими сюжет истории. Эксперименты показали эффективность предлагаемого метода для автоматического создания историй из набора сюжетных фраз.

Ключевые слова: генерация текстов; стратегия декодирования; GPT

1 Introduction

Natural language generation (NLG) is one of the important areas of computational linguistics. It aims to produce plausible and readable text in a human language. In recent years, the use of large-scale pre-trained language models (PLMs), in particular transformer-based PLMs [21], has shown promising results, allowing generating more diverse and fluent texts. Modern neural network models such as GPT-3 [2] can create texts that are difficult to distinguish from texts written by a human.

NLG technologies are crucial in many applications such as dialogue and question-answering systems, story generation, advertising, marketing, product and service reviews.

Controllable Text Generation is a problem actively explored in NLG. This is the task of generating texts that meet certain control constraints set by a human [16]. Sentiment, keywords, events, etc. can be considered as such constraints. For example, when generating a story, it is important to control the storyline and the ending.

There are two types of control over text generation models: soft and hard control. The aim of soft control is, e.g., to provide the desired sentiment or topic of the generated text. Hard control requires ensuring that the text contains explicit constraints, e.g., certain keywords. Figure 1 shows an example of hard controllable text generation, where the story is generated according to the keywords provided by the storyline and the order in which they appear [25].

Storyline	needed → money → computer → bought → happy
Generated story	John needed a computer for his birthday. He worked hard to earn money. John was able to buy his computer. He went to the store and bought a computer. John was happy with his new computer.

Figure 1: Example of controllable story generation with hard control

Many existing controllable generation methods [5], [8], [25] require the creation of training corpora and the implementation of a training procedure that is labor intensive and time consuming. This paper overcomes this problem by developing a plug-and-play method applicable to any large-scale PLM. Currently, there are not enough studies on the controllable text generation in Russian, so the proposed method is tested on Russian language models and text corpora.

The idea of the method is to generate a set of short sequences of words that provide a coherent transition from the prompt to the guide phrase, and then estimate the probability of following the guide phrase after each generated sequence and choose the most probable sequence. This method is plug-and-play, i.e. it can be used with any autoregressive model. The experiments carried out on generating stories from a set of events that make up the plot of a story prove the effectiveness of the proposed method for creating texts from a set of plot phrases.

The contribution of the paper is as follows:

- we offer *MaxProb* – a method of controllable text generation that generates stories in accordance with a user-specified sequence of guide phrases that make up the plot of the story;
- we apply the method to the Russian language;
- we form a text corpus containing stories with extracted storylines;
- we experiment with story generation to confirm the effectiveness of the proposed method.

2 Previous work

This section discusses the existing methods of controllable text generation that can be applied to the problem of story generation, which is of primary research interest. Automated story generation is the problem of mechanically selecting a sequence of events or actions that meet a set of criteria and can be told as a story [11]. Each story has a story world, interacting characters, and objects. The complexity of the story generation task is to generate a coherent and fluent story that is much longer than the user-specified prompt.

Controllable generation methods can be classified into three categories [26]: fine-tuning, retraining or refactoring, post-processing. Fine-tuning PLMs on a specialized data set is the main way to interact with models. Methods of this type fine-tune some or all of the model parameters to create texts that satisfy certain constraints. Early work on controllable story generation used convolutional and recurrent neural networks. Fan et al. [6] used a two-stage hierarchical approach. At the first stage, using the convolutional neural network, a premise, which determined the structure of the story, was generated. Then the premise was converted into a text passage using the seq2seq model. Yao et al. [25] used the RAKE algorithm [18] to build a storyline for each story from the corpus at the training stage using the most important words. After the storyline was generated, the seq2seq model converted it into text.

Reinforcement learning can be used for controllable story generation. For example, Tambwekar et al. [20] developed a reward-shaping technique that produces intermediate rewards at all different time-steps, which are then back-propagated into a language model in order to guide the generation of plot points towards a given goal.

Later, pre-trained language models based on the Transformer architecture began to be used for controllable generation. The prompt-based approach became widespread. Li and Liang [12] proposed a method called “prefix tuning” that freezes the parameters of the PLM and performs error backpropagation to optimize a small continuous task-specific vector called “prefix”. A similar P-tuning method [10] differs from prefix tuning in that it does not place a prompt with the “prefix” in the input, but constructs a suitable template composed of the continuous virtual token, which is obtained through gradient descent.

Retraining or refactoring involves changing the architecture of the language model or retraining a model from scratch. This approach is limited by the insufficient amount of labeled data and the high consumption of computing resources. One of the first models in this direction was CTRL [8]. The model was trained on a set of control codes. Zhang et al. [27] proposed POINTER, an insertion-based method for hard-constrained text generation, which involves preserving of specific words.

Cho et al. [4] proposed Story Control via Supervised Contrastive learning model to create a story conditioned on genre. The model learns conditional probability distribution by supervised contrastive objective, combined with log-likelihood objective.

Methods based only on using a decoder are called post-processing. Such methods require less computational resources. A representative of this group of methods is PPLM [Dathathri et al., 2020], which first trains an attribute discriminant model and then uses it to guide language model to generate the text with corresponding topic or sentiment. This group also includes the Keyword2Text method [15], which can be applied to an existing autoregressive language model without additional training. The idea of the method is to shift the output distribution of the language generation model to the semantic space of a given guide word in the word2vec or GloVe vector space. A similar idea is used in [22], but the difference is that the score function of the autoregressive language model is modified with the score function of another language model from the family of autoencoding models rather than with the cosine similarity to the target keyword.

Yang et al. [24] developed the Re3 framework to automatically generate longer stories of over two thousand words. Re3 first creates a structured plan, setting and characters by prompting GPT-3 with a premise. Then Re3 injects contextual information from both the plan and current story state into new GPT-3 prompt to generate new story passages.

In this paper, we propose a post-processing method that implements a decoding strategy based on heuristics. The difference from previous works [15], [22] lies in the fact that at each generation step for small sequences of tokens, the probability of following the guide phrase is estimated. The method is based on the idea that choosing a sequence of tokens, after which the probability of following the guide phrase is maximum, will induce the model to generate text, shifting its content to the guide phrase.

3 Controllable text generation

In this paper, we consider conditional probabilistic models for which the probability of the output text $X = \{x_1, \dots, x_n\}$ can be factorized by tokens:

$$P(X) = \prod_{i=1}^{n} P(x_i | x_{<i}), \quad (1)$$

where x_i denotes the i-th output token, and $x_{<i}$ denotes previous tokens $x_1, \dots, x_{i-1}$.

In accordance with formula (1), the goal of conditional text generation can be formulated as follows:

$$P(X|C) = \prod_{i=1}^{n} P(x_i | x_{<i}, C), \quad (2)$$

where C denotes the control conditions and X is the generated text, which complies with the control conditions.

While generating, sequences of natural language units (symbols, words, or sentences) are decoded from the probability distribution P. The decoding strategy plays an important role. At each time step, it selects tokens from the probability distribution over a model vocabulary. Beam search [14] and nucleus sampling [7] are examples of known decoding strategies.

Generative language models such as GPT learn to predict the next token in a given sequence of tokens. Text generation is a natural application for such models. However, when predicting the next token of a sequence, they are not able to take into account the context following it, which is supposed to be the content of the generated text.

In this study, we propose the *MaxProb* method, which at each generation step determines the most probable sequence of tokens for logically linking the prompt and the guide phrase that should be used in the text. The idea of the method is based on using intrinsic knowledge of a pre-trained language model to evaluate the token sequences and select the appropriate sequence for a coherent transition to the guide phrase. The proposed method can be applied to any autoregressive language model.

Let us consider the sequence $X = \{x_1, \dots, x_{i-1}, x_i, x_{i+1}, \dots, x_{i+k}, t, \dots, t_m\}$. For a given prompt $X_{1:i-1} = \{x_1, \dots, x_{i-1}\}$ and a guide phrase $T = \{t_1, \dots, t_m\}$ theoretically it is possible to find the connecting sequence $X_{i:i+k} = \{x_i, x_{i+1}, \dots, x_{i+k}\}$ using exhaustive search of tokens from the model vocabulary. However, such search has an exponential dependence on the length of the connecting sequence and is not applicable in practice. Therefore, in order to reduce the number of variants we propose a heuristic technique for generating and evaluating connecting sequences (Fig. 2).

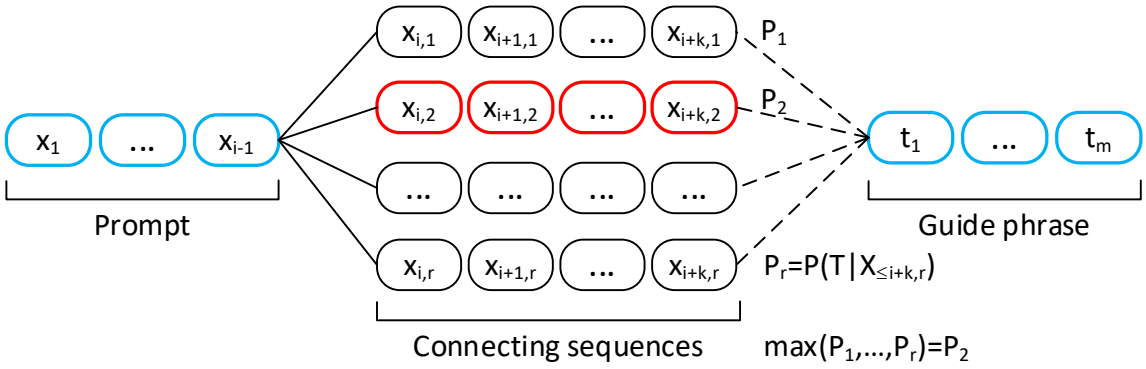


Figure 2: *MaxProb* method scheme

First, as continuations of the prompt $X_{1:i-1}$, r different sequences of tokens of length $k+1$ are generated using some decoding strategy. Next, for each of the r sequences, the probability of following the guide phrase T after it is determined by the formula:

$$P(X_{i:i+k}|X_{1:i-1}, T) = P(T|X_{\leq i+k}) = \prod_{j=1}^{m} P\left(t_j | t_{<j}, X_{\leq i+k}\right). \quad (3)$$

Further, at the current generation step, a sequence is selected for which the probability (3) is maximum, and the sequences of length $k + 1$ are repeatedly generated. In order to fulfill the condition of the explicit presence of the guide phrase in the text, after the generation of a given number of tokens is completed, this phrase can be inserted in the position in the text where it had the maximum probability for the entire generation time. After the phrase is inserted, the generation can continue towards the next guide phrase.

Formula (3) makes it possible to estimate the probability of following the guiding phrase for each connecting sequence of tokens, but does not evaluate their semantic similarity. There may be cases where semantic similarity is more important than the likelihood of following the guide phrase. To assess the similarity of the connecting sequence and the guide phrase, it is proposed to use the Jaccard coefficient:

$$K_J = \frac{C}{A + B - C}, \tag{4}$$

where A is the set of words in normal form from the prompt, B is the set of words in normal form from the guide phrase, C is the set of common words for the prompt and the guide phrase.

Taking into account formulas (3) and (4) for connecting sequences, the average score, which establishes a balance between the two measures, can be determined by the formula:

$$Score_{x_{i:i+k}} = w_{prob} P_{norm} + w_J K_J, \tag{5}$$

where w_{prob}, w_J are weight coefficients, P_{norm} is the normalized probability of following the guide phrase.

Thus, at each time step, the proposed method allows selecting the most logical sequence of tokens for linking the prompt and the guide phrase, based on the knowledge of the generative model itself.

As an example of how the method works, let us consider a text at some *i*-th generation step and a guide phrase separated by a sequence of unknown tokens, for example, of length 3 (Fig. 3). In the figure, the prompt for the autoregressive model is highlighted in blue, and the guide phrase is highlighted in orange. The connecting sequence is marked with labels <x₁><x₂><x₃>.

Однажды в лесу, около речки, сидел мальчик с бабушкой. Вдруг в это время из-за **<x₁><x₂><x₃>** волк напал на ребенка

Once in the forest, near the river, a boy was sitting with his grandmother. Suddenly, at this time, **<x₁><x₂><x₃>** the wolf attacked the child

Score	***P***	***K_J***	**<x₁><x₂><x₃>, Russian**	**<x₁><x₂><x₃>, English**
0.944	3.20E-11	0.200	кустов вышли волки	wolves came out from behind the bushes
0.226	6.50E-12	0.200	поворота вышел волк,	a wolf came out from around the corner,
0.105	1.90E-13	0.100	дерева на поляну	from behind a tree to a clearing
0.100	4.60E-18	0.100	дерева выскочило	from behind a tree jumped out
0.100	9.30E-19	0.100	деревьев вышел лев,	a lion came out from behind the trees,
0.100	5.70E-22	0.100	деревьев вышли три	from behind a tree appeared three
0.100	3.00E-24	0.100	деревьев показалась большая	from behind a tree appeared a large
0.052	2.80E-13	0.100	деревьев выскочили	from behind the trees jumped out
0.048	1.70E-13	0.100	поворота леса вышел	from around the corner of the forest came out
0.044	9.40E-15	0.100	поворота речки выскочил	out of the turn of the river, jumped out

Figure 3: Example of prompt and connecting sequences at the *i*-th generation step

The prompt is an input of the autoregressive model. With some decoding strategy, such as top-k sampling, r different sequences of 3 tokens <x_1><x_2><x_3> are generated. For them, the probabilities of following the guide phrase P and the Jaccard coefficients K_J are calculated. The calculated values are averaged by formula (5). The sequences of tokens are sorted in descending order of *Score*, and the sequence with the highest value of the average score is selected. The selected sequence is attached to the prompt, and the generation process continues until the specified number of tokens is generated.

4 Text corpus

To conduct experiments, a text corpus[1] was formed from fairy tales in Russian with extracted storylines. The corpus is made up of fairy tales placed on nukadeti.ru[2] with a length of no more than 5000 characters. In total, the training corpus contains 562 fairy tales.

In each fairy tale, plot phrases were singled out, i.e. phrases that determine the main events in the story, the storyline. To do this, first, in each fairy tale keywords and phrases were selected, using the methods *yake*[3] [3], *rakun*[4], *frake*[5], *textrank*[6], *rutermextract*[7], *keybert*[8] methods. Each method selected 15 keywords and phrases. The *yake* and *rutermextract* methods showed the best quality, so their results were used in the next stage to compose plot phrases.

The *yake* and *rutermextract* methods were selected out of six methods manually. The main problems with other methods were the following. The top keywords and phrases of the *rakun* and the *keybert* were very often parts of each other, they intersected, i.e. were parts of one longer phrase. So, the number of sentences with these selected keywords was very low and the plot could not be built out of them.

The *frake*'s results often contained just single words and it was very difficult to understand from which sentences they were selected (if they repeated several times).

The problem of *textrank* was that it didn't pay attention to sentence segmentation – many selected phrases were parts of two neighbor sentences.

Further, plot phrases were extracted from fairy tales according to the following algorithm:

1. Events were found. Events are syntactically related triples <object, action, object> (for example, "старуха, испекла, колобок" – "old woman, baked, bun"). The objects were selected from a set of keywords, and the actions was determined from the parse tree as nodes, syntactically associated with the objects. The stanza library[9] was used to make the syntax parsing of the sentences.

2. The most important events found were selected from the found events. Each selected event was assigned a weight obtained by summing the weights of the keywords extracted by the *yake* and *rutermextract* methods separately.

3. From the selected important events, a plot phrase was formed, determined by a 4-element set (o_1, v, o_2, m), where v is a verb, o are objects related to the verb, m is a modifier, prepositional object, or indirect object. Prepositions are possible before o and m. An example of an event: "grooves in the forest spilled into whole streams", where "spilled" is v, "grooves" and "streams" are o, "forest" is m ("канавки в лесу разлились в целые ручьи", v – "разлились", o – "канавки", "целые ручьи", m – "лесу").

For each of the two methods for extracting keywords, their own plot phrases were formed, the number of which, depending on the fairy tale, varied from 0 to 26. Figure 4 shows the distribution of the number of plot phrases extracted using the *yake* and *rutermextract* methods.

[1] https://github.com/icecreamz/MaxProb.
[2] https://nukadeti.ru.
[3] https://github.com/LIAAD/yake.
[4] https://github.com/SkBlaz/rakun.
[5] https://github.com/cominsys/FRAKE.
[6] https://github.com/JRC1995/TextRank-Keyword-Extraction.
[7] https://github.com/igor-shevchenko/rutermextract.
[8] https://github.com/MaartenGr/KeyBERT.
[9] https://stanfordnlp.github.io/stanza.

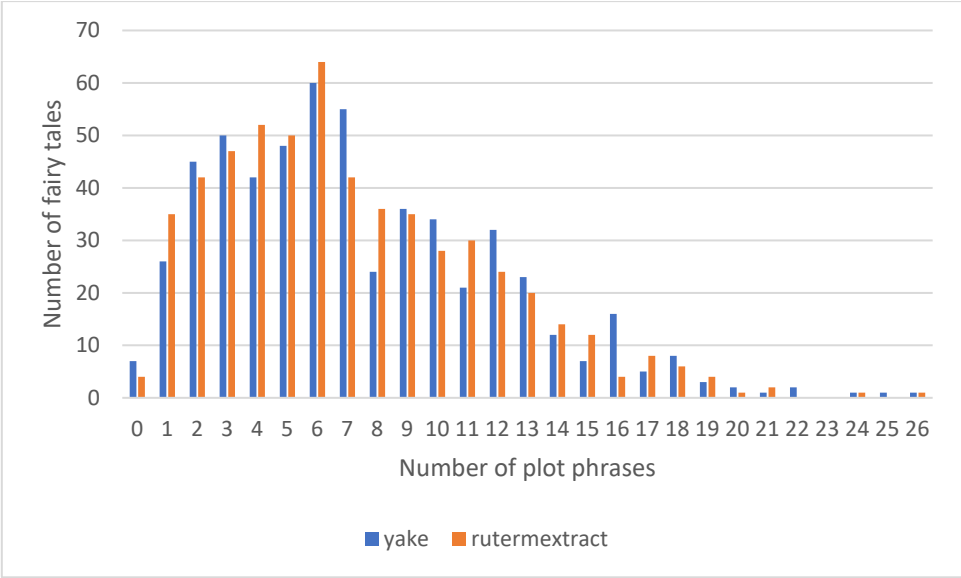


Figure 4: Distribution of the number of plot phrases

The number of sentences in fairy tales varied from 4 to 139. The distribution of the number of sentences is shown in Fig. 5.

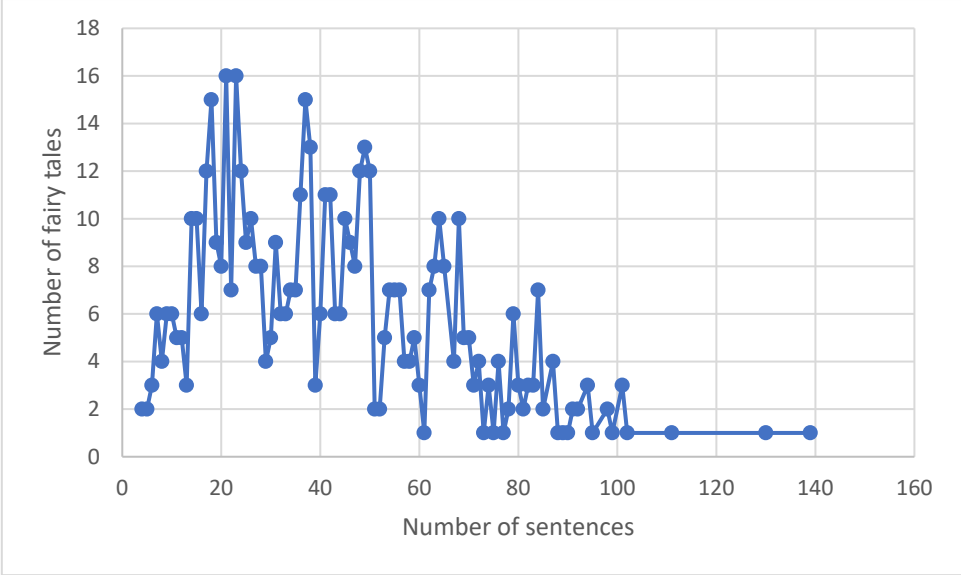


Figure 5: Number of sentences in fairy tales

Since the number of resulting plot phrases should correlate with the length of the tale, the plot was assembled from the selected phrases according to the following algorithm:

1. The minimum number of phrases in the plot is 1, the maximum is the rounded-up value of the logrhyme to base 2 of the number of sentences n in the text: $\lceil log_2 n \rceil$.
2. If the *yake* method returned the number of plot phrases in the above range, these phrases were taken in order as a plot.
3. If the *yake* method produced fewer plot phrases, and the *rutermextract* method yielded enough, then the *rutermextract* phrases were taken in order as a plot.
4. If both methods returned the number of phrases less than the minimum value, their results were combined without repetitions in the order of the sentences in the text.

5. If the *yake* method produced more plot phrases than the maximum allowable in accordance with point 1, then a part of the fragments with maximum weights was taken for the required amount.

Table 1 shows the distribution of the number of phrases in the plot in the training corpus. The first column contains the number of phrases in the plot, the second – the number of fairy tales with such a number of phrases, the third – the share of the total number of fairy tales in the training corpus, i.e., from 562 fairy tales.

A test corpus of 25 plots was also formed. The distribution by the number of plot phrases in the test corpus is proportional to the distribution in the training corpus and is given in the fourth column of Table 1.

# Plot phrases	# Fairy tales in the training corpus	Share of the total number of fairy tales, %	# Fairy tales in the test corpus
1	31	5.46	1
2	48	8.45	2
3	53	9.33	2
4	56	9.86	3
5	107	18.84	5
6	185	32.57	8
7	80	14.08	4
8	2	0.35	0

Table 1: Distribution of the number of phrases in the plot

Table 2 shows statistics on the number of tokens received using the ruGPT-3 Large tokenizer in fairy tales of training corpus, depending on the number of plot phrases.

# Plot phrases	Minimum number of tokens	Maximum number of tokens	Average number of tokens
1	28	900	230.9
2	85	400	238.9
3	115	1,015	344.3
4	128	752	308.9
5	212	950	476.4
6	406	1,283	796.0
7	757	1,503	1,150.1
8	1,555	1,897	1,726.0

Table 2: Number of tokens

5 Experimental Setup

Keywords used in plot events were extracted from texts using the *yake* and *rutemextract* libraries. The initial word forms for calculating the Jaccard coefficient were determined using the pymorphy2 library [9]. Text generation experiments were carried out using the ruGPT-3 Large[10] language model (760 million parameters), which is the Russian-language version of the GPT-2 model [17].

In the experiments, fairy tales were generated according to a given sequence of events that determines the plot of the fairy tale. The top-*k* sampling decoding strategy with parameter $k = 10$ was used as a decoding strategy in MaxProb to obtain connecting sequences of tokens.

The values of the weight coefficients in formula (5) were determined empirically based on the analysis of the generated connecting sequences. The coefficients took the values $w_{prob} = 0.9$ and $w_J = 0.1$. The probability of following the guide phrase turned out to be more significant, and due to the w_J coefficient, the connecting sequence that was closest in content to the guide phrase was ranked first.

[10] https://huggingface.co/sberbank-ai/rugpt3large_based_on_gpt2.

The length of connecting sequences was 3 tokens. Experiments were also carried out for windows ranging in size from 1 to 15 tokens. According to the results of the experiments, a small window of connecting sequences had a better effect on shifting the content of the generated text towards the plot phrase than a large window. With a large window size, suitable short sequences of words, most likely followed by a guide phrase, could be missed, and as a result, the content of the generated text deviated significantly from the content of the guide phrase.

The maximum length of the generated fairy tale (in tokens) depended on the number of plot phrases and was equal to the average number + 10% of the tokens (see Table 2).

The proposed method was compared with three methods of controllable text generation:

1. Inserting key phrases in a prompt (PromptLearn).

When conducting experiments using the PromptLearn method, the ruGPT-3 Large model was fine-tuned with 80% of the tales from the training corpus for three epochs. The prompt with size up to 1024 tokens was used as input data for the model:

"Plot: {plot phrase 1}, {plot phrase 2}, …, {plot phrase n}.\n
Text: {the text of fairy tale}"

For each tale, the number of plot phrases ranged from 1 to 8. To generate fairy tales, sampling was used with parameters $p = 0.95$ and $k = 50$. The length of the generated fairy tale was chosen similarly to MaxProb.

2. Few-shot learning (FewShotLearn).

The ruGPT-3 Large model was also used to apply the FewShotLearn method. The prompt was used as input for the model:

"Compose text with keywords:\n
Plot: {plot phrase 1}, {plot phrase 2}, …, {plot phrase n}.\n
Text: {the text of fairy tale} ###
Plot: {plot phrase 1}, {plot phrase 2}, …, {plot phrase n}.\n
Text: {the text of fairy tale}"

The number of fairy tales input to the model depended on the estimated maximum length of the generated text so that the total input sequence fit into 2048 tokens allowed for the model. The range of the number of input training examples is from 1 to 5, most often 3. When generating texts, the same parameters as for PromptLearn were used. The length of the generated fairy tale was chosen similarly to MaxProb.

3. Constrained beam search (ConstrainedBS).

ConstrainedBS was used as the baseline of controlled generation. Plot phrases were tokenized and used as a list of restrictions. The generation was carried out using the ruGPT-3 Large model. The prompt "Однажды" ("Once") was used as an input of the model. The number of beams varied from 7 to 10 to generate different stories. A prohibition on the repetition of 3-grams was also established. The length of the generated fairy tale was chosen similarly to MaxProb.

The quality of the generated texts was evaluated using automatic and human-centric evaluation methods. Four measures were used for automatic evaluation [13], [23], [28]:

- perplexity (PPL) – is a metric to measure how well the language probability model predicts a sample. It is usually calculated as the exponential mean of the negative log-probability per token in the language model. We calculated perplexity using the ruGPT-3 Medium[11] language model (350 million parameters);
- repetition (Rep) evaluates the proportion of repeated 4-grams in the text, where the tokens belong to the vocabulary of the ruGPT-3 Large model;
- Word Inclusion Coverage (Cov) shows the percentage of plot words included in the generated text. Plot and generated words are lemmatized;
- self-BLEU-5 evaluates the syntactic diversity of a given set of texts. It is defined as the average overlap between all generated texts.

[11] https://huggingface.co/sberbank-ai/rugpt3medium_based_on_gpt2.

Three measures were used for human-centric evaluation:

- coherence – whether the story is consistent in terms of causal relationships in the context;
- relevance – the story corresponds to the plot, the events in the story unfold in accordance with the storyline;
- interestingness – how the user likes the story, whether it is interesting.

6 Results and discussion

Table 3 shows the statistical characteristics of the generated texts, calculated using the GEM-metrics library[12]:

- Avg length – the average length of texts (in words);
- Vocab size – the number of different words;
- Distinct-n – the ratio of distinct n-grams over the total number of n-grams.

Generation methods	Avg length	Vocab size	Distinct-1	Distinct-2	Distinct-3
ConstrainedBS	447	3,149	0.11	0.49	0.85
FewShotLearn	158	1,998	0.19	0.57	0.77
PromptLearn	430	3,608	0.13	0.50	0.77
MaxProb	497	3,015	0.10	0.41	0.70

Table 3: Statistical characteristics of generated texts

Analyzing Table 3, you can see that the FewShotLearn method, on average, generated fairy tales 3 times shorter than the other three methods. It should be noted that when generating longer tales, the first tale was often interrupted and a new tale began.

Table 4 shows the average values of perplexity, repetition, word inclusion coverage, and self-BLEU-5 measures calculated for fairy tales generated from 25 storylines of test corpus. For each storyline, two fairy tales were generated. A total of 50 tales were generated by each method.

Additionally, the scores were also calculated for the base model ruGPT-3 Large. The ruGPT-3 Large model was preliminarily fine-tuned on the training corpus of fairy tales with the addition of the prefix “Текст: ” (“Text: ”) to the beginning of each fairy tale, which was then used as a prompt during generation. The experiments used the strategy of decoding top-k sampling with the parameter $k = 10$.

Generation methods	↓ PPL ± Std	↓ Rep, %	↑ Cov, %	↓ Self-BLEU-5
ruGPT-3	**5.3 ± 1.5**	26.43	20.07	0.028
ConstrainedBS	6.8 ± 2.5	**0.61**	80.86	0.094
FewShotLearn	9.9 ± 6.1	16.40	43.49	**0.014**
PromptLearn	6.8 ± 1.7	14.82	71.32	0.032
MaxProb	7.0 ± 1.4	18.33	**99.54**	0.063

Table 4: Automatic quality scores for generation methods

The values of the Cov measure in Table 4 show that the MaxProb method ensures that more than 99% of the words from the storyline events appear in the text. The texts generated by this method met the requirement of matching the storyline to the best extent. The smallest number of words from the storyline appeared in the texts generated by the FewShotLearn method and is 43.49%. In such texts, the required characters and events were rare. This is largely due to the relatively short length of the generated tales.

The values of the Rep measure for the FewShotLearn, PromptLearn, and MaxProb methods are quite close to each other and vary from 14.82% to 18.33%. The ConstrainedBS method has a Rep value close to zero as a result of setting the prohibition on the repetition of 3-grams, otherwise the generation was often reduced to repetitions of words. Repeatability values do not suggest a significant superiority of

[12] https://github.com/GEM-benchmark/GEM-metrics.

one method over others. Notably, controllable generation methods reduced the repeatability value compared to the ruGPT-3 base model.

The lowest PPL value among controllable generation methods was obtained for PromptLearn and ConstrainedBS and is 6.8. The MaxProb method showed a 0.2 higher average PPL, but it has a lower standard deviation, i.e. provides a more stable level of perplexity. For the FewShotLearn method, perplexity and standard deviation were the highest. It is known, that a lower perplexity value corresponds to a better model. The increase in perplexity compared to the base ruGPT-3 model indicates that the control process is “unnatural” for the model. This causes the model to be more "surprised" by the tokens observed in the text.

The self-BLEU-5 measure has the lowest value for FewShotLearn. The texts generated by this method turned out to be the most syntactically diverse. The variety of PromptLearn is at the level of the basic ruGPT-3 model. The least varied texts are for the ConstrainedBS method.

To calculate human-centric measures, the generated texts were evaluated by three annotators for coherence, relevance, and interestingness. The assessment was carried out on a 5-point Likert scale (1 – the worst, 5 – the best). For all the methods, only the generated sequence was evaluated, without prompt. Inter-annotator agreement was measured using the Spearman coefficient [1]. The value of this coefficient for the “coherence” criterion was 0.54, “relevance” – 0.87, “interestingness” – 0.59. The values, which are greater than 0.5 indicate high annotator agreement [19].

Table 5 shows the average scores of coherence, relevance and interestingness.

Generation methods	**↑ Coherence**	**↑ Relevance**	**↑ Interestingness**
ConstrainedBS	1.65	2.91	1.56
FewShotLearn	2.23	1.63	2.25
PromptLearn	**2.62**	2.23	**2.82**
MaxProb	2.20	**4.89**	2.74

Table 5: Human-centric quality scores for generation methods

The coherence scores for all methods turned out to be low, less than 3 points. The low coherence is due to the quality of the ruGPT-3 base model, which was used in the experiments. The PromptLearn method turned out to be the best in terms of coherence, the MaxProb method more often violated the coherence, and ConstrainedBS generated practically incoherent texts. However, MaxProb almost always ensured that all events from the storyline appeared in the text, as evidenced by a high relevance score. Despite the lowest coherence, the texts with MaxProb were slightly less interesting than with the PromptLearn method, but were more interesting than with FewShotLearn.

Figure 6 shows the parallel coordinates visualization of all calculated measures.

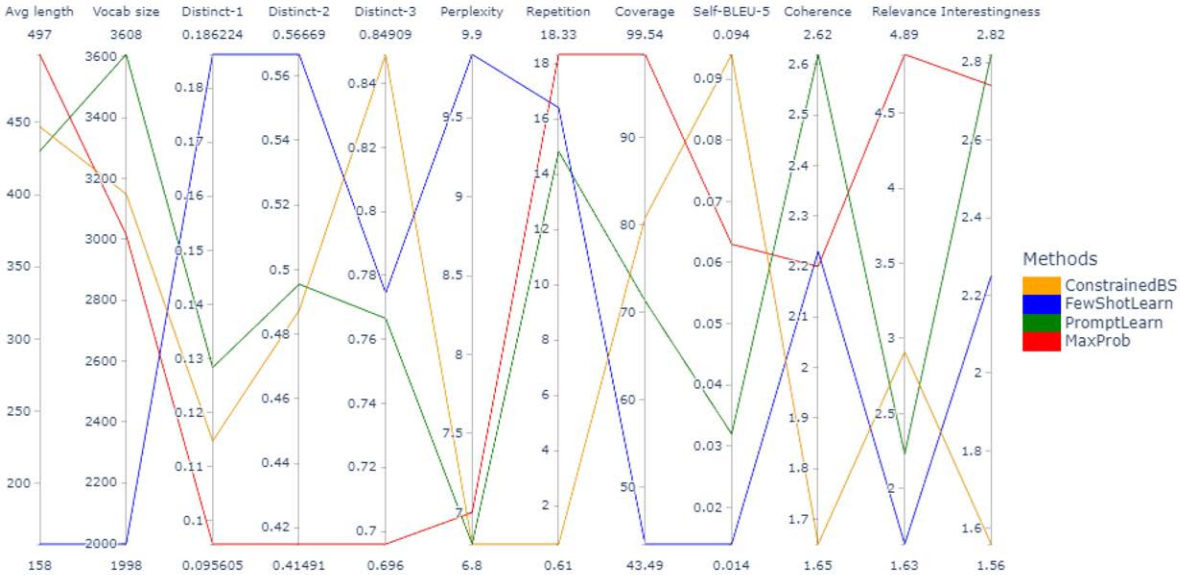


Figure 6: The parallel coordinates visualization of the measures

Let us give a specific example of the MaxProb method (Fig. 7). For the guide phrase “the cat ate sour cream” (“кот съел сметану”) for some *i*-th step, the text “An old woman had a cat, whom she loved very much and called: Ko-ko-ko. The cat loved” (“У одной старушки был кот, которого она очень любила и которого звала: Ко-ко-ко. Кот очень любил”). At the *i*-th step, using the decoding strategy top-*k* sampling, the connecting sequences of three tokens were obtained, shown in Fig. 7. For each sequence, the probabilities of following the guide phrase P by formula (3), the Jaccard coefficients K_J by formula (4) and the average values of *Score* by formula (5) are calculated. According to the results of the *i*-th step, the sequence “milk with bread,” (“молоко с хлебом,”) was chosen, which has the highest average *Score*.

У одной старушки был кот, которого она очень любила и которого звала: Ко-ко-ко. Кот очень любил **<x₁><x₂><x₃>** кот съел сметану

An old woman had a cat, whom she loved very much and called: Ko-ko-ko. The cat loved **<x₁><x₂><x₃>** the cat ate sour cream

Score	***P***	K_J	**<x₁><x₂><x₃>, Russian**	**<x₁><x₂><x₃>, English**
0.900	2.50E-10	0.111	молоко с хлебом,	milk with bread,
0.121	5.90E-12	0.143	старушку,	old woman,
0.113	3.60E-12	0.143	, чтобы его	, to be
0.105	1.30E-12	0.143	свою кошку и	his cat and
0.104	1.20E-12	0.143	ее, да	her, yes
0.102	6.60E-13	0.143	ее и не	her and not
0.101	1.60E-13	0.143	, когда его	, when he
0.100	6.40E-15	0.143	ее, Она	her, She
0.100	1.40E-15	0.143	эту старушку	this old woman
0.066	6.20E-12	0.125	молоко, и,	milk, and,

Figure 7: Connecting sequences and their scores on the *i*-th step of generation

Table 6 shows the connecting sequences for steps $i + 1$ through $i + 5$. The sequences that received the highest *Score* value are highlighted in blue at each step. These sequences were chosen as the most probable ones and added to the prompt.

№	**Step** $i + 1$	**Step** $i + 2$	**Step** $i + 3$	**Step** $i + 4$	**Step** $i + 5$
1	а еще больше	- сметану	. И вот	однажды кот съел	всю сметану
2	и, когда	любил сме- тану,	. А еще	однажды, когда	сметану и
3	а хлеб -	со сметаной	с молоком.	однажды вече- ром кошка	сметаны,
4	а больше всего	ел сметану,	, но молоко	однажды утром старушка	столько сме- таны,
5	поэтому, как	- с молоком,	, и поэтому	он как-то	все сметанное
6	но не любил,	с молоком,	, поэтому каж- дый	он любил сме- тану	все сметаны
7	и поэтому он	любил, когда	, которая была	он, чтобы	все молоко,
8	но он не	со сливочным	и хлеб.	, когда он	всё молоко,
9	и если молоко	с капустой,	. Поэтому ба- бушка	, однажды кот	всё, что
10	но молока в	с сыром,	. Кот ел	, как-то	целый хлеб и

№	Step $i+1$	Step $i+2$	Step $i+3$	Step $i+4$	Step $i+5$
1	and even more	– sour cream	. And then	one day the cat ate	all the sour cream
2	and, when	loved sour cream,	. And then	one day, when	sour cream and
3	and bread –	with sour cream	with milk.	Once in the evening the cat	sour cream,
4	and most of all	ate sour cream,	, but milk	Once in the evening the old woman	so much sour cream,
5	that's why, how	– with milk,	, and that's why	he once	all of sour cream
6	he didn't liked,	with milk,	, that's why every	he liked sour cream	all sour cream
7	and that's why he	liked, when	, which was	he, to	all milk,
8	but he didn't	with creamy	and bread.	, when he	all milk,
9	and if milk	with cabbage,	. That's why the old woman	, once the cat	all, that
10	but milk in	with cheese,	. The cat ate	, once	whole bread and

Table 6: Connecting sequences on steps $i+1$, …, $i+5$ of generation: Russian (top) and English (bottom) versions

As a result, after $i+5$ steps, the text was generated: "An old woman had a cat, whom she loved very much and called: Ko-ko-ko. The cat loved milk with bread, and even more – sour cream. And then one day the cat ate all the sour cream". This example demonstrates that choosing a sequence after which the probability of a guide phrase is maximum induces the generative model to lead the text to the required phrase. At the same time, the connecting sequence may not contain the guide phrase in an explicit form, but be close to it in meaning due to synonyms.

7 Conclusion

The proposed *MaxProb* method allows generating stories in accordance with a user-specified sequence of guide phrases that determines the plot of the story. Guide phrases describe some of the key events in the story and consist of several words. The method uses a generative language model to estimate the probability of following a guide phrase after various short sequences of tokens generated by the model. The method selects the sequence with the highest probability, prompting the model to shift the content of the text towards the guide phrase. Experiments carried out using the Russian-language corpus of fairy tales with extracted storylines showed that the proposed method provides a high proportion of story words (more than 99% in Cov) and phrases (4.89 points in Relevance) in the text. In terms of text quality (PPL measure and interestingness), the method is comparable to the PromptLearn fine-tuning method, but it does not require creating a training corpus and the executing of a time-consuming training procedure.

Ethical considerations

The proposed method helps to control the content of automatically generated text according to the user's needs. Note that large language models, including the one used in the proposed ruGPT-3 method, generate texts similar to texts written by a person. However, it is not guaranteed that the generated texts are factually correct. They may contain false or fictitious information that may mislead the non-expert reader. When using plot phrases containing factually incorrect information, the generation will be based on false content and, therefore, will lead to the creation of inaccurate texts. Like any tool, it can be used for negative purposes. Content control can lead to the creation of fake text for the purpose of deception, disinformation or propaganda. We hope that our method will be used for positive purposes, like helping writers to create fairy tales in accordance with a given plot. Placing such methods in the public domain will help develop countermeasures to detect them.

Acknowledgements

This work was supported by Russian Science Foundation, project № 23-21-00330, https://rscf.ru/en/project/23-21-00330/.

References

[1] Amidei J., Piwek P., Willis A. Agreement is overrated: A plea for correlation to assess human evaluation reliability // Proceedings of the 12th International Conference on Natural Language Generation. – 2019. – P. 344–354.
[2] Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J. et al. Language models are few-shot learners // Advances in Neural Information Processing Systems. – 2020. – Vol. 33. – P. 1877–1901.
[3] Campos R., Mangaravite V., Pasquali A., Jorge A., Nunes C., Jatowt A. YAKE! Keyword Extraction from Single Documents using Multiple Local Features // Information Sciences Journal. – 2020. – Vol. 509. – P. 257–289.
[4] Cho J., Jeong M., Bak J., Cheong Y.-G. Genre-controllable story generation via supervised contrastive learning // Proceedings of the ACM Web Conference 2022. – 2022. – P. 2839–2849.
[5] Dathathri S., Madotto A., Lan J., Hung J., Frank E., Molino P., Yosinski J., Liu R. Plug and play language models: A simple approach to controlled text generation // Computing Research Repository. – 2020. – arXiv:1912.02164. – Access mode: https://arxiv.org/abs/1912.02164.
[6] Fan A., Lewis M., Dauphin Y. Hierarchical neural story generation // Computing Research Repository. – 2018. – arXiv:1805.04833. – Access mode: https://arxiv.org/abs/1805.04833.
[7] Holtzman A., Buys J., Du L., Forbes M., Choi Y. The curious case of neural text degeneration // Proceedings of the 8th International Conference on Learning Representations. – 2020. – P. 1–16.
[8] Keskar N.S., McCann B., Varshney L., Xiong C., Socher R. CTRL – A Conditional Transformer Language Model for Controllable Generation // Computing Research Repository. – 2019. – arXiv:1909.05858. – Access mode: https://arxiv.org/abs/1909.05858.
[9] Korobov M. Morphological Analyzer and Generator for Russian and Ukrainian Languages // Analysis of Images, Social Networks and Texts. – 2015. – P. 320–332.
[10] Lester B., Al-Rfou R., Constant N. The Power of Scale for Parameter-Efficient Prompt Tuning // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. – 2021. – P. 3045–3059.
[11] Li B., Lee-Urban S., Johnston G., Riedl M. O. Story generation with crowdsourced plot graphs // Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence. – 2013. – P. 598–604.
[12] Li X. L., Liang P. Prefix-Tuning: Optimizing Continuous Prompts for Generation // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. – 2021. – P. 4582–4597.
[13] Lin B.Y., Zhou W., Shen M., Zhou P., Bhagavatula C., Choi Y., Ren X. CommonGen: A constrained text generation challenge for generative commonsense reasoning // Findings of the Association for Computational Linguistics: EMNLP 2020. – 2020. – P. 1823–1840.
[14] Meister C., Vieira T., Cotterell R. If beam search is the answer, what was the question? // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. – 2020. – P. 2173–2185.
[15] Pascual D., Egressy B., Meister C., Cotterell R., Wattenhofer R. A Plug-and-Play Method for Controlled Text Generation // Findings of the Association for Computational Linguistics: EMNLP 2021. – 2021. – P. 3973–3997.
[16] Prabhumoye S., Black A.W., Salakhutdinov R. Exploring Controllable Text Generation Techniques // Proceedings of the 28th International Conference on Computational Linguistics. – 2020. – P. 1–14.
[17] Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I. Language models are unsupervised multitask learners // OpenAI blog. – 2019. – Vol. 1(8). – Access mode: https://openai.com/blog/better-language-models.
[18] Rose S., Engel D., Cramer N., Cowley W. Automatic keyword extraction from individual documents // Text Mining: Applications and Theory. – 2010. – P. 3–20.
[19] Rosenthal J.A. Qualitative descriptors of strength of association and effect size. Journal of social service Research. – 1996. – Vol. 21(4). – P. 37–59.
[20] Tambwekar P., Dhuliawala M., Martin L.J., Mehta A., Harrison B., Riedl M.O. Controllable Neural Story Plot Generation via Reward Shaping // Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19. International Joint Conferences on Artificial Intelligence Organization. – 2019. – P. 5982–5988.
[21] Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is All you Need // Proceedings of the 31st Conference on Neural Information Processing Systems (NeurIPS). – 2017. – Vol. 30. – P. 6000–6010.
[22] Vychegzhanin S., Kotelnikov E. Collocation2Text: Controllable Text Generation from Guide Phrases in Russian // Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue-2022" – Issue 21. – P. 564–576.

[23] Welleck S., Kulikov I., Roller S., Dinan E., Cho K., Weston J. Neural text generation with unlikelihood training // Proceedings of the 8th International Conference on Learning Representations. – 2020. – P. 1–18.
[24] Yang K., Tian Y., Peng N., Klein D. Re3: Generating longer stories with recursive reprompting and revision // Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022). – 2022. – P. 4393–4479.
[25] Yao L., Peng N., Weischedel R., Knight K., Zhao D., Yan R. Plan-and-Write: Towards Better Automatic Storytelling // Proceedings of the AAAI Conference on Artificial Intelligence. – 2019. – Vol. 33(01). – P. 7378–7385.
[26] Zhang H., Song H., Li S., Zhou M., Song D. A Survey of Controllable Text Generation using Transformer-based Pre-trained Language Models // Computing Research Repository. – 2022. – arXiv:2201.05337. Access mode: https://arxiv.org/abs/2201.05337.
[27] Zhang Y., Wang G., Li C., Gan Z., Brockett C., Dolan B. POINTER: Constrained Progressive Text Generation via Insertion-based Generative Pre-training // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). – 2020. – P. 8649–8670.
[28] Zhu Y., Lu S., Zheng L., Guo J., Zhang W., Wang J., Yu J. Texygen: A benchmarking platform for text generation models // The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval. – 2018. – P. 1097–1100.

The prosody of the Russian question

Yanko T. E.
Institute of Linguistics, Russian
Academy of Sciences /
1 bld. 1 Bolshoy Kislovsky Lane,
125009 Moscow,
tanya_yanko@list.ru

Abstract

The analysis of Russian interrogative prosody is based on a model of a question as consisting of the two components: the illocutionary proper component and the illocutionary improper component. The illocutionary improper component includes the data for information retrieval. The illocutionary proper component can be formed both by segmental means of expression (by an interrogative word or a particle) or solely by prosody (as in Russian yes-no questions). The prosody of Russian questions having the interrogative words or the interrogative particle *li* is highly variable, whereas the prosody of Russian yes-no questions expressed by prosody is stable. The latter is the Russian rising accent, which has a rise on the tonic syllable of the accent-bearer followed by a fall on the post-tonics if any. The illocutionary improper component can be located sentence initially and carry a specific falling accent (namely, a late fall). A specific type of a question with the interrogative proper component omitted is recognized. Such questions carry a late fall, or a falling-rising accent on the accent-bearer. The analysis is exemplified by the frequency tracings of the sound sentences taken from the Russian National Corpus and other open sources. As the instrument for verifying the acoustic data, we used the computer system Praat. The paper is illustrated throughout with pitch contours of sound records.

Keywords: sound speech; interrogative sentences; prosody; illocutionary force
DOI: 10.28995/2075-7182-2023-22-554-565

Просодическая модель речевого акта вопроса

Янко Т. Е.
Институт языкознания РАН /
125009, Москва,
Большой Кисловский пер. 1 стр. 1
tanya_yanko@list.ru

Аннотация

Анализ просодии русского вопросительного предложения основан на иллокутивной модели вопроса как состоящего из двух компонентов: собственно иллокутивного и несобственно иллокутивного. Несобственно иллокутивный компонент вопроса содержит условия для поиска информации слушающим. Собственно иллокутивный компонент может быть сформирован как с помощью вопросительного слова или вопросительной частицы *ли*, так и чисто просодически. Просодия вопросов, имеющих сегментные средства выражения иллокутивной силы, весьма вариативна, в то время как просодия русских *да-нет*-вопросов, имеющих суперсегментные средства выражения значений, стабильна. Это подъем частоты основного тона с падением на заударных, если они есть (ИК-3, по Е.А. Брызгуновой). Несобственно иллокутивный компонент вопроса может предшествовать собственно иллокутивному и тогда он несет падение с поздним таймингом (ИК-2, по Е. А. Брызгуновой). Выделяется особый тип вопросов с опущенным собственно вопросительным компонентом. Такие вопросы несут падение ИК-2 или нисходяще-восходящую кривую частоты основного тона типа ИК-4. Анализ иллюстрируют примеры, полученные из звучащего подкорпуса Русского Национального корпуса и из других открытых источников. В качестве инструмента для верификации слуховых гипотез использована компьютерная система анализа устной речи Praat.

Ключевые слова: звучащая речь; вопросительное предложение; просодия; иллокутивная сила

1 Введение

В работе анализируется просодическая структура русских вопросительных предложений. Особенность плана выражения вопроса состоит в том, что в русском языке используются как чисто просодические средства выражения иллокутивной силы вопроса (в *да-нет*-вопросах), так и сегментные средства в виде вопросительных слов и вопросительной частицы *ли*, а также в виде средств формирования особого вида вопросов с *не правда ли*?; *не так ли*?; *так*?; *да*? (или tag-questions в английской терминологической традиции). В вопросах с сегментными средствами выражения иллокутивной силы просодия играет формирующую роль, отделяя вопрос от других речевых актов в потоке речи и один компонент речевого акта от другого. Иначе, при присутствии в предложении сегментных средств выражения иллокутивной силы означающим просодии можно считать установление границ речевого акта и границ компонентов речевого акта. Соответственно, при выражении иллокутивной силы сегментные средства мажорируют суперсегментные. При сегментном способе выражения иллокутивной силы просодические показатели формирования границ речевого акта и компонентов речевого акта вариативны. При суперсегментном способе выражения иллокутивной силы просодические показатели стабильны. Так, в разделе 3 показано, что собственно иллокутивный компонент русского *да-нет*-вопроса маркируется подъемом частоты основного тона (f_0) на ударном слоге по типу ИК-3 (Е. А. Брызгунова [2]) и что использование этого средства не зависит от расположения акцентоносителя вопроса в предложении и способа просодического оформления несобственно вопросительного компонента. Просодия же вопроса с вопросительным словом зависит от взаимного расположения собственно иллокутивного и несобственно иллокутивного компонентов и фактора активации информации в сознании собеседников, а также от свободного выбора говорящего.

Просодия функционирует не автономно, она определенным образом накладывается на сегментный материал предложения, она функционирует в комбинации с линейной структурой предложения и принципами выбора словоформ — носителей просодически релевантных изменений f_0. В предложениях с одинаковой лексико-синтаксической структурой и единообразной кривой изменения f_0, но различными словоформами — носителями просодических пиков, просодические структуры различны.

Просодия русского вопросительного предложения неоднократно была предметом рассмотрения [3, 4, 6, 8, 9, 12]. Опираясь на эти результаты, мы, однако, видим новизну настоящей работы в принципиальном разделении просодических и сегментных средств выражения вопросительности, в анализе роли просодии в предложениях с сегментным выражением иллокутивной силы, а также в разработке иллокутивной модели русского вопроса, которая коротко охарактеризована в разделе 2, см. также [19]. В эту модель встроено предложенное здесь описание просодии вопроса.

В работе использованы данные звучащего подкорпуса Национального корпуса русского языка [10] и малого рабочего корпуса автора [МРК], собранного из открытых источников. В качестве таких источников как наиболее богатых вопросительными предложениями дискурсивных жанров использованы материалы пресс-конференций, интервью, допросов и художественных фильмов. Обращение к таким источникам связано с тем, что подавляющее большинство звучащих корпусов в большей степени, чем материал диалога и полилога, отражают материал нарратива, практически не содержащего вопросительных предложений. Для верификации слуховых гипотез мы пользуемся системой анализа звучащей речи Praat [1].

2 Иллокутивная модель русского вопроса

Мы выделяем в вопросе два иллокутивных компонента: собственно иллокутивный (собственно вопросительный) и несобственно иллокутивный (несобственно вопросительный). Собственно вопросительный компонент соответствует тому, что спрашивается, а несобственно вопросительный — тому, о чем спрашивается. В вопросе *Еда для маленькой где*? [10] компонент *еда для маленькой* — это несобственно иллокутивный компонент, а *где* — собственно иллокутивный. В этом вопросе несобственно иллокутивный и собственно иллокутивный компонент разделены т.н. иллокутивным швом. С понятием иллокутивного шва ср. такие связанные с сегментацией речевого акта на релевантные компоненты понятия, как просодический шов (prosodic breaks) в работе

[7] и в цитированной там литературе, а также ритмико-синтаксический барьер [13] и коммуникативный барьер [14: 390]. Означаемое иллокутивного шва — это граница между двумя компонентами, составляющими речевой акт, а означающее — маркирование каждого иллокутивного компонента отдельной интонационной конструкцией (ИК) в духе Е. А. Брызгуновой и несущей ее словоформы, или словоформы-акцентоносителя. В анализируемом вопросе — это два падения типа ИК-2 (знак \\ после словоформы-акцентоносителя), по Е. А. Брызгуновой [2], и соответствующие им словоформы-акцентоносители: *маленькой* и *где*? Между компонентами, разделенными иллокутивным швом, есть или возможна пауза. ИК, несомые акцентоносителями вопроса, идентичны: *Еда для маленькой*\\ *где*?\\, т.е. обратной адаптации акцентов здесь нет. Предложения, имеющие иллокутивный шов, мы будем называть иллокутивно расчлененными. *Да-нет*-вопрос *А телефончик Ватикана*\\ *не подскажете//*? [10] тоже иллокутивно расчлененный.

В вопросе же *Где*/- *проще установить фонетические соответствия**?* [10] с начальным подъемом (знак /- после вопросительного слова *где*), относительно ровной плато-фазой между началом и исходом вопроса и падением типа ИК-1 в финале (знак \) иллокутивного шва нет. Здесь мы исходим из гтпотезы о том, что подъем на *где* семиотически нерелевантен, он только предваряет конечное падение на акцентоносителе вопроса словоформе *соответствия*.

Другие типы речевых актов также подвержены иллокутивному расчленению, ср. повествоательное предложение *Одно из наших прав собственности/ — это право на природную ренту*\ [10] и императив *А вы челюсть*\\ *ему вправьте*\\ [10]. Подъем типа ИК-3 (знак /) маркирует тему, отчлененную от ремы, акцентоноситель которой, в свою очередь, несет падение ИК-1 [2]. В императиве акцентоноситель препозитивного несобственно иллокутивного компонента несет падение типа ИК-2, собственно иллокутивный компонент также несет падение ИК-2. Обратной адаптации акцентов в императиве нет. Несобственно иллокутивный компонент в иллокутивно расчлененных предложениях предшествует собственно иллокутивному. В вопросах и императивах инициальное расположение несобственно иллокутивного компонента — это результат линейно-акцентного преобразования, состоящего в вынесении несобственно иллокутивного компонента в начальную позицию (ср. аналогичное противопоставление в [4: 243]).

В дефолтных (исходных), т.е. таких, где линейно-синтаксическая структура вносит в семантическую структуру предложения минимальный вклад, вопросы (и императивы) имеют несобственно иллокутивный компонент в позиции после собственно иллокутивного. В дефолтном же повествовательном предложении, в отличие от вопросов и императивов, несобственно иллокутивный компонент (тема) предшествует собственно иллокутивному (реме). Расчленение компонентов речевого акта на фрагменты, не равные собственно и несобственно иллокутивным компонентам, также возможно, но на нем мы здесь не останавливаемся.

Понятие коммуникативной парадигмы как класса предложений с единой лексико-синтаксической структурой, но различными линейными структурами было впервые предложено в работе И.И. Ковтуновой [5], затем в виде понятия линейно-акцентного преобразования, применяемого к исходному члену парадигмы, развито в работе Е.В. Падучевой [11]. Впоследствии анализ линейно-акцентных преобразований речевого акта сообщения получил развитие во многих работах. В этой работе понятие линейно-акцентного преобразования применяется к анализу вопросов на примере вынесения в препозицию несобствнно вопросительного компонента.

В вопросах с *не так ли*?, *не правда ли*?, *так*?, *да*? несобственно иллокутивный компонент дефолтно предшествует собственно иллокутивному, и такие вопросы всегда иллокутивно расчлененные: *Скверная погода*\\, *не правда*\\ *ли*? [10]. Просодия русских tag-questions вариативна, здесь приведен лишь вариант с двумя нисходящими конструкциями ИК-2.

Вопросы, полученные из иллокутивно расчлененных путем отсечения собственно вопросительного компонента, который реконструируется из контекста, иллокутивно нерасчлененные: *Значит, вы точно будете на Чёрных Камнях*\\? [10]. Предполагается, что в этом вопросе опущено *не так ли*?. Мы называем такие вопросы эллиптическими.

В разделе 3 ниже предложено описание просодии *да-нет*-вопросов, в разделе 4 — вопросов с вопросительным словом, в разделе 5 — вопросов с *не так ли?* и *так?*, в разделе 6 — эллиптических вопросов. Вопросы с частицей *ли*, кроме вопросов с *не так ли* и *не правда ли*, здесь не рассматриваются: их просодия проанализирована в [18]. Просодия вопроса с *ли* как вопроса с сегментным средством выражения иллокутивной силы вариативна. Просодия вопросов может в до-

полнение к маркированию иллокутивных компонентов включать указание на дискурсивную незавершенность, если вопросы задаются целой серией или если говорящий после задания вопроса объясняет, почему этот вопрос задается. Описание просодии вопросов в контексте дискурсивной незавершенности дано в [17-18]. Отдельной (исследованной [3:398;15: 46-67]) проблемой служат композиции компонентов вопроса с контрастом и эмфазой. Эту задачу мы здесь оставляем в стороне.

3 Просодия *да-нет*-вопросов

В русских иллокутивно нерасчлененных *да-нет*-вопросах иллокутивная сила выражается с помощью интонационной конструкции ИК-3. Акцентоноситель такого вопроса выбирается в соответствии с объемом информации, подвергаемой верификации. Так, в вопросе 1) генерал, привлекая внимание дамы, интересуется, не хотела ли бы она служить под его началом. Соответственно, акцентоноситель вопроса — *хотели:*

1) Вы не хотели/ бы служить в десантных войсках [10].

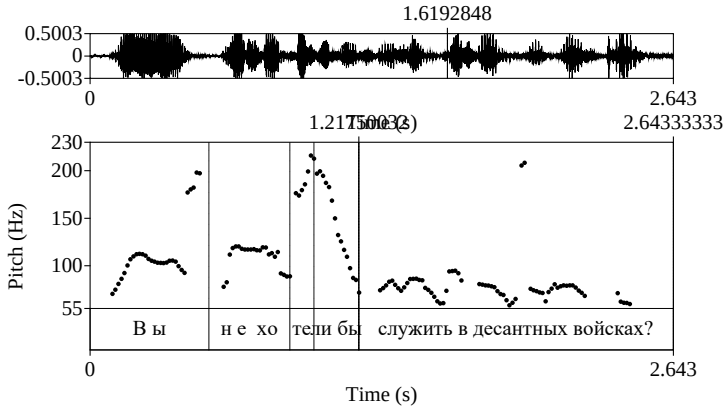


Рисунок 1: График изменения f_0 (нижняя панель) в примере 1)

Словоформа *хотели* несет подъем *да-нет*-вопроса ИК-3. Такое маркирование вопроса — хорошо известный факт, мы приводим его для полноты картины. В вопросе 2) верификации требует информация, соответствующая именной группе *партия оловянного солдатика* в целом. (В угловые скобки помещается необходимый для анализа материала контекст). Акцентоноситель такой группы — несогласованное определение *солдатика*. О выборе акцентоносителя в компонентах речевых актов с разнообразной лексико-синтаксической структурой см. [16].

2) <Что тебе остается только?> Партия оловянного солдатика/? [10].

Линейная структура примеров 1) и 2) дефолтная.

В иллокутивно расчлененных *да-нет*-вопросах с инверсией просодическая структура характеризуется падением типа ИК-2 на акцентоносителе препозитивного несобственно иллокутивного компонента и подъемом типа ИК-3 на акцентоносителе собственно иллокутивного компонента:

3) *Спать\\ не жестко/*? [МРК].

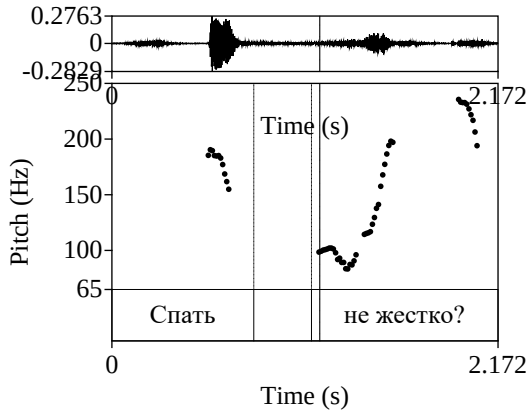


Рисунок 2: График изменения f_0 в примере 3).

В примере 3) наблюдается падение ИК-2 на единственной словоформе препозитивного несобственно вопросительного компонента и подъем на ударном слоге постпозитивного собственно вопросительного компонента *не жестко*. Между несобственно вопросительным и собственно вопросительным компонентом имеется пауза. Дефолтная линейно-акцентная структура для соответствующей лексико-синтаксической структуры: *Не жестко/ спать*? с единственным подъемом ИК-3 на словоформе *жестко*.

Отдельного анализа требуют *да-нет*-вопросы с *или*. Рассмотрим дизъюнктивные группы с двумя членами. Просодия *да-нет*-вопроса различает строгую дизъюнкцию и нестрогую дизъюнкцию. При строгой дизъюнкции производится выбор только одной возможности из двух, при нестрогой — предполагается, что запросу удовлетворят обе возможности, а также каждая из них. Пример 4) иллюстрирует строгую дизъюнкцию, пример 5) — нестрогую. В 4) носитель ИК-3 — словоформа *могут* (первый дизъюнктивный член); в 5) — словоформа *белым* (второй дизъюнктивный член).

4) <*Как вы считаете?*>. *Могут/ или пугают*\\? [10].

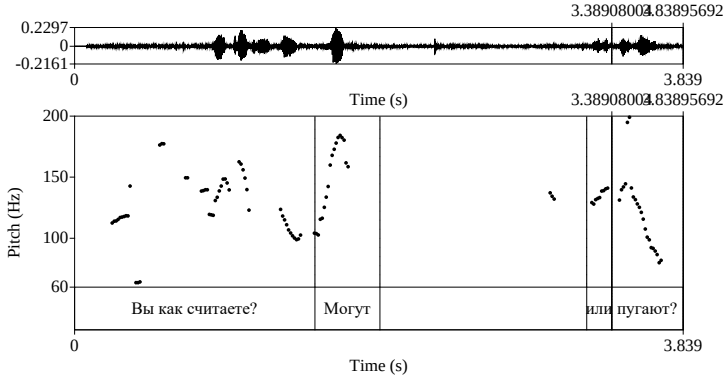


Рисунок 3: График изменения f_0 в примере 4)

5) *<Для чего?> Помогать красным или белым/? <Нет, грабить>* [10].

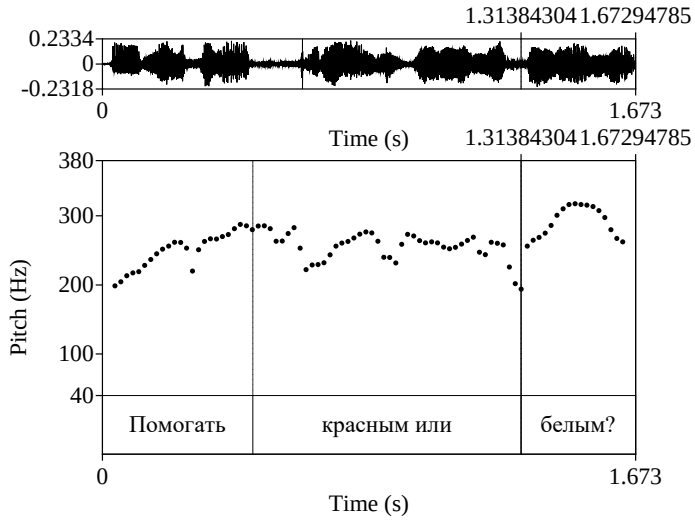


Рисунок 4: График изменения f_0 в примере 5)

В примере 4) наблюдается два релевантных движения f_0: подъем ИК-3 на акцентоносителе первого дизъюнктивного члена, который совпадает с самим этим членом, и падение ИК-2 — на акцентоносителе второго. Здесь действует принцип обратной адаптации: *да-нет*-вопрос маркируется подъемом, а второй член дизъюнкции несет нисходящий тон. В вопросе же 5) имеется только одна релевантная интонационная конструкция — ИК-3. Акцентоноситель принадлежит группе второго члена дизъюнкции. Дизъюнкция здесь подчинена инфинитиву *помогать*. В соответствии с дефолтным принципом выбора акцентоносителя в подчиненной дизъюнктивной группе акцентоносителем становится второй член дизъюнкции словоформа *белым*.

Таким образом, в *да-нет*-вопросах с *или* подъем, маркирующий вопрос, в случае строгой дизъюнкции, включающей два дизъюнктивных члена, расположен на первой дизъюнктивной группе, а в случае нестрогой дизъюнкции — на втором.

4 Просодия вопросов с вопросительным словом

Дефолтная просодия русского вопроса с вопросительным словом характеризуется падением ИК-1 на акцентоносителе вопроса, который расположен в исходе предложения. Начало предложения несет подъем на препозитивном вопросительном слове. Достигнутая высота f_0 снижается к исходу вопроса. Начальный подъем мы не считаем коммуникативно релевантным. Это компенсаторный подъем начала вопроса, который в исходе завершается конструкцией ИК-1, и между началом и концом релевантных изменений f_0 не наблюдается:

6) *Где/- проще установить фонетические соответствия*\? [10].

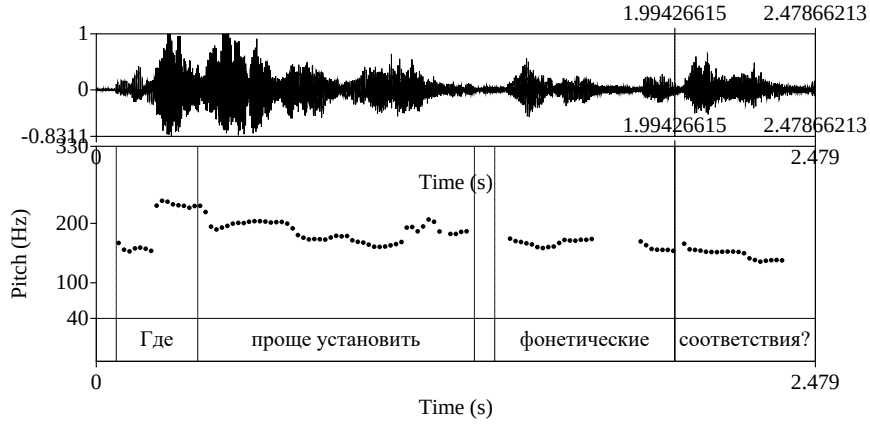


Рисунок 5: График изменения f_0 в примере 6)

При уходе с начальной позиции вопросительное слово (при отсутствии дополнительных условий, о которых будет сказано ниже) получает ИК-2, ср. падение f_0 на *где* на Рисунке 6:

7) *Ты где*\\? [10].

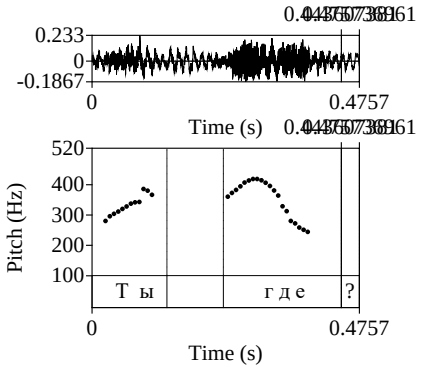


Рисунок 6: График изменения f_0 в примере 7)

Поясним также, что вопросительное слово в начальной позиции, как и в не-начальной, в определенных условиях (т.е. при наложении дополнительных значений, например контраста, линейно-просодических преобразований, а также под влиянием фактора активации информации в сознании коммуникантов) тоже может быть носителем ИК-2. Иначе говоря, просодическая структура вопроса, как в вопросе 6), характеризует только дефолтные, или исходные, структуры вопроса. Кроме того, вопросительное слово может нести ИК-2, если оно служит единственной словоформой вопроса.

Если в вопросе 7) несобственно вопросительный компонент *ты* имеет клитическую (словесно и коммуникативно атоническую) форму, то в вопросе 8) препозитивный несобственно иллокутивный компонент *фуражка моя*, как и собственно иллокутивный компонент *где*, несет падение ИК-2. Обратной адаптации акцентов здесь нет. Вопрос иллокутивно расчленен на два компонента, каждый из которых имеет свой акцентоноситель и свою модель изменения f_0:

8) *Фуражка*\\ *моя где*\\? [10].

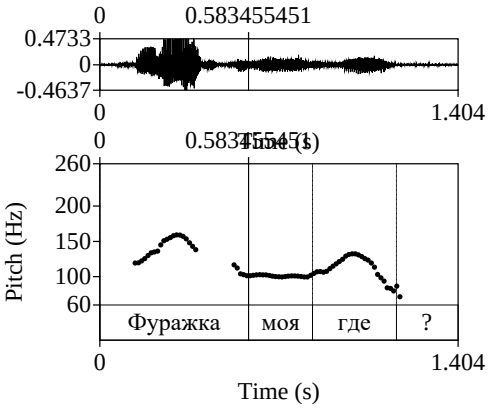


Рисунок 7: График изменения f_0 в примере 8)

Дефолтная структура вопроса 7): *Где моя фуражка*\?

В вопросах, контекст которых предполагает, что будет задан вопрос с *где* (*кто*, *когда*, *зачем*), вопросительное слово имеет атоническую форму даже в позиции не начала предложения:

9) *Паспорт*\\ *мой где*? [10].

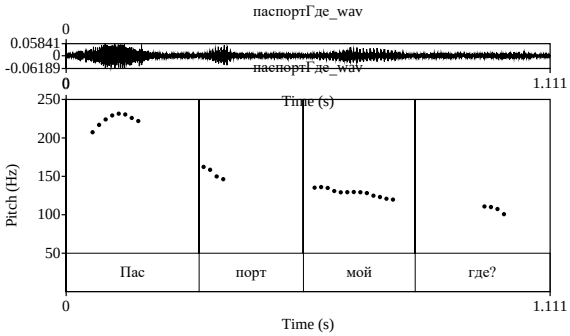


Рисунок 8: График изменения f_0 в примере 9)

В вопросе 9) единственным акцентоносителем служит словоформа *паспорт*, несущая падение ИК-2. На то, что говорящий должен задать вопрос с *где*, указывают его движения, обследующие карманы: говорящий находится в поиске, неизвестно только, что он ищет. Этот пример (и другие) говорят о том, что при вербализации вопросительного слова действует параметр активации информации в сознании собеседников: если ясно, что чего-то не хватает, вопросительное слово *где* (и другие вопросительные слова в соответствующем контексте) могут клитизироваться. Аналогичный принцип действует и при выборе акцентоносителя ремы сообщения, где словоформы, соотносимые с известной информацией, лишаются своего права на роль акцентоносителя, при том, что они имеют это право в соответствии с иерархией синтаксических приоритетов. Дефолтная иллокутивная структура для примера 9): *Где мой паспорт*\?. Она аналогична той, что представлена примером 6). Таким образом, в вопросах с вопросительным словом просодическая структура чувствительна к активации информации в сознании слушателя.

5 Просодия русских tag-questions

В русские tag-questions мы включаем вопросы с собственно вопросительным компонентом *не так ли*?, *не правда ли*?, *так*?, *ведь так*?, *да*?, *верно*?, *правда*?. Собственно вопросительный компонент в отсутствие частицы *ли* артикулируется с подъемом ИК-3 на *так*, *да* и *верно*, а просодия собственно вопросительного компонента с *ли* вариативна. Собственно иллокутивный компонент с *ли* может нести ИК-2, ИК-3 и ИК-4; ср. просодию вопросов с *ли* [18]. В то же время акцентоноситель достаточно автономного несобственно иллокутивного компонента независимо от просодической реализации собственно вопросительного компонента также может нести ИК-2, ИК-3 и ИК-4. Анализ звучащих данных дает большое разнообразие пар интонационных конструкций в русских tag-questions. Наиболее часто встречающиеся комбинации: ИК-4 — ИК-4, ИК-2 — ИК-2 и ИК-2 — ИК-3. Рассмотрим вопрос 10) с *не так ли*. Здесь несобственно вопросительный компонент (акцентоноситель — словоформа *близка*) несет падение ИК-2, а собственнно вопросительный компонент (акцентоноситель — словоформа *так*) — подъем ИК-3:

10) *Но окончательная победа близка\\, не так/ ли*? [10].

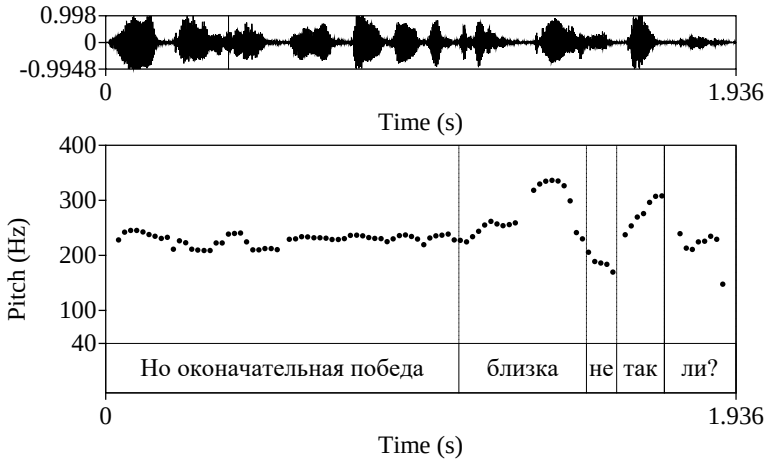


Рисунок 9: График изменения f_0 в примере 10)

Выявленные ограничения на сочетаемость ИК в иллокутивных компонентах русских tag-questions состоят в следующем: вопрос без *ли* имеет ИК-3 или ИК-4 (но не ИК-2) на собственно иллокутивном компоненте; ИК-4 в собственно вопросительном и несобственно вопросительном компонентах, как правило, используются парой. Примеры на все возможные комбинации ИК в двух компонентах мы здесь не приводим.

6 Просодия эллиптических вопросов

Вопросы, которые расчленены иллокутивным швом на два компонента — собственно иллокутивный и несобственно иллокутивный, могут подвергаться иллокутивному эллипсису, при котором сохраняется только несобственно иллокутивный компонент вопроса (Sic!). Если исходить из того, что компонент, оставшийся после эллиптического усечения, может сохранять свою исконную просодию, следует предположить, что вопросы с отсутствующим собственно вопросительным компонентом несут ИК-2 или ИК-4. Теоретически возможное опущение собственно вопросительного компонента в tag-question с собственно вопросительным компонентом, несущим ИК-3, мы не рассматриваем, так как в этом случае результат эллипсиса не отличим от *да-нет*-вопроса. Обратимся к примеру 11).

11) *Вы хотите что-то написать*\\? [10].

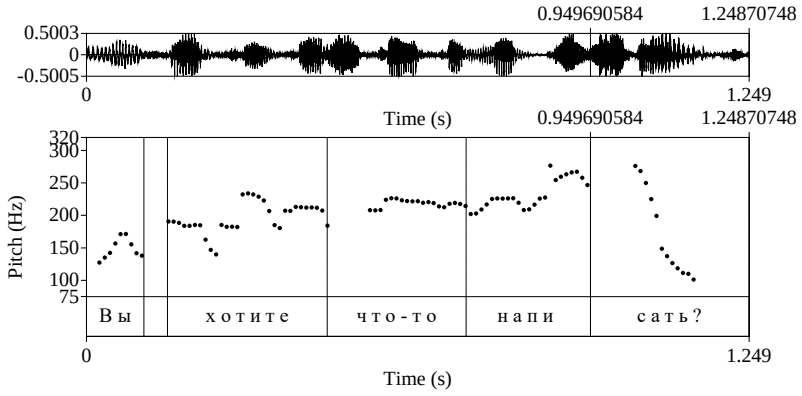


Рисунок 10: График изменения f_0 в примере 11)

В предложении 11) единственным релевантным движением f_0 оказывается падение ИК-2 на словоформе *написать*. Здесь нет ни вопросительного слова, ни вопросительной частицы, ни восходящей просодии. На каком основании мы считаем это предложение вопросительным? Прежде всего, мы полагаемся на мнение экспертов — создателей Мультимедийного подкорпуса НКРЯ, подготовивших транскрипты звучащих текстов и квалифицировавших этот и другие аналогичные примеры как вопросы. Далее, в исходе таких предложений расположен носитель ИК-2, а не ИК-1, что соответствовало бы стандарту сообщения. Кроме того, правым контекстом таких предложений, как предложение 11), служит подтверждение или опровержение догадки говорящего, если при восстановлении эллипсиса реконструируется tag-question (ср. пример 12)), или ответ на вопрос с *где,* (*как, почему*), ср. примеры 13) и 14):

12) — *Значит, вы точно будете на Черных Камнях*\\? — *Да-да, на Черных Камнях* [10].

13) — *А четвертая*\\ *школа*? — *Четвертая школа здесь. Вот за этим домом* [10].

14) — *А парк*\/ *возле него? — А парк … ходят туда только с колясками гулять* [10].

Эллиптические вопросы в виде запроса на подтверждение выводов и догадок говорящего имеют сентенциальную синтаксическую форму и несут ИК-2 на акцентоносителе вопроса, ср. примеры 11)-12). Они восходят к tag-questions. Эллиптическим характером таких вопросов объясняется нисходящий акцент, формирующий вопрос: он унаследован от несобственно вопросительного компонента в tag-question, ср. пример 10). Эллиптические вопросы, восходящие к расчлененным вопросам с вопросительным словом, могут нести как ИК-2 (пример 13)), так и ИК-4 (пример 14)). Формирование вопроса — это хорошо изученная в литературе функция ИК-4 [3-4]. Эти вопросы имеют синтаксическую форму групп: именной, инфинитивной, числовой. Таким образом, исходя из возможности постановки ИК-4 в эллиптическом вопросе можно сделать вывод о том, что просодия различает синтаксическую структуру эллиптического вопроса: ИК-4 (а также ИК-2) — для группы, ИК-2 — для сентенциальной формы.

7 Заключение

Анализ просодии вопросов демонстрирует следующее.

(1) В вопросе представлены собственно вопросительный и несобственно вопросительный компоненты. Препозиция несобственно вопросительного компонента служит результатом специального линейно-акцентного преобразования, при котором несобственно иллокутивный компонент

получает статус зачина предложения и маркируется просодически. Если сравнить иллокутивную структуру вопроса с иллокутивной структурой сообщения, обратит на себя внимание различное дефолтное расположение собственно и несобственно иллокутивных компонентов: в сообщении в отличие от вопроса несобственно иллокутивный компонент (тема) предшествует собственно иллокутивному компоненту (реме).

(2) Иллокутивная сила русского *да-нет*-вопроса (без вопросительной частицы *ли*) маркируется ИК-3, которая фиксируется на акцентоносителе собственно вопросительного компонента. Если несобственно вопросительный компонент предшествует собственно вопросительному, первый маркируется падением ИК-2 на акцентоносителе несобственно вопросительного компонента. Таким образом, при формировании просодии *да-нет*-вопроса с препозицией несобственно вопросительного компонента действует принцип обратной адаптации акцентов.

(3) В вопросах с сегментным маркированием иллокутивной силы просодия играет сегментирующую роль, отделяя собственно иллокутивный компонент от несобственно иллокутивного и речевой акт от соседних речевых актов в потоке дискурса. Просодия тогда не служит выражению иллокутивной силы. При препозиции несобственно вопросительного компонента в вопросах с вопросительным словом акцентоноситель несобственно вопросительного компонента несет падение ИК-2, как и акцентоноситель собственно вопросительного компонента. Принцип обратной адаптации акцентов здесь не действует.

(4) Предпринятый ранее анализ вопросов с частицей *ли*, говорит о том, что просодия вопросов с *ли*, крайне вариативна [18]. Как собственно вопросительный, так и препозитивный несобственно вопросительный компонент вопроса с *ли* могут нести ИК-2, ИК-3 и ИК-4 каждый и формировать различные комбинации интонационных конструкций в паре «несобственно вопросительный — собственно вопросительный компонент».

(5) Русские tag-questions всегда иллокутивно расчлененные. Они формируют любые двойные комбинации из множества {ИК-2, ИК-3, ИК-4} при маркировании пары «несобственно вопросительный — собственно вопросительный компонент».

(6) При эллиптическом опущении собственно вопросительного компонента иллокутивно расчлененного вопроса образуется эллиптический вопрос, который маркируется ИК-2 или ИК-4 на акцентоносителе.

(7) В вопросе с постпозитивным вопросительным словом действует принцип клитизации (коммуникативной “безударности”) вопросительного слова, если его значение известно из контекста.

References

[1] Boersma P., Weenink D. Praat: Doing phonetics by computer. Version 6.3.08. Online: Praat: doing Phonetics by Computer (uva.nl), 2023 (accessed date: 15.01.2023).

[2] Bryzgunova E. A. Intonation [Intonatsija], Russian Grammar [Russkaja grammatika]. Vol. 1, Nauka, Moscow, 1982. P. 103-118.

[3] 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.

[4] Kobozeva I. M. (2005) An essay in characterizing lexical-syntactic, semantic, and pragmatic properties of interrogative dialogical turns in terms of features [Opyt razrabotki priznakovoj bazy dlja harakteristiki leksikosintaksicheskih, semanticheskih i pragmaticheskih svojstv voprositel'nyh replik], Proceedings of the International Conference “Dialogue’2005” [Trudy mezhdunarodnoj konferencii «Dialog 2005»]. 2005. P. 238–244.

[5] 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. Pr. S. 132-194.

[6] Kodzasov S. V., Bonch-Osmolovskaja A. A., Zaharov L. M., Kobozeva I. M., Krivnova O. F. Data base ‘Intonation of Russian Dialogue’: interrogative sentences [Baza dannyh «Intonacija russkogo dialoga»: voprositel'nye repliki], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ‘Dialogue 2005’. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii ‘Dialog 2005’]. Issue 10. [Vyp. 10]. Moscow: RGGU Publ. 2005. P. 245-247.

[7] Krivnova O. F. The depth of prosodic breaks in spoken text (experimental data) [Glubina prosodicheskih shvov v zvuchashhem tekste (jeksperimental'nye dannye)], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference ‘Dialogue 2015’ [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii ‘Dialog 2015’. Issue 14 (21) [Vyp. 14 (21)]: Vol. 1 [T.1]. Moscow: RGGU Publ. 2015. [M.: Izdatel'stvo RGGU, 2015]. P. 338-351.

[8] Knyazev S.V., Dyachenko S.V. (2023) Melodic contour of yes-no question in Western middle-Russian dialect with akan'je. Part II: Pskov dialect [Melodicheskij kontur obshhego voprosa v zapadnom srednerusskom akajushhem govore. Chast' I: Seligero-torzhkovskie govory]. Lomonosov Philology Journal. Series 9. Philology, 2023, no. 2, pp. 44–60

[9] Post M. Spoken corpora of spontaneous speech as a source to study polar question intonation in Russian dialects, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2022' [Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii 'Dialog 2022']. Vyp. 21 (28), M.: RGGU, 2022. P. 477-488.

[10] Russian National Corpus [Nacional'nyj korpus russkogo jazyka]: <www.ruscorpora.ru>.

[11] Paducheva E.V. 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. 1984. P. 25-31.

[12] Rathcke T. V. A perceptual study on Russian questions and statements. Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung der Universität Kiel (AIPUK), 2006. 37. pp. 51-62.

[13] Zaliznjak A.A. To the study of the language of birch barks, Yanin V. L., Zaliznyak A. A. Novgorod letters on birch bark. From the excavations of 1984-1989 [K izucheniju jazyka berestjanyh gramot, Janin V. L., Zaliznjak A. A. Novgorodskie gramoty na bereste. Iz raskopok 1984-1989 g.g.] M.: Nauka, 1993. P. 191-319.

[14] Zimmerling A.V. Word order systems of Slavic languages in a typological aspect [Sistemy porjadka slov slavjanskih jazykov v tipologicheskom aspekte] M.; Jazyki slavjanskoj kul'tury. 2013 P. 390.

[15] Yanko T. E. Kommunikativnye strategii russkoy rechi [The communicative strategies of the Russian speech]. Moscow, Yazyki slavyanskoi kul'tury, 2001.

[16] Yanko T. E. Accent placement principles in Russian, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue' (2011). [Komp'juternaja lingvistika i intellektual'nye tehnologii]. Issue 10. [Vyp. 10]. Moscow: RGGU Publ. 2011. P. 288-301.

[17] Yanko T.E. Imperatives, vocatives, and questions in coherent discourse: the prosodic markers of incompleteness in the Russian spoken speech corpora [Rechevye akty v strukture svjaznogo diskursa: pokazateli nezavershennosti po dannym korpusov zvuchashhej rechi, Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii Dialog 2018]. Issue17 (14) [Vyp. 17 (24)], M.: RGGU Publ. [Izdatel'stvo RGGU], 2018. P. 791-802.

[18] Yanko T.E. The Russian *li* questions prosody [Prosodija voprosov s chasticej *li*], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2019' [Komp'juternaja lingvistika i intellektual'nye tehnologii. Po materialam ezhegodnoj Mezhdunarodnoj konferencii 'Dialog 2019']. Vyp. 18 (25), M.: M.: RGGU Publ. [Izdatel'stvo RGGU], 2019. P. 715-725.

[19] Yanko T.E. (in press) The illocutionary structure of the Russian questions: meaning and expression [Illokutivnaja struktura russkogo voprositel'nogo predlozhenija: znachenija i sredstva ih vyrazhenija], Proceedings of V. V. Vinogradov's Institute of the Russian language [Trudy Instituta russkogo jazyka im. V.V. Vinogradova].

Parallel corpus as a tool for semantic analysis: The Russian discourse marker *stalo byt'* (‘consequently’)

Anna A. Zalizniak
Institute of Linguistics of the RAS /
Moscow, B. Kislovskiy 1
anna.a.zalizniak@gmail.com

D. O. Dobrovol’skij
Russian Language Institute of the RAS /
Moscow, Volkhonka 18/2
Institute of Linguistics /
Moscow, B. Kislovskiy 1
Stockholm University /
10691 Stockholm Sweden
dobrovolskij@gmail.com

Abstract

The article examines the semantics of the Russian discourse marker *stalo byt’*, using the data obtained by analyzing translational correspondences extracted from parallel corpora of the Russian National Corpus (RNC). Typically, this discourse marker is an indicator of inferential evidentiality, by which the speaker marks the fact that the given statement is a conclusion made by the speaker on the basis of the information they received and accepted as true by default. In addition, *stalo byt’* has two secondary types of usage – “rhetorical” and “narrative” – where the basic semantics of this discourse marker is subject to certain modifications. One of the key points of analysis is the reconstruction of semantic mechanisms providing the actual semantics of *stalo byt’*.

Keywords: semantics, Russian, parallel corpus, translation, discourse markers, evidentiality, epistemic assessment, semantic shift

DOI: 10.28995/2075-7182-2023-22-566-578

Параллельный корпус как инструмент семантического анализа: русское *стало быть*[1]

Зализняк Анна А.
Институт языкознания РАН /
Москва, Большой Кисловский пер. 1
anna.a.zalizniak@gmail.com

Добровольский Д. О.
Институт русского языка РАН /
Москва, Волхонка 18/2
Институт языкознания РАН / Москва,
Большой Кисловский пер. 1
Стокгольмский университет /
10691 Стокгольм, Швеция
dobrovolskij@gmail.com

Аннотация

В статье исследуется семантика русского дискурсивного слова *стало быть* с использованием данных, полученных путем анализа переводных соответствий, извлекаемых из параллельных корпусов НКРЯ. Демонстрируется, что в своем основном типе употребления это слово представляет собой показатель инференциальной эвиденциальности, при помощи которого говорящий маркирует тот факт, что вводимое им утверждение – это умозаключение, сделанное им на основании полученной им информации и по умолчанию принимаемое за истинное. Кроме того, у *стало быть* имеется два производных типа употребления – «риторическое» и «нарративное», – где базовая семантика этого дискурсивного слова подвергается определенным модификациям. Также реконструируется путь формирования этой единицы и анализируются семантические механизмы, обеспечивающие ее актуальную семантику.

Ключевые слова: семантика, русский язык, параллельный корпус, перевод, дискурсивные слова, эвиденциальность, эпистемическая оценка, семантический переход

[1] Работа выполнена при поддержке РНФ, грант № 22-18-00586.

1 «Монофокусный» метод семантического анализа

Появление параллельных корпусов текстов открыло новые возможности семантического анализа. В частности, был разработан «монофокусный» метод контрастивного корпусного исследования, в рамках которого сопоставление оригинального текста с его переводом используется как инструмент выявления неочевидных компонентов значения той единицы одного из этих двух языков, которая является объектом анализа (находится «в фокусе»). Эти неочевидные компоненты значения могут быть обнаружены как среди вариантов («моделей») перевода интересующей нас единицы на другой язык, так и среди «стимулов» ее появления при переводе с другого языка. Такой метод корпусного анализа применялся в течение последних лет для исследования семантики ряда языковых единиц, в том числе, дискурсивных слов, в работах [Сичинава 2014; Зализняк 2015; Шмелев 2015; Добровольский, Зализняк 2021, 2022; Добровольский, Левонтина 2015; Dobrovol'skij, Pöppel 2022] и др.; наиболее эффективным он оказывается при исследовании «лингвоспецифичных» единиц (т.е. таких, которые не имеют однозначного переводного эквивалента). К таким единицам безусловно относится русское дискурсивное слово *стало быть*[2].

Очевидно, что использование того или иного эквивалента – это выбор переводчика, который может оказаться более или менее удачным. Однако, если исключить из рассмотрения (достаточно редкие) случаи откровенно ошибочного перевода, все случаи неточного перевода, при котором какие-то компоненты значения оригинала утрачиваются или, наоборот, возникают лишние, с точки зрения анализа интересующей нас единицы оригинала представляют не меньшую, а возможно и большую ценность, чем случаи достаточного точной передачи ее смысла, поскольку позволяют эксплицировать скрытые семантические компоненты.

В данной статье излагаются результаты применения этого метода к анализу русского дискурсивного слова *стало быть*. Эта единица представляет интерес, во-первых, потому что она имеет достаточно сложную семантику, которая до сих пор не была предметом специального анализа, во-вторых, потому что единица с аналогичной внутренней формой отсутствует в других – по крайней мере европейских – языках, в том числе в близкородственных славянских.

В словаре МАС [IV, 255] словосочетание *стало быть* помещено под ромбом в статье «стать» с толкованием-перифразой *значит, следовательно*. Действительно, основным для этой единицы следует считать значение умозаключения, которое делает говорящий на основании каких-то данных. В РГ-80 *стало быть* названо «союзным аналогом», оформляющим причинно-следственные отношения – наряду с такими словами как *потому* (*и*), *поэтому, значит, следовательно, итак, таким* образом (§ 3152), а также «специализированным коррелятом» условного союза, который называет следствие, вывод, умозаключение – наряду со словами *следовательно, значит, выходит* и *знать* (§ 3018-3020). В словаре [Морковкин (ред.) 2022] у *стало быть* различаются три значения на основании его синтаксической функции: вводное слово, союз и соотносительное слово.

Однако, как семантика дискурсивной единицы *стало быть*, так и прагматические условия ее появления требуют дальнейшего исследования. В частности, само выражаемое ею значение умозаключения весьма специфично, и эта специфика может быть выявлена при помощи, в том числе, анализа типов переводных эквивалентов.

Анализ проводился на материале немецкого, английского, французского, итальянского и испанского параллельных корпусов НКРЯ для обоих направлений перевода[3]; частично использовался также материал ряда славянских параллельных корпусов и основного корпуса НКРЯ.

[2] В соответствии с определением термина «дискурсивное слово», введенным в оборот в работах группы Д. Пайара (см. [Баранов и др. 1993; Киселева, Пайар (ред.) 1998]) и ставшим общепринятым, данная категория включает в том числе неоднословные единицы.

[3] Немецкий подкорпус (1245 вхождений *стало быть*), английский (1193), итальянский (245), французский (234), испанский (83); всего – 3000 примеров.

2 *Стало быть*: путь формирования дискурсивного слова

До середины XVIII в. словосочетание *стало быть* употребляется как свободное, и означает ‘начало иметь место <некоторое положение дел>’; встречается в контексте словоформы, входящей в состав сказуемого (ср. (1), (2))[4]:

(1) И как немного от болезни ***стало быть*** *лехче*, не дожидаясь совершенного выздоровления, съехал на галеру. [А. М. Макаров (ред.). Гистория Свейской войны (Поденная записка Петра Великого) (1698-1721)]

(2) По притчинѣ такой перемѣны полковникъ, которой при насъ провожатымъ былъ, предлагалъ намъ, что водою ѣхать ***стало быть*** *не можно*, и для того надобно пообождать сухаго пути <…> [А. Л. Леонтьев. Путешествие китайского посланника к калмыцкому Аюке-хану (1762)]

Значение умозаключения появляется в конце XVIII в., когда выражение *стало быть* начинает употребляться в контексте самостоятельной предикативной единицы, в результате чего оно выходит из состава предложения и становится вводным ср.:

(3) Естьли они меня погубят, они в том и будут отвечать богу, и коли они променяли деньги на человека, то, ***стало быть***, *им деньги милее дочери* <…> [П. А. Плавильщиков. Бобыль (1790)]

После 1800 г. на употребление словосочетания *стало быть* в старом значении ‘начало иметь место <некоторое положение дел>’ в НКРЯ имеются лишь единичные примеры. В подавляющем большинстве случаев *стало быть* употребляется в новом значении умозаключения, ср.:

(4) Поблагодарили А. П. Бунину за ея ко мнѣ письмо и стихи; скажи ей, что я всегда то сдѣлаю, что могу, и если не сдѣлаю, такъ ***стало быть*** *невозможно было*. [А. С. Шишков. Письма жене (1813-1814)]

В современном дискурсивном слове *стало быть* от первоначального *стало быть P*, т.е. ‘наступило новое состояние P’ сохраняется компонент ‘переход в новое состояние’: у говорящего возникло некоторое новое представление – результат интерпретации только что полученной информации. Идея наступления новой ситуации в мире трансформируется в идею возникновения нового представления о мире в сознании говорящего (т.е. происходит «субъектификация» по [Traugott 1982]). Это представление не является в полном смысле знанием, но является убеждением, которое по умолчанию говорящий готов принять за истинное.

Путь возникновения компонента логического следствия в семантической эволюции *стало быть* от значения ‘перехода в состояние’ к современному значению умозаключения может быть реконструирован следующим образом. Исходное значение указывает на временную последовательность ситуаций: имело место ситуация S1, в некоторый момент наступила ситуация S2. В ходе диахронической семантической эволюции отношение последовательности во времени преобразуется в отношение логического следствия, устанавливаемого говорящим (очевидно, по известному принципу post hoc ergo propter hoc), ср. [Traugott 1982: 258] о переходе значения следования во времени в значение логического следствия в англ. *since, so, then, thence, hence, therefore, consequently* (ср. рус. *следовательно*), а также [Traugott, König 1991; Heine, Kuteva 2002: 275] об англ. *since*. Тот же процесс имел место в рус. *поэтому*, исп. *pues*, и др., ср. семантический переход ID7461 в Базе данных семантических переходов (www.datsemshift.ru).

Другим источником значения умозаключения являются указательные наречия с исходным значением образа действия (ср. рус. *таким образом*); см. семантический переход ID7464. Такого рода словосочетания являются одним из типов «стимулов» появления *стало быть* в русских переводах.

[4] В примерах *стало быть* и его переводной эквивалент выделяются п/ж курсивом, а вводимый *стало быть* фрагмент – светлым курсивом.

Согласно статистике НКРЯ[5], частота употребления дискурсивного слова *стало быть* в основном корпусе начинает резко возрастать в 30-е годы XIX в., достигает максимума около 1870 г. (ок.160 ipm) и далее начинает постепенно убывать; после 2000 г. скорость убывания несколько увеличивается, и к концу периода 2000–2021 частотность *стало быть* оказывается ок. 1 ipm (см. рис. 1). В устном корпусе пик частотности приходится на середину XX в. (ок. 132 ipm), после 1980 г. она начинает убывать, и к концу 2010-х гг. она оказывается также ок. 1 ipm (см. рис. 2). Скорость убывания частотности *стало быть* в разговорной речи в XXI в. возрастает; так, из 104 вхождений в период 2000–2019 на период 2010–2019 приходится всего 12, из них девять принадлежат людям старше 50-ти, два примера – это исполнение произведений XIX в., и лишь один принадлежит 34-летнему человеку, и это «нарративное» (см. ниже) *стало быть*. Аналогичную тенденцию демонстрируют также данные основного корпуса НКРЯ за период 2000–2021 (см. рис. 3). Приведенные данные свидетельствуют о том, что на сегодня единица *стало быть* является одновременно разговорной и уходящей, т.е. она присутствует преимущественно в разговорной речи старшего поколения, а также в литературных текстах, воспроизводящих речь разных эпох.

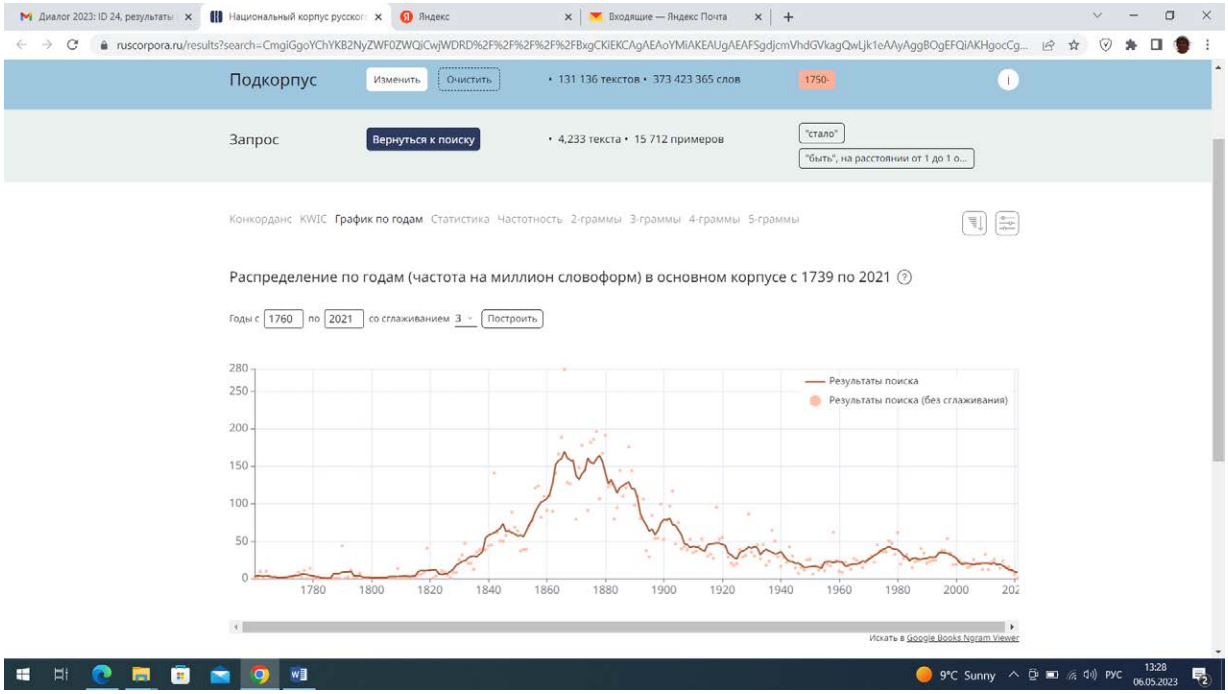


Рис. 1: График распределения *стало быть* по годам
в основном корпусе НКРЯ за период с 1760 по 2021 гг.

[5] Дата просмотра 05.05.2023.

Рис. 2: График распределения *стало быть* по годам
в устном корпусе НКРЯ за период с 1930 по 2019 гг.

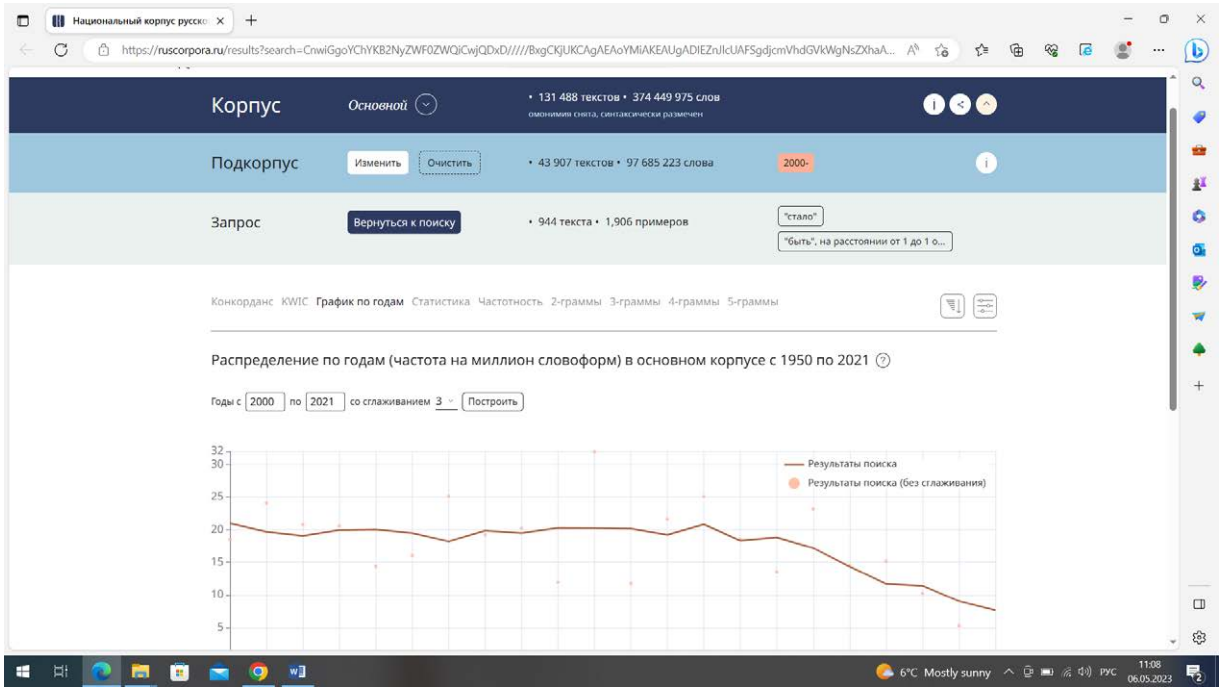


Рис. 3: График распределения *стало быть* по годам
в основном корпусе НКРЯ за период с 1950 по 2021 гг.

3 *Стало быть* как показатель инференциальной эвиденциальности

Наш главный тезис состоит в том, что стало быть является в русском языке показателем инференциальной эвиденциальности (inferential evidentiality, по [van der Auwera, Plungian 1998: 85]: “the subtype that identifies the evidence as based upon reasoning”; см. также [Aikhenvald 2004] и др.), причем особого рода. В своем основном типе употребления это дискурсивное слово вводит

утверждение, которое представляет собой умозаключение, сделанное говорящим на основании каких-то данных и по умолчанию принимается им за истинное. Этим *стало быть* отличается от слов *видимо, по-видимому* и *похоже* (также обычно относимых к показателям инференциальной эвиденциальности, ср. напр. [Козинцева 2007]), представляющих собой вывод в форме предположения, которое может оказаться неверным[6]. *Стало быть* отчетливо противопоставлено также показателю эпистемической необходимости *должно быть*, который также вводит предположение, обладающее большой степенью уверенности. Ср. (5) и (6):

(5) Он не подходит к телефону – ***должно быть**, он спит.*

(6) Он не подходит к телефону – ***стало быть**, он спит.*

В первом случае это предположение (которое может оказаться неверным), во втором – окончательное суждение, принимаемое говорящим за истинное. Второе предложение уместно только в том случае, если говорящий откуда-то знает, что то, что он спит, является единственной возможной причиной того, что он не подходит к телефону. И наоборот, в предложении (7) (по мотивам известного эпизода из «Семнадцати мгновений весны») неуместно было бы употребление *должно быть*, поскольку говорящему известно наличие безусловной связи между этими двумя ситуациями.

(7) На подоконнике стоит цветок – ***стало быть**, явочная квартира обнаружена.*

В этом отношении *стало быть* сближается со словами *значит* (в наибольшей степени), а также *следовательно*, отличаясь от последнего, в частности, менее рациональным характером процедуры вывода.

Напомним, что умозаключение вида 'имеет место наблюдаемый факт А, следовательно имеет место ненаблюдаемый факт В' может представлять собой обращение причинно-следственного отношения 'факт В является причиной факта А', т.е. ментальную операцию, восстанавливающую ненаблюдаемую причину по ее наблюдаемому следствию[7]. Так, в примере (6) факт 'он спит' реально является причиной факта 'он не подходит к телефону'. При этом как *стало быть*, так и *значит* (и, безусловно, *следовательно*) могут вводить также следствие, ср.: *Деньги кончились, следовательно/стало быть/значит мы никуда не едем*: отсутствие денег – причина, отказ от поездки – следствие. Ср. также *Год был засушливый, следовательно/стало быть/значит урожай будет плохой*, и т.п. Но *стало быть* и *значит* могут вводить также восстанавливаемую причину – в отличие от *следовательно*, для которого такое употребление не то что исключено, но нехарактерно, ср.: *Он не подходит к телефону, стало быть/значит/?следовательно он спит*. Возможность такой перестановки причины и следствия характерна для большинства коннекторов типа нем. *also*, фр. *donc*, итал. *dunque, quindi*, исп. *pues, entonces* и т.п., выступающих как наиболее частотные (ок. 55 %) переводные эквиваленты для русского *стало быть*.

Однако процедура умозаключения не обязательно опирается на причинно-следственное отношение: это могут быть просто два обстоятельства, которые связаны между собой таким образом, что имеют место одновременно – в силу устройства обсуждаемого фрагмента мира или какой-то конвенции, ср. *Сегодня воскресенье, стало быть/значит/следовательно магазин закрыт.* Это могут быть также два разных способа назвать одно и то же или интерпретировать некоторый наблюдаемый факт; в этом случае могут появляться эквиваленты типа 'то есть', 'иначе говоря', 'означать', ср. (8)–(10):

[6] Различие между *стало быть,* с одной стороны, и *видимо, по-видимому, похоже* – с другой, состоит, во-первых, в степени уверенности говорящего во водимым этими словами утверждении. Во-вторых, *стало быть*, в отличие от остальных вышеупомянутых слов, вводит умозаключение, сделанное только что: это всегда новая для говорящего информация. В-третьих, в *стало быть* компонент умозаключения находится в фокусе. При этом основанием для вывода для всех обсуждаемых слов могут служить данные любого рода, не обязательно зрительные, см. подробнее [Зализняк 2021].

[7] В этом случае такое наблюдаемое (реальное, онтологическое) следствие выступает в роли «логической причины», ср. [Богуславская, Левонтина 2004: 86], а также противопоставление «реальная vs. логическая каузация» в [Иорданская, Мельчук 1996: 200-201].

(8) Все для вас, для крестьян; ***стало быть***, *и для тебя*. [Иван Гончаров. Обломов (1849-1858)]
Tutto per voi, per i contadini; ***cioè*** *anche per te*. [Ivan Goncarov. Oblomov (Argia Michettoni)]

(9) Сам Т. Манн в истории создания своего романа говорит об этом так: «Побольше шутливости, ужимок биографа <…>, ***стало быть***, *глумления над самим собой*, чтобы не впасть в патетику <…>». [Михаил Бахтин. Проблемы поэтики Достоевского (1963)]
Th. Mann lui-même parle ainsi de la genèse de son roman: “Il faut davantage de plaisanteries, de grimaces de la part du biographe <…>, ***autrement dit***, *de persiflage de soi-même* pour ne pas tomber dans l'emphase <…>.” [Mikhaïl Bakhtine. La poétique de Dostoïevski (Isabelle Kolitcheff, 1970)]

(10) Verlangt hundert Dollar dafür. ***Das heißt***, *er gibt es für achtzig*. Scheint mir billig zu sein. [Erich Maria Remarque. Schatten im Paradies (1965-1970]
Просит за нее сто долларов. Отдаст, ***стало быть***, *за восемьдесят*. По-моему, дешево. [Эрих Мария Ремарк. Тени в раю (Л. Б. Черная, В. Котелкин, 1971)]

При этом вывод, формулируемый при помощи *стало быть*, все же не обязательно является полноценным утверждением; в нем может присутствовать элемент неуверенности, поскольку, как говорящий отчасти сам понимает, знание им фактов, на которые он опирается, может быть неполным, да и сама логика умозаключения может оказаться ошибочной. О неполной «надежности» утверждения, вводимого *стало быть*, свидетельствует возможность перевода модальными наречиями и эпистемическим ‘должен’, ср.:

(11) Оружие в розыске не числилось, ***стало быть***, *до этих убийств при совершении преступлений не использовалось*. [Александра Маринина. Шестерки умирают первыми (1995)]
Die Waffe befand sich nicht in der Sachfahndung, ***offenbar*** *war sie, bevor die Morde begangen wurden, noch nicht zu kriminellen Zwecken benutzt worden*. [Alexandra Marinina. Mit verdeckten Karten (Natascha Wodin, 2003)]

(12) – Какого убитого? – спросил человек и поглядел исподлобья… – Тут вот на улице, три дня, как его убили… – Ага, ***стало быть***, *юнкер или офицер*… [Михаил Булгаков. Белая гвардия (1924)]
– Quel mort? demanda l'homme en. les regardant par en dessous. – Dans la rue, tout près d'ici, il a été tué il y a trois jours… – Ah! ah! *Un junker ou un officier* ***probablement***… [Mikhail Boulgakov. La Garde Blanche (Claude Ligny, 1970)]

(13) А уж если она повышает голос, ***стало быть***, *пожар полыхает вовсю*. [Александра Маринина. Шестерки умирают первыми (1995)]
Und wenn sie anfing zu schreien***, dann mußte*** *höchste Alarmstufe angesagt sein*. [Alexandra Marinina. Mit verdeckten Karten (Natascha Wodin, 2003)]

(14) Зато если тебе скажут все, что ты просишь, ***стало быть***, *все в порядке*, твой адрес на контроле не находится. [Александра Маринина. Шестерки умирают первыми (1995)]
Wenn man dir die Auskunft gibt, ist ***höchstwahrscheinlich*** *alles in Ordnung,* dann brauchen wir uns keine Sorgen mehr zu machen. [Alexandra Marinina. Mit verdeckten Karten (Natascha Wodin, 2003)]

Отдельно отметим характерное использование – в качестве как «стимула», так и «модели» перевода – нем. частицы *wohl* с исходным значением ‘хорошо’, что также свидетельствует о субъективном – и, тем самым, не полностью надежном характере эпистемической оценки, выражаемой *стало быть*. Ср.:

(15) *Eure Mutter versteht* ***wohl*** *keine zu backen*? [Erwin Strittmatter. Tinko (1954)]
Стало быть, *твоя мать их печь не умеет*? [Эрвин Штриттматтер. Тинко (Вс. Розанов, 1956)]

(16) Куда ж вы? ***Стало быть***, *нет дома чаю*? [Ф. М. Достоевский. Бесы (1872)]
Wo wollen Sie denn hin? *Sie haben* ***wohl*** *keinen Tee im Hause*? [Fëdor Dostoevskij. Die Dämonen (Hermann Röhl, 1920)]

Именно эта «эпистемическая неопределенность», возможность колебания между полной и неполной уверенностью определяет лингвоспецифичность русского *стало быть*.

4 Типы употребления *стало быть* в современном языке

Мы предлагаем различать три основных типа употребления *стало быть*, которые обозначим условно как «основное», «риторическое» и «нарративное».

1. **Основное**: показатель инференциальной эвиденциальности, т.е. это вывод, который делает говорящий из известных фактов и который он принимает, с определенной долей сомнения, за истинный.

1.1. В утвердительном предложении в монологе или диалоге. Рассмотрим следующий пример.

(17) Однако неожиданно возле него столкнулись две женщины, и одна из них, востроносая и простоволосая, закричала над самым ухом поэта другой женщине так: – Аннушка, наша Аннушка! С Садовой! Это ее работа! Взяла она в бакалее подсолнечного масла, да литровку-то о вертушку и разбей! Всю юбку изгадила… Уж она ругалась, ругалась! А он-то, бедный, ***стало быть*** *поскользнулся да и поехал на рельсы*… [Михаил Булгаков. Мастер и Маргарита (1929-1940)]

При переводе этого фрагмента на другие языки реализуются две разные стратегии. Во французском и испанском переводах модальный оператор вообще опущен («модель перевода» – ZERO): в переводе отражен лишь компонент уверенности говорящего в истинности вывода, т.е. второе утверждение представлено в виде факта, ср.:

(18) <…> Et l'autre, le malheureux, **ZERO** *il a glissé là-dessus* et il s'est retrouvé sur les rails… [Mikhaïl Boulgakov. Le Maître et Marguerite (p 1) (Claude Ligny, 1968)]

(19) <…>¡ y ese pobrecito que **ZERO** *se resbala* y a la vía…! [Mijaíl Bulgákov. El maestro y Margarita (Amaya Lacasa Sancha, 1967)]

А в английском, немецком и итальянском переводах использован показатель вероятностной оценки: предикат 'должен' в эпистемическом значении или модальное наречие 'видно', т.е. отражен тот факт, что второе утверждение представляет собой всего лишь умозаключение говорящего, которое при взгляде со стороны может оказаться неверным, ср.:

(20) <…> And he, poor man, ***must*** *have slipped* and – right on to the rails… [Mikhail Bulgakov. Master and Margarita (Richard Pevear, Larissa Volokhonsky, 1979)]

(21) <…> Der arme Kerl ***muß*** *ausgerutscht sein* und auf die Schienen gefallen. [Michail Bulgakow. Der Meister und Margarita (Thomas Reschke, 1968)]

(22) <…> E lui, poverino, ***si vede che*** *è scivolato* ed è andato a finire sulle rotaie… [Mikhail Bulgakov. Il Maestro e Margherita (p 1) (Vera Dridso, 1967)]

Что Аннушка разлила подсолнечное масло – это известный говорящему факт (то, что она испачкала юбку и как ругалась – говорящий, очевидно, наблюдал). Что Берлиоз *поскользнулся* на этом масле *и поэтому* упал на рельсы – вывод, сделанный говорящим и представляющийся ему истинным – что маркируется в русском оригинале словом *стало быть*.

В переводах этого предложения на славянские языки в основном использованы те же две стратегии: в болгарском – ZERO, в польском, чешском словенском и сербском – модальные наречия эпистемической оценки ('вероятно'); в белорусском и украинском использованы наиболее близкие к русскому *стало быть* показатели умозаключения.

1.2. В вопросе, требующем ответа: говорящий выражает свою уверенность в правильности сделанного вывода на основании имеющихся в его распоряжении данных, но допускает их неполноту; задавая вопрос, говорящий ожидает от собеседника, который, по его предположению,

обладает большей информацией, подтверждения правильности сделанного вывода (при этом по форме предложение со *стало быть* по форме может быть как вопросительным, так и утвердительным)[8]. Так, в примере (23) маркируя при помощи *стало быть* вывод, сделанный на основании информации, полученной из предшествующей реплики диалога (ср. «жара африканская, редкая в наших широтах»), говорящий дает понять, что не полностью уверен в правильности сделанного им вывода (ср. «если не ошибаюсь»), и предлагает собеседнику подтвердить этот вывод.

(23) – Притом жара африканская, редкая в наших широтах. <…> – ***Стало быть*** *сами из России будете*, если не ошибаюсь. – Из Белокаменной. [Борис Пастернак. Доктор Живаго (1945-1955)]

Соответственно, *стало быть* может появляться в переводе на месте показателей эпистемической оценки, ср. 'я полагаю' в (24), форма конъюнктива в (25), модальный глагол в эпистемическом значении в (26):

(24) – Deberá disculparle, no conoce a nadie. – *Su guardaespaldas,* ***supongo***. – En efecto. [Eduardo Mendoza. La verdad sobre el caso Savolta (1975)]
– Вы должны извинить его, он пока никого не знает. – ***Стало быть***, *это ваш телохранитель*? – Вот именно. [Эдуардо Мендоса. Правда о деле Саволты (Николай Любимов, 1985)]

(25) Cuestión de celos, probablemente. – ¿O ***sea*** *que hay otro*? – Digo yo… [Eduardo Mendoza. La verdad sobre el caso Savolta (1975)]
Из ревности, наверное. – ***Стало быть***, *есть кто-то еще*? – Возможно. [Эдуардо Мендоса. Правда о деле Саволты (Николай Любимов, 1985)]

(26) „Meine Stadtwohnung." „Dann ***dürfte*** der *Ihr Autoschlüssel sein*?" „So ist es." [Siegfried Lenz. Fundbüro (2003)]
– От моей городской квартиры. – А этот, ***стало быть***, *от машины*? – Так оно и есть. [Зигфрид Ленц. Бюро находок (Г. М. Косарик, 2004)]

В вопросе, требующем ответа, на месте *стало быть* одновременно с показателем умозаключения могут появляться показатели эпистемической оценки, ср.:

(27) Ihr habt ***also wahrscheinlich*** *das Haus selber gebaut* oder es sehr umgestaltet? [Adalbert Stifter. Der Nachsommer (1857)]
Вы, ***стало быть***, *сами и построили дом* или изрядно его перестроили? [Адальберт Штифтер. Бабье лето (С. К. Апт, 1999)]

(28) – Позвольте! – вдруг воскликнула она, – какого Берлиоза? это, что в газетах сегодня… – Как же, как же… – Так *это,* ***стало быть***, *литераторы за гробом идут*? – спросила Маргарита и вдруг оскалилась. – Ну, натурально, они! [Михаил Булгаков. Мастер и Маргарита (ч. 2) (1929-1940)]
<…> – Mais ***alors***, *ce sont* ***sans doute*** *des écrivains qui suivent son enterrement*? <…> [Mikhaïl Boulgakov. Le Maître et Marguerite (p 2) (Claude Ligny, 1968)]

Появление в переводе, помимо показателей вывода (типа нем. *also* и фр. *alors*), еще и показателей эпистемической оценки (типа нем. *wahrscheinlich* и фр. *sans doute*) является свидетельством того, что русское выражение *стало быть* включает компонент неуверенности в правильности сделанного вывода.

[8] *Стало быть* обычно употребляется в общем вопросе, представляющем собой запрос на подтверждение высказанной говорящим гипотезы. В тех редких случаях, когда оно использовано в частновопросительном предложении (т.е. содержащем вопросительное слово, ср.: *– Стало быть, сколько дней, вы полагаете, не тронется ваша икра? – спросил он.* (В.А. Каверин. Открытая книга)) *стало быть* указывает на умозаключение, которое говорящий предлагает сделать собеседнику.

Специально отметим, что вводимое *стало быть* умозаключение обычно делается на основании данных, полученных *только что*: из предшествующей реплики собеседника в диалоге (ср. примеры выше) или из собственного наблюдения (ср. (29)).

(29) Тут Маргарита замерла, потому что узнала вдруг этого Майгеля. Он несколько раз попадался ей в театрах Москвы и в ресторанах. «Позвольте… – подумала Маргарита, – он, ***стало быть***, *что ли, тоже умер*?» [Михаил Булгаков. Мастер и Маргарита (ч. 2) (1929-1940)]

2. **Риторическое**: когда говорящий только *делает вид* что вводимое при помощи *стало быть* высказывание – это результат его умозаключения. Причины использования такой дискурсивной стратегии могут быть различными. В частности, этот эффект возникает в ответе на вопрос, представляющем собой безоговорочное подтверждение высказанной собеседником гипотезы; при этом говорящий *как бы* ссылается на то, что это из чего-то с необходимостью следует.

(30) – На нее напоролися? – чмокал Перхуша. – ***Стало быть***, *на нее*. [Владимир Сорокин. Метель (2010)]

Такое риторическое *стало быть* может быть, в частности, использовано при отказе отвечать на вопрос об основании высказанного только что мнения, ср.:

(31) – Неделю у нас прожил; кроме хорошего, ничего от него не видали, – сказала она. – Обходительный, умный, справедливый. – Почем вы это все узнали? – ***Стало быть***, *узнала*. [Л. Н. Толстой. Хаджи-Мурат (1896-1904]

Риторическое *стало быть* может также употребляться для выражения иронии, в качестве маркера условного принятия на веру сделанного собеседником утверждения, ср. появляющиеся в последующем тексте *допустим и это* и *предположим даже* в примере (32); ср. также характеристику «иронически» в (33):

(32) – Ну, хорошо, – ответил мастер, – ведьма так ведьма. Очень славно и роскошно! Меня, ***стало быть***, *похитили из лечебницы*! тоже очень мило. Вернули сюда, допустим и это… Предположим даже, что нас не хватятся, но скажи ты мне ради всего святого, чем и как мы будем жить? [Михаил Булгаков. Мастер и Маргарита (ч. 2) (1929-1940)]

(33) – Ты серьезно уверена в том, что мы вчера были у сатаны? – Совершенно серьезно, – ответила Маргарита. – Конечно, конечно, – иронически заметил мастер, – теперь, ***стало быть***, *налицо вместо одного сумасшедшего двое*! И муж и жена. [Михаил Булгаков. Мастер и Маргарита (ч. 2) (1929-1940]

3. **Нарративное**: *стало быть* как маркер возвращения к прерванному сюжету, используя который говорящий дает понять, что дальнейший рассказ как бы следует из сказанного ранее, *как если бы* сказанное ранее было причиной того, о чем идет речь сейчас. Так, в примере (34) *стало быть* в этой функции маркера возврата к прерванной собеседником линии изложения говорящим своего сюжета появляется в русском переводе на месте *del resto, dovete sapere* (что-то вроде *а еще вы должны знать следующее*).

(34) И какие огромные куски они отхватывали! Я никогда не думал, что рыбы еще прожорливее, чем маленькие мальчики… Они съели мою морду, мою шею и гриву, мою кожу на ногах, мою шкуру на спине <…> – Отныне, – сказал покупатель с отвращением, – я, с божьей помощью, никогда не буду есть рыбы! <…> – Я вполне разделяю ваше мнение, – ответил Деревянный Человечек и засмеялся. – ***Стало быть***, когда рыбы съели ослиную кожу, в которую я был обернут с головы до ног, они, натурально, *наткнулись на кости*… [Carlo Collodi. Pinocchio (1883) | Карло Коллоди. Приключения Пиноккио (Э. Казакевич, 1959)]

В (35) героиня говорит о приходе красных, потом отвлекается на историю своей семьи (выделено фигурными скобками); возвращение к прерванному рассказу маркируется дискурсивным словом *ну вот*, а *стало быть* использовано в качестве дополнительного маркера той же дискурсивной функции.

(35) Когда стали мы, то есть, наши красные, к ихнему главному городу белому подходить, этот самый Комаров министр посадил маменьку со всей ихнею семьей в особенный поезд литерный и приказали увезть, ведь маменька были пуганые и без них не смели шагу ступить. {А про меня он даже не знал, Комаров. Не знал, что я такая есть на свете. Маменька меня в долгой отлучке произвели и смертью обмирали, как бы кто об том ему не проболтался. Он ужасть как того не любил, чтобы дети, и кричал и топал ногами, что это одна грязь в доме и беспокойство. Я, кричал, этого терпеть не могу.} Ну вот, ***стало быть***, как стали подходить красные, *послали маменька за сторожихой Марфой на разъезд Нагорную*, это от того города в трех перегонах. [Борис Пастернак. Доктор Живаго (1945-1955)]

Дальнейший путь семантической эволюции нарративного *стало быть* – десемантизация, «выветривание», аналогичное тому, которое претерпело превратившееся в «слово-паразит» слово *значит*, с той же исходной семантикой вывода на основании каких-то данных.

5 Заключение

Итак, русское *стало быть* в своем основном типе употребления представляет собой показатель инференциальной эвиденциальности: при помощи этого слова говорящий маркирует тот факт, что вводимое им утверждение – это умозаключение, сделанное им на основании какой-то (полученной только что) информации. Это умозаключение принимается говорящим за истинное – с возможной оговоркой, обусловленной тем, что полной уверенности в истинности информации, послужившей источником сделанного умозаключения, у него может не быть; в этом случае говорящий обычно использует *стало быть* в вопросительном высказывании, имеющем целью получить от собеседника подтверждение правильности своего вывода. Таким образом, в рамках основного типа употребления *стало быть* осциллирует между убежденностью и сомнением. Наличие именно этих двух составляющих значения *стало быть* подтверждается проведенным анализом его переводных эквивалентов в пяти языках, которые распределяются следующим образом: ок. 60% использованных профессиональными переводчиками эквивалентов составляют языковые единицы, маркирующие умозаключение (прежде всего, коннекторы, но также некоторые другие средства выражения идеи логического следования, в том числе, вставка причинного союза в предшествующую клаузу, а также единицы с исходным значением ‘так, таким образом’ и др.), ок. 7% составляют единицы со значением (высокой) оценки вероятности – модальные наречия и модальные глаголы в эпистемическом значении, ок. 6% – языковые единицы со значением «передаваемого смысла» или тождества, ок. 2% приходится на различного рода описательные конструкции и другие средства, выбранные переводчиком для передачи значения *стало быть* в определенном контексте, что также свидетельствует о лингвоспецифичности обсуждаемого русского дискурсивного слова. И, наконец, примерно в 25% случаев обнаруживается «нулевая» эквивалентность, т.е. русское *стало быть* опускается в переводе с русского или, наоборот, возникает как бы «из ничего» в переводе на русский. В таких случаях идея следования/умозаключения либо остается невыраженной, либо наоборот эксплицируется.

Кроме того, у *стало быть* имеется два производных типа употребления, которые мы обозначили как «риторическое» и «нарративное», где базовая семантика этого дискурсивного слова подвергается определенным модификациям.

Благодарности

Авторы благодарят анонимных рецензентов за конструктивные замечания, которые были по возможности учтены в окончательной версии статьи.

Литература

[1] Баранов А.Н., Плунгян В.А., Рахилина Е.В. Путеводитель по дискурсивным словам русского языка. — М.: Помовский и партнеры, 1993.
[2] Богуславская О.Ю., Левонтина И.Б. Смыслы 'причина' и 'цель' в естественном языке // Вопросы языкознания, 2004. №2, с. 68–88.
[3] Добровольский Д. О., Зализняк Анна А. Параллельный корпус как инструмент семантического анализа: немецкий модальный глагол *sollen* // Труды международной конференции «Корпусная лингвистика-2021». — СПб.: Скифия-принт, 2021, с. 209–218.
[4] Добровольский Д. О., Зализняк Анна А. Эвиденциальность и эпистемическая оценка в значении немецких глаголов *sollen* и *wollen* (по данным немецко-русского параллельного корпуса) // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог 2022». Вып. 21 (28). — М.: РГГУ, 2022, с. 132–140.
[5] Добровольский Д.О., Левонтина И.Б. Модальные частицы и идея актуализации забытого (на материале параллельных корпусов) // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2015». Вып. 14 (21). — М.: РГГУ, 2015, с. 106–117.
[6] Зализняк Анна А. Лингвоспецифичные единицы русского языка в свете контрастивного корпусного анализа // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2015». Вып. 14 (21). — М.: РГГУ, 2015, с. 651–662.
[7] Зализняк Анна А. Дискурсивные слова *видимо* и *по-видимому*: актуальная и диахроническая семантика. // Компьютерная лингвистика и интеллектуальные технологии. По материалам международной конференции «Диалог». Вып. 20 (27). — М.: РГГУ, 2021, с. 720–728.
[8] Иорданская Л.Н., Мельчук И.А. К семантике русских причинных предлогов // Московский лингвистический журнал. Т. 2. — М.: РГГУ, 1996.
[9] Киселева К., Пайар Д. (ред.). Дискурсивные слова русского языка. Опыт контекстно-семантического описания. Сб. статей. — М.: Метатекст, 1998.
[10] Козинцева Н.А. Косвенный источник информации в высказывании (на материале русского языка) // Храковский В.С. (ред.) Эвиденциальность в языках Европы и Азии. — СПб.: Наука, 2007, с. 37–46.
[11] МАС — «Малый академический словарь» = Словарь русского языка в четырех томах. 3-е, стереотип. изд. — М.: Русский язык, 1985–1988.
[12] Морковкин В.В. (ред.), Большой универсальный словарь русского языка. В 2-х томах. — М.: АСТ-Пресс, 2022.
[13] РГ-80 — Русская грамматика: В 2-х т. / Гл. ред. Н.Ю. Шведова. — М.: Наука, 1980.
[14] Сичинава Д.В. Использование параллельного корпуса для количественного изучения лингвоспецифичной лексики // Язык, литература, культура: Актуальные проблемы изучения и преподавания. Вып. 10. — М.: Макс Пресс, 2014, с. 37–44.
[15] Шмелев А.Д. Русские лингвоспецифичные лексические единицы в параллельных корпусах: возможности исследования и «подводные камни» // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог 2015» Вып. 14 (21). — М.: РГГУ, 2015.
[16] Aikhenvald, A.Y. Evidentiality. — Oxford: Oxford University Press, 2004.
[17] Dobrovol'skij D., Pöppel L. Russian constructions with *nu i* in parallel corpora // Mellado Blanco, Carmen (ed.). Productive Patterns in Phraseology and Construction Grammar. A Multilingual Approach. — Berlin, Boston: de Gruyter, 2022, pp. 191–213.
[18] Heine B., Kuteva T. World Lexicon of Grammaticalization. — Cambridge: Cambridge Univ. Press, 2002.
[19] Traugott E.C. From propositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization // W.P. Lehmann, Y. Malkiel (eds.). Perspectives on Historical Linguistics. — Amsterdam, Philadelphia, 1982, pp. 245–271.
[20] Traugott E.C., König E. The semantics-pragmatics of grammaticalization revisited // E.C. Traugott, B. Heine (eds.). Typological Studies in Language 19: 1991, pp. 189–218.
[21] van der Auwera J., Plungian V. Modality's semantic map // Linguistic Typology, 2, 1998, pp. 79–124.

References

[1] Aikhenvald, A.Y. Evidentiality. — Oxford: Oxford University Press, 2004.

[2] Baranov A.N., Plungian V.A., Rakhilina E.V. Guide to Russian discourse words [Putevoditel' po diskursivnym slovam russkogo yazyka]. — Moscow: Pomowski & Partner, 1993.

[3] Boguslavskaya O.Yu., Levontina I.B. The meanings of 'cause' and 'goal' in natural language [Smysly 'prichina' i 'tsel'' v estestvennom yazyke] // Voprosy jazykoznanija, 2004. No. 2, pp. 68–88.

[4] Dobrovol'skij D.O., Levontina I.B. Modal particles and the actualization of forgotten details (based on the materials of parallel corpora) [Modal'nye chastitsy i ideya aktualizatsii zabytogo (na materiale parallel'nykh korpusov)] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2015". Issue. 14 (21) [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog 2015"]. — Moscow: RGGU, 2015, pp. 106–117.

[5] Dobrovol'skij D., Pöppel L. Russian constructions with *nu i* in parallel corpora // Mellado Blanco, Carmen (ed.). Productive Patterns in Phraseology and Construction Grammar. A Multilingual Approach. — Berlin, Boston: de Gruyter, 2022, pp. 191–213.

[6] Dobrovol'skij D. O., Zalizniak Anna A. Parallel corpus as a tool for semantic analysis: German modal verb *sollen* [Parallel'nyj korpus kak instrument semanticheskogo analiza: nemetskiy modal'nyj glagol *sollen*] // Proceedings of the International Conference "Corpus Linguistics-2021" [Trudy mezhdunarodnoy konferentsii "Korpusnaya lingvistika-2021"]. — St. Petersburg: Scythia-print, 2021, pp. 209–218.

[7] Dobrovol'skij D. O., Zalizniak Anna A. Evidentiality and epistemic modality in the semantics of the German verbs *sollen* and *wollen* (based on the data from the German-Russian parallel corpus) [Evidentsial'nost' i epistemicheskaya otsenka v znachenii nemetskikh glagolov *sollen* i *wollen* (po dannym nemetsko-russkogo parallel'nogo korpusa)] // Computational Linguistics and Intellectual Technologies. Proceedings of the International Conference "Dialogue 2022". Issue 21 (28) [Komp'yuternaya lingvistika i intellektual'nye tekhnologii "Dialog 2022". Vyp. 21 (28)]. — Moscow: RGGU, 2022, pp, 132–140.

[8] Heine B., Kuteva T. World Lexicon of Grammaticalization. — Cambridge: Cambridge Univ. Press, 2002.

[9] Iordanskaja L.N., Mel'čuk I.A. On the semantics of Russian causal prepositions [K semantike russkikh prichinnykh predlogov] // Moskovskiy lingvisticheskiy zhurnal. Vol. 2. — Moscow: RGGU 1996.

[10] Kiseleva K, Paillard D. (eds.). Russian discourse words. Towards contextual and semantic description. Collection of articles [Diskursivnye slova russkogo yazyka. Opyt kontekstno-semanticheskogo opisaniya. Sb. statey]. — Moscow: Metatext, 1998.

[11] Kozintseva N.A. Indirect source of information in the utterance (based on Russian data) [Kosvennyj istochnik informatsii v vyskazyvanii (na materiale russkogo yazyka)] // Khrakovsky V.S. (ed.) Evidence in the languages of Europe and Asia [Evidentsial'nost' v yazykakh Evropy i Azii]. — St. Petersburg: Nauka, 2007, pp. 37–46.

[12] MAS — Dictionary of the Russian language in four volumes [Slovar' russkogo yazyka v chetyrekh tomakh]. 3rd ed. — Moscow: Russkiy yazyk, 1985–1988.

[13] Morkovkin V.V. (ed.). A large universal dictionary of the Russian language. In 2 volumes [Bol'shoy universal'nyj slovar' russkogo yazyka. V 2 tomakh]. — Moscow: AST-Press, 2022.

[14] RG-80 — Russian grammar [Russkaya grammatika]: In 2 volumes. N.Yu. Shvedova (ed.). — Moscow: Nauka, 1980.

[15] Shmelev A.D. Russian language-specific lexical units in parallel corpora: research opportunities and "pitfalls" [Russkie lingvospetsifichne leksicheskie edinitsy v parallel'nykh korpusakh: vozmozhnosti issledovaniya i «podvodnye kamni»] // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2015". Issue. 14 (21) [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog 2015"]. — Moscow: RGGU, 2015.

[16] Sichinava D.V. The use of a parallel corpus for the quantitative study of language-specific vocabulary [Ispol'zovanie parallel'nogo korpusa dlya kolichestvennogo izucheniya lingvospetsifichnoy leksiki] // Language, Literature, Culture: Actual Problems of Study and Teaching [Yazyk, literatura, kul'tura: Aktual'nye problemy izucheniya i prepodavaniya]. Issue 10. — Moscow: Maks Press, 2014, pp. 37–44.

[17] Traugott E.C. From propositional to textual and expressive meanings: Some semantic-pragmatic aspects of grammaticalization // W.P. Lehmann, Y. Malkiel (eds.). Perspectives on Historical Linguistics. — Amsterdam, Philadelphia, 1982. 245–271.

[18] Traugott E.C., König E. The semantics-pragmatics of grammaticalization revisited // E.C. Traugott, B. Heine (eds.). Typological Studies in Language 19: 1991, pp. 189–218.

[19] van der Auwera J., Plungian V. Modality's semantic map // Linguistic Typology, 2, 1998, pp. 79–124.

[20] Zalizniak Anna A. Russian language-specific units in the light of contrastive corpus analysis // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2015". Issue 14 (21) [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog 2015"]. — Moscow: RGGU, 2015, pp. 651–662.

[21] Zalizniak Anna A. Russian discourse markers *vidimo* and *po-vidimomu* ('apparently'): synchronic and diachronical semantics // Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialogue 2021". Issue 20 (27). [Komp'yuternaya lingvistika i intellektual'nye tekhnologii: Po materialam ezhegodnoy Mezhdunarodnoy konferentsii "Dialog 2021"]. — Moscow: RGGU, 2021, pp. 720–728.

RUSSIAN PREDICATIVES AND FREQUENCY METRICS

Anton Zimmerling
Pushkin State Russian Language
Institute / Institute of Linguistics,
Russian Academy of Science
fagraey64@hotmail.com

Abstract

This paper introduces five metrics for measuring the frequencies of dative predicatives in Russian.A dative predicative is a word or multiword expression licensing the dative-predicative-structure, where the semantic subject of the non-agreeing non-verbal predicate is marked by the dative case. I measure the frequencies of the predicatives in the contact position <-1;1> with the same-clause dative subject pronouns in 1Sg (*m*-metrics) and 3Sg (*e*-metrics). The *m*-metrics is applied for retrieving a list of dative predicatives from a corpus. I argue that for each large text collection there is a minimal *m*-value confirming that an item belongs to the core of the dative-predicative structure. The *m/e* score makes up the third metrics that shows whether an element is oriented towards the use in the 1st person or not. Basing on the *m*-metrics, I retrieved 3 lists of predicatives in the subcorpus of 2000 – 2021 texts included in the Russian National Corpus. The A list includes 87 items with $m \geq 10$, the B list includes 44 items with $m \geq 50$, the C list includes 24 items with $m \geq 100$. 72-79% of items in each list have an *m/e* value ≥ 1,25. A linguistic interpretation of this result is that for each list of dative predicatives it is true that the majority of its elements are autoreferential expressions oriented towards the use in the 1st person present indicative tense in the direct speech. The fourth metrics shows the total number of occurrences of a word or multiword expression in the corpus (*N*). I argue that the *N* score must be measured before POS tagging, and lemmatization. The fifth and the last metrics is the *m/N* score. The RNC data suggest an inverse correlation between the score of an item in the context specific for dative-predicative structures (*m*) and its overall frequency in the corpus (*N*). This effect is explained by the regular homonymy of high frequent predicatives with high frequent adverbials and parenthetical expressions.

Keywords: corpus grammar, frequency dictionaries, lexicon, dative predicatives
DOI: 10.28995/2075-7182-2023-22-579-589

РУССКИЕ ПРЕДИКАТИВЫ В ЗЕРКАЛЕ СТАТИСТИКИ

Циммерлинг А. В.
Государственный институт русского
языка им. А.С.Пушкина / Институт
языкознания РАН
fagraey64@hotmail.com

Аннотация

В статье предлагается пять метрик для создания частотного словаря дативных предикативов в русском языке. Дативный предикатив определяется как элемент, допускающий дативно-предикативную структуру, где семантический субъект несогласуемого неглагольного предиката оформляется дат.п. Ранжирование предикативов производится по числу предложений дативно-предикативной структуры в выборке по запросу предикатив + субъектное местоимение 1 л.ед.ч. *мне* в контактной позиции на расстоянии <-1;1> (*m*-метрика) и предикатив + субъектные местоимения 3л. ед.ч. *ему/ей* в той же позиции (*e*-метрика). Словарь предикативов строится на основе *m*-метрики. Для каждой большой коллекции текстов имеется минимальное значение *m*, подтверждающее, что данный элемент принадлежит ядру класса дативных предикативов. Отношение *m/e* используется как третья метрика. Она указывает на то, ориентирован ли элемент на употребление в 1л. в

режиме речи. С помощью *m*-метрики было получено три списка в подкорпусе текстов 2000 – 2021 гг. в НКРЯ. Список А содержит 86 единиц с $m \geq 10$, список В — 44 единицы с $m \geq 50$, список С — 24 единицы с $m \geq 100$. 72-79% элементов каждого списка имеют значение $m/e \geq 1{,}25$. Этот результат подтверждает, что большинство элементов каждого списка ориентированы на употребление в 1 л. ед.ч. презенса индикатива в прямой речи. Четвертая метрика указывает общее число вхождений слова или словосочетания в корпус (*N*). Значение *N* подсчитывается до лемматизации и определения части речи. Отношение *m/N* является пятой метрикой. Данные НКРЯ указывают на обратную зависимость между числом употреблений в контексте, характерном для дативно-предикативной конструкции (*m*), и общим числом вхождений в корпус (*N*). Этот эффект объясняется тем, что наиболее частотные предикативы связаны отношениями регулярной омонимии с высокочастотными наречиями и вводными словами.

Ключевые слова: корпусная грамматика, словарь, дативные предикативы, конструкции

1. Introduction

I discuss the procedure of measuring the frequencies of a productive grammatical construction the elements of which do not make a single lexical class but represent special predicative uses of words from different parts of speech and multiword expressions linked with syntactic structures imposing non-trivial conditions on agreement and case-marking.

The baseline hypothesis is that the majority of Russian predicatives with the dative case-marking on the subject argument are autoreferential expressions including a link to the speaker, who is the source of information about the internal state experienced by him/her at the moment of speech. The aims of the study is to check this hypothesis and to establish, whether the autoreferentiality effects arise due to the inherent lexical features of Russian dative predicatives or are modeled in syntax.

2. Dative-predicative structures and their diagnostics

Russian has a productive class of predicatives licensing syntactic structures, where the animate semantic subject of a non-agreeing non-verbal element is marked with the dative case, hence — dative-predicative structures (DPS). The relation between DPS sentences and word classes is a puzzle. One the one hand, Russian grammar does not require that the dative slot of any predicative or verb is realized overtly. One the other hand, occasional combinations of a predicative with the dative argument do not prove that it is part of the DPS lexicon. The lexicon of a grammatical construction is a list of lexical items regularly used in this construction by all or most speakers. However, with Russian DPS predicatives one must measure the frequencies of the sentences with a filled dative slot, cf. *X-у было стыдно признавать ошибку* ‘X was ashamed to admit his/her mistake’, not just the hits of the lemma *стыдно* or the collocation *стыдно признавать* ‘ashamed to admit smth’. The word *стыдно* in contrast to *грустно* ‘sad’, ‘sadly’, *холодно* ‘cold’, ‘coldly’ belongs to the minority of predicatives that lack side-uses as adverbials. The preceding research provides no instructions how to get the ratio of the relevant DPS uses from the total number of hits of items like *стыдно* or *грустно*. Some DPS predicatives are idiomatic multiword expressions, cf. *X-у все равно* ‘X does not care’.

2.1. The syntax

The role of the dative element can be explained differently. According to [9: 151], most types of Russian sentences can be expanded by the position of the animate dative participant. On this account, it is a free ‘determinant’ or in conventional terms, adjunct, therefore the dative slot does not constrain any class of predicates. This prediction is wrong, since the DPS construction is selective and blocks the combinations that cannot be interpreted as standard designations of internal states experienced by an animate subject. Although Russian authors sporadically produce weird sentences like ??*Нам гневно делается* (Anthony of Sourozh, 1992) ‘we get angry’, lit. *‘to us becomes wrathfully’, ??*Морозно мне* (M.Ancharov, 1989) ‘I feel freezingly cold’, lit. *‘to me is chilly’, they are rejected by the majority of speakers according to [14] and have low frequency in text corpora[1]. Under the alternative ap-

[1] Note that *морозно* and *гневно* are equally marginal as DPS items, although *морозно* ‘It is frosty’, ‘It is chilly

proach, the dative element is semantic subject and the class of DPS predicatives consists of elements capable of describing internal states [8]. This analysis predicts that dative arguments switch the lexical meaning of the predicatives. This is likely for the physical sensations, cf. *Сегодня холодно* 'It is cold today' ⇒ *Мне холодно* 'I am cold', *здесь темно* 'It is dark here', *Мне темно здесь* ⇒ 'It is dark for me'. Without the dative argument *холодно* or *темно* normally describe ambient characteristics, while with the filled dative slot they describe the reactions of an experiential subject, cf. [5; 6]. With the predicatives of interpretation, which do not describe the sensations or affections directly but interpret them in some way, cf. *важно* 'important' the switch is less evident, cf. ***(Мне)*** *важно закончить работу сегодня* 'It is important (**for me**) to finish the work today'. If DPS predicatives make up a lexical class, one needs a list of non-verbal non-agreeing elements with a valency on the animate dative argument [2: 83]. However such lists can only be retrieved in the experiment or corpus study, where approval rates or frequency scores are measured.

2.2. Autoreferentiality

DPS sentences express the meaning of internal davidsonian states[2], i.e. spatiotemporal situations with an animate priority argument [10; 11: 273][3]. This meaning is not unique for Russian DPS sentences, cf. [13: 424-431]. However, the dative case-marking adds a special quality: DPS items are oriented towards the use in the 1Sg in the direct speech, while other types of Russian predicatives sharing the taxonomic meaning of davidsonian states with them normally cannot be used in this context. While it is standard to say *мне*$_{\text{DAT}}$ *грустно* 'I am sad', *мне*$_{\text{DAT}}$ *дурно* 'I feel bad' sentences like $^{??}$*я*$_{\text{NOM}}$ *сейчас навеселе*, int. 'I am tipsy now', **я*$_{\text{NOM}}$ *без чувств*, int. 'I am losing my senses', 'I faint' are awkward. A plausible explanation of this asymmetry is that the majority of Russian DPS predicatives are autoreferential expressions: the speaker himself/herself is the source of information about his/her internal state of feeling bad or sad in the interval including the moment of speech [18]. Meanwhile, Russian predicatives with nominative case-marking on the subject, cf. *навеселе*, *без чувств* are oriented towards describing the experience of other people. The autoreferentiality effect gives a clue for retrieving dative predicatives from a corpus. DPS sentences are copular structures with a slot for the BE-auxiliary or less frequent auxiliaries like *стать*, *сделаться* 'become'. The contact position of a predicative and the 1Sg subject dative pronoun *мне* roughly corresponds to the context of the present indicative, where the overt BE-auxiliary is missing in Russian. Although the search queries PRED + "мне" in the contact position <-1; 1> do not exclude the examples, where an overt auxiliary is found to the left or the right from the search window, cf. ***было***$_{\textbf{AUX.PST}}$ *<мне грустно>* ~ *<грустно мне>* ***было***$_{\textbf{AUX.PST}}$ **'**I was sad', the preceding research indicates that the majority of hits retrieved by such queries indeed patterns with autoreferential contexts in the present indicative tense [16].

2.3. The lexicon

The DPS construction is characteristic of several European languages. The volume of the class of DPS predicatives was measured via a double sociolinguistic and corpus study for Russian [14] and Bulgarian [15]. These authors checked a set of 422 stimuli for Russian. They argue that most Russian

---

outdoors' is a standard impersonal predicative describing the state of weather. The Russian National Corpus (RNC) totals 2143 hits of *гневно*, 2135 of which represent the uses as a non-predicative adverbial and just 8 (0,38%) pattern with agreeing adjectives or predicatives. From 497 hits of *морозно,* 439 (88,4%) pattern with impersonal predicatives.

[2] The cover term *состояния* 'states' used in the Russian studies, is vague. The term 'davidsonian states' is a tribute to Donald Davidson, who defined states as static spatiotemporal situations that exist during a time interval [3]. Internal <davidsonian> states have a priority experiential argument [12; 13: 429 - 431].

[3] In Davidson's account, spatiotemporality is a definitional property: it is assumed that every process and every external or internal state, cf. *The sun is rising. X is in London. X is sad* takes place in some locus, irrespective of the fact, whether the predicate combines with a locative phrase or framing adverbial. An anonymous reviewer suggests that Russian sentences like *Я видел, как ему жаль птичку (*в темной комнате)* should be described as Kimian states, i.e. predicates lacking spatial features [7]. However, *X-у жаль птичку* 'X feels sorry for the bird' describes the feeling of *X* that holds during some time and not the result of *Y*-s observation. Moreover, internal states, e.g., the feeling of being sad, happy, sorry, etc. cannot be observed from outside, though *Y* via some kind of practical reasoning can reconstruct the situation, where X is sad or happy, basing on the external symptoms of sadness or happiness.

speakers have over 200 DPS predicatives in their active vocabulary, but only one part of it is shared. In the variable part, Russian speakers typically select quasi-synonymic DPS items corresponding to generalized lexical meanings like 'X does not care', 'X is delighted', 'X is disgusted', etc. The same test of stimuli was checked on RNC. The search was restricted with one dedicated context — the contact position of the predicative and the 1Sg dative subject pronoun *мне* in the window <-1;1>. The retrieved samples proved large enough to range 400 – 500 items. The authors conclude that high frequent DPS items always have a high approval rate, while DPS items with a high approval rate generally are high frequent, with the exception of some predicatives describing ontologically rare situations, cf. *X-у по колено* 'X is up to his knees', *X-у было по щиколотку* 'X was up to his ankles'. This effect was presumably due to the design of the experiment: the speakers had no difficulties with reconstructing the situations, where such DPS items were appropriate, but the corresponding contexts in the RNC were rare.

I adopt the method of retrieving DPS sentences by narrowing the search with the 1st person contexts and introduce several new metrics for ranging DPS predicatives. In order to eliminate the diachronic factor and make the input data homogeneous, I focus on 2000 – 2021 texts included in the RNC[4]. I also measure the scores of negative and non-negative DPS items on a separate basis and make other adjustments in the set of stimuli. The DPS lexicon in [13; 16: 248] was grouped into 15 thematic classes labeled 'physical sensations'(Class 1), 'modalities' (Class 2), 'affections' (Class 3), 'moral attitudes' (Class 4), '(in)convenience' (Class 5), '(im)pertinence' (Class 6), 'internal need' (Class 7), 'complience' (Class 8), 'difficulty of execution' (Class 9), '(in)disposition' (Class 10), 'general evaluations' (Class 11), '(ir)relevance' (Class 12), '(in)efficiency' (Class 13), 'sensory and intellectual responses' (Class 14), 'parametric features' (Class 15). I adopt this classification and add new items, where appropriate.

3. The frequency dictionary of Russian DPS predicatives

3.1. M-metrics

The lists of DPS predicatives are built by *m*-metrics, which tells the number of confirmed DPS clauses in the syntactic corpus assembled by the query "STIMULUS" + "*мне*" in the window <-1; 1>. The stimulus must be identified as a DPS predicative and the dative pronoun must be the same clause element acting as its semantic subject. The DPS sentences are copular structures that bring about several formal conditions, notably the absence of agreement and the nominative NP that could act as agreement controllers, see below **3.2**.

I take the list of DPS stimuli in [14; 17: 254-255] and adjust it to the tasks of present study. The set of 478 stimuli checked in the 2017 experiment included fillers and obsolete words that went into disuse in the second half of the XX century or earlier. I eliminate all low frequent items from the 2017 set and check the upper part of stimuli starting with $m \geq 10$. The main RNC corpus had 159 such items in 2017. The 2000 – 2021 corpus is smaller. Setting the lower limit at $m \geq 10$, we retrieved 87 DPS predicatives. By lifting the limit up to $m \geq 44$, we get a second list containing 44 DPS items. Setting the limit at $m \geq 100$ leaves us with 24 most frequent DPS items. These lists are referred to as A87, B44 and C24. The maximal *m* score is attested by НАДО ($m = 1402$). The syntactic corpus linked with A87 contains 9619 DPS sentences[5]. The mean expected score m_{87} is 9619/87 = 110, 56. The syntactic corpus linked with the shortest list, C24 contains 7322 DPS sentences. That means that the 24 most frequent DPS predicatives (27, 6%) give 76,1% of DPS sentences.

3.2. The stimuli

The combinations with the free negation *не* were treated as separate entries, if the non-negative expression is used as a DPS predicative: the examples with НЕ НАДО, НЕ НУЖНО, etc. were subtracted from the samples with НАДО, НУЖНО. We considered all spelling variants like НЕ ВАЖНО

[4] 43 928 texts, 98 023 229 words (11.2022).

[5] The requirement that the predicative and its subject are realized overtly and assume a contact position makes each sentence in the syntactic corpus unique. The duplication across samples is excluded. The duplication within a sample is only possible if the RNC search engine returns one and the same text fragment twice.

~ НЕВАЖНО. The A87 list contains 20 items with negation, the most frequent of them being НЕ НАДО (m=334), НЕ НУЖНО (125) and НЕ ЖАЛКО (64). Comparative forms were treated as separate entries, cf. ЛУЧШЕ (*m*=121), ЛЕГЧЕ (89), and ПРОЩЕ (53). The samples with the spelling variants –ЕЕ/-ЕЙ were merged, cf. ИНТЕРЕСН-ЕЕ/-ЕЙ (18). The optative combination ХОРОШО БЫ 'It would be nice' (10) was considered a separate entry different from ХОРОШО 'good' (176). The corresponding examples were subtracted from the scores of the positive forms.

The A87 list includes 12 multiword expressions, 5 of them are also contained in B44 and the upper 3 — in C24, cf. ВСЕ РАВНО (312), НЕ ДО Z-а (60), БЕЗ РАЗНИЦЫ (19), ТАК И НАДО (19) , НЕ ПО СЕБЕ (15), and НЕ ПОД СИЛУ (10). The idioms ВСЕ РАВНО 'X does not care' and ТАК И НАДО 'X deserved it' are treated as separate entries; the score of ТАК И НАДО is subtracted from the score of НАДО. The insertion of the subject dative pronoun into the idiom ТАК *мне* И НАДО was considered an idiosyncratic option equivalent to the contact position of the dative pronoun: otherwise this idiom should be excluded.

No filters were applied to sort out gross expressions. The colloquial words ПОФИГ (*m*=16) and ПО ФИГУ ~ ПОФИГУ (18) were considered separate entries. I substituted the predicate variable in the idiom *X-у* ***Z-ать*** *на Y-а* 'X **does not care** about Y' with the infinitives of physiological verbs: ПЛЕВАТЬ (*m*=135), НАПЛЕВАТЬ (76) и НАСРАТЬ (17) made it to the A87 list.

3.3. Syntactic disambiguation and nominative expressions

Russian DPS sentences are usually analyzed as structures blocking NPs in the nominative case both in the subject [8] and in the object position [15]. A different approach is outlined in [1: 305-308]. Non-adjectival predicates like *X-у не под силу* 'it is beyond X's reach' are an issue, since they license both DPS sentences, cf. *X-у не под силу решить эти задачи* 'To solve these tasks is beyond X's reach' and dative-nominative structures like *X-у* ***эти***$_{\text{NOM}}$ ***задачи***$_{\text{NOM}}$ *не под силу* '**These tasks** are beyond X's reach'. I adopt the mainstream approach and exclude the sentences with a nominative subject from the syntactic DPS corpus. This decision only has a minor effect on A87, since dative-nominative structures are infrequent in the samples derived by the *m*-metrics.

The sentences with a dative pronoun and a noun/NP from the class *лицо* 'face', *признание* 'confession' in the nominative-accusative are two-way ambiguous. If the nominative analysis is taken, the ambiguous predicate head is recognized as an agreeing short adjective in the neutrum singular form, cf. (1a-b). If the accusative analysis is taken, the predicate is recognized as a DPS item, cf. (2a-b).

(1) a. мне плохо видно$_{\text{ADJ.NOM.SG}}$ ее **лицо**$_{\text{NOM.SG.N}}$.
'I can't see her face clearly', lit. 'Her face is badly visible to me.'

b. Мне плохо видна$_{\text{ADJ.NOM.F}}$ ее **шея**$_{\text{NOM.SG.F}}$.
'I can't see her neck clearly', lit. 'Her neck is badly visible to me.'

(2) a.Мне плохо видно$_{\text{PRED}}$ ее **лицо**$_{\text{ACC.SG.N}}$.
'I can't see her face clearly.'

b. Мне плохо видно$_{\text{PRED}}$ их **лица**$_{\text{ACC.PL}}$.
'I can't see their faces clearly.'

Another kind of ambiguity is caused by the pronominal expressions *это* 'this', *все это* 'all this'. If they fill in the valency of an active or passive verb, they must be considered referential pronouns/DPs in the accusative or nominative case, cf. (3a). If they lack strong referential properties and refer to the situation as a whole without referring to any of its parts, they are caseless expressions that do not take the subject or object positions, cf. (4a).

(3) a. **Все это**$_{\text{NOM.SG.N}}$ мне куплено$_{\text{PRT.PASS.NOM.SG.N}}$
'All this has been bought for me.'

b. **Все эти вещи**$_{\text{NOM.PL}}$ мне куплены$_{\text{PRT.PASS.NOM.PL}}$.
All these things have been bought for me.'

(4) a. **Все это** мне грустно$_{PRED}$.
‘All this is sad to me’,

b. *Все эти вещи мне грустны.
int. * ‘All these things are sad for me.’

3.4. E-metrics

The same set of 87 stimuli was checked with the dative pronouns *ему* ‘3Sg.Dat.M’ and *ей* ‘3Sg.Dat.F’ in the contact position in the window <-1; 1>. The number of the confirmed DPS clauses is called *e*-metrics. The *e*-metrics provides a tool for checking autoreferentiality. The syntactic corpus built via the *e*-metrics for A87 contains 5434 DPS sentences and is ca. 1,8 times smaller compared to the corpus assembled by the *m*-metrics. The mean expected value e_{87} is 5434/87 = 61, 31. Another index showing the frequency drop in the *e*-corpus is the number of the DPS items fitting to the minimal values for C24, B44 and A87: there are only 11 predicatives in the C*11 list ($e \geq 100$), 31 predicatives in the B*31 list ($e \geq 50$) and 68 predicatives in the A*68 list ($e \geq 10$). The shrinking is most pronounced with high frequent DPS items, where C*11 exports 10 DPS items from C24 and lifts one item from B44, ДОСТАТОЧНО ($m = 79$, $e = 101$). All B*31 items, with the exception of УДОБНО$_1$ ($m = 34$, $e = 80$) are contained in B44 and all A*68 items are contained in A87. The last result is trivial, since A87 per definition lacks items with $m < 10$. The first two ones are not: they show that just 2 DPS items from 87 swap their positions in the mid-range and high-range lists.

3.5. Thematic classes

The thematic classes of the DPS lexicon are distributed evenly in our data. The largest list, A87 includes 12 classes from 15, only Classes 7 <‘internal need’>, 10 < ‘(in)disposition’> and 13 < ‘(in)efficiency’> are missing, since they lack frequent DPS predicatives with $m \geq 10$. B44 also lacks Classes 8 <complience> and 15 <‘parametric features’>. The shortest list, C24 retains 8 different classes but drops Classes 5 < ‘(in)convenience’> and 6 <‘<im>pertinence’ >.

Tab. 1. The coverage of the DPS construction in Russian (2000 – 2021).

List	m	Retained classes	Missing classes
A87	≥ 10	1, 2, 3, 4, 5, 6, 8, 9, 11, 12, 14, 15	*7, *10, *13
B44	≥ 50	1, 2, 3, 4, 5, 6, 9, 11, 14	*7, *10, *13, *, *8, *15
C24	≥ 100	1, 2, 3, 4, 9, 11, 12, 14	*7, *10, *13, *15,*8, *15, *5, *6

These figures confirm that Modern Russian has high frequent DPS predicatives in most thematic classes and uses them in diverse ontological situations.

3.6. Semantic disambiguation

A87 includes a pair of DPS items that are treated as homonyms, since they represent different thematic classes: *X-y* ПЛОХО$_1$ (Class 1, $m = 49$), cf. *Мне внезапно стало плохо* ‘I suddenly felt badly’ vs *X-y* ПЛОХО$_2$ (Class 11, $m= 149$), cf. *Ей было плохо жить со свекровью* ‘It was bad for her to live with her mother-in-law’. Their profiles can only be kept apart after semantic disambiguation. ПЛОХО$_2$, is also part of B44 and C24. Semantic disambiguation is relevant for *X-y* УДОБНО$_1$ (Class 5, $m = 34$), cf. *Я попыталась лечь, как мне удобно* ‘I tried to lie down as comfortably as I could’, НЕУДОБНО$_2$ (Class 4, $m = 40$), cf. *Неудобно мне как-то стало* ‘I felt kind of awkward’, НЕЛОВКО$_2$ ‘Class 4, $m = 39$’, cf. *Мне неловко об этом писать* ‘I am embarrassed to write about this’, where the homonymic predicatives are low frequent elements that do not make it to A87. The items (X-y) М*А*ЛО ‘X does not have enough’ ($m = 51$, $e= 96$) and (X-y) МАЛ*О* ‘Something is too small for X’ are pronounced differently but spelled in the same way, therefore the samples with М*А*ЛО must be checked for the casual hits of МАЛ*О*.

3.7. The m/e metrics and its application

The *m/e* score serves as the third metrics. It is applied after the lists of frequent DPS items are retrieved by the *m*-metrics. With low *m* scores > 10 and comparably low *e* scores, the fluctuations of the *m/e* score are not significant. With high and mid-frequent DPS items, it makes sense to measure both the individual profiles of DPS predicative and the general characteristics of the lists. Let us assume that a DPS predicative is autoreferential, if *m/e* ≥ 1, 25, i.e. if the uses in the 1st person singular are at least 25% more frequent compared to the uses of the 3rd person singular in the same position. The mean expected score for the A87 list m_{87}/e_{87} = 1,79 exceeds this level with a margin, but it is difficult to interpret this result without ranging the elements of each list on the basis of their individual *m/e* scores. Let us introduce a distinction of mildly non-autoreferential vs strictly non-autoreferential expressions. A DPS predicative is mildly non-autoreferential, if 1 ≤ *m/e* < 1, 25 and strictly non-autoreferential, if *m/e* < 1.

Tab. 2. Autoreferential DPS items in the Russian National Corpus (2000 – 2021).

	m/e	A87, *m* ≤ 10	B44, *m* ≤ 50	C24, *m* ≤ 100
+ Autoreferential	***m/e* ≥ 1,25**	**71,27%**	**72, 728%**	**79, 17%**
Mildly-non-autoreferential	1, 0 ≤ *m/e* < 1, 25	12,64%	13,636%	12,5%
-Autoreferential	***m/e* < 1**	**16,09%**	**13, 636%**	**8, 33%**

Tab. 2 shows that the share of the autoreferential DPS items increases with their frequency. More precisely, the C24 list containing the items with *m* ≤ 100 has just 2 strictly non-autoreferential items, ЛУЧШЕ (*m/e* = 0,85) and НЕОБХОДИМО (*m/e* = 0,87) and 19 autoreferential items (79,17%). Meanwhile, there is no contrast between A87 and B44: lifting the low *m* value from 10 to 50 leaves the percentage of the autoreferential items at the same level (71,3% — 72,7%). The *m/e* scores in A87 are in the range 0, 4 ≤ *m/e* ≤ 21. It makes sense to exclude the low frequent elements to get a more balanced picture[6].

3.8. The N-metrics and lemmatization

The *N*-metrics gives the number of hits of a word or multiword expression in a corpus. I argue that the *N* score must be measured before POS tagging and lemmatization. Almost all DPS items have regular homonyms predicted by their morphology. The largest group of homonyms is adjectival words with the – *o*-final, historically — short adjectives in Nom-Acc.Sg.N. Many of them, cf. *грустно* 'sad', 'sadly' are used in parallel as agreeing adjectives, adverbials and non-agreeing predicatives. Some items have a fourth side-use as parenthetical elements, cf. *видно* 'it is seen' ∨ or 'visible' ∨ 'apparently'. An *–o-* item can be tagged either as adverbial (ГРУСТНО$_{ADV}$) or as part of the adjectival paradigm (ГРУСТНО$_{ADJ}$). The latter decision depends on two factors: a) the existence of the adjectival lemma in the dictionary and/or the instruction confirming that the ГРУСТНО$_{ADJ}$ is used in the agreeing position; b) the (in)ability of the parser to recognize the agreement controller. The RNC parser occasionally fails to lemmatize *–o*-items correctly. I provide two illustrations. In (5) the parser failed to recognize the substantivized form *смешное* 'funny', 'what is funny' as the agreement controller and wrongly tagged *грустно* as an adverbial. In (6) the parser wrongly analyzed the non-argument expression *все это* 'all this' as an agreeing subject and tagged *грустно* as an adjective.

(5) Печальное$_{ADJ.SG.N}$ нам смешно$_{ADJ.SG,N}$, а **смешное$_{ADJ.SG.N}$** грустно$_{ADJ.SG.N}$ (A.Morozov, 1985-2001).
'What is sad is funny to us, and **what is funny** is sad.'

(6) Как-то грустно$_{PRED}$ мне$_{1SG.DAT}$ **все это** (A.Terekhov, 1997 – 2001)
'Somehow I feel sad about **all this**.'

[6] E.g., ДУРНО 'X feels badly' occurs in the 2000 – 2021 texts only 397 times but provides 15 autoreferential contexts (*m*=15) without a single example with the 3rd person singular subject pronoun in the contact position.

The deep syntactic annotation of DPS predicatives in the contact position with the subject dative pronoun makes the lemmatization of the *-o*-items in the remaining part of the corpus redundant. What matters is not the POS tags and lemmas of the elements homonymic to the DPS predicatives, but the share of the DPS hits in the sample derived by the *m*-metrics vs the raw data containing the total score of hits for the whole set of homonyms including the tested DPS item. RNC provides the ipm estimates for all words and collocations, but splits the data into different lemmas. This is unhappy with comparative forms. E.g., the search item *хуже* 'worse' returns back the lemmas ПЛОХОЙ, ПЛОХО, ХУЖЕ and even ХОРОШО (the antonym of ПЛОХО). The search item *лучше* 'better' returns back 7 lemmas, including exotic suggestions like ВСЕМИЛОСТИВШЕ (the second frequent lemma!). Similar issues arise in all cases, where the spelling varies.

3.9. The m/N metrics

The *m/N* score is the fifth metrics. It shows the proportion of the confirmed DPS hits in the syntactic sub-corpus built via the *m*-metrics vs the total score of all elements identic with or homonymic to the corresponding DPS predicative. I call this set 'quasi-homonymic list'. It is irrelevant for the *m/N* score whether the elements of this list are real homonyms, as, e.g. in the pair НАДО$_1$ 'necessary' vs НАДО$_2$ 'above', diverged uses of the same underlying morphological form, cf. *грустно* 'sad', 'sadly' or DPS uses outside the *m* context. A pair or tuple of quasi-homonymic lists is called 'quasi-homonymic hyperset'.

I checked two hypotheses: A) The number of DPS hits in the 1st person contexts feeds on the score of quasi-homonyms and increases proportionally; B) some elements are more specialized in the DPS construction than other elements. The hypothesis A) makes wrong predictions. The situation at the poles of the *N* scale resembles the inverse correlation between *N* and the *m/N* score. The highest frequent element, МОЖНО (N = 121490) has one of the lowest *m/N* scores (0,0022), despite a high *m* score (265). The second most frequent element, ЯСНО (N = 112008) has the lowest *m/N* score (0,0005). Meanwhile, the elements with the highest *m/N* scores, НАСРАТЬ (0,2394), ПО ФИГУ (0,2195) and ПОФИГ (0,1441) have the lowest *N* scores: НАСРАТЬ occurs only 71 times, ПО ФИГУ — 81 times and ПОФИГ — 111 times.

In the mid-range, there is neither a gradual decline nor a gradual increase of the *m/N* score with the rise of *N*. We dropped all low frequent elements with $N < 1000$, the two highest frequent elements with $N > 100000$, two elements with highest *m* score and set the *m* limit at $m \geq 30$. The trimmed list contains 48 items in the range $30 \leq m \leq 496$, $1025 \leq N \leq 46602$. The same or nearly the same *m* value is reached by the DPS items with very different *N* scores, cf. ХОРОШО (m = 176, N = 46602, m/N = 0,0038) with СТЫДНО (m = 175, N = 3076, m/N = 0,0568). This negative result hints that the hypothesis B) is correct. To explain the *m*/*N* scores, one has to consider the individual profiles of the items like ХОРОШО and СТЫДНО. In this pair, СТЫДНО is more specialized in the DPS construction and the expectancy of the 1st person use with a subject pronoun in the contact position for this item is almost 15 times higher compared to ХОРОШО.

The cross-comparison of negative and non-negative DPS items and their quasi-homonyms provides a tool for checking the hypothesis B). There are 13 such pairs in A87. In 3 of them the negation does not constrain the number of syntactic contexts: (НЕ) НАДО, (НЕ) ЖАЛЬ, and (НЕ) НУЖНО. These 6 items lack adverbial side-uses. The same holds for the pair (НЕ) ИЗВЕСТНО, but the non-negative member occurs here in a wider set of contexts. In 3 pairs — (НЕ) ТРУДНО, (НЕ) СТРАШНО and (НЕ) ЖАЛКО — the negative member lacks regular adverbial side-uses, while the non-negative member retains them. Finally, in 6 pairs adverbial uses are attested with both members of the quasi-synonymic hyperset. In all 13 pairs, the negative member is significantly less frequent. The baseline hypothesis is that the *m/N* score increases in the context of negation, since the negative members are expected to be less frequent and more specialized in the predicative function[7]. However, the absence or presence of adverbial uses does not predict that the negative member has an increased or decreased

[7] Almost all hits of НЕ СТРАШНО, НЕ ЖАЛКО and НЕТРУДНО tagged by the RNC engine as adverbials are actually non-agreeing predicatives. The sole example of the genuine adverbial use is weird: *Трудный, неприемлемый для нас человек, сыгранный с легкостью,* ***нетрудно****, ненапряженно, -- это и по-особому назидательный случай в практике сцены* (N.Berkovskij, 1990 – 2000).

m/N score: each subgroup includes both pairs of the type δ ($m/N_{NON\text{-}NEG} - m/N_{NEG}$) > 0 and pairs of the type δ ($m/N_{NON\text{-}NEG} - m/N_{NEG}$) < 0.

Tab. 3. Negative and non-negative DPS items in RNC, 2000-2021.

Without negation	*N*	*m/N*	With negation	*N*	*m/N*	*δ*
I. No adverbial side-uses with both members						
ЖАЛЬ	4606	*0,0486*	НЕ ЖАЛЬ	177	**0,1242**	0,0756
НАДО	78872	0,0192	НЕ НАДО	11828	**0,0282**	0,009
НУЖНО	35580	**0,0345**	НЕ НУЖНО	4145	0,03	-0,0045
ИЗВЕСТНО	15192	**0,0326**	НЕИЗВЕСТНО	4938	0,0141	-0,0185
II. No regular adverbial side-uses with the negative member						
СТРАШНО	7301	0,0036	НЕ СТРАШНО	778	**0,0411**	0,0375
ТРУДНО	14455	**0,0235**	НЕТРУДНО	1453	0,0151	-0,0084
ЖАЛКО	3482	**0,0459**	НЕ ЖАЛКО	711	0,09	-0,0441
III. Regular adverbial side-uses with both member						
ПОНЯТНО	12042	0,0053	НЕПОНЯТНО	4153	**0,0202**	0,0149
ИНТЕРЕСНО	11856	0,0231	НЕИНТЕРЕСНО	1230	**0,0349**	0,0118
ХОРОШО	46602	0,0036	НЕХОРОШО	1259	**0,015**	0,0114
ПРИЯТНО	5157	**0,0337**	НЕПРИЯТНО	1576	0,031	-0,0027
ВАЖНО	10792	***0,0093***	НЕВАЖНО	3616	0,0006	-0,0087
ЛЕГКО	14148	**0,0446**	НЕЛЕГКО	1229	0,0044	-0,0402

The pairs, where the *m/N* decreases in the context of negation, can have some hidden property, e.g. the high initial *m/N* score by the non-negative member. However, this does not explain the increase on НЕ ЖАЛЬ, despite ЖАЛЬ has a high *m/N* score (0,0486) and the slight decrease on НЕВАЖНО, despite ВАЖНО has a low *m/N* score (0,0486).

4. General discussion and conclusions

There are two kinds of data — the frequencies of specific elements associated with the described grammatical construction and general properties associated with the lists of DPS predicative representing the upper part of the frequency dictionary. The ranks of specific predicatives, with the possible exception of the 2-3 most frequent items (НАДО, НУЖНО, ИЗВЕСТНО) depend on the chosen corpus. Meanwhile, the orientation towards the 1st person contexts in the direct speech and the type of meaning indicating that the speaker himself/herself is the source of information about his/her internal state are general features of the Russian DPS construction and its lexicon. There are reasons to think that these features are only minimally text-dependent. One needs a corpus that is large enough to range a list of predicatives and has 1st person contexts. Since a vast majority of Russian DPS predicatives is autoreferential, the lists of the predicatives can be retrieved via the *m*-metrics, which serves two purposes: 1) it gives the number of confirmed DPS clauses with overt subject pronouns in the syntactically annotated corpus assembled by the search query "STIMULUS" + "*мне*" in the window <-1; 1>; 2) it provides a ranging of mid-frequent and high-frequent DPS items.

For each text collection, there is a minimal *m* value, which tells apart regular DPS items from occasional combinations with a dative pronoun. A control list can be retrieved via the *e*-metrics, which provides a second syntactic corpus with confirmed DPS hits in the 3rd person contexts with 3rd person singular subject pronouns in the contact position. The positive *m/e* score confirms that the predicative is entrenched in the DPS construction: ca. 71— 79% of mid- and high-frequent DPS items have the *m/e* scores ≥ 1, 25. The share of non-autoreferential predicatives with the *m/e* score < 1 is minimal in the list containing the most frequent items with *m* > 100.

Russian DPS predicatives always have homonyms. The score of all homonyms (*N*) provides the background for the frequency dictionary. The score *m/N* shows the expectation of finding a DPS construction in the 1st context with a subject pronoun. There is no general formula predicting the *m/N* ratio

for each item, at least in the RNC. This negative result is in accord with the baseline hypothesis that Russian DPS sentences represent a highly idiomatic grammatical construction that does not borrow its elements from the general lexicon but creates it in the dedicated syntactic contexts.

There are several ways of implementing the applied procedure in corpus studies, grammatical theory and cross-language comparison: 1) the retrieved dictionary can be checked on other corpora of Russian; 2) the frequency metrics can be applied for the description of other Russian constructions with an animate priority argument; 3) the statistic profile of the Russian DPS construction and the relevant features '± syntactic animacy', '± autoreferentiality'underlying it can be compared to the characteristics of similar dative constructions in the world's languages.

Acknowledgments

This research has been supported by the Russian Science Foundation, project no. 22-18-00528 "Clausal connectives in sentence and discourse: Semantics and grammaticalization paths".

References

[1] Apresjan Ju. D. Sintaksičeskie priznaki leksem [The syntactic features of lexemes], *Russian linguistics*. 1985. Vol. 19, 2-3. P. 289 – 317.

[2] Bonč-Osmolovskaja A. Kvantitativnye metody v diahroničeskih korpusnyh issledovanijah, Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2015'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2015']. Issue 14. [Vyp. 10]. Moscow: RGGU Publ. 2015. P. 80-94.

[3] Davidson D. The individuation of events, D.Davidson (ed.), Essays on actions and events. Oxford: Clarendon Press, 1980. P. 163-180.

[4] Ivanova E., Zimmerling A. Shared by All Speakers? Dative predicatives in Bulgarian and Russian, *Bulgarian Language and Literature*. 2019, LXI, 4. P. 353–363.

[5] Kustova G.I. Tipy infinitivnyx konstrukcij s predikativami (po dannym Nacional'nogo korpusa russkogo jazyka [The types of infinitive constructions with predicatives (according to the Russian National Corpus], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2021'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2021']. Issue 20. [Vyp. 20]. Moscow: RGGU Publ. 2021. P. 456-463.

[6] Kustova G.I. Semantičesjue tipy infinitivnyh konstrukcij russkih predikativov [The semantic types of infinitive constructions with Russian predicatives], S.Koeva, E.Yu.Ivanova, J.Tisheva and A.Zimmerling (eds.), Ontologija na situaciite za săstojanie – lingvistično modelirane. Săpostavitelno izsledvane za bălgarski i ruski [The ontology of stative situations – linguistic modeling. A contrastive study of Bulgarian and Russian]. Professor Marin Drinov publ, Sofia. 2022. P. 246–279.

[7] Maienborn C. On Davidsonian and Kimian states, Comorovski, I., K. von Heusinger (eds.). Existence. Semantics and Syntax. Dordrecht: Springer. 2007. P. 107–130.

[8] Pospelov N.S. V zaščitu kategorii sostojanija [In defence of the category of state], *Voprosy jazykoznanija* [*Issues in linguistics*]. 1955, 2. P. 55 – 65.

[9] Švedova N.Y. Russkaja grammatika [Russian grammar]. In 2 vols. Vol. 2. Nauka, Moscow, 1982.

[10] Yanko T.E. Kommunikativnyj status russkih benefaktivnyh konstrukcij [The communicative status of Russian benefactive constructions], *Moscowskij lingvističeskij žurnal* [Moscow linguistic journal], 1996.

[11] Yanko T.E. Kommunikativnye strategii russkoj reči [The communicative strategies in Russian speech . Yazyki slavyanskoi kul'tury, Moscow. 2021.

[12] Zaliznjak Anna A. Issledovanija po semantike predikatov vnutrennego sostojanija [Investigations in the semantics of inner state predicates]. Otto Sagner, München, 1992.

[13] Zaliznjak Anna A. Mnogoznačnost' v jazyke i sposoby ee predstavlenija [Polysemy and its representations]. Jazyki slavjanskoj kul'tury, Moscow. 2006.

[14] Zimmerling A. Russkie predikativy v zerkale eksperimenta i korpusnoj grammatiki [Russian predicatives in the perspective of the sociolinguistic experiment and corpus grammar], Computational Linguistics and Intellectual Technologies. Papers from the Annual International Conference 'Dialogue 2017'. [Komp'juternaja lingvistika i intellektual'nye tehnologii. Trudy mezhdunarodnoj konferencii 'Dialog 2017']. Issue 16. [Vyp. 16]. Moscow: RGGU Publ. 2017. P. 466-481.

[15] Zimmerling A. Predikativy i predikaty sostojanija v russkom jazyke [Predicatives and the predicates of state in Russian], *Slavistična revija*. 2018, 1. P. 45–64.

[16] Zimmerling A. Avtoreferentnost' i klassy predikativnyh slov [Autoreferentiality and predicative classes], V.V.Kazakovskaja, M.B.Voejkova (eds.), Problemy funkcional'noj grammatiki.Otnošenie k govorjaščemu v semantike grammatičeskih kategorij [The issues in functional grammar. The speaker-oriented grammatical categories]. Jazyki slavjanskoj kul'tury, Moscow. 2020. P. 23-58.
[17] Zimmerling A. Ot integral'nogo k aspektivnomu [From integral frameworks to aspective descriptions]. Aletheia, Sankt-Peterburg and Moscow. 2021a.
[18] Zimmerling A. Primary and secondary predication in Russian and the SLP: ILP distinction revisited, V. Warditz (ed.), Russian Grammar: System – Language Usage - Language Variation. Peter Lang, Frankfurt a.M. et al. 2021b. P. 543–560.

Russian National Corpus [Nacional'nyj korpus russkogo jazyka]: <www.ruscorpora.ru>.

Abstracts

RECEIPT-AVQA-2023 CHALLENGE

Begaev A., **Orlov E.**, Budapest, Hungary

In this work, we introduce a new challenging Document VQA dataset, named Receipt AVQA, and present the results of the associated RECEIPT-AVQA-2023 shared task. Receipt AVQA is comprised of 21,835 questions in English over 1,957 receipt images. The receipts contain a lot of numbers, which means discrete reasoning capability is required to answer the questions. The associated shared task has attracted 4 teams that have managed to beat an extractive VQA baseline in the final phase of the competition. We hope that the published dataset and promising results of the contestants will inspire further research on understanding documents in scenarios that require discrete reasoning.

CONSTRUCTING A SEMANTIC CORPUS FOR RUSSIAN: SEMONTOCOR

Boguslavsky I. M.[1,2], **Dikonov V. G.**[1], **Inshakova E. S.**[1], **Iomdin L. L.**[1], **Lazursky A. V.**[1], **Rygaev I. P.**[1], **Timoshenko S. P.**[1], **Frolova T. I.**[1], [1]A. A. Kharkevich Institute for Information Transmission Problems, Moscow, Russia; [2]Universidad Politécnica de Madrid, Madrid, Spain

The SemOntoCor project focuses on creating a semantic corpus of Russian based on linguistic and ontological resources. It is a satellite project with regard to a semantic parser (SemETAP) being developed, the latter aiming at producing semantic structures and drawing various types of inferences. SemETAP is used to annotate SemOntoCor in a semi-automatic mode, whereupon SemOntoCor, when reaching sufficient maturity, will help create new parsers and other semantic applications. SemOntoCor can be viewed as a further step in the development of SynTagRus with its several layers of annotation. SemOntoCor builds on top of the morpho-syntactic annotation of SynTagRus and assigns each sentence a Basic Semantic Structure (BSemS). BSemS represents the direct layer of meaning of the sentence in terms of ontological concepts and semantic relations between them. It abstracts away from lexico-syntactic variation and in many cases decomposes lexical meanings into smaller elements. The first phase of SemOntoCor consists in annotating a Russian translation of the novel "The Little Prince" by Antoine de Saint-Exupery (1532 sentences, 13120 tokens).

PSEUDO-LABELLING FOR AUTOREGRESSIVE STRUCTURED PREDICTION IN COREFERENCE RESOLUTION

Bolshakov V.[1,2], **Mikhaylovskiy N.**[1,3], [1]NTR Labs; [2]BMSTU, Moscow, Russia; [3]Higher IT School, Tomsk State University, Tomsk, Russia

Coreference resolution is an important task in natural language processing, since it can be applied to such vital tasks as information retrieval, text summarization, question answering, sentiment analysis and machine translation. In this paper, we present a study on the effectiveness of several approaches to coreference resolution, focusing on the RuCoCo dataset as well as results of participation in the Dialogue Evaluation 2023. We explore ways to increase the dataset size by using pseudo-labelling and data translated from another language. Using such technics we managed to triple the size of dataset, make it more diverse and improve performance of autoregressive structured prediction (ASP) on coreference resolution task. This approach allowed us to achieve the best results on RuCoCo private test with increase of F1-score by 1.8, Precision by 0.5 and Recall by 3.0 points compared to the second-best leaderboard score. Our results demonstrate the potential of the ASP model and the importance of utilizing diverse training data for coreference resolution.

LIGHT COREFERENCE RESOLUTION FOR RUSSIAN WITH HIERARCHICAL DISCOURSE FEATURES

Chistova E. V., **Smirnov I. V.**, FRC CSC RAS, Moscow, Russia

Coreference resolution is the task of identifying and grouping mentions referring to the same real-world entity. Previous neural models have mainly focused on learning span representations and pairwise scores for coreference decisions. However, current methods do not explicitly capture the referential choice in the hierarchical discourse, an important factor in coreference resolution. In this study, we propose a new approach that incorporates rhetorical information into neural coreference resolution models. We collect rhetorical features from automated discourse parses and examine their impact. As a base model, we implement an end-to-end span-based coreference resolver using a partially fine-tuned multilingual entity-aware language model LUKE. We evaluate our method on the RuCoCo-23 Shared Task for coreference resolution in Russian. Our best model employing rhetorical distance between mentions has ranked 1st on the development set (74.6% F1) and 2nd on the test set (73.3% F1) of the Shared Task. We hope that our work will inspire further research on incorporating discourse information in neural coreference resolution models.

PARTITIVE GENITIVE IN RUSSIAN: DICTIONARY AND CORPUS DATA

Chuikova O. Iu., Herzen State Pedagogical University of Russia, St. Petersburg, Russia

The paper aims at comprehensive analysis of the verbs compatible with the partitive genitive object. Based on the Dictionary of Russian Language, the list of perfective verbal lexemes that are able to take the genitive object is compiled and semantic features that unite these verbs are revealed. The features are divided into two groups: aspectually relevant features and aspectually irrelevant features. The corpus-based analysis of the use of the verbs that take both genitive and accusative objects makes it possible to identify features that increase the likelihood of certain object case-marking.

BIMODAL SENTIMENT AND EMOTION CLASSIFICATION WITH MULTI-HEAD ATTENTION FUSION OF ACOUSTIC AND LINGUISTIC INFORMATION

Dvoynikova A. A., **Karpov A. A.**, St. Petersburg Federal Research Center of the Russian Academy of Sciences, Saint-Petersburg, Russia

This article describes solutions to couple of problems: CMU-MOSEI database preprocessing to improve data quality and bimodal multitask classification of emotions and sentiments. With the help of experimental studies, representative features for acoustic and linguistic information are identified among pretrained neural networks with Transformer architecture. The most representative

features for the analysis of emotions and sentiments are EmotionHuBERT and RoBERTa for audio and text modalities respectively. The article establishes a baseline for bimodal multitask recognition of sentiments and emotions – 63.2% and 61.3%, respectively, measured with macro F-score. Experiments were conducted with different approaches to combining modalities – concatenation and multi-head attention. The most effective architecture of neural network with early concatenation of audio and text modality and late multi-head attention for emotions and sentiments recognition is proposed. The proposed neural network is combined with logistic regression, which achieves 63.5% and 61.4% macro F-score by bimodal (audio and text) multi-tasking recognition of 3 sentiment classes and 6 emotion binary classes.

INTRODUCTION MODEL IN RUSSIAN «PEAR REPORTAGES»: THE ROLE OF COMMON GROUND

Fedorova O. V. , Lomonosov Moscow State University, Moscow, Russia

In this study, the peculiarities of the character introduction in the genre of live reportage were studied. The participants were 25 students oh the Lomonosov Moscow State University. Speech production was elicited by means of the “Pears Film” by W. Chafe. Different types of the collective common ground were considered. It turned out that, unlike narratives of other genres, the chronological scale is more important for the introduction than the status scale. It was also shown that the collected reportages from the point of view of the introduction peculiarities are more similar to classical retellings than to the sports reportages.

FOREGROUND AND BACKGROUND IN RUSSIAN SIGN LANGUAGE NARRATIVES: THE ROLE OF ASPECT AND ACTIONALITY

Filimonova E. V., Russian State University for the Humanities; Institute of linguistics, Russian Academy of Sciences, Moscow, Russia

The paper explores the role of aspect and actionality in foregrounding and backgrounding of clauses in Russian Sign Language narratives. Corpus study shows similarities to functions of aspectual markers and actionality in spoken languages. Besides grammatical markers and predicate types, non-manual marking and prosodic features of verbal sign can contribute to clause foregrounding and backgrounding.

MULTIMODAL DISCOURSE TREES IN FORENSIC LINGUISTICS

Galitsky B. A.[1], **Ilvovsky D. A.**[2], **Goncharova E. F.**[2,3], [1]Knowledge Trail Inc., San Jose, CA, USA; [2]NRU HSE; [3]AIRI, Moscow, Russia

We extend the concept of a discourse tree (DT) in the discourse representation of text towards data of various forms and natures. The communicative DT to include speech act theory, extended DT to ascend to the level of multiple documents, entity DT to track how discourse covers various entities were defined previously in computational linguistics, we now proceed to the next level of abstraction and formalize discourse of not only text and textual documents but also various kinds of accompanying data. We call such discourse representation Multimodal Discourse Trees (MMDTs). The rational for that is that the same rhetorical relations that hold between text fragments also hold between data values, sets and records, such as Reason, Cause, Enablement, Contrast, Temporal sequence. MMDTs are evaluated with respect to the accuracy of recognition of criminal cases when both text and data records are available. MMDTs are shown to contribute significantly to the recognition accuracy in cases where just keywords and syntactic signals are insufficient for classification and discourse-level information needs to be involved.

INCREMENTAL TOPIC MODELING FOR SCIENTIFIC TREND TOPICS EXTRACTION

Gerasimenko N.[1,2], **Chernyavskiy A.**[3], **Nikiforova M.**[1], **Ianina A.**[4], **Vorontsov K.**[2,4], [1]Sberbank, [2]MSU Institute for Artificial Intelligence, [3]National Research University Higher School of Economics, [4]Moscow Institute of Physics and Techonology (MIPT)

Rapid growth of scientific publications and intensive emergence of new directions and approaches poses a challenge to the scientific community to identify trends in a timely and automatic manner. We denote trend as a semantically homogeneous theme that is characterized by a lexical kernel steadily evolving in time and a sharp, often exponential, increase in the number of publications. In this paper, we investigate recent topic modeling approaches to accurately extract trending topics at an early stage. In particular, we customize the standard ARTM-based approach and propose a novel incremental training technique which helps the model to operate on data in real-time. We further create the Artificial Intelligence Trends Dataset (AITD) that contains a collection of early-stage articles and a set of key collocations for each trend. The conducted experiments demonstrate that the suggested ARTM-based approach outperforms the classic PLSA, LDA models and a neural approach based on BERT representations. Our models and dataset are open for research purposes.

FINE-TUNING TEXT CLASSIFICATION MODELS FOR NAMED ENTITY ORIENTED SENTIMENT ANALYSIS OF RUSSIAN TEXTS

Glazkova A., University of Tyumen, Tyumen, Russia

The paper presents an approach to named entity oriented sentiment analysis of Russian news texts proposed during the RuSentNE evaluation. The approach is based on RuRoBERTa-large, a pre-trained RoBERTa model for Russian. We compared several types of entity representation in the input text, and evaluated strategies for handling class imbalance and resampling entity tags in the training set. We demonstrated that some strategies improve the results of pre-trained models obtained on the dataset presented by the organizers of the evaluation.

ASPECT-BASED ARGUMENT GENERATION IN RUSSIAN

Goloviznina V. S., **Fishcheva I. N.**, **Peskisheva T. A.**, **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

The paper explores the argument generation in Russian based on given aspects. An aspect refers to one of the sides or property of the target object. Five aspects were considered: "Safety", "Impact on health", "Reliability", "Money", "Convenience and comfort". Various approaches were used for aspect-based generation: fine-tuning, prompt-tuning and few-shot learning. The ruGPT-3Large model was used for experiments. The results show that traditionally trained model (with fine-tuning) generates 51.6% of the arguments on given aspects, with the prompt-tuning approach – 33.9%, and with few-shot learning – 10.6%. The model also demonstrated the ability to generate arguments on new, previously unknown aspects.

RUSENTNE-2023: EVALUATING ENTITY-ORIENTED SENTIMENT ANALYSIS ON RUSSIAN NEWS TEXTS

Golubev A. A.[1], **Rusnachenko N. L.**[2], **Loukachevitch N. V.**[1], [1]Lomonosov Moscow State University, [2]Bauman Moscow State Technical University, Moscow, Russia

The paper describes the RuSentNE-2023 evaluation devoted to targeted sentiment analysis in Russian news texts. The task is to predict sentiment towards a named entity in a single sentence. The dataset for RuSentNE-2023 evaluation is based on the Russian news corpus RuSentNE having rich sentiment-related annotation. The corpus is annotated with named entities and sentiments towards these entities, along with related effects and emotional states. The evaluation was organized using the CodaLab competition framework. The main evaluation measure was macro-averaged measure of positive and negative classes. The best results achieved were of 66% Macro Fmeasure (Positive+Negative classes). We also tested ChatGPT on the test set from our evaluation and found that the zero-shot answers provided by ChatGPT reached 60% of the F-measure, which corresponds to 4th place in the evaluation. ChatGPT also provided detailed explanations of its conclusion. This can be considered as quite high for zero-shot application.

FREQUENCY DYNAMICS AS A CRITERION FOR DIFFERENTIATING INFLECTION AND WORD FORMATION (IN RELATION TO RUSSIAN ASPECTUAL PAIRS)

Gorbova E. V., independent researcher, **Chuikova O. Iu.**, Herzen State Pedagogical University of Russia

The paper reports the results of the critical evaluation of the quantitative approach to the distinction between inflection and word formation through the analysis of the trends in the frequency of word forms. The possibility of such analysis is provided by voluminous corpus data and tools for visualizing these trends. Both theoretical foundations of the proposed approach and the results of the pilot study of its applying to Russian aspectual triplets were considered. These cast doubt on the validity of distinguishing between inflection and word formation based on the trends in the frequency of word forms as a reliable tool used to reveal the unity or difference of lexical semantics and thus to define textual units as belonging to the same or different language units.

COMPUTER-ASSISTED DETECTION OF TYPOLOGICALLY RELEVANT SEMANTIC SHIFTS IN WORLD LANGUAGES

Gruntov I., Institute of Linguistics, Moscow, Russia, **Rykov E.**, HSE University, Moscow, Russia

The paper contains the description of a semi-authomatic method for the detection of typologically relevant semantic shifts in the world's languages. The algorithm extracts colexified pairs of meanings from polysemous words in digitised bilingual dictionaries. A machine learning classifier helps to separate those semantic shifts that are relevant to the lexical typology. Clustering is applied to group similar pairs of meanings into semantic shifts.

VAGUE REFERENCE IN EXPOSITORY DISCOURSE: MULTIMODAL REGULARITIES OF SPEECH AND GESTURE

Iriskhanova O.[1, 2], **Kiose M.**[1, 2], **Leonteva A.**[1, 2], **Agafonova O.**[1], [1]Moscow State Linguistic University; [2]Institute of Linguistics RAS, Moscow, Russia

The paper looks into the vague reference expressed in speech and gesture distribution in expository discourse. The research data are the monologues of 19 participants with total length of 2 hours 38 minutes. In these monologues, the use of vague reference (expressed in placeholders and approximators, with total amount of 2528) and functional gesture types (deictic, representational, pragmatic and adaptors, with total amount of 2309) was explored, with the aim of identifying the regular patterns of speech and gesture distribution and co-occurrence. The multimodal regularities include 1) the proportional frequency of four gesture types use equal to 6.8 / 14.4 / 28.7 / 50.1, which manifests overall distribution of co-speech gesture in expository discourse, 2) the significant difference in co-speech gesture use with placeholders and approximators which manifests itself in the use of three gesture types, adaptors, representational and pragmatic gestures, 3) the individually maintained significant difference in co-speech gesture use with placeholders and approximators which manifests itself in adaptors. These regularities can serve as predictors for identifying the specifics of vague reference in multimodal expository discourse.

A NEW DATASET FOR SENTENCE-LEVEL COMPLEXITY IN RUSSIAN

Ivanov V.[1, 2], **Elbayoumi M. G.**[2], [1]Kazan Federal University, Kazan, Russia; [2]Innopolis University, Innopolis, Russia

Text complexity prediction is a well-studied task. Predicting complexity sentence-level has attracted less research interest in Russian. One possible application of sentence-level complexity prediction is more precise and fine-grained modeling of text complexity. In the paper we present a novel dataset with sentence-level annotation of complexity. The dataset is open and contains 1,200 Russian sentences extracted from SynTagRus treebank. Annotations were collected via Yandex Toloka platform using 7-point scale. The paper presents various linguistic features that can contribute to sentence complexity as well as a baseline linear model.

THE PROBLEM OF LINGUISTIC MARKUP CONVERSION: THE TRANSFORMATION OF THE COMPRENO MARKUP INTO THE UD FORMAT

Ivoylova A. M.[1], **Dyachkova D. S.**[1], **Petrova M. A.**[2], **Michurina M. A.**[1], [1]RSUH; [2]A4 Technology, Moscow, Russia

The linguistic markup is an important NLP task. Currently, there are several popular formats of the markup (Universal Dependencies, Prague Dependencies, and so on), which are mostly focused on morphology and syntax. Full semantic markup can be found in the ABBYY Compreno model. However, the structure of the format differs significantly from the models mentioned above. In the given work, we convert the Compreno markup into the UD format, which is rather popular among NLP researchers, and enrich it with the semantical pattern.

Compreno and UD present morphology and syntax differently as far as tokenization, POS-tagging, ellipsis, coordination, and some other things are concerned, which makes the conversion of one format into another more complicated. Nevertheless, the conversion allowed us to create the UD-markup containing not only morpho-syntactic information but also the semantic one.

KNOWLEDGE TRANSFER BETWEEN TASKS AND LANGUAGES IN THE MULTI-TASK ENCODER-AGNOSTIC TRANSFORMER-BASED MODELS

Karpov D., **Konovalov V.**, MIPT, Dolgoprudny, Russia

We explore the knowledge transfer in the simple multi-task encoder-agnostic transformer-based models on five dialog tasks: emotion classification, sentiment classification, toxicity classification, intent classification, and topic classification. We show that these models' accuracy differs from the analogous single-task models by ~0.9%. These results hold for the multiple transformer backbones. At the same time, these models have the same backbone for all tasks, which allows them to have about 0.1% more parameters than any analogous single-task model and to support multiple tasks simultaneously. We also found that if we decrease the dataset size to a certain extent, multi-task models outperform singletask ones, especially on the smallest datasets. We also show that while training multilingual models on the Russian data, adding the English data from the same task to the training sample can improve model performance for the multi-task and single-task settings. The improvement can reach 4–5% if the Russian data are scarce enough. We have integrated these models to the DeepPavlov library and to the DREAM dialogue platform.

ATTENTION-BASED ESTIMATION OF TOPIC MODEL QUALITY

Kataeva V., **Khodorchenko M.**, ITMO University, St Petersburg, Russia

Topic modeling is an essential instrument for exploring and uncovering latent patterns in unstructured textual data, that allows researchers and analysts to extract valuable understanding of a particular domain. Nonetheless, topic modeling lacks consensus on the matter of its evaluation. The estimation of obtained insightful topics is complicated by several obstacles, the majority of which are summarized by the absence of a unified system of metrics, the one-sidedness of evaluation, and the lack of generalization. Despite various approaches proposed in the literature, there is still no consensus on the aspects of effective examination of topic quality. In this research paper, we address this problem and propose a novel framework for evaluating topic modeling results based on the notion of attention mechanism and Layer-wise Relevance Propagation as tools for discovering the dependencies between text tokens. One of our proposed metrics achieved a 0.71 Pearson correlation and 0.74 φ_k correlation with human assessment. Additionally, our score variant outperforms other metrics on the challenging Amazon Fine Food Reviews dataset, suggesting its ability to capture contextual information in shorter texts.

FOREGROUNDING AND ACCESSIBILITY EFFECTS IN THE GAZE BEHAVIOR OF THE READERS WITH DIFFERENT COGNITIVE STYLE

Kiose M.[1,2], **Rzheshevskaya A.**[1], **Izmalkova A.**[3,1], **Makeev S.**[4], [1]Moscow State Linguistic University; [2]Institute of Linguistics RAS; [3]Higher School of Economics; [4]Lomonosov Moscow State University, Moscow, Russia

This paper explores accessibility effects in the gaze behavior of readers with different cognitive style, impulsive and reflective, as mediated by graphological and linguistic foregrounding in the discursive acts in 126 areas of interest (AOIs). The study exploits 1890 gaze behavior probes available at open access Multimodal corpus of oculographic reactions MultiCORText. We identified that while graphological foregrounding makes initial or final components of discursive act more accessible for the impulsive readers, reflective readers also observe the components within the act. Linguistic foregrounding produces higher access with impulsive readers in case the linguistic form is visually focalized (phonological foregrounding and parallel structures); meanwhile, with reflective readers this is the information density appearing in elliptical and one-component sentences which maintains higher access.

TOWARDS A RUSSIAN MULTIMEDIA POLITENESS CORPUS

Klokova K.[1], **Krongauz M.**[2], **Shulginov V.**[1,2], **Yudina T.**[1], [1]MIPT, [2]HSE, Moscow, Russia

Communication involves an exchange of information as well as the use of linguistic means to begin, sustain, and end conversations. Politeness is seen as one of the major language tools that facilitate smooth communication. In English, politeness has been an area of great interest in pragmatics, with various theories and corpus annotation approaches used to understand the relationship between politeness and social categories like power and gender, and to build Natural Language Processing applications. In Russian linguistics, politeness research has largely focused on lexical markers and speech strategies. This paper introduces the ongoing work on the development of the Russian Multimedia Politeness Corpus and discusses an annotation framework for oral communicative interaction, with an emphasis on adapting politeness theories for discourse annotation. The proposed approach lies in the identification of frames that encompass contextual information and the selection of relevant spatial, social, and relational features for the markup. The frames are then used to describe standard situations, which are marked by typical intentions and politeness formulae and paraverbal markers.

AN EXPERIMENTAL STUDY OF ARGUMENT EXTRACTION FROM PRESUPPOSITIONAL CLAUSES IN RUSSIAN

Knyazev M., Institute for Linguistic Studies, Russian Academy of Sciences, Saint Petersburg, Russia; HSE University, Saint Petersburg, Russia; Lomonosov Moscow State University, Moscow, Russia

The paper discusses two acceptability rating studies testing wh-interrogative and relative extractions of arguments from *čto*-clauses of presuppositional predicates like *žalet'* 'regret', as contrasted with nonpresuppositional predicates like *nadejat'sja* 'hope' and nominalized (*to čto*) clauses. The results show a difference in extraction between bare and nominalized clauses but no difference between presuppositional and nonpresuppositional clauses, raising potential doubts about the analysis of presuppositional clauses as DPs with a silent D.

COLLABORATIVE CONSTRUCTIONS IN RUSSIAN CONVERSATIONS: A MULTICHANNEL PERSPECTIVE

Korotaev N. A., Institute of Linguistics RAS; Russian State University for the Humanities, Moscow, Russia

The talk provides a multichannel description of how interlocutors co-construct utterances in conversation. Using data from the "Russian Pears Chats & Stories", I propose for a tripartite sequential scheme of collaborative constructions. When the scheme is fully realized, its first step not only includes the initial component of the construction, but also presupposes that the first participant makes a request for a co-operative action; the final component of the construction is provided by the second participant during the

second step; while the third step consists of the first participant's reaction. On each step, the participants combine vocal and non-vocal resources to achieve their goals. In some cases, non-vocal phenomena provide an essential clue to what is actually happening during co-construction, including whether the participants act in a truly co-operative manner. I distinguish between three types of communicative patterns that may take place during co-construction: "Requested Cooperation", "Unplanned Cooperation", and "Non-realized Interaction". The data suggest that these types can be influenced by the way the knowledge of the discussed events is distributed among the participants.

FACT-CHECKING BENCHMARK FOR THE RUSSIAN LARGE LANGUAGE MODELS

Kozlova A., **Shevelev D.**, **Fenogenova A.**, SberDevices, Moscow, Russia

Modern text-generative language models are rapidly developing. They produce text of high quality and are used in many real-world applications. However, they still have several limitations, for instance, the length of the context, degeneration processes, lack of logical structure, and facts consistency. In this work, we focus on the fact-checking problem applied to the output of the generative models on classical downstream tasks, such as paraphrasing, summarization, text style transfer, etc. We define the task of internal fact-checking, set the criteria for factual consistency, and present the novel dataset for this task for the Russian language. The benchmark for internal fact-checking and several baselines are also provided. We research data augmentation approaches to extend the training set and compare classification methods on different augmented data sets.

TEXT COMPLEXITY AS A NON-DISCRETE VALUE: RUSSIAN L2 TEXT COMPLEXITY DATASET ANNOTATION BASED ON ELO RATING SYSTEM

Laposhina A. N., Pushkin State Russian Language Institute, Moscow, Russia

The task of assessing text complexity for L2 learners can be approached as either a classification or regression problem, depending on the chosen scale. The primary bottleneck in such research lies in the limited availability of appropriate data samples. This study presents a combined approach to create a dataset of Russian texts for L2 learners, placed on a continuous scale of complexity, involving expert pairwise comparisons and the Elo rating system. For this pilot dataset, 104 texts from Russian L2 textbooks, TORFL tests, and authentic sources were selected and annotated. The resulting data is useful for evaluation of the automated models for assessing text complexity.

WHOSE WORD? PROBLEMS OF LEXICOGRAPHIC REPRESENTATION OF IDEOLOGICALLY MARKED WORDS (THE LEXICON OF THE RUSSIAN-UKRAINIAN CONFLICT)

Levontina I. B., **Shmeleva E. Ya.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

The article deals with the problems of presenting ideologically marked words in the dictionary. It is based on the analysis of the words that appeared in the Russian language or received new meanings during the Russian-Ukrainian conflict. The difficulty of the lexicographic representation of such words is that their evaluative potential is mobile, for example, offensive nicknames can be assimilated by "offended" ones and become neutral words. Ideologically marked words can either exist in the lexicon for a long time or be quickly replaced by other lexical units. Therefore, in the interpretation of ideologically marked words, it is advisable to indicate the approximate time of their existence. In addition to temporary indicators, in the dictionary entry of such words, it is necessary to indicate whose word it is, that is, on whose behalf an assessment is given to a person or event. Since we believe that explanatory dictionaries should contain not only common names, but also proper names, the article also discusses geographical names.

PARAMETER-EFFICIENT TUNING OF TRANSFORMER MODELS FOR ANGLICISM DETECTION AND SUBSTITUTION IN RUSSIAN

Lukichev D.[1 2], **Kryanina D.**[1], **Bystrova A.**[1], **Fenogenova A.**[3], **Tikhonova M.**[1, 3], [1]HSE University; [2]Sber; [3]SberDevices, Moscow, Russia

This article is devoted to the problem of Anglicisms in texts in Russian: the tasks of detection and automatic rewriting of the text with the substitution of Anglicisms by their Russian-language equivalents. Within the framework of the study, we present a parallel corpus of Anglicisms and models that identify Anglicisms in the text and replace them with the Russian equivalent, preserving the stylistics of the original text.

DISAMBIGUATION IN CONTEXT IN THE RUSSIAN NATIONAL CORPUS: 20 YEAS LATER

Lyashevskaya O. N.[1, 2], **Afanasev I. A.**[1, 3], **Rebrikov S. A.**[1, 4], **Shishkina Y. A.**[1, 5], **Suleymanova E. A.**[6], **Trofimov I. V.**[6], **Vlasova N. A.**[6], [1]HSE University; [2]Vinogradov Russian Language Institute RAS; [3]MTS AI; [4]Kurchatov Institute; [5]Moscow Institute of Physics and Technology, Moscow, Russia; [6]A. K. Ailamazyan Program Systems Institute of RAS, Pereslavl-Zalessky, Russia

An updated annotation of the Main, Media, and some other corpora of the Russian National Corpus (RNC) features the part-of-speech and other morphological information, lemmas, dependency structures, and constituency types. Transformer-based architectures are used to resolve the homonymy in context according to a schema based on the manually disambiguated subcorpus of the Main corpus (morphology and lexicon) and UD-SynTagRus (syntax). The paper discusses the challenges in applying the models to texts of different registers, orthographies, and time periods, on the one hand, and making the new version convenient for users accustomed to the old search practices, on the other. The reannotated corpus data form the basis for the enhancement of the RNC tools such as word and n-gram frequency lists, collocations, corpus comparison, and Word at a glance.

MULTIMODAL HEDGES FOR COMPANION ROBOTS: A POLITENESS STRATEGY OR AN EMOTIONAL EXPRESSION?

Malkina M. P.[1], **Zinina A. A.**[2, 3, 1], **Arinkin N. A.**[2, 3], **Kotov A. A.**[2, 3], [1]MSLU; [2]Kurchatov Institute, [3]RSUH, Moscow, Russia

We examine the use of multimodal hedges (a politeness strategy, like saying *A kind of!*) by companion robots in two symmetric situations: (a) user makes a mistake and the robot affects user's social face by indicating this mistake, (b) robot makes a mistake, loses its social face and may compensate it with a hedge. Within our first hypothesis we test the politeness theory, applied to

robots: the robot with hedges should be perceived as more polite, threat to its social face should be reduced. Within our second hypothesis we test the assumption that multimodal hedges, as the expression (or simulation) of internal confusion, may make the robot more emotional and attractive. In our first experiment two robots assisted users in language learning and indicated their mistakes by saying *Incorrect!* The first robot used hedges in speech and gestures, while the second robot used gestures, supporting the negation. In our second experiment two robots answered university exam questions and made minor mistakes. The first robot used hedges, while the second robot used addressive strategy in speech and gestures, e. g. moved its hand to the user and said *That's it!* We have discovered that the use of hedges as the politeness strategy in both situations makes the robot *comfortable to communicate with*. But robot with hedges looks more *polite* only in the experiment, where it affects user's social face, and not when the robot makes mistakes. However, the usage of hedges as an emotional cue works in both cases: the robot with hedges seems to be *cute* and *sympathy provoking* both when it attacks user's social face or loses its own social face. This spectrum of hedge usage can demonstrate its transition from an expressive cue of a negative emotion (nervousness) to a marker of speaker's friendliness and competence.

AUGMENTATION METHODS FOR SPELLING CORRUPTIONS

Martynov N., **Baushenko M.**, **Abramov A.**, **Fenogenova A.**, SberDevices, Moscow, Russia

The problem of automatic spelling correction is vital to applications such as search engines, chatbots, spellchecking in browsers and text editors. The investigation of spell-checking problems can be divided into several parts: error detection, emulation of the error distribution on the new data for model training, and automatic spelling correction. As the data augmentation technique, the adversarial training via error distribution emulation increases a model's generalization capabilities; it can address many other challenges: from overcoming a limited amount of training data to regularizing the training objectives of the models. In this work, we propose a novel multi-domain dataset for spelling correction. On this basis, we provide a comparative study of augmentation methods that can be used to emulate the automatic error distribution. We also compare the distribution of the single-domain dataset with the errors from the multi-domain and present a tool that can emulate human misspellings.

AUTOCORRELATIONS DECAY IN TEXTS AND APPLICABILITY LIMITS OF LANGUAGE MODELS

Mikhaylovskiy N.[1,2], **Churilov I.**[2], [1]Higher IT School, Tomsk State University, Tomsk, Russia; [2]NTR Labs, Moscow, Russia

We show that the laws of autocorrelations decay in texts are closely related to applicability limits of language models. Using distributional semantics we empirically demonstrate that autocorrelations of words in texts decay according to a power law. We show that distributional semantics provides coherent autocorrelations decay exponents for texts translated to multiple languages. The autocorrelations decay in generated texts is quantitatively and often qualitatively different from the literary texts. We conclude that language models exhibiting Markovian behavior, including large autoregressive language models, may have limitations when applied to long texts, whether analysis or generation.

NAMED ENTITY-ORIENTED SENTIMENT ANALYSIS WITH TEXT2TEXT GENERATION APPROACH

Moloshnikov I.[1], **Skorokhodov M.**[1], **Naumov A.**[1], **Rybka R.**[1,2], **Sboev A.**[3,1], [1]NRC "Kurchatov Institute"; [2]Russian Technological University "MIREA"; [3]National Research Nuclear University "MEPhI", Moscow, Russia

This paper describes methods for sentiment analysis targeted toward named entities in Russian news texts. These methods are proposed as a solution for the Dialogue Evaluation 2023 competition in the RuSentNE shared task. This article presents two types of neural network models for multi-class classification. The first model is a recurrent neural network model with an attention mechanism and word vector representation extracted from language models. The second model is a neural network model for text2text generation. High accuracy is demonstrated by the generative model fine-tuned on the competition dataset and CABSAR open dataset. The proposed solution achieves 59.33 over two sentiment classes and 68.71 for three-class classification by f1-macro.

"PEARS ARE BIG GREEN": GESTURES WITH CONCRETE OBJECTS

Nikolaeva Y. V., Lomonosov Moscow State University, Interdisciplinary Scientific and Educational School "Preservation of the World Cultural and Historical Heritage", Moscow, Russia

The paper examines hand gestures when referring to inanimate referents. The aim of the study was to explore which factors determine the features of a gesture within the framework of modes of representation. Four main types of modes of representation were considered: drawing or shaping the form of the referent, acting, pointing, and presentation (PUOH); in addition, a new category of beat gestures was added.

As a result, it was shown that communicative dynamism or other referent characteristics such as control of the object or its inferability from the previous context do not fully determine the use of gestures with the referent. As an alternative hipothesis, we propose a notion of gesture information hierarchy, where discursive factors, such as previous mentions of the referent and the introduction or change of the protagonist along with the way an object is used determines the form of the gesture.

RUSSIAN CONSTRUCTICON 2.0: NEW FEATURES AND NEW PERSPECTIVES OF THE BIGGEST CONSTRUCTICON EVER BUILT

Orlov A. V.[1], **Butenko Z. A.**[1,2], **Demidova D. A.**[2], **Starchenko V. M.**[1], **Rakhilina E. V.**[1,3], **Lyashevskaya O. N.**[1,3], [1]HSE University, Moscow, Russia; [2]UiT The Arctic University of Norway, Tromsø, Norway; [3]Vinogradov Institute for Russian language RAS, Moscow, Russia

Russian constructicon is an open-access linguistic database containing detailed descriptions of over 3,800 Russian grammatical constructions. In this paper we present a new, enlarged and updated version of Russian Constructicon (RusCxn) as well as new trajectories of development which were opened for the resource after the update. Since its first release, RusCxn, has undergone many significant changes. Our team has expanded the number of constructions present in the database 1,5 times, introduced new meta-information features such as glosses, significantly reworked the architecture and the design of Russian Constructicon's website, and improved the search facilities. The above-mentioned changes not only make RusCxn more attractive and convenient-to-use, but they can also greatly facilitate typological research in the field of Construction Grammar and improve the mapping between constructicography-orinented resources for different languages.

LINGUISTIC ANNOTATION GENERATION WITH CHATGPT: A SYNTHETIC DATASET OF SPEECH FUNCTIONS FOR DISCOURSE ANNOTATION OF CASUAL CONVERSATIONS

Ostyakova L.[1,2], **Petukhova K.**[2], **Smilga V.**[2], **Zharikova D.**[2], [1]HSE University; [2]Moscow Institute of Physics and Technology, Moscow, Russia

This paper is devoted to examining the hierarchical and multilayered taxonomy of Speech Functions, encompassing pragmatics, turn-taking, feedback, and topic switching in open-domain conversations. To evaluate the distinctiveness of closely related pragmatic classes, we conducted comparative analyses involving both expert annotators and crowdsourcing workers. We then carried out classification experiments on a manually annotated dataset and a synthetic dataset generated using ChatGPT. We looked into the viability of using ChatGPT to produce data for such complex topics as discourse. Our findings contribute to the field of prompt engineering techniques for linguistic annotation in large language models, offering valuable insights for the development of more sophisticated dialogue systems.

POLY-PREDICATION IN INFORMAL MONOLOGICAL DISCOURSE (ACCORDING TO «WHAT I SAW» CORPUS)

Panysheva D., Russian State University for the Humanities, Moscow, Russia

The article discusses the relationship between the mode of discourse and quantitative metrics of poly-predication. Based on the material of the corpus "What I Saw", oral and written versions of stories are compared according to the relative frequency of polypredicative constructions and the representation of certain types of polypredication, the features of semantics and grammatical labeling of such structures are described. Using the nonparametric Wilcoxon criterion, the absence of statistical significance between the density of poly-predication in the oral and written parts of the corpus is proved.

RUSSIAN ADDITIVE MARKERS *TAKŽE* AND *TOŽE*: A SYNCHRONIC AND DIACHRONIC PERSPECTIVE

Pekelis O. E., Russian State University for the Humanities/Moscow, Russia; HSE University/Moscow, Russia

It is well known that Russian additive markers *takže* and *tože* differ in terms of information structure: the scope of *takže* is focus, while the scope of *tože* is topic. Based on data of several corpora of Russian, this paper shows that in modern Russian, *takže* and tože are opposed on other language levels as well, namely syntactically (in terms of word order), lexically (a variant of *takže* that is synonymous with *tože* including at the level of the information structure, is going out of use), stylistically and as far as their involvement in grammaticalization processes is concerned (*takže* but not *tože* developed into a coordinate conjunction and a discourse marker). However, as evidenced by Russian National Corpus data, most of these contrasts were absent or less pronounced in the Russian language of the 18th-19th centuries. Thus, in the last two centuries *takže* and *tože* evolved toward their consistent differentiation.

THE COBALD ANNOTATION PROJECT: THE CREATION AND APPLICATION OF THE FULL MORPHO-SYNTACTIC AND SEMANTIC MARKUP STANDARD

Petrova M. A.[1], **Ivoylova A. M.**[2], **Bayuk I. S.**[1], **Dyachkova D. S.**[2], **Michurina M. A.**[2], [1]A4 Technology; [2]RSUH, Moscow, Russia

The current paper is devoted to the Compreno-Based Linguistic Data (CoBaLD) Annotation Project aimed at creating text corpora annotated with full morphological, syntactic and semantic markup. The first task of the project is to suggest a standard for the full universal markup which would include both morphosyntactic and semantic patterns. To solve this problem, one needs the markup model, which includes all necessary markup levels and presents the markup in a format convenient for users. The latter implies not only the fullness of the markup, but also its structural simplicity and homogeneity. As a base for the markup, we have chosen the simplified version of the Compreno model, and as data presentation format, we have taken Universal Dependencies.

At the second stage of the project, the Russian corpus with 400 thousand tokens (CoBaLD-Rus) has been created, which is annotated according to the given standard. The third stage is devoted to the testing of the new format. For this purpose, we have held the SEMarkup Shared Task aimed at creating parsers which would produce full morpho-syntactic and semantic markup. Within this task, we have elaborated neural network-based parser trained on our dataset, which allows one to annotate new texts with the CoBaLD-standard. Our further plans are to create fully annotated corpora for other languages and to carry out the experiments on language transfers of the current markup to other languages.

HALF-MASKED MODEL FOR NAMED ENTITY SENTIMENT ANALYSIS

Podberezko P., **Kaznacheev A.**, **Abdullayeva S.**, **Kabaev A.**, MTS AI, Moscow, Russia

Named Entity Sentiment analysis (NESA) is one of the most actively developing application domains in Natural Language Processing (NLP). Social media NESA is a significant field of opinion analysis since detecting and tracking sentiment trends in the news flow is crucial for building various analytical systems and monitoring the media image of specific people or companies.

In this paper, we study different transformers-based solutions NESA in RuSentNE-23 evaluation. Despite the effectiveness of the BERT-like models, they can still struggle with certain challenges, such as overfitting, which appeared to be the main obstacle in achieving high accuracy on the RuSentNE-23 data. We present several approaches to overcome this problem, among which there is a novel technique of additional pass over given data with masked entity before making the final prediction so that we can combine logits from the model when it knows the exact entity it predicts sentiment for and when it does not. Utilizing this technique, we ensemble multiple BERTlike models trained on different subsets of data to improve overall performance. Our proposed model achieves the best result on RuSentNE-23 evaluation data and demonstrates improved consistency in entity-level sentiment analysis.

PROSODIC PORTRAIT OF THE RUSSIAN CONNECTOR PRICHOM IN THE MIRROR OF THE MULTIMEDIA CORPUS

Podlesskaya V. I., Institute of linguistics, Russian Academy of Sciences; Russian State University for the Humanities, Moscow, Russia

Based on data from the multimedia subcorpus of the Russian National Corpus, the paper addresses prosodic features of discourse fragments introduced by the connector *prichom* 'and besides'. The data of instrumental and perceptual analysis show that the fragment with *prichom* has communicative-prosodic autonomy: firstly, it has an internal thematic structure with an obligatory rheme and an optional theme; and secondly, there is a prosodic break before this fragment. The autonomy of the fragment introduced by *prichom* is preserved in

a variety of contexts: (i) both in cases where this fragment is a complete clause and when it is a fragmented clause; (ii) both in those cases when the previous fragment is prosodically realized as final (projecting no continuation), and when it is realized as non-final (projecting continuation); (iii) both in those cases when the fragment introduced by *prichom* is an element of the main narrative chain, and when it is inserted parenthetically inside another fragment. In addition to the above, a fragment with *prichom* can form a separate turn in the conversation. Thus, the detected prosodic features of the fragment with *prichom* make it possible to objectify the idea earlier expressed in the literature (Kiselyova 1971, Vinogradov 1984, Inkova 2018, inter alia): that structures with *prichom* are built in two "communicative steps", or that they are used to express "concomitance established at the level of speech acts". Clauses connected by the relationship of syntactic subordination quite often lose their prosodic autonomy (Podlesskaya 2014 a, b), and vice versa, clauses in coordinated constructions tend to retain prosodic autonomy. Therefore, the prosodic autonomy of the components of the construction with *prichom*, retained in various contexts, speaks in favor of its coordinated status, while a number of syntactic tests proper speak of the opposite.

HWR200: NEW OPEN ACCESS DATASET OF HANDWRITTEN TEXTS IMAGES IN RUSSIAN

Potyashin I.[1], **Kaprielova M.**[1,2], **Chekhovich Y.**[1,2], **Kildyakov A.**, **Seil T.**[1], **Finogeev E.**[1], **Grabovoy A.**[1,2], [1]Antiplagiat; [2]FRC CSC RAS, Moscow, Russia

Handwritten text image datasets are highly useful for solving many problems using machine learning. Such problems include recognition of handwritten characters and handwriting, visual question answering, near-duplicate detection, search for text reuse in handwriting and many auxiliary tasks: highlighting lines, words, other objects in the text. The paper presents new dataset of handwritten texts images in Russian created by 200 writers with different handwriting and photographed in different environment. We described the procedure for creating this dataset and the requirements that were set for the texts and photos. The experiments with the baseline solution on fraud search and text reuse search problems showed results of results of 60% and 83% recall respectively and 5% and 2% false positive rate respectively on the dataset.

SIMPLE YET EFFECTIVE NAMED ENTITY ORIENTED SENTIMENT ANALYSIS

Sanochkin L.[1,2], **Bolshina A.**[1], **Cheloshkina K.**[1,2], **Galimzianova D.**[1,2], **Malafeev A.**[1,2], [1]MTS AI, [2]HSE, Moscow, Russia

Sentiment analysis, i.e. the automatic evaluation of the emotional tone of a text, is a common task in natural language processing. Entity-Oriented Sentiment Analysis (EOSA) predicts the sentiment of entities mentioned in a given text. In this paper, we focus on the EOSA task for the Russian news. We propose a text classification pipeline to solve this task and show its potential in such tasks. Moreover, in general, EOSA implies labeling both named entities and their sentiment, which can require a lot of annotator labour and time and, thus, presents a major obstacle to the development of a production-ready EOSA system. To help alleviate this, we analyse the potential of applying an Active learning approach to EOSA tasks. We demonstrate that by actively selecting instances for labeling in EOSA the annotation effort required for training machine learning models can be significantly reduced.

IS IT POSSIBLE TO MAKE THE RUSSIAN PUNCTUATION RULES MORE EXPLICIT?

Shmelev A., Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper deals with some issues related to the Russian punctuation rules and their account in computer checkers and correctors (both “analytic” and “synthetic”). It also discusses variation of punctuation. The paper offers a critical assessment of reference books devoted to punctuation and makes special reference to certain verbs of propositional attitude and their parenthetical use (in particular, *dumat’* ‘to think,’ *videt’* ‘to see,’ and *slyshat’* ‘to hear). It claims that the inherent characteristics of the verbs under consideration influence the punctuation, and therefore every verb deserves a detailed description (lexicographic portrait). In particular, *videt’* and *slyshat’* behave quite differently when used as parenthetical verbs. A step towards making the punctuation rules more explicit may consist in providing an index of words mentioned in the rules together with a subject index.

THE ROLE OF INDICATORS IN ARGUMENTATIVE RELATION PREDICTION

Sidorova E., **Akhmadeeva I.**, **Kononenko I.**, **Chagina P.**, A.P. Ershov Institute of Informatics Systems, Siberian Branch, Russian Academy of Sciences, Novosibirsk, Russia

The article presents a comparative study of methods for argumentative relation prediction based on a neural network approach. The distinctive feature of the study is the use of argumentative indicators in the preparation of the training sample. The indicators are generated based on the discourse marker dictionary. The experiments were carried out using an annotated corpus of scientific and popular science texts, including 162 articles available on the ArgNet-Bank Studio web platform. A set of all argumentative relations is described by internal connections of arguments and include the conclusion and the premise. In the first stage of training set construction, fragments of text that included two consecutive sentences were examined. In the second stage, indicators were retrieved from the corpus texts and, for each indicator, statements presumably corresponding to the premise and conclusion of the argument were extracted. In total, 4.2 thousand indicator-based training contexts and 13.6 thousand pairs of sentences were obtained from the corpus with annotation of the presence of an argumentative relation. Based on this training sample, four classifiers were built: without indicators, with marking indicators in sentences using tags, taking into account segmentation of text based on indicators, with segmentation and tags. The results of the experiments on argumentative relation prediction are presented.

TEXT VQA WITH TOKEN CLASSIFICATION OF RECOGNIZED TEXT AND RULE-BASED NUMERICAL REASONING

Surkov V. O., **Evseev D. A.**, Moscow Institute of Physics and Technology, Dolgoprudny, Russia

In this paper, we describe a question answering system on document images which is capable of numerical reasoning over extracted structured data. The system performs optical character recognition, detection of key attributes in text, generation of a numerical reasoning program, and its execution with the values of key attributes as operands. OCR includes the steps of bounding boxes detection and recognition of text from bounding boxes. The extraction of key attributes, such as quantity and price of goods, total etc., is based on the BERT token classification model. For expression generation we investigated the rule-based approach and the T5-base model and found that T5 is capable of generalization to expression types unseen in the training set. The proposed architecture of the question answering system utilizes the structure of independent blocks, each of which can be enhanced or replaced while keeping other components unchanged. The proposed model was evaluated in the Receipt-AVQA competition and on FUNSD dataset.

SCALAR STRUCTURE FOR *POLU- HALF*

Tatevosov S. G., Lomonosov Moscow State University Interdisciplinary School "Preservation of the World Cultural and Historical Heritage", Moscow, Russia, **Kisseleva X. L.**, Vinogradov Russian Language Institute of the Russian Academy of Sciences, Moscow, Russia

This paper explores restrictions on the distribution of *polu-* 'half' in combination with adjectival stems in Russian. Relying on the literature on degree semantics, we analyze *polu-* as a degree modifier that specifies the degree to which the adjective maps an individual as ½ of the maximal degree. This correctly predicts that *polu-* can only combine with upper closed scales. We argue that unlike *half* in English, *polu-* does not require a scale be lower closed.

TEXT SIMPLIFICATION AS A CONTROLLED TEXT STYLE TRANSFER TASK

Tikhonova M., HSE University, SberDevices, Moscow, Russia, **Fenogenova A.**, SberDevices, Moscow, Russia

The task of text simplification is to reduce the complexity of the given piece of text while preserving its original meaning to improve readability and understanding. In this paper, we consider the simplification task as a subfield of the general text style transfer problem and apply methods of controllable text style to rewrite texts in a simpler manner preserving their meaning. Namely, we use a paraphrase model guided by another style-conditional language model. In our work, we perform a series of experiments and compare this approach with the standard fine-tuning of an autoregressive model.

AN ATTEMPT TO DETERMINE A PREPOSITION AND DELIMIT THE CLASS OF DERIVED PREPOSITIONS IN RUSSIAN

Uryson E., Russian Language Institute RAS, Moscow, Russia

The object of the paper are Russian words traditionally described as derived prepositions. The problem is that there is no formal definition of preposition in theoretical or applied linguistics. Non-derivative, or primitive prepositions are given in grammar by the closed list, so strictly speaking there is no need to define this class of words. However. we must have criteria for determining derived prepositions. I suggest a set of necessary conditions that a preposition must satisfy. I demonstrate that so called adverbial prepositions in Russian do not satisfy them and should be described as adverbs. Similarly, some Russian verbal prepositions, and some Russian denominative prepositions should not be described as prepositions.

ESTIMATING COGNITIVE TEXT COMPLEXITY WITH AGGREGATION OF QUANTILE-BASED MODELS

Veselov A. S., Lomonosov Moscow State University, Moscow, Russia, **Eremeev M. A.**, New York University, New York, USA, **Vorontsov K. V.**, Moscow Institute of Physics and Technology, Moscow, Russia

In this paper, we introduce a novel approach to estimating the cognitive complexity of a text at different levels of language: phonetic, morphemic, lexical, and syntactic. The proposed method detects tokens with an abnormal frequency of complexity scores. The frequencies are taken from the empirical distributions calculated over the reference corpus of texts. We use the Russian Wikipedia for this purpose. Ensemble models are combined from individual models from different language levels. We created datasets of pairs of text fragments taken from social studies textbooks of different grades to train the ensembles. Empirical evidence shows that the proposed approach outperforms existing methods, such as readability indices, in estimating text complexity in terms of accuracy. The purpose of this study is to create one of the important components of the system of recommendation of scientific and educational content.

MAXPROB: CONTROLLABLE STORY GENERATION FROM STORYLINE

Vychegzhanin S. V., **Kotelnikova A. V.**, **Sergeev A. V.**, **Kotelnikov E. V.**, Vyatka State University, Kirov, Russia

Controllable story generation towards keywords or key phrases is one of the purposes of using language models. Recent work has shown that various decoding strategies prove to be effective in achieving a high level of language control. Such strategies require less computational resources compared to approaches based on fine-tuning pre-trained language models. The paper proposes and investigates the method *MaxProb* of controllable story generation in Russian, which works at the decoding stage in the process of text generation. The method uses a generative language model to estimate the probability of its tokens in order to shift the content of the text towards the guide phrase. The idea of the method is to generate a set of different small sequences of tokens from the language model vocabulary, estimate the probability of following the guide phrase after each sequence, and choose the most probable sequence. The method allows evaluating the consistency of the token sequence for the transition from the prompt to the guide phrase. The study was carried out using the Russian-language corpus of stories with extracted events that make up the plot of the story. Experiments have shown the effectiveness of the proposed method for automatically creating stories from a set of plot phrases.

THE PROSODY OF THE RUSSIAN QUESTION

Yanko T. E., Institute of Linguistics, Russian Academy of Sciences, Moscow, Russia

The analysis of Russian interrogative prosody is based on a model of a question as consisting of the two components: the illocutionary proper component and the illocutionary improper component. The illocutionary improper component includes the data for information retrieval. The illocutionary proper component can be formed both by segmental means of expression (by an interrogative word or a particle) or solely by prosody (as in Russian yes-no questions). The prosody of Russian questions having the interrogative words or the interrogative particle li is highly variable, whereas the prosody of Russian yes-no questions expressed by prosody is stable. The latter is the Russian rising accent, which has a rise on the tonic syllable of the accent-bearer followed by a fall on the post-tonics if any. The illocutionary improper component can be located sentence initially and carry a specific falling accent (namely, a late fall). A specific type of a question with the interrogative proper component omitted is recognized. Such questions carry a late fall, or a falling-rising accent on the accent-bearer. The analysis is exemplified by the frequency tracings of the sound sentences taken from the Russian National Corpus and other open sources. As the instrument for verifying the acoustic data, we used the computer system Praat. The paper is illustrated throughout with pitch contours of sound records.

PARALLEL CORPUS AS A TOOL FOR SEMANTIC ANALYSIS: THE RUSSIAN DISCOURSE MARKER STALO BYT' (*CONSEQUENTLY*)

Zalizniak Anna A. , Institute of Linguistics of the RAS, Moscow, Russia, Dobrovol'ski jD. O. , Russian Language Institute of the RAS; Institute of Linguistics of the RAS, Moscow, Russia

The article examines the semantics of the Russian discourse marker *stalo byt',* using the data obtained by analyzing translational correspondences extracted from parallel corpora of the Russian National Corpus (RNC). Typically, this discourse marker is an indicator of inferential evidentiality, by which the speaker marks the fact that the given statement is a conclusion made by the speaker on the basis of the information they received and accepted as true by default. In addition, *stalo byt'* has two secondary types of usage—"rhetorical" and "narrative"—where the basic semantics of this discourse marker is subject to certain modifications. One of the key points of analysis is the reconstruction of semantic mechanisms providing the actual semantics of *stalo byt'*.

RUSSIAN PREDICATIVES AND FREQUENCY METRICS

Zimmerling A. V., Pushkin State Russian Language Institute; Institute of Linguistics, Russian Academy of Science, Moscow, Russia

This paper introduces five metrics for measuring the frequencies of dative predicatives in Russian. A dative predicative is a word or multiword expression licensing the dative-predicative-structure, where the semantic subject of the non-agreeing non-verbal predicate is marked by the dative case. I measure the frequencies of the predicatives in the contact position <−1;1> with the same-clause dative subject pronouns in 1Sg (*m*-metrics) and 3Sg (*e*-metrics). The *m*-metrics is applied for retrieving a list of dative predicatives from a corpus. I argue that for each large text collection there is a minimal *m*-value confirming that an item belongs to the core of the dative-predicative structure. The *m/e* score makes up the third metrics that shows whether an element is oriented towards the use in the 1st person or not. Basing on the *m*-metrics, I retrieved 3 lists of predicatives in the subcorpus of 2000–2021 texts included in the Russian National Corpus. The A list includes 87 items with $m \geq 10$, the B list includes 44 items with $m \geq 50$, the C list includes 24 items with $m \geq 100$. 72–79% of items in each list have an *m/e* value ≥ 1,25. A linguistic interpretation of this result is that for each list of dative predicatives it is true that the majority of its elements are autoreferential expressions oriented towards the use in the 1st person present indicative tense in the direct speech. The fourth metrics shows the total number of occurrences of a word or multiword expression in the corpus (*N*). I argue that the *N* score must be measured before POS tagging, and lemmatization. The fifth and the last metrics is the *m/N* score. The RNC data suggest an inverse correlation between the score of an item in the context specific for dative-predicative structures (*m*) and its overall frequency in the corpus (*N*). This effect is explained by the regular homonymy of high frequent predicatives with high frequent adverbials and parenthetical expressions.

Авторский указатель

Author Index

Научное издание

Компьютерная лингвистика и интеллектуальные технологии

По материалам ежегодной
международной конференции «Диалог»

Выпуск 22, 2023

Ответственный за выпуск **А. В. Ульянова**
Вёрстка **К. А. Климентовский**